Depth attention for scene understanding

Zongwei Wu

THESE DE DOCTORAT DE L'ETABLISSEMENT UNIVERSITE BOURGOGNE FRANCHE-COMTE

PREPAREE A L'UNIVERSITE DE BOURGOGNE

Ecole doctorale nº 37

Sciences Physiques pour l'Ingénieur et Microtechniques

Doctorat de : Instrumentation et Informatique de l'image

Par

M. WU Zongwei

Depth Attention for Scene Understanding

Thèse présentée et soutenue à Dijon, le 21/11/2022

Composition du Jury :

M. CHEN Liming	Ecole Centrale de Lyon	Président
M. WOLF Christian	Naver Labs Europe	Rapporteur
M. THOME Nicolas	Sorbonne University	Rapporteur
M. PICARD David	École des Ponts ParisTech	Examinateur
M. DEMONCEAUX Cédric	Université Bourgogne Franche-Comté	Directeur de thèse
M. STOLZ Christophe	Université Bourgogne Franche-Comté	Co-directeur de thèse
M. ALLIBERT Guillaume	Université Côte d'Azur	Encadrant

Université Bourgogne Franche-Comté
École Doctorale SPIM

Doctoral Thesis

Depth Attention for Scene Understanding

Zongwei Wu

Thèse soutenue publiquement le 21/11/2022,
devant le jury composé de:

Encadrants:
Guillaume Allibert, Associate Professor at I3S, Université Côte d'Azur
Christophe Stolz, Associate Professor at ImViA, Université Bourgogne Franche-Comté
Cédric Demonceaux, Professor at ImViA, Université Bourgogne Franche-Comté

Rapporteurs:
Christian Wolf, Principal Scientist at Naver Labs Europe
Nicolas Thome, Professor at ISIR, Sorbonne University
Examinateurs:
David Picard, Senior Research Scientist at IMAGINE, École des Ponts ParisTech
Liming Chen, Professor at LIRIS, Ecole Centrale de Lyon

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Instrumentation and image processing
(Instrumentation et informatique de l'image)

in the

ImViA
EA 7535
Université de Bourgogne

UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ
ÉCOLE DOCTORALE SPIM

Abstract

ImViA

EA 7535

Université de Bourgogne

Doctor of Philosophy

Depth Attention for Scene Understanding

by Zongwei Wu

Deep learning models can nowadays teach a machine to realize a number of tasks, even with better precision than human beings. Among all the modules of an intelligent machine, perception is the most essential part without which all other action modules have difficulties in safely and precisely realizing the target task under complex scenes. Conventional perception systems are based on RGB images which provide rich texture information about the 3D scene. However, the quality of RGB images highly depends on environmental factors, which further influence the performance of deep learning models. Therefore, in this thesis, we aim to improve the performance and robustness of RGB models with complementary depth cues by proposing novel RGB-D fusion designs.

Traditionally, pixel-wise concatenation with addition and convolution is the widely applied approach for RGB-D fusion designs. Inspired by the success of attention modules in deep networks, in this thesis we analyze and propose different depth-aware attention modules and demonstrate our effectiveness in basic segmentation tasks such as saliency detection and semantic segmentation. First, we leverage the geometric cues and propose a novel depth-wise channel of attention. We merge the fine-grained details and the semantic cues to constrain the channel attention into various local regions, improving the model discriminability during the feature extraction. Second, we investigate the depth-adapted offset which serves as a local but deformable spatial attention for convolution. Our approach forces the networks to take more relevant pixels into account with the help of depth prior. Third, we improve the contextualized awareness within RGB-D fusion by leveraging transformer attention. We show that transformer attention can improve the model robustness against feature misalignment. Last but not least, we focus on fusion architecture by proposing an adaptive fusion design. We learn the trade-off between early and late fusion with respect to the depth quality, yielding a more robust manner to merge RGB-D cues for deep networks. Extensive comparisons on the reference benchmarks validate the effectiveness of our proposed methods compared to other fusion alternatives.

Acknowledgements

I would like to first express my sincere gratitude to my supervisors, Guillaume Allibert, Christophe Stolz, and Cédric Demonceaux, for their help and guidance during this thesis. They are always available for any discussion, scientific or not, at any time. Three-year Ph.D. is a short journey that I clearly remember the first day when I met each of them, in Dijon, Le Creusot, or Antibes. Three-year Ph.D. is also a long journey that we have experienced so many unforgettable moments together, joyous or depressed. All these experiences contribute to this thesis and make this journey unique and memorable. My thanks to all my supervisors are endless, especially to Cédric for his trust and confidence starting from the very beginning.

Life in Bourgogne is less fun compared to big cities such as Paris, especially during the first two years in the small Le Creusot. However, the warm colleagues from the Vibot team changed everything. Fabrice and Lew, with whom I chatted a lot, witnessed my growing up from a newbie to a more "mature" doctoral researcher. There are many Olivier(s) in the lab that indeed frustrated me a lot at the beginning. However, all of them are very kind and welcomed me with great kindness. Christophe, Hamid, and Nathalie, along with the university staff in the IUT, also helped me a lot in dealing with all the materials or administrative work. I also had the possibility to welcome Renato and Carlos as new colleagues on the team. We had delicious lunches and dinners together, as well as the tasty horchata. We are indeed from different cultures, but we get along so well together.

Of course, I will not forget my dearest friends. Marc, Thibault, Daniel, Ahmed, and Theo, all of them made me feel at home in Le Creusot. Yonglin and Yao, who joined the team slightly later, brought our Mandarin into the team! I could have never imagined meeting better friends than you in a delocalized town in France! After two years in the town, I moved to Dijon to join the main site of ImViA. Thanks to Cyrille, Dom, Fan, and all the other friends who took care of me.

Very or even extremely lucky, at the end of the three years, I got the chance to visit CVL, ETH Zurich led by Luc. I was warmly welcomed and hosted by Danda. I cannot forget the long discussions that we had and his patience in helping me even with very tiny details. He also tried his best to make sure of my integration into the BIWI group. During the visiting stay, I had the chance to make many new friends. Among them, I want to express my deepest thanks to Guolei, Shuo, Qin, Dengping, Hao, Yulun, Kai, and Christos who took care of me and guided me so much in all aspects. I would also like to thank Christine, Christina, and Kris who helped me through all the administrative work. A special thanks to Christine who took care of my demand and satisfied my curiosity on the meaning of BIWI - vision science.

Eventually, the visiting stay in Zurich comes to an end quickly, which also means the end of my doctoral study. I would like to express my gratitude to the members of the jury for their kindness in participating in my Ph.D. defense committee. Many thanks to Christian and Nicolas for accepting to be the "rapporteur". I can imagine the quantity of work that they have spent on carefully reading my manuscript. Their constructive remarks contribute to the final version of my doctoral work. I would also thank David and Liming for being the "examinateur" and president of the jury, respectively. Thanks for being present for my defense and for your valuable comments which clarified and opened new doors for my future research.

Unfortunately, my family, my parents and grandparents, could not be present during my defense. I would express my deepest gratitude to them for their support and unrequited love. Finally, I would like to express my earnest love to my girlfriend *Zhuyun* who is always there with me, supports me, and helps me in many aspects. It will not be possible to achieve this thesis without the help of all of them.

To my first, cutest, and most precious pet Babo who left us too early ...

Contents

List of Figures

List of Tables

List of Abbreviations

CNN	**C**onvolutional **N**eural **N**etwork
RGB-D	**R**ed **G**reen **B**lue **D**epth
SOTA	**S**tate **O**f **T**he **A**rt
MLP	**M**ulti-**L**ayer **P**erceptron
NLP	**N**atural-**L**anguage **P**rocessing
GAP	**G**lobal-**A**verage **P**ooling
GMP	**G**lobal-**M**aximum **P**ooling
SOD	**S**alient-**O**bject **D**etection

Chapter 1

Introduction

1.1 Context and Motivation

Computer vision algorithms aim to provide machines with the capability to understand the 3D scene. Traditionally, the input for computer vision algorithms is the 2D RGB images (R for red, G for green, and B for blue). Thanks to the development of image processing algorithms, it becomes possible to detect the contour of the object with Sobel or Canny detectors or to compute the key features with SIFT method. Despite the rich 2D features on the image plane, it is still challenging to explore the 3D information and further include them in image processing.

For human beings, visual perception is realized through binocular vision, i.e., we combine visual information measured from two eyes. Inspired by neurological observation, researchers develop a stereo vision that requires a series of images as input. By comparing features or key points of the same scene from at least two images, 3D information such as depth can be better extracted by analyzing the correspondences of objects from different camera poses.

Despite the plausible results achieved by stereo vision on modeling geometry, the requirement of multiple images of the same scene from the input side limits its popularity compared to monocular images, mainly due to "redundant" data acquisition and data storage. These phenomena can also be noticed by analyzing the size of the existing public datasets. Currently, the largest stereoscopic dataset is the InStereo [33, 86] with 2K images for indoor scenes and Holopix [66] with 50K images for the in-the-wild scenario, while for monocular images, the largest publicly available dataset is the ImageNet [29] which contains more than 14 million images. Therefore, one question is naturally raised: is it possible to leverage geometric cues in a monocular image?

Researchers on RGB-D images and sensors have provided a positive answer to the question. Recently, with the development of 3D sensors such as RGB-D Kinect, Radar, and Lidar, depth images can be obtained from the input side at a more affordable cost. Another approach to obtain the depth is by the mean of monodepth estimation. It is

worth noting that the recent development of large labeled datasets and AI (artificial intelligence) makes it possible to estimate depth from a single image thanks to the significant data prior, which has been regarded as an ill-posed problem for several decades. Indeed, RGB and depth images are complementary to one another. The prior (RGB) contains rich photometric information and is sensitive to color changes, while the latter (depth) contains rich geometric cues and can improve the awareness of scale changes and out-of-the-plane rotation. Taking advantage of both modalities as input, computer vision algorithms can achieve superior performance on scene understanding.

Starting from the year 2011, deep neural networks [59, 133] have brought a revolution in the field of computer vision. Different from early works based on handcraft features, deep networks adopt a gradient-based learning strategy, i.e., backpropagation, to find the optimal parameters for the encoder-decoder architectures. Since then, deep learning methods for computer vision tasks have drawn great attention. Hundreds of thousands of deep networks have been proposed and have almost dominated all the vision tasks, even surpassing human beings in many applications. It is worth noting that most works, especially those milestones such as different VGG [133], ResNet [59], and ViT [32] backbones, are trained and tested with RGB images as input. As discussed in previous paragraphs, this is mainly due to the tremendous visual color data produced daily and the existing large RGB benchmarks for pretraining. Inspired by the development of depth sensors and different estimation methods, this thesis seeks to discover an efficient manner to improve the RGB baseline performance with complementary depth awareness.

1.2 Complementary Modalities to RGB Image

Creating intelligent and effective sensing systems is a major challenge nowadays. Conventionally, most sensing systems are based on a simple RGB camera. One typical example is the regular consumer cameras equipped on most smartphones. After several decades of development, nowadays there exist different types of RGB cameras with all kinds of prices, sizes, and functionalities. Since RGB cameras can provide rich textual information, including all the contours, they can help humans to realize complex tasks together with software algorithms.

However, it is also challenging for such a sensor to operate optimally under all conditions and all the time. For example, similar to the human being, while the lighting condition is unsatisfactory, i.e., during the night, in the tunnel, rainy and foggy, the obtained RGB will be under low-quality. Therefore, RGB cameras can no more provide informative clues on the contours as before. Occupancy is another factor. The

RGB image can only provide textual features of the nearest object while being agnostic of the occluded objects which are hidden beyond the scene.

Consequently, monitoring environmental factors such as weather conditions and occupancy are becoming increasingly important and challenging for scene understanding, especially for autonomous driving, drone, and robots. Therefore, one promising solution is to create perception systems with multiple sensors, especially sensors providing a complementary modality. For example, depth sensors can help to provide the object distance and geometric cues on the object boundaries. Thermal images can contribute to facilitating the scene understanding through specific infrared imaging, yielding a robust manner to deal with low-lightening conditions. Event cameras can provide extreme accurate cues on moving objects and become advantageous for visual perception in dynamic scenes. Among all the sensors, depth modalities are so far the most developed systems together with RGB images, especially in autonomous driving with the additional awareness of the 3D scene.

1.3 Depth Acquisition

There exist three popular sensors to acquire depth information: Radar, Lidar, and depth camera. Radar and LiDAR provide 3D point clouds, while depth cameras provide a 2D depth image. In the following sections, we briefly review the pros and cons of each sensor.

1.3.1 RaDAR

Radar (radio detection and ranging) has been widely used in military applications since this kind of sensor can precisely locate and track the object's position and moving speed. The radar consists of a transmitter and a receiver. The transmitter sends radio waves in a targeted direction, which are further reflected once reach a measurable object. The reflected waves are sent back to the receiver. Based on the returned signal, algorithms are able to provide informative cues about the target object.

Radar technologies have been explored for driving systems by Mercedes-Benz in 1999. Different from a depth camera, a radar system is more sensitive and can provide more information on moving objects. Furthermore, RaDARs can provide objects with severe occlusion, which is not the case with a depth camera. Another advantage of radar is its robustness against unsatisfactory visibility and noise. However, the RaDAR sensor can only provide a limited number of points on the objects, which is significantly less informative compared to the depth sensor in terms of scene understanding.

1.3.2 LiDAR

LiDAR (light detection and ranging) works in a similar way as RaDAR. The main difference is that LiDAR utilizes laser lights, while Radar is based on radio waves. Therefore, in a commercial application such as an autonomous vehicle, LiDAR can see farther objects in the scene, while it is not possible for radar. Nowadays, LiDAR has a detection range of more than 100 meters with extremely high precision.

LiDAR can be regarded as a 3D scan that provides geometric information about the 3D scene. The density of point clouds depends on the number of lasers, also known as LiDAR channels. For example, the commonly used Velodyne-16 sends 16 lasers. To provide accurate 3D information about the environment, the LiDAR sensor requires a real-time computation with hundreds of thousands of points. Therefore, LiDAR sensor requires more computation power compared to camera and radar, which also yields a higher price for LiDAR sensors.

1.3.3 Depth Camera

Recent depth or range cameras often measures object distance by using Time-of-Flight (TOF) remote sensing technologies. The most successful product is the Microsoft Kinect camera. Specifically, they first illuminate the scene and the measured objects with controlled patterns of dots, i.e., infrared light or LED. Then they compute the time that the reflected light takes to travel between the object and the camera. The flight time is directly proportional to the distance between the camera and the measured object. This Time-of-Flight measurement is carried out independently by each pixel of the camera, thus making it possible to obtain a complete 3D image of the measured object. Depth cameras can use in both indoor and outdoor scenes. The price of an effective depth camera is inexpensive, which has drawn great interest for different applications such as drones and industrial robots.

Among the depth image and the point cloud, in this thesis, we are particularly interested in fusing RGB images with depth images. The major reason is that depth images can be treated as 2D data, while point clouds provided by RaDAR or LiDAR are in 3D form. The latter requires more computational cost due to the additional channel. Another reason is that depth images are dense. Each pixel on the image contains a valuable 3D cue. However, while we project the point clouds obtained by RaDAR or LiDAR on the image plane, the obtained map is sparse, yielding several holes. These kinds of images require additional processing such as depth completion to obtain the dense map. Indeed, a single frame RGB-D image sees only the unoccluded objects of the 3D world. However, these objects are the most crucial and important obstacles to analyze for both human beings and intelligent machines.

1.4 Scope and Challenges

RGB-D fusion can be applied to different tasks. During this thesis, we are particularly interested in two main segmentation applications: semantic segmentation and salient object detection which are two of the most basic topics for computer vision.

- **Semantic Segmentation** aims to label each pixel within the input image. It classifies a number of classes and separates each class from the rest. Different from object proposal which outputs a bounding box prediction [52,58], semantic segmentation can detect objects that cover a wide range of areas in the image at a pixel level, making it possible to detect irregularly shaped objects cleanly. Because of this precise detection, semantic segmentation can be applied in a variety of industries that require accurate scene understanding. One typical and popular application of semantic segmentation is for the robot in both indoor and outdoor scenes. For the indoor scene, the cleaning robot and the robotic arm are the two typical applications of semantic segmentation. The former (cleaning robot) needs to identify objects such as the floor and other obstacles in order to find the best path to accomplish the mission, and the latter (robotic arms) requires perfectly localizing the target object for grasping. For the outdoor scene, autonomous vehicles and drones are two other typical applications of semantic segmentation. The objectives of both applications are the same, i.e., the perfect, safe, and autonomous control of robots in a complex and unknown environment. Therefore, accurate perception is highly required to have a robust representation of a complex and unstructured environment for obstacle avoidance.

  Despite the plausible results achieved by deep networks, existing RGB models [9,126,198] are more sensitive to color changes rather than geometric differences, mainly due to the lack of depth input from the input side. Therefore, while dealing with scenes where objects share the same color, state-of-the-art models can fail to accurately separate. Sometimes this can be also challenging for human beings. For example, as shown in Fig 1.1, there exist the bath towel, sofa, wall, cat, and rabbit. While it is easy to separate objects such as bath towels and sofas, however, it is extremely challenging to distinguish the cat and the rabbit due to the same visual appearance. The sub-optimal light condition is also challenging for robotics applications such as autonomous driving [50], especially during the night, rainy, and foggy scenes. Therefore, it is beneficial and essential to profit from other modalities such as depth to improve the performance and robustness of deep neural networks against inferior conditions. The additional depth cues should contribute to better dealing with different scales and generating clearer

boundaries, and even being able to calibrate the RGB image when the visual appearance is sub-optimal.

- **Salient object detection** seeks to segment image contents that visually attract human attention the most. It shares the similar idea of key feature detectors such as SIFT. The main difference is that saliency detection outputs regions of interest, while SIFT predicts pixels of interest. Saliency detection can be regarded as an extreme case of semantic segmentation where there are two labels, i.e., salient and non-salient. Studies have shown that salient objects are always characterized by uniqueness, focus, and objectness, which makes them distinctive from both local and global surroundings. Saliency detection can be applied in various applications such as image cropping, web image filtering, medical image processing, image search, and so on. Recent researches also show that saliency can be coupled with object detection [134] and video object segmentation [155]. Similar to semantic segmentation, RGB salient object models have difficulties performing well under several challenging conditions, such as low-lighting conditions or similar appearance between foreground and background. One way to address these issues is to employ depth cues [37], which are naturally complementary to RGB images with spatial information. Different from semantic segmentation which requires an accurate separation of all objects within the image, saliency detection only focuses on the visually most attractive parts. Therefore, for saliency detection, we are essentially interested in leveraging depth cues within local salient regions instead of all pixels.

Since depth cues can contribute to the scene understanding, how to efficiently explore the geometry along with RGB images has become a vital research topic for RGB-D models. We argue that the basic assumption of RGB-D fusion is that these modalities contain both heterogeneous and homogeneous information. Since RGB and depth images describe almost the same scene (with slight differences in the field of view due to the sensor specification if the depths are acquired rather than estimated), they share similar information at the semantic level. However, there is also a considerable difference between these two modalities. RGB images contain rich visual appearance information such as color and intensity, while depth maps are more sensitive to geometric changes, such as occlusion, scale changes, and out-of-the-plane rotation. Despite the plausible results achieved by recent fusion methods, it is still unknown how to efficiently and effectively fuse RGB-D features. Specifically, it can be noticed that most existing methods process RGB and depth features separately and fuse them through addition or concatenation. Therefore, these methods are agnostic of information redundancy.

FIGURE 1.1: Based on the RGB image, can you segment my Babo and his cat friend? What happens if you have additional cues such as (pseudo) Depth?

Furthermore, RGB and depth may contain unsatisfactory features from the input side. For RGB images, there may exist several local regions with unsatisfactory light conditions or blurred objects. For depth maps, the measurement can be uncertain and inaccurate. When the depths are measured from sensors, the accuracy can be affected by environmental factors such as object distance, object texture, etc. While the depths are computed from stereo images or monocular depth estimation, the accuracy is highly dependent on the quality of estimation methods.

Finally, while the depth is registered from the camera, the RGB-D sensor setups require a full calibration of the 2D-3D systems, such as perfect and ideal extrinsic calibration and timestamp synchronization. However, in practice, these exigences are always hard to achieve. Proposing an efficient and robust fusion method with respect to sensor misalignment has become a vital research topic nowadays.

1.5 Contributions

In this thesis, we seek to propose new fusion designs to address the aforementioned issues. We briefly summarize our major contributions as follow:

1.5.1 Depth-wise channel attention

Existing saliency works often adopt channel attention to emphasize the attentive features for both RGB and depth modalities. However, the vanilla channel attention [64,121,161] is agnostic of fine-grained cues since the first step of channel attention is to squeeze the spatial resolution. Thus, despite the auxiliary depth information, it is still challenging for models equipped with vanilla channel attention to distinguish objects with similar appearances but at distinct camera distances. Therefore, from a new perspective, we propose a granularity-based attention RGB-D saliency detection.

Specifically, we leverage the Otsu thresholding algorithm to first generate various local regions according to the granularity [87, 107]. These regions can be considered as distinct local spatial attention. Then for each region, we apply local channel attention by masking out the others. Therefore, we improve the vanilla channel attention with a better awareness of multi-granularity properties from geometric priors. This approach can be regarded as a depth-wise operation. Similar to depth-wise convolution, we split the input feature into different parts with respect to depth. Then we spatially constrain attention around the different local regions. Finally, we merge them together to form the locally-enhanced output. We extensively validate the effectiveness of the proposed challenging RGB-D benchmarks. Our fusion design can improve saliency detection in several challenging scenarios where the state-of-the-art approaches fail, notably in cases where multiple objects with similar appearances but at distinct camera distances.

1.5.2 Depth-guided spatial attention

We observe that pixels sharing the same semantic label tend to share the same depth similarity, and more specifically, the 3D planarity. Despite the plausible result achieved by deep neural networks, especially the convolutional ones, the fixed size and shape of the convolutional kernel limit its capability to model contextualized awareness according to the geometry [154]. Therefore, we introduce a new convolutional neural network that leverages the depth and planarity priors to deform the sampling positions for basic convolutional operators, i.e., convolution and pooling. Specifically, instead of applying the convolution on the 2D image plane, we first back project the conventional 2D sampling position to the 3D space to create the sampling point cloud with the help of depth information and intrinsic parameters. Among these 3D points, we use the mean square least method to output the estimated plane coefficients. Based on these coefficients, we generate a depth-adapted planar grid, whose projection on the 2D image forms the depth-guided deformable sampling position. This deformation plays the role of local depth attention to improve the discriminability of RGB features. We demonstrate through two tasks, i.e., semantic segmentation and saliency detection, the generalization capability of such fusion design. Compared to other RGB-D fusion alternatives [65, 154, 176], we show that depth as offset can better leverage the geometric cues to improve the baseline performance.

1.5.3 Cross-modal transformer attention

Recently, transformer networks have led to another revolution in the computer vision society. Initially designed for NLP (natural language processing) tasks, the transformer has shown its capability in modeling long-range dependencies to process contextualized awareness for input sequences. Starting from 2020, transformer attention has been applied in various vision tasks and rapidly superior performance compared to CNN networks in various applications [17, 32, 98, 218]. Compared to convolution, the transformer is built upon global attention with inter key-query correlation. We observe that by extending the inter key-query correlation to cross-modal key-query correlation, transformer attention suggests a natural way to aggregate RGB-D features. Inspired by this observation, we propose to first extract modality-specific features and then aggregate them through transformer attention. Our key idea is to leverage transformer attention to improve the scene understanding with enhanced awareness of visual differences and geometric cues, respectively. To enable position awareness and leverage locality into our transformer fusion, we propose a semantic-aware position encoding generator built upon convolutions. We process a modality-specific sequence as input and generate a category-aware position encoding. We aim to spatially constrain the attention around the neighboring area to better segment objects. Extensive comparisons on RGB-D indoor benchmark datasets have shown the superior performance and robustness of our network compared to pixel-wise fusion counterparts.

1.5.4 Layer-Wise Attention for RGB-D fusion

Most existing RGB-D fusion works can be roughly grouped into categories: early fusion, middle fusion, and late fusion. Early fusion methods concatenate RGB images and depth maps from the input side and process the mixed RGB-D features through deep networks. Late fusion methods always first extract RGB and depth features separately through parallel encoders and then merge them at the semantic level. Different from late fusion, middle fusion works merge RGB and depth cues at each level during feature extraction to form multi-scale shared features. Despite the plausible results achieved by previous fusion works, existing works often require a fixed handcraft design, which cannot be adapted to different inputs. To address this issue, we seek to design an adaptive fusion network that can automatically switch from different fusion designs according to the inputs. Our intuition is that good quality depth should contain rich geometric or low-level features which correlate well with stemming layers of a deep network. Therefore, in such a case, the early fusion is more suitable to merge multi-modal features. However, while the depth map is unsatisfactory, it becomes hard to explore the low-level features. Therefore, a middle or late fusion should be preferred so that RGB-D cues are more merged at deep semantic space

instead of stemming layers. To achieve such a goal, we propose layer-wise attention which learns the trade-off between early and late fusions, depending upon the provided depth quality. We show through the RGB-D saliency task that such fusion avoids the negative influence of the spurious depths while being opportunistic when high-quality depths are provided. We expect to validate the effectiveness and the generalization capability of such design for other tasks, such as semantic segmentation and object detection, in a similar setting of RGB-D inputs.

1.6 Organization

This thesis dissertation is divided into six different chapters:

- Chapter 2 introduces a short history of the most related CNN, attention, and transformer works in computer vision. We also present different depth representations including both 3D and 2D data. We highlight the advantages of RGB and depth fusion on the 2D image and provide an overview of RGB-D fusion milestones.
- Chapter 3 discusses the integration of fine-grained details into the vanilla channel attention to form the granularity-aware attention. Additionally, we explore the different cross-modal attention fusion designs for saliency detection.
- Chapter 4 explores how to use the depth to deform the RGB sampling position to be adapted to the perspective effect. We show that the depth as offset, in other words as local spatial attention, can significantly improve the baseline performance for different vision tasks.
- Chapter 5 shows how to leverage the transformer attention for RGB-D fusion. We show that transformer attention is more robust to feature misalignment compared to pixel-wise hard associations. Furthermore, we introduce a novel learnable positional encoding that is modality-specific and can leverage rich spatial cues from hierarchical features.
- Chapter 6 presents the layer-wise attention for RGB-D fusion with respect to the input depth quality. We also improve conventional spatial attention with superior robustness against feature misalignment. The proposed mechanisms allow us to efficiently exploit the multi-model inputs while being robust against low-quality depths.
- Chapter 7 concludes this thesis and discusses some future perspectives on the presented works.

Chapter 2

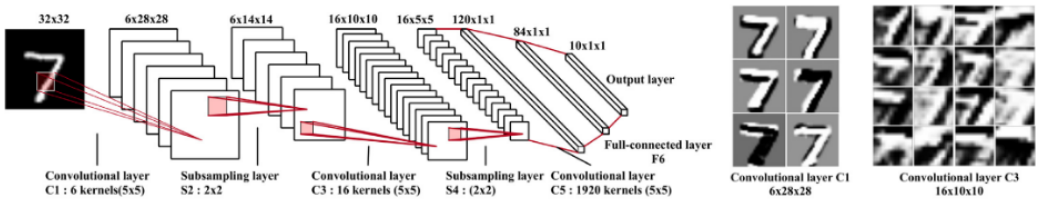
Development of Deep Neural Networks: a Brief Literature Review

Deep learning is a very large topic and it is quasi-impossible to review all the details such as convolution, pooling, relu, batch normalization, dropout, MLP, etc. It is also impossible to view all different data augmentations, weight initialization, loss fusion, gradient descent, etc. In this manuscript, we assume that the readers have already gained basic notions of deep learning. Therefore, in this section, only several milestones in the deep learning area will be briefly reviewed, including CNN backbones, attention modules, and transformer networks. The objective is to provide an overview of the deep network's history, with a zoom on the development of different attention models. Also, note that we do not introduce the comparison with our proposed approaches. Detailed comparisons can be found in each of our proposed methods.

2.1 Convolution Neural Network

When we talk about deep neural networks, convolutional ones are just unavoidable. Initially designed for image classification, convolutional neural networks (CNN) have dominated the computer vision society for almost ten years. The final objective of CNN is to extract the characteristics of each image by compressing them with different layers of convolution. The input image passes through a succession of filters and creates a new matrix with a smaller resolution but with a higher channel dimension. These new matrices, also known as feature maps, contain therefore more semantic cues of the images and can contribute to the scene understanding.

The basic operator of CNN is convolution, which is a simple mathematical operation widely used for image processing and recognition. The basic idea of convolution is to add neighboring pixels to each element of the image, weighted by the kernel elements. A classical convolution can be denoted as:



(a) LeNet-5 network (b) Learned features

FIGURE 2.1: Illustration of LeNet-5 network, a convolutional neural network at an early stage. The image is from [53].

$$\mathbf{y}(p) = \sum_{p_n \in \mathbf{R(p)}} \mathbf{w}(p_n) \cdot \mathbf{x}(p + p_n), \tag{2.1}$$

where $\mathbf{w}$ is the weight matrix. $\mathbf{R(p)}$ is the grid for point p. Physically it represents a local neighborhood on an input feature map, which conventionally has a regular shape with certain dilation 1, such that :

$$\mathbf{R(p)} = a\vec{u} + b\vec{v} \tag{2.2}$$

where $(\vec{u}, \vec{v})$ is the pixel coordinate system of input feature map and $(a, b) \in (\Delta d \cdot \{-1, 0, 1\})^2$. In image processing, different forms of convolution kernel have been proposed and have shown great advances in image blurring, sharpening, edge detection, and others. These kernels have pre-defined handcraft weights such as identity matrix, gaussian filters, Sobel kernels, Canny kernels, etc.

In the deep learning area, convolution also plays an important role. Different from the previous pre-defined weights, deep learning aims to find the optimal weights for convolution kernel by using gradient descent. Gradient descent is an optimization algorithm that finds the minimum of any convex function by gradually converging towards the minimum. For example, for supervised learning where the ground truth is known, gradient descent is used to minimize the cost function, which is indeed a convex function (for example the mean squared error). Since in the thesis, our objective is NOT to propose novel optimization methods, in the following paragraphs we simply and briefly explain the main logic of gradient descent for background understanding.

The first step of gradient descent is to start from a random initial value (a random kernel weight) and then we measure the value of the slope with this initialization. The slope in mathematics is computed as the derivative of the loss function.

Once we obtain the derivative, the next step is to define how much we progress in the direction of the slope which descends. This distance is termed Learning Rate, which could be translated as learning speed. This operation results in modifying the value of the parameters (kernel weights) of our model.

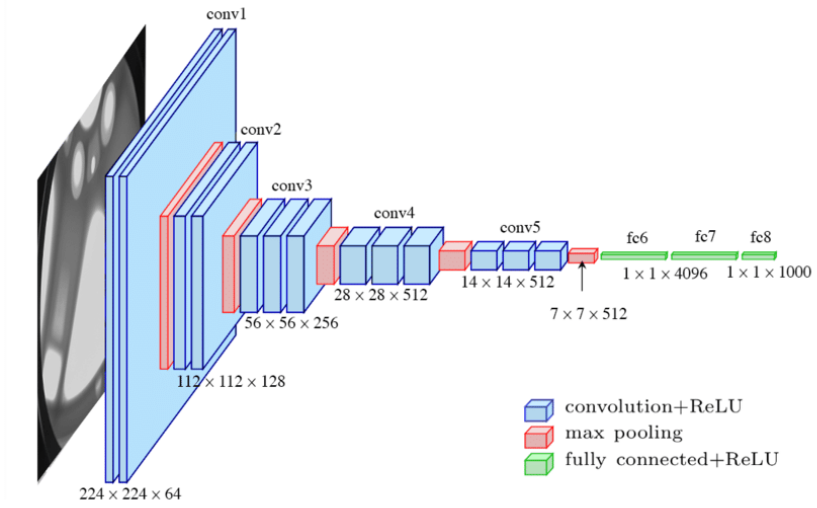


FIGURE 2.2: Illustration of VGG-16 networks which extract the features from the input and class the image into a class. The image is from [105].

By repeating these two steps in a loop, the gradient descent is therefore an iterative algorithm that makes it possible to find the ideal value for the learned convolutions. It is worth noting that a large value of the learning rate can contribute to the fast convergence of the learning models. However, the final results may not be optimum since several local minima can be missed. On the other hand, a small learning rate may be stacked in local minima, while being agnostic to the global minima. How to find the best learning rate is yet an open question for researchers in the field.

Once the basic convolution operator is unveiled, it becomes easier to better understand a deep convolutional neural network (CNN) which can be regarded as a simple combination of different layers of convolution as shown in Figure 2.1. In this section, we briefly review several milestones that are related to the thesis.

In 2014, VGG nets [133] have been proposed which can be regarded as the first very deep Convolutional Networks. This work shows that the depth of neural networks can significantly affect the performance and comes up with two models: VGG-16 and VGG-19, where 16 and 19 stand for the number of convolutional layers contained in each model. An example of VGG-16 is shown in Figure 2.2. Different from previous works, VGG nets replace large-size kernels with a combination of 3×3 kernels one after the other. This design contributes to reducing the model complexity for convolution and makes a deep network possible with respect to a limited GPU size. However, despite the uniform design for different layers and the appealing performance, VGG nets are heavy and time-consuming during training, i.e., the model size of VGG-16 is around 533Mb. Additionally, VGG nets have shown that while the number of layers increases, the performance of the model also increases. However, in practice, a simple

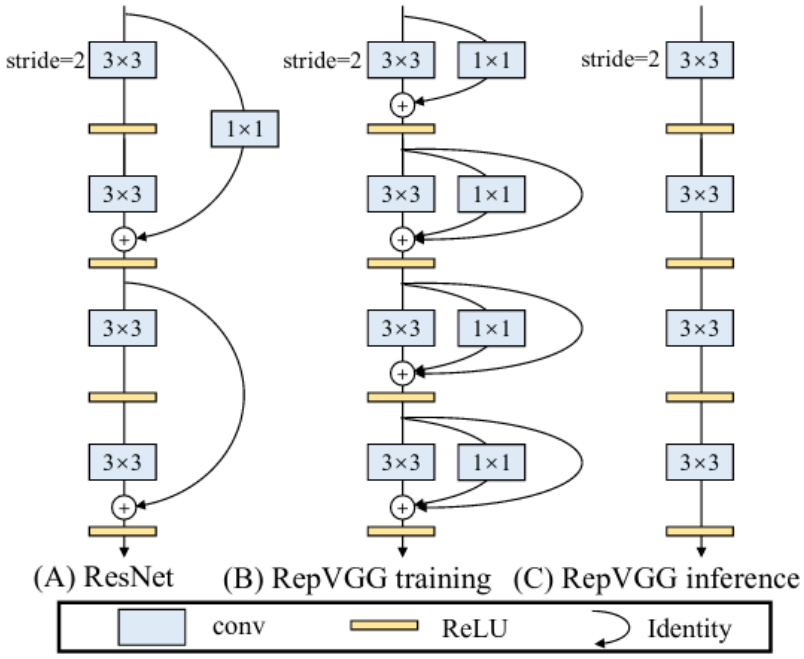


FIGURE 2.3: VGG nets with the residual connection. The image is from [31].

increase in layers does not necessarily lead to better performance. This can be also figured out by the fact that there only exist VGG 16 and 19 but not VGG 50 or VGG 100.

In fact, one of the main ideas of deep neural networks is based on back-propagation. A neural network uses a loss function to express the precision in each node. The back-propagation uses gradient descent by the chain rule which computes the gradient of the loss function one layer at a time with respect to the other weights in the network. This happens in reverse through the neural network, hence named the back-propagation. After the process, nodes with a high error will have less weight than those with a lower error. The back-propagation or the gradient descent will pass through all the nodes from the end to the start. Therefore, while the network is too deep, in the case of VGG nets, the back-propagation will become minimal, yielding a gradient vanishing and making weight changes at the stemming layers very small.

To address this issue, in 2016 ResNet [59] showed that a simple identity function can perfectly deal with the gradient vanishing since the local gradient becomes 1 instead of 0. Therefore, even with networks with a significant number of layers, the gradient can be back-propagated without decreasing in value. With the help of the identity function or the residual connection, ResNet comes up with different variations with more convolution layers such as ResNet-50, ResNet-101, and ResNet-152. After ResNet, a number of other works have been published in order to propose a powerful CNN backbone such as Res2Net [47], ResNext [174] etc. A recent work

RepVGG [31] inspires by the success of ResNet and, as shown in Figure 2.3, proposes a new architecture with residual connection to make VGG networks great again.

2.2 Self Attention

In the field of deep learning, the development of CNN has shown great achievement in various computer vision tasks such as object detection and semantic segmentation. These applications are usually built on top of CNN backbones. In previous sections, we have briefly reviewed several models such as VGG and ResNet. The effectiveness of such backbones has been fully verified and is widely used in various computer vision tasks. It can be seen that a deep neural network always contains a contracting path and an expansive path. The contracting path, also known as the encoder, extracts high-level features from the input image. For CNN networks, the contracting path consists of different layers of convolutions with respect to different backbones, e.g., VGG and ResNet. The core computation is the convolution operator, which learns feature maps from the input feature map through the convolution kernel. Essentially, convolution can be regarded as a feature fusion of a local region, which includes spatial and inter-channel feature fusion. Formally, let an input image I with size $I \in \mathbb{R}^{C \times H \times W}$, the contracting path outputs a high-level feature map x with size $x \in \mathbb{R}^{c \times h \times w}$ by jointly fusing spatial and channel cues within the convolutional kernel. As suggested in previous works, features in the stemming layers retain higher spatial resolution, while features in the latter layers have a smaller spatial resolution but retain more semantic details:

$$c > C, \quad h < H, \quad w < W \tag{2.3}$$

For the convolution operation, a large part of the work is to improve the receptive field, that is, to integrate more feature fusion in space, or to extract multi-scale spatial information. For feature fusion along channel direction, the convolution operation basically fuses all channels of the input feature map by default. However, during the development, one question is naturally raised: are all the spatial and channel information important? Specifically, taking the feature map x as example, $x \in \mathbb{R}^{c \times h \times w}$. Should each pixel contribute equally to the output when $h \times w$ is significant, i.e., 1920×1080? Similarly, should each channel contribute equally to the output when c is significant, i.e., 2048 for ResNet-101? To tackle these issues, different self-attention modules, especially channel and spatial attention modules, have been proposed. In the following paragraphs, we review the several milestones for self-attention modules.

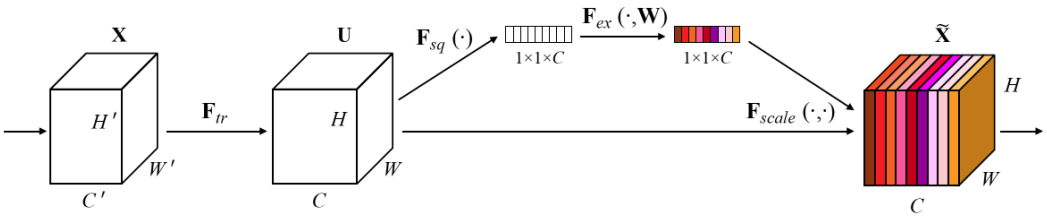


FIGURE 2.4: Illustration of SENet. The image is from [64]

2.2.1 Channel Attention

2.2.1.1 Squeeze and Excitation Network

Squeeze and Excitation Network (SENet) [64] is the pioneering work for channel attention. The idea of SENet is to pay attention to the relationship between channels, hoping that the model can automatically learn the importance of different channel features. To this end, SENet proposes the Squeeze-and-Excitation (SE) module, as shown in Figure 2.4:

Specifically, let an input feature map $x \in \mathbb{R}^{c \times h \times w}$, SE block first adopts the Squeeze step which computes the average presentation for each channel with the help of average pooling which squeezes the spatial dimension, forming a vector v with size $v \in \mathbb{R}^{c \times 1 \times 1}$. This vector is then fed into a multi-layer perceptron (MLP) to compute an attention map. The principle of this structure is to enhance the important features and weaken the unimportant features by controlling the size of the scale so that the extracted features can focus solely on channel cues rather than spatial cues. Finally, the attention map is multiplied with the feature map x, yielding a highlighted presentation with enhanced channel attention.

It is worth noting that there are many algorithms for computing a global representation for the Squeeze step. SENet uses the simplest averaging method which averages the information of all pixels into one value. This choice is made because the final re-calibration weight is applied to the entire channel, and the spatial representation must be computed based on the overall information of the channel. In addition, SENet aims to study the correlation between channels instead of the spatial distribution. Therefore, global average pooling can mask the spatial distribution information, yielding a more accurate re-calibration weight for channels.

The excitation part is implemented with 2 MLP. The first MLP compresses c channels into c/r channels to reduce the amount of computation. The second MLP restores reduced features back to c channels. r refers to the compression ratio for the channel dimension. The MLP aims to exploit the correlation between channels to learn a meaningful re-calibration weight. In fact, the squeeze output of a mini-batch sample cannot be directly used as the re-calibration weight since the last should be trained

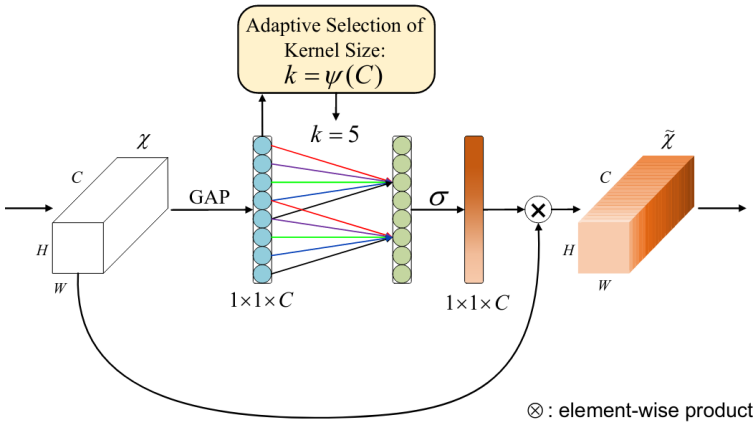


FIGURE 2.5: Illustration of details on efficient channel attention. The image is from [121].

based on the entire data set instead of based on a single batch. Therefore, an MLP is necessary to learn a global presentation.

Following SEnet [64], several studies have been proposed which aim to improve SE blocks by capturing more complex and more informative channel dependencies or even incorporating additional spatial attention. Although these methods achieve higher performance, they often bring extra and heavy computational costs over the plain version. In 2020, researchers propose a novel and effective channel attention module, ECAnet, with limited additional learning parameters.

2.2.1.2 Efficient Channel Attention

The original SENet is built upon 2 MLP and then uses a sigmoid function to generate the channel weight given the input features. The two MLP layers are designed to capture nonlinear cross-channel interactions, which include dimension reduction to control model complexity. Although this strategy has been widely used in other channel attention modules, the authors of SEnet experimentally show that reduction also brings side effects to channel attention prediction, and makes the captured dependencies between all channels inefficient. Therefore, as shown in Figure 2.5, researchers propose an efficient channel attention module for deep CNNs, named ECAnet [121] standing for efficient channel attention, which avoids dimension reduction and effectively captures the information of cross-channel interactions.

Specifically, after channel-level global average pooling without dimension reduction, ECA captures local cross-channel interaction information by considering each channel and its k neighboring channels, which ensures both the model efficiency and computation cost. To achieve this goal, ECA applies 1D convolution of size k, where the convolution kernel size k represents the coverage of local cross-channel interactions,

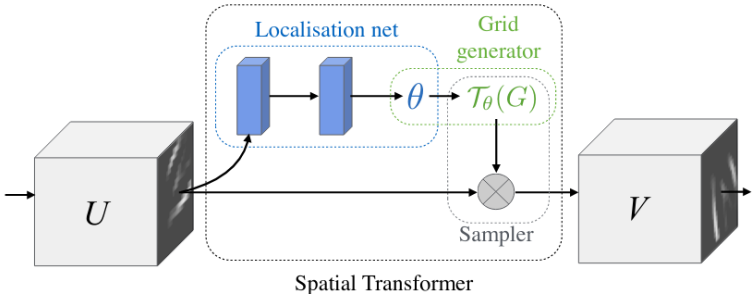


FIGURE 2.6: Illustration of details on spatial transformer network. The image is from [68].

i.e., how many neighboring channels near the target channel participate in the channel's coverage. This kernel size k is proportional to the channel dimension. Compared with backbone models, deep CNNs with ECA modules (called ECA-net) introduce few extra parameters and almost negligible computation, while bringing performance gains.

Apart from these two widely used channel attention, there are a lot of other works aiming to improve the squeeze part and/or the excitation part as shown in Figure 2.5. We encourage readers to refer to the survey [54] for more details on the channel attention modules.

In addition to channel dimension, spatial resolution is also important for computer vision tasks, especially for object localization. Therefore, in the following section, we will review another type of attention work, i.e., spatial attention, to better enhance the response at the pixel-level.

2.2.2 Spatial Attention

2.2.2.1 Spatial Transformer Network

While channel cues are important for semantic understanding, spatial cues can contribute to precise object location. In fact, for computer vision tasks, we hope that the deep network can achieve a certain invariance to the changes in object pose or position. Therefore, we can learn the deep model from a limited number of images and generalize the knowledge to other scenarios where objects are in different poses and positions. The traditional CNN uses convolution and pooling operations to achieve translation invariance at a certain level. However, this invariance is only true at the image plane, i.e., objects translating along the image axis. While the objects translate at the direction of depth, i.e., the normal of the image, the previous invariance is no more validate. Therefore, conventional CNN is not invariant to geometric transformations such as rotation, distortion, scale changes, etc.

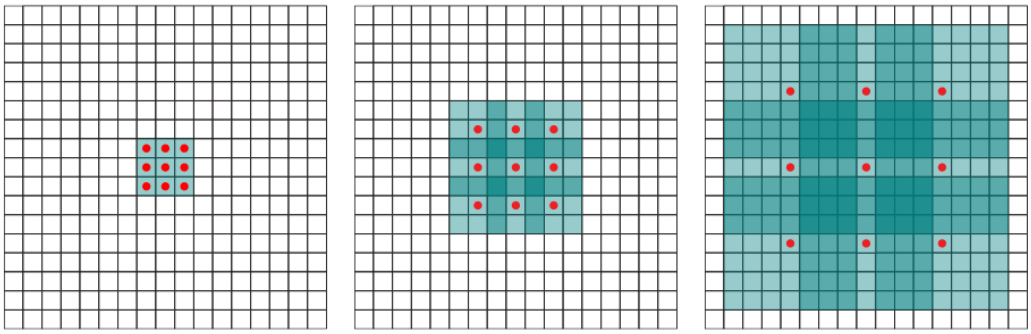


FIGURE 2.7: Illustration of details on dilated convolution. The image is from [184].

Therefore, a number of works tackle this axis and propose different approaches aiming to enhance spatial awareness to better handle the geometric changes. At an early stage, STN [68] attempts to learn the spatial transformations from the input image. The derivable STN does not need redundant annotations, and can adaptively learn the spatial transformation methods for different data. It can not only perform a spatial transformation on the input but also can be inserted into any layer of the existing network as a complementary module to realize the spatial transformation of different feature maps. Finally, the network model learns invariance to translation, scale transformation, rotation, and more common distortions, which also makes the model perform better on many benchmark datasets.

As shown in Figure 2.6, each ST module consists of a Localization net, Grid generator, and Sample. The localization net determines the parameter θ which stands for the transformation required by the input image/feature. Grid generator aims to find the mapping matrix $\mathcal{T}(\theta)$ between output and input features through θ. Finally, the Sampler applies the mapping matrix to the input features. For more details, we refer readers to the original paper [68].

2.2.3 Dilated Convolution

While STN adds an additional module to explicitly model the transformation, another research direction is to implicitly integrate the spatial awareness into the convolution operation, e.g., the pioneering work Dilated Convolution. Dilated/Atrous Convolution, by name, injects holes into the standard convolution map to increase the reception field. Compared with the original normal convolution, the dilated convolution has one more hyper-parameter called dilation rate, which refers to the number of intervals of the kernel (e.g. normal convolution is dilatation rate 1) as shown in Figure 2.7. With the help of an increased receptive field, the neural network can yield better invariance to the geometric changes within the enlarged sampling position.

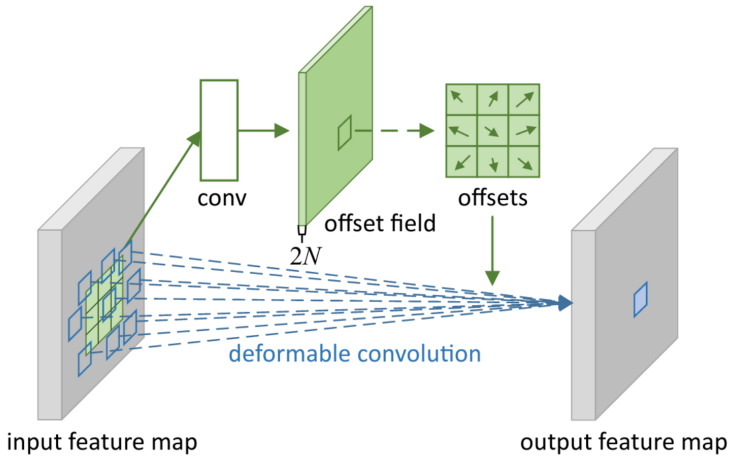


FIGURE 2.8: Illustration of details on deformable convolution. The image is from [28].

Following the original dilated convolution, there exist several improved versions to adjust the limitation of the original work. For example, from the enlarged dilation, it can be seen that this design is used to obtain long-ranged information. However, this design is more suitable for the segmentation of large objects and may yield an unsatisfactory result for small objects. How to deal with the relationship between objects of different sizes at the same time is the key to designing a dilated convolution network. There exist several works which aim to address the above-mentioned issue and can be found in [54].

2.2.4 Deformable Convolution

Sharing the same idea as Dilated convolution, Deformable Convolution [28] is proposed in 2017 which can be regarded as a more general representation of convolution. It proposes two new modules to improve the deformation modeling capabilities of CNNs, called "deformable convolution" and "deformable ROI pooling", both of which are based on adding extra offsets in the module for spatial sampling positions. These offsets are learned during the training and do not require additional supervision as shown in Figure 2.8. These new modules can easily replace common modules of existing CNNs and use backpropagation for end-to-end training, resulting in deformable convolutional neural networks. Formally, the deformable convolution can be formulated as follow:

$$\mathbf{y}(p) = \sum_{p_n \in \mathbf{R(p)}} \mathbf{w}(p_n) \cdot \mathbf{x}(p + p_n + \Delta p_n) \tag{2.4}$$

The original method has shown great performance on semantic segmentation and object detection and is widely applied in different tasks such as saliency detection, tracking, etc.

2.2.5 Non-local Attention

Through recent research, it can be seen that one trend of spatial attention is to better leverage contextualized information, i.e., dilated convolution with a large squared receptive field or deformable convolution with a more malleable form. Recently, non-local attention has drawn great research interest in the computer vision community. In deep neural networks, capturing long-range dependencies is crucial. For sequential data (e.g. speech, language), loop operations are the main solution for long-range dependency modeling. For image data, conventionally long-range dependencies are modeled by stacking deep convolutions to form large receptive fields.

Both convolution and loop operations deal with local neighborhoods in space or time. Therefore, long-range dependencies can only be captured when these operations are repeated, i.e, gradually propagating signals in the text/image. One severe limitation of such a repeating process is, e.g., computationally inefficient. Therefore, in 2018, researchers propose non-local [156] operations as an efficient, simple, and general component for capturing long-range dependencies in deep neural networks as shown in Figure 2.9. The proposed non-local operation is a generalization of the classical non-local mean operation in computer vision. Intuitively, the non-local operation computes the response of a position as a weighted sum of features at all positions in the input feature map.

According to the original paper, using non-local operations has several advantages. Firstly, non-local operations compute the correlation between any two pixels/position within the feature map, regardless of their euclidean distance. This makes non-local attention more suitable and powerful to model long-range dependencies compared to previous work. Secondly, through the empirical result, with only a few layers (e.g. 5 layers), non-local operations are efficient and achieve significantly better performance compared to the baseline. Finally, the non-local operation can be easily combined with any existing deep neural network.

It is worth noting that non-local attention [156] is the first tentative to apply global attention in computer vision tasks. It is similar to the self-attention or transformer module for machine translation [147]. However, it focuses on 2D input, i.e., image. Another major difference is that non-local attention is more like single-head attention which serves as a complement module for the CNN backbone, while the original transformer is multi-head attention and serves as the backbone to extract the long-range

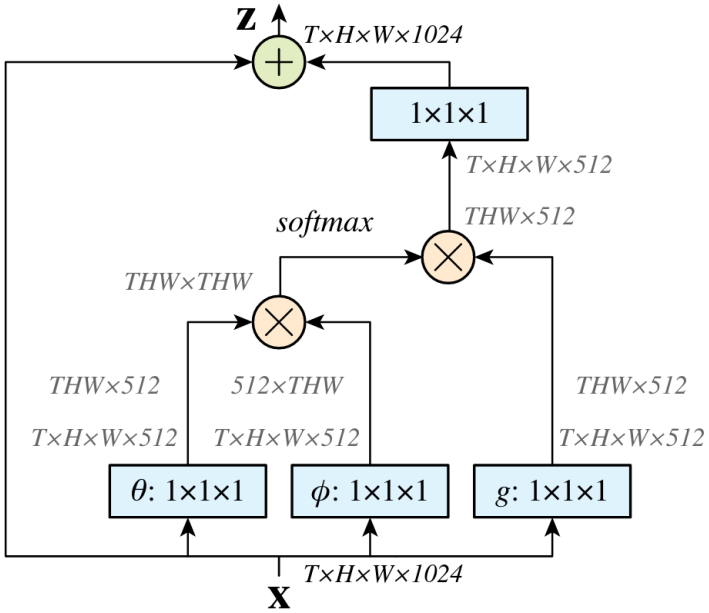


FIGURE 2.9: Illustration of details on non-local attention. The image is from [156].

dependencies. Inspired by the success of transformers in NLP, researchers designed a transformer backbone for computer vision tasks, termed ViT, which will be discussed more in detail in the following section.

2.2.6 Transformer

2.2.6.1 Self and Cross Attention

Following the idea of self-attention, in 2020 ViT [32] is proposed to treat an image as a series 16×16 words. In other words, ViT first applies transformer attention in computer vision tasks, which has challenged the CNN empire.

Initially, transformer attention [147] is designed for NLP (natural language processing) tasks. The basic idea is to compute the correlation between a given query (e.g., a target word in the output sentence) and certain key elements (e.g., source words in the input sentence). The correlation between query and key elements provides an attention map that prioritizes the most important words in the sentences. Where both key and query elements are from the same source, it is called self-attention which analyzes the intra-sentence relations. While the elements are from different sources, it is called cross-attention which analyzes the relations between different sentences. Figure 2.10 shows the sketch of transformer attention.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{2.5}$$

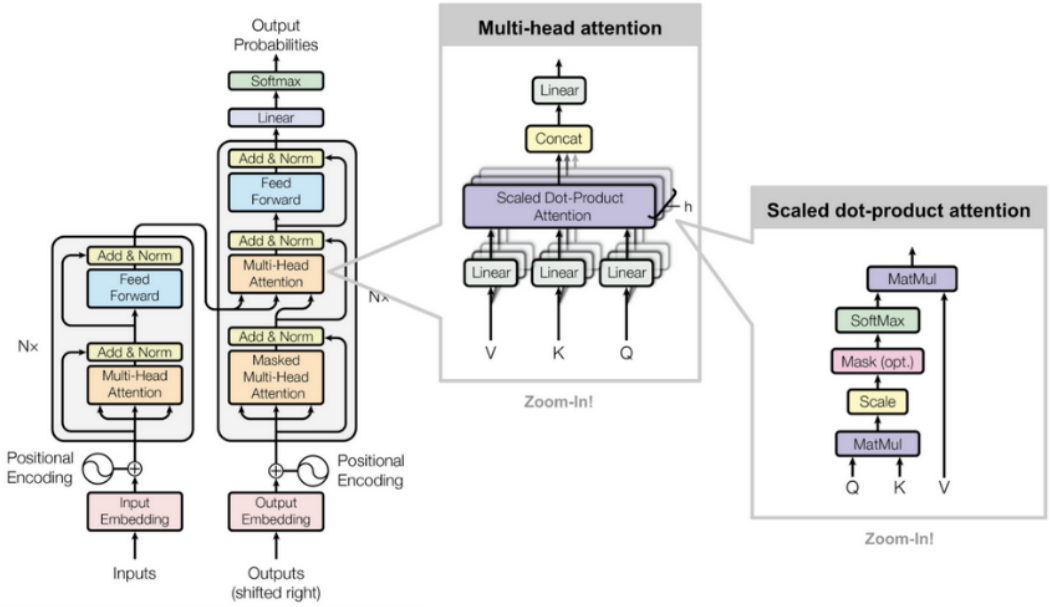


FIGURE 2.10: Illustration of a transformer attention. The image is from online source[1] basing on [147].

If we take a deeper look at the design, the non-local operation is actually a variant of self-attention, which can be understood as a small part of the Transformer structure. Therefore, from this perspective, there is no essential difference between the two, and they both establish long-distance dependencies. From a functional point of view, non-local attention or self-attention is equivalent to the encoder part of the Transformer, which only does feature extraction. However, in addition to self-attention, the transformer also includes the decoder part, i.e., cross-attention, which can be used for reasoning. Therefore, the Non-local algorithm does not jump out of the standard practice of CNN, while the Transformer can directly predict the result, and the entire pipeline is more refreshing and clean, i.e., without various decoder modules.

Nowadays, transformers [32, 98, 147] have achieved leading performance in various vision tasks. It is worth noting that the transformer is initially designed to process 1D signals, i.e., languages which can be extremely long. Therefore, capturing long-range dependencies is the most crucial part of NLP models. However, for CV tasks, images are the most common input. While understanding an image, the local correlation is also important since one pixel is always linked to neighboring pixels. This is also the reason why CNN has achieved great performance since its birth. Hence, directly extending transformer models to CV applications may yield sub-optimal performance. While this bottleneck can be solved with a large training dataset, reducing

[1] https://lilianweng.github.io/posts/2020-04-07-the-transformer-family/

the training/inference time remains an imperative topic for both research and industrial applications. Furthermore, adding the locality-awareness is also important for transformer networks [24, 84, 97, 181].

2.2.6.2 Complexity Reduction

Despite the efficiency and great performance of transformers in vision tasks, the computational cost is O(N2) is also an ineligible issue for deep neural networks, especially in cases where the input data is a high-resolution image. In fact, during the initialization stage of transformer attention, the weights are assigned to all feature pixels and they are almost equal. This means that the network needs to learn what are the most meaningful/informative locations through the image, and these locations should be sparse instead of dense. What's more, an original transformer requires a high computational cost when calculating attention weights since the network needs to compute the correlation between one pixel (query) and all others (keys). This leads to a quadratic relationship with the number of feature pixels. Therefore, it is difficult to apply Transformers to high-resolution features.

To tackle these issues, recent works aim to explore new forms of Transformers to reduce the computational cost. In this section, we briefly review several milestone works which are related to this thesis.

Deformable DETR One idea is to select a set of keys instead of the whole feature map. In other words, since we want to learn sparse spatial positions, why not use the deformable convolution set? However, deformable convolution also lacks relational modeling capabilities. Nevertheless, this is what the transformer attention is best at. Hence, in 2021, researchers propose Deformable DETR [218], which contains the advantages of both deformable convolution and transformer. Specifically, for each query, before focusing on all spatial positions (all positions are used as keys), Deformable DETR only focuses on more meaningful positions that the network considers to contain more local information (less and a fixed number of positions are used as keys), alleviating the problem of large feature maps. Technically, during the implementation process, the input feature map is fed to a linear map and outputs 3 features. The first 2 features encode the offset of the sampling and determine which keys should be found for each query, and the last feature contains the contribution of keys. Instead of computing the correlation between key and value, Deformable DETR only normalizes the contribution of the found keys, yielding a significantly lighter computational cost compared to the original form (10x faster compared to DETR).

Swin Transformer Another group of works aims to constrain the global attention to a series of local attention and finally merge them. One of the most exciting works must

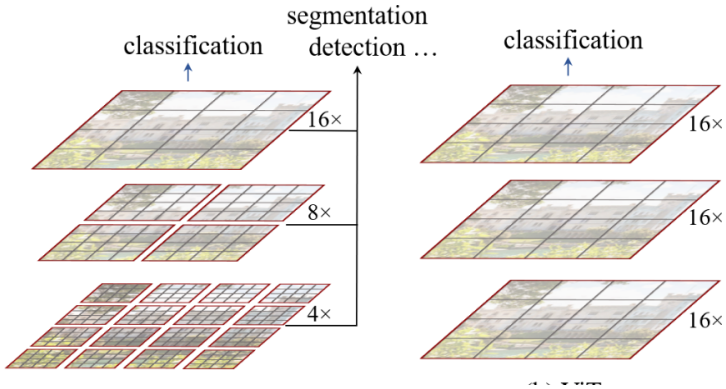


FIGURE 2.11: Comparison between Swin Transformer (left) and original ViT (right). The image is from [98].

be the best paper of ICCV 2021, i.e., Swin Transformer [98]. The biggest contribution of Swin Transformer is to propose a backbone that can be widely used in all computer vision fields, and most of the common hyperparameters in CNN networks can also be manually adjusted in Swin Transformer, such as the number of network blocks, the number of layers, the size of the input image, etc.

Before the Swin Transformer, networks such as ViT used a small and fixed-sized image as input, i.e., 224 for ViT. Therefore, for an input image, the first step is to resize the resolution to fit the requirement. This strategy will undoubtedly lose a lot of information. Unlike previous works, the input of the Swin Transformer is the original size of the image. In addition, Swin Transformer uses the most commonly used hierarchical network structure as in CNN. It is worth noting that the receptive field of Swin Transformer is similar to a CNN network: while the network level deepens, the receptive field of nodes is also expanding. The hierarchical structure of Swin Transformer also gives it the ability to perform segmentation or detection tasks with structures such as FPN [89] and U-Net [126].

The details of the window attention can be found in Figure 2.11. The basic idea of Swin Transformer is to constrain the self-attention to local windows. Therefore, the number of relative queries is limited to a fixed receptive field (Window), which can reduce the amount of calculation and introduce locality prior. Another important design of Swin Transformer is the shifted window. Unlike traditional sliding windows, the design of non-overlapping windows is more friendly to hardware implementation, resulting in faster actual running speed. As shown in Figure 2.11, in the sliding window design, different points use different neighborhood windows to calculate the relationship, which is not hardware friendly. In the non-overlapping windows used by Swin Transformer, the points in the unified window will use the same neighborhood for calculation, which is more speed-friendly. Practical tests show that the non-overlapping

window method is about 2 times faster than the sliding window method. Another meaning is that the shift operation is performed in two consecutive layers. In the L layer, the window partition starts from the upper left corner of the image, and in the L+1 layer, the window partition moves half a window to the lower right. This design ensures that there can be information exchange between non-overlapping windows.

2.2.6.3 Conditional Positional Encoding: CPVT

For transformer attention, since the self-attention operation is permutation-invariant, positional encodings (PE) are required to explicitly encode the positional information of the tokens in the sequence. The ViT model uses the learned fixed-size positional embedding, but when the image input size changes, the positional embedding needs to be interpolated to adapt to the change in the number of input tokens, which will result in performance loss. Therefore, researchers propose CPVT [24] to learn and constrain the positional encoding by convolution layers. The solution of CPVT is to introduce a convolution with zero-padding to implicitly encode the positional information (PEG), thereby eliminating the need for explicit positional embedding. The key point is that the CPVT model can adaptively fit with the spatial resolution.

2.2.6.4 Joint Convolution and Transformer: ACmix

Recent research has shown that Transformers are limited to model local awareness, which is in the meantime the strength of convolutional networks. One recent research direction is therefore combining transformer with CNN [98, 108]. Swin Transformer has shown a great example by introducing several local properties to transformer design. Different from Swin Transformer, another research direction [108] aims to explicitly joint the advantages of CNN and transformer together.

We have discussed in the previous sections that convolution and transformer have similar computational mechanisms. In previous works, convolution and self-attention are two powerful techniques for representation learning and are often considered as two different mechanisms. In the CVPR22 paper ACmix [108], the authors demonstrate that most of the computations in both paradigms are actually realized by the same operations, demonstrating a strong intrinsic relationship between them. Specifically, the author splits both convolution and self-attention into two stages. In the convolution operation, a traditional convolution with a kernel size of $k \times k$ can be decomposed into $k \times k$ individual 1×1 convolutions. As for the self-attention module, 1×1 convolutions are used to generate query, key, and value. Then the attention weights are computed following the conventional self-attention. By doing so, the first stage of both convolution and self-attention contains similar operations. This makes it possible to combine these two seemingly different paradigms and form the proposed

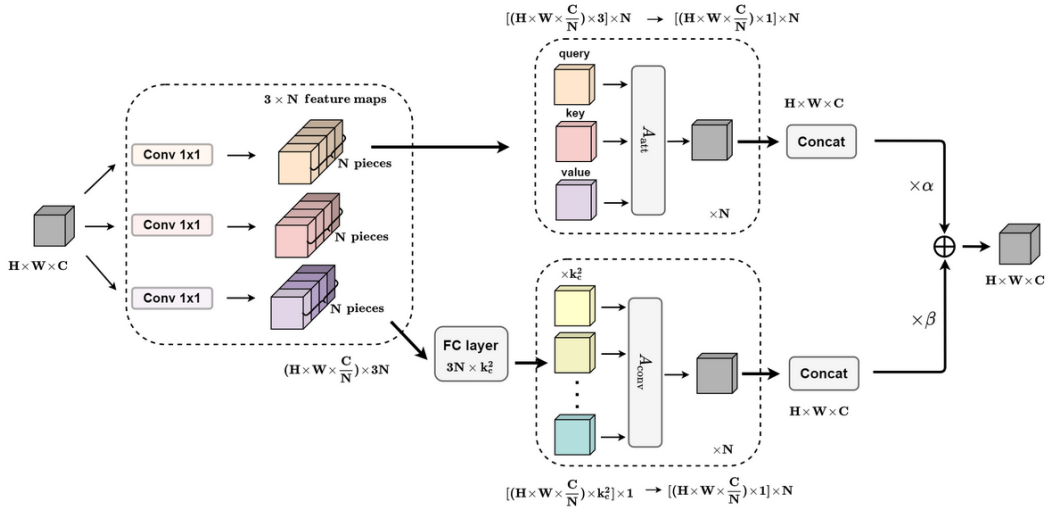


FIGURE 2.12: Illustration of details on ACmix. The image is from [108].

ACmix, which enjoys the benefits of local and global attention while having minimal computational overhead compared to pure convolution or self-attention. The details can be found in the Figure 2.12 and the original paper [108].

Inspired by the success of both channel and spatial attention, several researchers propose to jointly apply both attention to the baseline, forcing the network to learn both semantic and resolution cues.

2.3 Joint Channel-Spatial Attention

2.3.1 Convolutional Block Attention Mechanism

We have discussed in the previous section the importance of the attention module in the channel and spatial direction. It can be noticed that the presented works only focus on one dimension, with few networks explicitly tackling both dimensions. Therefore, following the same motivation but with another perspective, CBAM [161] proposes a joint channel and spatial attention module, aiming to increase the response at the most informative regions and channels and suppress unnecessary features. To emphasize meaningful features in both spatial and channel dimensions, the authors sequentially apply channel and spatial attention modules to learn what to pay attention to and where to pay attention in the channel and spatial dimensions, respectively. This not only saves parameters and computing power but also ensures that each module can be integrated into the existing network architecture as a plug-and-play module.

Specifically, for channel awareness, the whole computation is similar to the original SEnet. The main difference is that during the squeeze part, in CBAM both global max pooling and average pooling are applied. For spatial awareness, CBAM computes

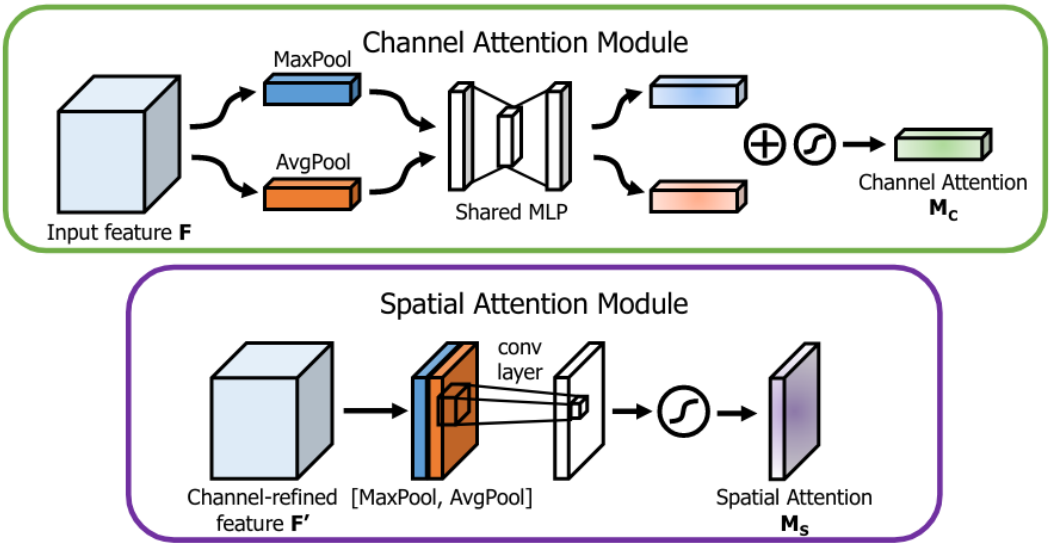


FIGURE 2.13: Illustration of a Channel and Spatial attention for CNN. The image is from [161].

the spatial attention map by applying average pooling and max pooling on the input feature map $x \in \mathbb{R}^{c \times h \times w}$ and obtains the attention map m with size $m \in \mathbb{R}^{2 \times h \times w}$. Different from channel attention which squeezes the spatial resolution, here the pooling methods are along the channel axis and therefore squeeze the channel dimension. Then the attention map is fed into a 2D convolution and a sigmoid function to reduce the fuse average-max pooled features and form the final attention map. Finally, the spatial attention is multiplied by the input feature map x.

Note that in the original paper [161], the authors suggest that channel attention and spatial attention can be combined in a parallel or sequential manner. But the authors found that combining sequentially and putting the channel attention at the front achieves better results.

2.3.2 Dual Attention

DANet [43] proposes an another manner to leverage both spatial and channel attention as shown in Figure 2.14. Specifically, the models combine two types of attention modules on top of traditional dilated FCNs, which model semantic inter-dependencies in spatial and channel dimensions, respectively. The Position Attention module selectively aggregates the features of each position by taking a weighted sum of the features of all positions. Similar features will be related to each other regardless of the euclidian distance. Meanwhile, the Channel Attention module selectively emphasizes interdependent channel maps by integrating relevant features in all channel maps. These two attention modules are exploited to capture global information in

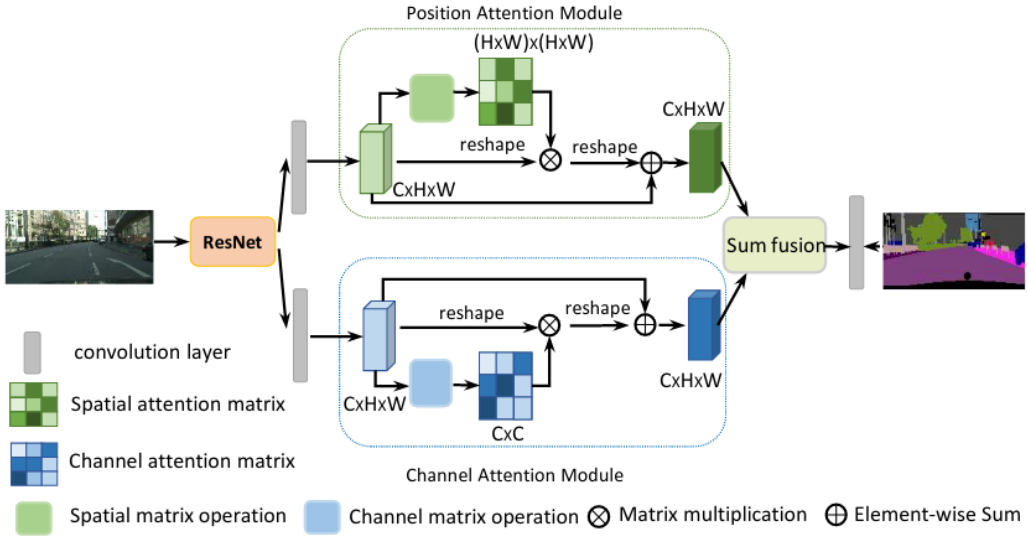


FIGURE 2.14: Illustration of details on dual attention. The image is from [43].

images. The dual attention has nowadays been exploited in various applications such as RGB-D semantic segmentation [207], image generation [142], etc.

The Position Attention Module and Channel Attention Module proposed by DANet are actually the same as the self-attention calculation method in Transformer. The difference is that one of the features involved in the calculation is a picture feature (2D) and the other is a word embedding (1D).

2.4 RGB-D Fusion

In previous sections, we briefly reviewed several milestones in the deep learning area. Most of these works are designed for RGB single images as input. While it is more trivial to extend these approaches to video tasks such as video semantic segmentation [139, 140], it is more challenging to extend these designs to multi-modal inputs, especially with RGB-D inputs. In this section, we will briefly review several RGB-D fusion designs so that the readers can have a global view of this field.

2.4.1 Depth as 3D Data

How to deal with complementary depth is a key research topic for RGB-D tasks. Different from 2D RGB images, RGB-D images provide additional cues on 3D geometry. Therefore, a straightforward motivation is to project the 2D pixels to form the 3D representations as shown in Figure 2.15. Among all the 3D representations, the most widely used ones are the voxel format and the point cloud format.

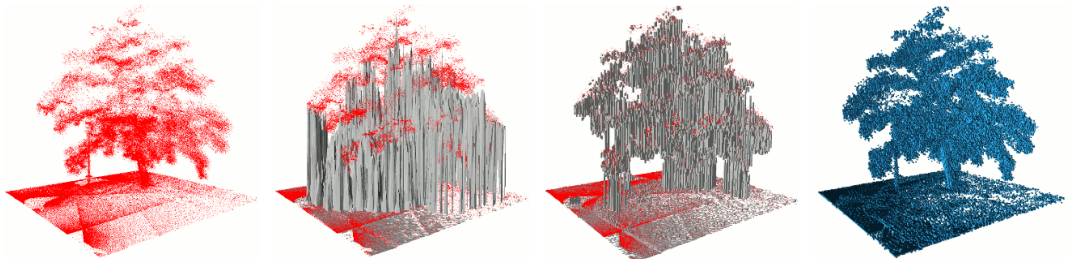


FIGURE 2.15: Illustration of different 3D representations. From left to right: **Point cloud**, elevation map, multi-level surface map, and **voxel** representation. The image is from [63].

Voxel can be regarded as a 3D format of the pixel. It describes the occupancy of a 3D grid. Once the 3D voxel is obtained from the depth sensor, a straightforward idea is to extend conventional 2D CNN to 3D CNN by adding the additional depth dimension. Previous works [103, 171] have shown that this method works well when dealing with applications such as shape recognition.

Despite the demonstrated success, it inefficiently consumes huge memory as data is often sparse on the 3D scene. In contrast to previous works, [116, 118] propose to directly use the point cloud representation. Point cloud data are often orderless. In other words, by reshaping the order of the point cloud, the object features remain the same. In addition, each point inside the point cloud is highly correlated with others. By deeply analyzing these characters, PointNet [118] is proposed to directly deal with point cloud input. This method has shown a great advantage in both computational cost and performance compared to 3D CNN with voxels and has become the standard to deal with 3D data.

2.4.1.1 2D-3D Fusion

The 3D geometric cues provided by the depth data can naturally complement the RGB input. Therefore, proposing a 2D-3D fusion module has drawn great attention. One typical work is the DenseFusion [148] for 6D pose estimation. The main idea is that since RGB data and point cloud data are heterogeneous data located in different feature spaces, so DenseFusion uses a heterogeneous network to process these two kinds of data separately while retaining the structure of the two kinds of data themselves. It proposes a dense pixel-level fusion method, which integrates the features of the RGB data and the features of the point cloud in a more suitable way.

Specifically, the first step is to pre-process the input RGB and point cloud data. DenseFusion realizes the semantic segmentation on the RGB image and extracts the point cloud corresponding to each mask. The cropped RGB regions, along with the corresponding point cloud are fed into the deep network.

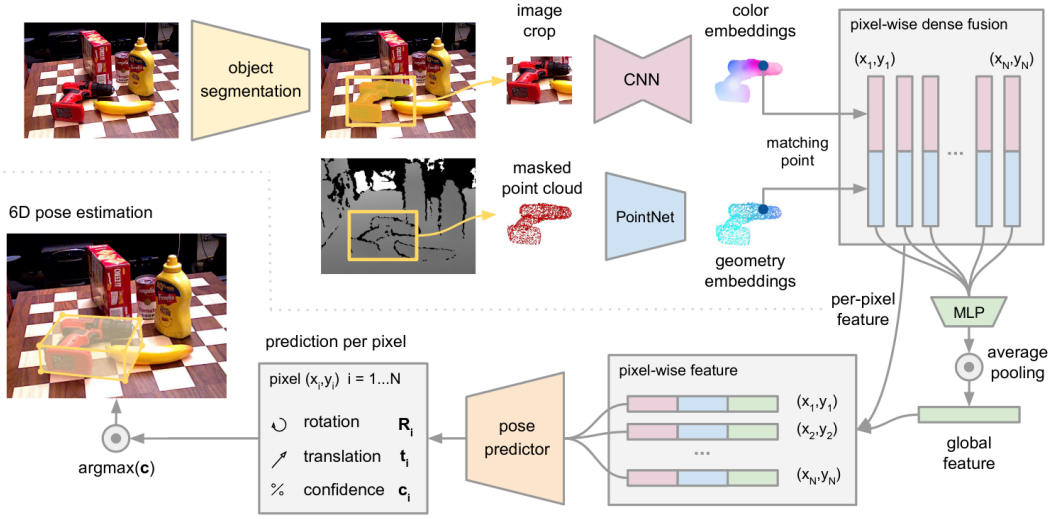


FIGURE 2.16: Illustration of DenseFusion pipeline. The image is from [148].

The deep network follows the conventional encoder-decoder design as shown in Figure 2.16. However, it has two branches. Firstly, it uses a fully convolutional network to project each pixel in the RGB crop to the color feature space, as well as the point cloud with a separate branch. Then, these two modalities are simply concatenated and fed into MLP for information integration, after which an average pooling is applied to obtain global features. Finally, the average feature, along with the modality-specific features, are concatenated to form the shared feature maps which are finally proceeded by the pose estimation decoder.

DenseNet shows a simple concatenation between different modalities can significantly boost the performance over the single-modal baseline. Following DenseNet, there are a lot of other 6D pose estimation models such as [62] with different and more complicated fusion modules.

2.4.2 RGB-D 2D Fusion

Despite the plausible results with 2D-3D fusion, 3D data are always processed by a 3D deep network, which is commonly heavier than 2D networks. Therefore, another group of research aims to explore depth as a 2D map and realize the fusion on the 2D plane. Since there exist a lot of works focusing on RGB-D 2D fusion, hence, in this section, we only review the most representative works. More detailed related works can be found in each chapter.

2.4.2.1 Depth as 2D map

Instead of processing the 3D data, an alternative is to consider depth as another 2D image complementary to the RGB image. Deep neural networks for paired 2D RGB-D

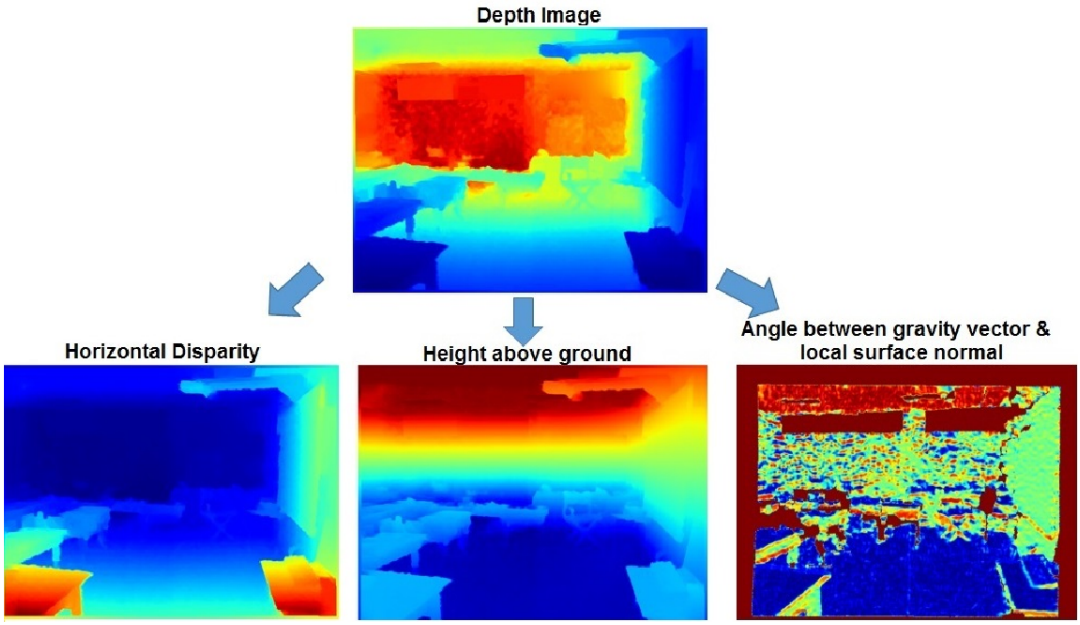


FIGURE 2.17: Illustration of encoded HHA from a depth image. The image is from [124].

images have attracted research interests for years and numerous improvements have been achieved. One straightforward idea is to use 2D depth as an additional input along with an RGB image. By deeply extracting depth features, networks can improve performance over the RGB baseline. This idea is widely used in tasks such as RGB-D Salient Object Detection.

However, for RGB-D semantic segmentation, researchers often adopt another representation of depth. At the early stage, [55] proposes to encode a depth map to a 3-channel HHA image, which refers to Horizontal disparity, Height above ground, and normal Angle. By using this method, we can encode the depth map to the HHA map which shares the same dimension as the RGB input. An example can be seen in Figure 2.17. Since then, the encoded HHA is widely used in RGB-D semantic segmentation tasks.

While the representations are different, the depth cue remains in the 2D dimension which is the same as the RGB input. Therefore, various fusion methods have been proposed to aggregate RGB and depth information. [217] thoroughly divides different fusion strategies into five categories as shown in Figure 2.18. More commonly, we can group different fusion works into three groups: early, middle, and late fusion. Early fusion often merges RGB-D images at stemming layers. Some of them even aggregate RGB-D images from the input and form 4-channel or 6-channel input. The advantage of such a strategy is the computational cost since after the early fusion, only one feature extraction is required at the semantic level. However, due to the imbalance between RGB and depth features, especially at the early stage, this design cannot

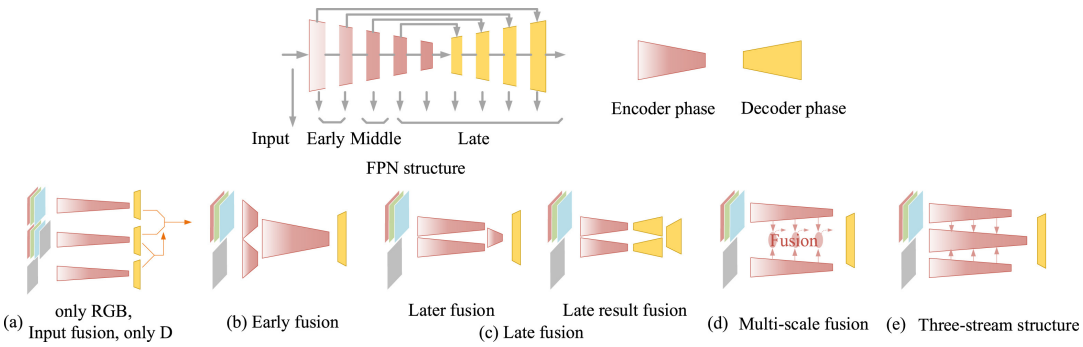


FIGURE 2.18: Illustration of conventionally applied fusion designs in the literature. The image is from [217].

fully leverage the multi-modal cues. Therefore, some other works propose to realize late fusion. [37] adopts an extremely late fusion that aggregates RGB-D cues at the output level. But most other late fusion works merge RGB-D cues at the semantic level, e.g., the output of the encoder. To further model the feature fusion at each level, a number of others works propose to fuse RGB-D features at each stage. This design is termed middle fusion. Each fusion strategy has its pros and cons. It is therefore hard to judge which one is better. Generally, experimental results show that middle fusion can yield better performance, but it is also the most time-consuming fusion strategy compared to its counterparts.

2.4.2.2 Pixel-wise Fusion

At the early stage, RGB-D features are simply aggregated with addition or concatenation convolution as shown in Figure 2.19. FuseNet [57] is one of the typical examples of incorporating the auxiliary depth information into the RGB encoder-decoder through simple addition. It first extracts multi-scale depth features and then simultaneously adds them to the RGB mainstream at each scale. The addition can also be replaced by concatenation. Commonly, researchers concatenate RGB-D features along the channel dimension. Therefore, the concatenation is always combined with an additional convolution to reduce the doubled channel size. Despite the plausible improvement compared to the RGB baseline, both addition, and concatenation fusion have several limits. Firstly, they do not take the noise of RGB-D images into account. Secondly, they assume that RGB-D features share different but completely complementary information. Therefore, they do not take feature redundancy into account. Finally, they assume that both modalities are well-aligned at the pixel-wise level.

2.4.2.3 Fusion with Self-Attention

To tackle the above-mentioned limits and be inspired by the success of attention modules, several works further explore the effectiveness of such designs for RGB-D fusion. For example, ACNet [65] first apply self-attention modules to self-calibrate

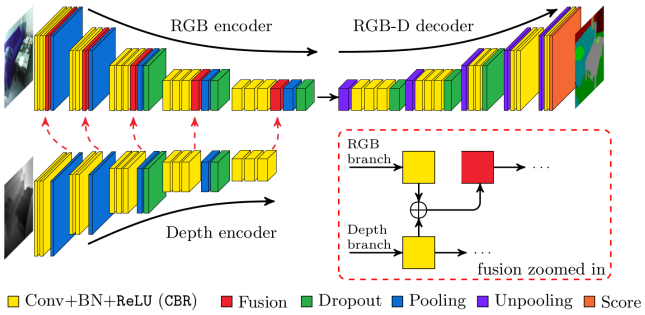


FIGURE 2.19: Illustration of pixel-wise fusion strategy with simple addition. The image is from [57].

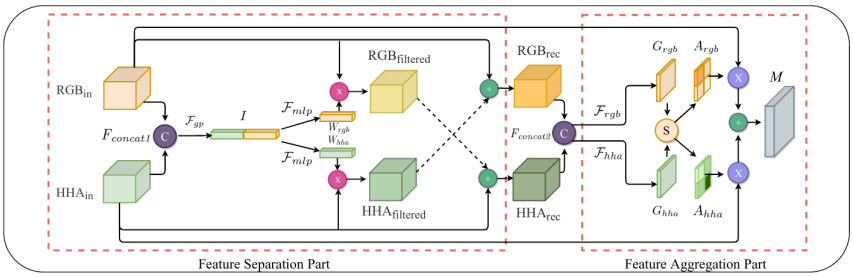


FIGURE 2.20: Illustration of SAGate pipeline and fusion details. The image is from [15].

RGB and depth features and then fuses them at each stage. The self-calibration step is exactly the same as the original channel attention proposed by SENet [64].

Sharing a similar idea, SA gate [15] further explicitly leverages spatial and channel cues to firstly calibrate modality in a bi-directional manner and further realize middle fusion. The pipeline is shown in Figure 2.20. In addition to the channel attention applied in ACNet, SAGate adopts both spatial and channel attention. Furthermore, SAGate explicitly leverages the cross-modal interaction to calibrate both RGB and depth features with cross-modal cues, while ACNet only adopts self-modal attention to improve the feature representation.

2.4.2.4 Fusion with Non-local Attention

In contrast to previous works based on self-attention (channel and spatial attention), several works explore non-local attention to better leverage the contextualized cues. D-CNN [154] is one of the pioneering works which integrates the depth distance into the weight function for convolutional operations. Specifically, it computes the depth distance between two pixels and uses this depth-aware weight to recalibrate the convolution and pooling. The details can be found in Figure 2.21.

Another line of work is to explore long-range attention for RGB-D fusion. The pioneering work is the CANet [207] which computes the cross-modal attention at the

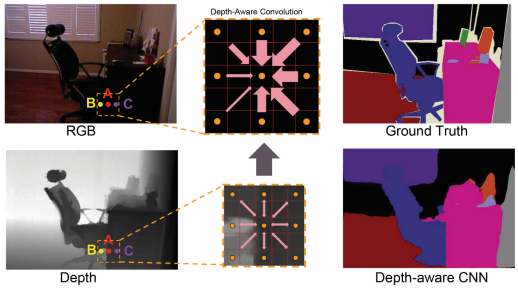
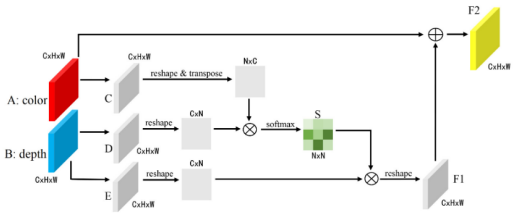
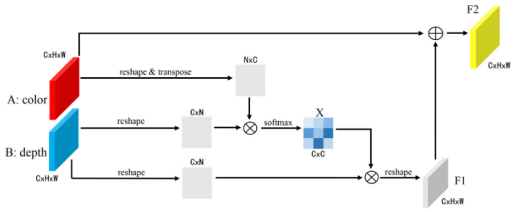


FIGURE 2.21: Illustration of depth-aware convolution. The image is from [154].



(a) Position co-attention fusion module



(b) Channel co-attention fusion module

FIGURE 2.22: Illustration of RGB-D fusion proposed by CANet. The image is from [207].

feature level. The details can be found in Figure 2.22. It can be seen that the idea is highly similar to Dual Attention [43] and non-local attention [156].

2.5 Summary

In this chapter, we reviewed several milestones during the development of Deep Learning models from CNN to Transformer via different attention modules. We also review various existing RGB-D fusion methods from pixel-wise fusion to non-local attention via self-attention. It can be seen that the development of RGB-D fusion methods is highly correlated to the development of RGB tasks, with around one year or two years gap. In the following chapters, we will present our contributions achieved during this these.

Chapter 3

RGB-D Salient Object Detection via Hierarchical Depth Awareness

RGB-D saliency detection aims to fuse multi modalities to accurately localize salient regions. Existing works often adopt attention modules for feature modeling, with few methods explicitly leveraging fine-grained details to merge with semantic cues. Thus, despite the auxiliary depth information, it is still challenging for existing models to distinguish objects with similar appearances but at distinct camera distances. In this chapter, from a new perspective, we propose a novel Hierarchical Depth Awareness network (HiDAnet) for RGB-D saliency detection. Our motivation comes from the observation that the multi-granularity properties of geometric priors correlate well with the neural network hierarchies. To realize multi-modal and multi-level fusion, we first use a granularity-based attention scheme to strengthen the discriminatory power of RGB and depth features separately. Then we introduce a unified cross dual-attention module for multi-modal and multi-level fusion in a coarse-to-fine manner. The encoded multi-modal features are gradually aggregated into a shared decoder. Further, we exploit a multi-scale loss to take full advantage of the hierarchical information. Extensive experiments on challenging benchmark datasets demonstrate that our HiDAnet performs favorably over the state-of-the-art methods by large margins.

3.1 Introduction

Salient object detection (SOD) aims to find the most prominent region inside an image that visually attracts human attention. Conventional SOD approaches only take color images as inputs. With deep learning models, RGB SOD has achieved significant success [30, 91, 172, 194, 202]. However, these models may result in unsatisfactory performance when dealing with complex scenes, e.g., low-contrast light or object occlusion.

Recent advanced RGB-D sensors provide accessibility to depth maps at low cost. The complementary geometric cues can contribute to scene understanding. In the

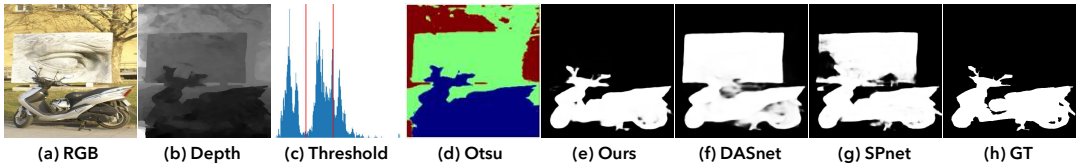


FIGURE 3.1: Motivation of our hierarchical depth awareness. (a) and (b) are the paired RGB-D inputs. (c) and (d) represent Multi-Otsu thresholding on depth histogram and the generated Otsu regions, respectively. Our approach takes full advantage of depth priors to improve the feature discriminatory power and obtain the saliency mask (e). Compared to two state-of-the-art (SOTA) RGB-D models (f) and (g), our method favorably yields results closer to the ground-truth mask (h).

literature, two main designs have been widely exploited, i.e., single-streaming schemes that combine RGB-D images from the input side [44, 189, 203] and multi-streaming network that extracts multi-modal features separately and combines them at semantic levels [38, 69, 100, 141, 187, 190, 195, 210]. Existing networks often directly extract semantic features through the deep network, with few methods fully explore the rich geometric priors provided by the depth map.

Previous works on channel attention [59, 64, 121, 161] have shown their effectiveness in emphasizing the attentive features among channels. A number of saliency detection works [38, 69, 190, 199] adopt channel attention to enhance multi-modal features. However, the first step of learning channel attention is to aggregate the spatial information of feature maps to construct a $1 \times 1 \times C$ vector by using global average pooling, where C is the number of channels. As a result, the foreground and background contribute equally to the output, which is not optimal to distinguish salient objects. Considering these issues, an intuitive motivation is to design local channel attention referring to depth priors in order to improve feature representation learning.

As shown in Fig. 3.1, while dealing with complex scenes, current state-of-the-art (SOTA) RGB-D models [199, 210] fail to extract the salient region due to similar visual appearance between the foreground and background (Fig. 3.1(f) and (g)). However, we observe that salient regions often share similar depth properties, i.e., a certain granularity of depth prior, that help to distinguish the salient objects from the background (Fig.3.1(b) and (d)). Inspired by this observation, we develop a local feature enhancement scheme with granularity-based attention (GBA) to improve saliency detection. Specifically, we propose to first generate various local regions according to the granularity via Otsu thresholding [87, 107]. These regions can be considered as distinct local spatial attention. Then for each region, we apply local channel attention to improve the feature discriminatory power. Fig. 3.1(c) and (d) illustrate such an example of the Otsu threshold values and granularity-aware masks,

respectively. We show that our approach can better reason about salient regions (Fig. 3.1(e)) that are closer to the ground truth (Fig. 3.1(h)).

We further introduce a cross dual-attention module (CDA) to learn channel and spatial attention from auxiliary modalities to improve the current streaming. The enhanced features are hierarchically fused for final saliency map generation. Besides, the same cross-interaction scheme is embedded to articulate features between encoders and decoders through a U-Net-like [126] architecture. We attentively mirror the multi-scale encoder features to preserve valuable geometric priors within each decoder. The encoded features are gradually fused to a shared decoder. Finally, we use a multi-scale loss on top of outputs from each decoder to optimize the saliency map. Concretely, our contributions are summarized as follows:

- We propose a novel granularity-based attention scheme that attends to fine-grained details in order to strengthen the feature discriminability of each modality.
- We design a new multi-modal and multi-level fusion scheme with a multi-scale loss to take full advantage of the network hierarchy.
- We extensively validate our HiDAnet on large-scale challenging benchmarks. Our approach performs favorably over SOTA models with large margins.

3.2 Related Work

There are extensive surveys [5, 26, 130, 153, 205, 209] of salient object detection in the literature. In this section, we briefly review related RGB-D saliency detection as follows:

Multi-Modal Fusion. The auxiliary depth map provides extra geometric clues in addition to visual appearance. To efficiently merge both modalities, several fusion methods have been proposed. A number of works [13, 44, 45, 188, 189, 203] directly concatenate the depth map with RGB images from the input side through a single-stream network. On the one hand, JLDCF and its successor [44, 45] explore the siamese design for saliency detection by concatenating RGB and depth images in an additional dimension with a joint learning scheme. DANet [203] forms a four-channel input and enhances the extracted features with a dual-attention mechanism learned from depth. [188, 189] propose the stochastic framework to analyze the uncertainty during human labeling and model the distribution of the saliency output. Different from previous works, [12, 13] attempt to address RGB-D SOD from the 3D point of view with a 3D convolutional neural network. The recent [208] leverages the depth cues to mimicks multi-view images and then fuse them to form the final output.

On the other side, multi-stream models [38, 69, 74, 100, 141, 187, 190, 195, 210] have achieved leading performances in RGB-D SOD. These models adopt two parallel encoders on different modalities, and the features are fused through different strategies. Several works [38, 186, 215] firstly enhance the depth features before fusing with RGB features. It is worth noting that a portion of the depth maps in existing saliency datasets are not of satisfactory quality. As discussed in [20, 37, 44, 170], the depth may contain measurement or estimation bias. Thus, DCF [69] designs a calibration module to improve the depth quality. [20, 67, 74, 170] propose a layer-wise attention to model the geometric contribution with respect to the network depth. [20] explores an additional backbone to learn the weighting scalar purely from depth. [170] analyzes the similarity between RGB and depth features to regular the depth contribution. Sharing the same motivation, [74] computes the reliability of each modality at each stage and then merges them through their reliability. Instead of learning the weighting scalar, [67] generates the weighting maps at each scale to calibrate the feature response. Similarly, [197] leverages bilateral attention to improve foreground-background features separately. Unlike these works, we first divide the feature map into several local regions with the help of depth granularity. The feature maps are further calibrated with different local attention to improve the feature discriminability. Compared to [67, 197], our fined-grained details are statically computed by maximizing the inter-class distance without learning parameters, leading to more reasonable and stable locally-calibrated areas.

There exist other works which only extract features from RGB input while the depth map only serves as supervision [70, 115, 199]. In this context, [71, 167] propose to leverage the pseudo-depth to guide the RGB learning. A2dele [115] further formulates depth supervision as a knowledge transfer problem. CoNet [70] and DASnet [199] propose a multi-task learning framework with an additional depth head together with the saliency branch. However, we argue that these methods cannot fully leverage the multi-modal cues during feature extraction. Instead, we propose a cross-interaction scheme to take full advantage of cross-modal cues. We benefit from the auxiliary modality to alleviate errors in the feature modeling (depth to RGB, and RGB to depth).

Multi-Level Fusion. U-Net with skip connections [126] has shown its effectiveness in pixel-level segmentation tasks. Several RGB-D SOD models [100, 109, 187, 210] equip this design for clearer boundary generation. [109] adopts the feature-wise addition. [100, 210] concatenate the encoder features with the decoder. [187] designs a dense connection between high-level features and the decoder. In this work, we exploit the contribution of attention modules for skip connections applied to SOD. It is

worth mentioning the success of skip connections can be mainly attributed to aggregation between the semantic features provided by the contracting path and fine-grained features from the expansion path. From a new perspective, we consider the encoder-decoder features as multi-modal features, and a unified cross-fusion scheme is applied to boost the performance.

Attention for Feature Enhancement. Attention methods such as transformer [147], CBAM [161], SEnet [64], DA [43], and ECA [121] have demonstrated their success in other vision tasks. A number of RGB-D saliency models also equip attention modules to extract attentive features from different modalities. VST [93] and TriTrans [100] adopt transformer [147] for saliency detection. [149,199,204] apply the SE module to compute modality-specific attention for feature calibration. Similarly, CDInet [187] designs a depth-induced channel attention to enhance RGB features. From another perspective, [193] deeply explores the spatial attention at different scales with the help of decoupled dynamic convolution. Sharing the same motivation, DFMnet [195] adopts a depth holistic attention on top of features with different resolutions. More recently, several works leverages both spatial and channel attention to jointly improve the feature representation. For example, BBSnet [38] applies the CBAM [161] on the depth map to improve the depth quality before fusion. [160] further improves the CBAM by highlighting spatial features. Sharing the same motivation, CMINet [190] applies the DA [43] on to lately merge RGB-D features. Different from previous works with bi-directional cross-modal attention, HAINet [77] explores the purified depth to improve the RGB features in turn.

Despite the proven effectiveness, previous channel attention schemes do not fully benefit from the geometric priors. For example, the same attention can be applied to both foreground and background. The rich geometric priors in the input depth map have rarely been discovered, which limits the performance of RGB-D saliency detection. DSA2F [141] introduces a depth-sensitive module with the help of the depth histogram. However, it computes the depth region with a fixed threshold for each input image and the attention scores are simply computed by a $Conv_{1\times1}$. In contrast, we propose to dynamically generate multi-granularity regions with the multi-Otsu method [87, 107]. The fine-grained details are further integrated with channel attention to enhance the feature discriminability for sharper edge generation.

3.3 Method

Fig. 3.2 presents the overall framework of our proposed HiDAnet. Note that the Otsu masks are generated from the depth map during the pre-processing. Firstly, RGB and

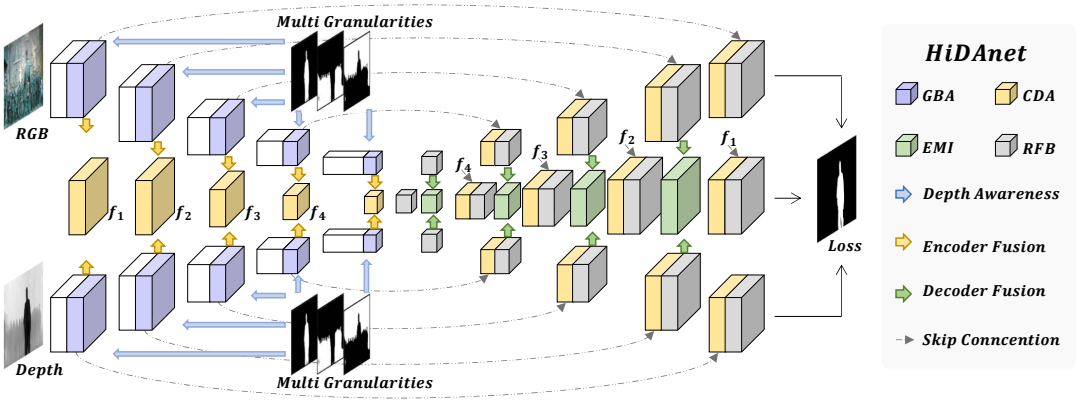


FIGURE 3.2: Architecture of our HiDAnet. Our network adopts U-Net-like design and consists of granularity-based attention (GBA Section 3.3.1), cross dual-attention module (CDA Section 3.3.2), and efficient multi-input fusion (EMI Section 3.3.3). RFB is the receptive field block from [94] for accurate object detection. White blocks denote the network backbone. Our granularity-based attention strengthens the discriminatory power of RGB and depth features separately. Our cross dual-attention module takes advantage of cross-domain cues to attentively realize multi-modal and multi-level fusion in a coarse-to-fine manner. Our efficient fusion scheme effectively models the shared information from each modality. The shared features are further improved with the skip connections for final saliency map generation. Best viewed in color.

depth maps are fed into two parallel encoders for feature extraction. For each individual encoder (RGB/Depth), we propose a granularity-aware module (GBA) with the help of input Otsu masks to enhance the discriminatory power, e.g., foreground and background. This module is naturally embedded into different levels of the encoder to correlate with the network hierarchies. With the enhanced features, we propose a unified fusion mechanism (CDA) for multi-modal and multi-level fusion. It enables a cross-domain interaction with both channel and spatial attention to learn the informative shared features in a coarse-to-fine manner. These features are later gradually aggregated into the shared decoder through the efficient multi-input fusion module (EMI). Lastly, we exploit a multi-level loss to take full advantage of the network hierarchies. Details of each component are presented in the following sections.

3.3.1 Feature Extraction with Granularity-Based Attention

We observe that the multi-granularity properties of geometric priors correlate well with the network hierarchies of saliency models. Inspired by this observation, we propose the granularity-based attention that aims to attentively combine the spatial attention mask with the conventional channel attention as shown in Fig. 3.3. For earlier layers, it strengthens the low-level representations to precisely localize the salient object with a sharp boundary. For deeper layers, it improves semantic abstraction and contributes to the identification of salient objects regardless of appearance variations.

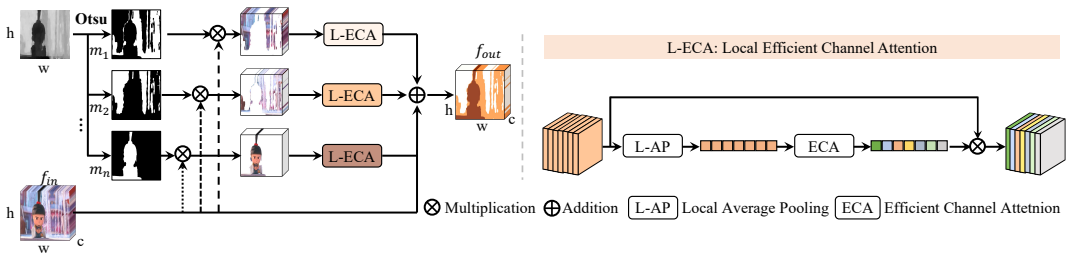


FIGURE 3.3: Diagram of the granularity-based attention. The depth awareness is encoded via local efficient channel attention (LECA). ECA is from [121].

Given the depth map D with its histogram H, we dynamically generate the fine-grained details. According to the value/distance within the depth map, we use the multi-Otsu thresholding algorithm [87, 107] to discretize the histogram H into several different regions. The vanilla Otsu algorithm [107] works for the bi-level thresholding, where pixels are divided into two classes, $C_0(d)$ with distance $[0, d]$ and $C_1(d)$ with distance $[d+1, 255]$ given the threshold $d \in \mathbb{N}$. It steps through all possible thresholds $d \in [0, 255]$ to find the threshold that minimizes the intra-class variance, which is defined as a weighted sum of variances of the two classes:

$$\sigma_w^2(d) = w_0(d)\sigma_0^2(d) + w_1(d)\sigma_1^2(d), \tag{3.1}$$

where σ and w stand for the variance and probability of each class. The probability distributions are computed as the number of pixels contained in the interval:

$$w_0(d) = \sum_{i=0}^{d} p(i); \quad w_1(d) = \sum_{i=d+1}^{255} p(i). \tag{3.2}$$

In this work, we use the extended multi-Otsu [87] to generate multiple thresholds. Assuming T random thresholds $(d_1, d_2, ..., d_T)$ dividing the depth into $T+1$ parts. Let (σ_i^2, w_i) be the variance and the pixels number of region i $(1 \leq i \leq T+1)$. The optimal values $\{d_1^*, d_2^*, ..., d_T^*\}$ are chosen by maximizing the inter-class variance:

$$\{d_1^*, d_2^*, ..., d_T^*\} = argmax\{\sigma_w^2(d_1, d_2, ..., d_T)\}, \tag{3.3}$$

where $\sigma_w^2 = \sum_{i=1}^{T+1} w_i\sigma_i^2$. To reduce the computational cost, we only generate the Otsu regions once during pre-processing and further resize them to fit the resolution of feature maps from different scales.

For the i^{th} region m_i, $(1 \leq i \leq T+1, i \in \mathbb{N}^*)$, we mask out the feature map f_{in} with element-wise multiplication to suppress the inactive area through $f_{in} \otimes m_i$. Then, channel attention is applied to improve the feature representation with local awareness. Compared to the vanilla channel attention [64, 121], we replace the global average

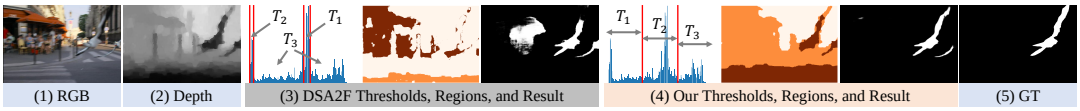


FIGURE 3.4: Visual comparison with the concurrent alternative. Different from DSA2F [141]. Our method maximizes the inter-class variance, leading to more accurate masks compared to DSA2F. We further explore the granularity cues via channel attention, yielding results closer to the ground truth (5).

pooling with the local average pooling that attends to the local details referring to geometric priors. Finally, the locally enhanced features are aggregated by a residual connection for the final output generation f_{out}. The overall process can be formulated as:

$$\begin{aligned} LECA(x) &= \sigma(Conv_{1d}(LAP(x))) \otimes x, \\ f_{out} &= \sum_{i=1}^{T+1} LECA(f_{in} \otimes m_i) + f_{in}, \end{aligned} \tag{3.4}$$

where $\sigma(\cdot)$ is the Sigmoid activation, $\otimes$ is the element-wise multiplication, and LAP denotes the local average pooling on each masked region. We provide more details on the differences between the proposed granularity-based attention and traditional channel attention in the ablation study Section 3.5 Tab. 3.5.

Remarks. Several previous works have proposed to explore depth prior in various manners such as the contrast in CPFP [200], the edge in CoNet [70], or the histogram in DSA2F [141]. Our approach resembles the DSA2F that both methods belong to threshold-based segmentation frameworks. However, one main difference is that we dynamically generate optimized masks with the Ostu algorithm, while DSA2F applies fixed thresholds on the $T + 1$ largest depth distribution modes that cannot adapt to different scenarios without handcraft adjusting. Fig. 3.4 illustrates the difference in the thresholds and regions. We observe that our approach computes more discriminative regions, yielding a more effective and robust manner to explore the depth prior. Moreover, since the Otsu algorithm optimizes the thresholds by maximizing inter-class variance, our generated masks are more robust to the depth noise compared to the concurrent work. Additionally, we leverage the granularity with channel attention, while DSA2F simply uses a $Conv_{1\times1}$ for local awareness. As shown in Fig. 3.4, by integrating the fine-grained details into the channel attention, we can reason about more accurate saliency regions closer to the ground truth. The quantitative comparison with [70, 141, 200] can be found in Section 3.4.3 Tab. 3.1. Our superior performance proves that we can better model the depth priors.

3.3.2 Encoder Fusion with Cross Dual-Attention Module

Previous studies [37, 115, 199] have affirmed the effectiveness of learning from two heterogeneous modalities for RGB-D SOD. Color images provide rich information in visual appearance while depth maps contain more spatial priors. Both modalities contribute to modulating homogeneous semantic information. Therefore, the objective of multi-modal learning is to efficiently fuse features with diverse information from different modalities. Similar to multi-modal features, multi-level features also contain both heterogeneous and homogeneous information: high-level features are richer in abstract semantic cues while low-level features are richer in fine-grained details. Thus, from a new perspective, we design a unified fusion scheme to make full use of cross-domain cues for both multi-modal and multi-level reasoning.

Assuming two paired multi-modal features f_x and f_y. We firstly build a transformation F_t to map the inputs $f_x, f_y \in \mathbb{R}^{C\times h\times w}$ to feature maps $f'_x, f'_y \in \mathbb{R}^{C'\times h\times w}$ with $C' = \frac{C}{2}$. Specifically, F_t is the combination of a 1×1 convolution which halves the channel size and a 3×3 convolution which is expected to activate the edge response:

$$\begin{aligned} f'_x &= F_t(f_x) = Conv_{3\times 3}(Conv_{1\times 1}(f_x)), \\ f'_y &= F_t(f_y) = Conv_{3\times 3}(Conv_{1\times 1}(f_y)). \end{aligned} \tag{3.5}$$

Once obtaining the lightweight representation, the next step is to aggregate features from different domains (RGB-D or encoder-decoder). We observe from Fig. 3.1 that the fine-grained details, such as relative boundary, facilitate the identification of salient objects. Simultaneously, in case it is difficult to distinguish objects at the same distance on the depth map, e.g., when distinguishing the motorbike from the street, the visual appearance becomes more reliable. Inspired by this observation, we aim to use heterogeneous clues to compensate for the single-domain streaming.

To this end, we propose a cross dual-attention fusion scheme as shown in Fig. 3.5. Specifically, from each input feature map, we learn the 1-D channel attention $M_c \in \mathbb{R}^{C'\times 1\times 1}$ to determine *what* information to be involved, and the 2-D spatial attention $M_s \in \mathbb{R}^{1\times h\times w}$ to determine *which* part to focus. We formally have the operations:

$$\begin{aligned} M_c(f') &= \sigma(MLP(GAP(f')) + MLP(GMP(f'))), \\ M_s(f') &= \sigma(Conv_{7\times 7}(Concat(CAP(f'), CMP(f')))), \end{aligned} \tag{3.6}$$

where $\sigma(\cdot)$ is the Sigmoid activation, MLP is the multi-layer perceptron, GAP and GMP are the global average and max pooling, respectively, and CAP and CMP are the average and max pooling across the channel, respectively. With the learned dual attention from separate feature maps, we enable a cross-domain interaction. In such

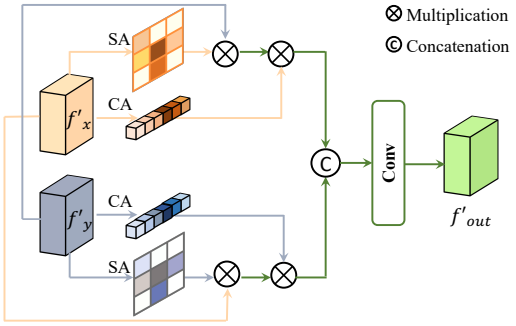


FIGURE 3.5: Encoder Fusion. We adopt a multi-scale multi-level fusion scheme with cross-domain supervision.

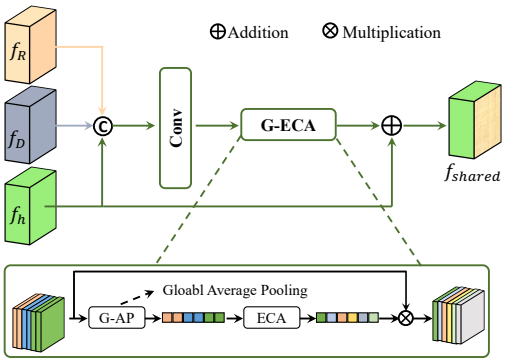


FIGURE 3.6: Decoder Fusion. We adopt efficient fusion with global pooling to attentively aggregate decoded features.

a way, we can alleviate the ambiguities in the domain-specific features. Finally, the cross-enhanced features are fed into concatenation and convolution to form the shared representation f'_{out}. The overall process can be formulated as:

$$\begin{aligned} f_x^{enh} &= M_s(f'_y) \otimes M_c(f'_y) \otimes f'_x, \\ f_y^{enh} &= M_s(f'_x) \otimes M_c(f'_x) \otimes f'_y, \\ f'_{out} &= Conv_{3\times3}(Concat(f_x^{enh}, f_y^{enh})), \end{aligned} \tag{3.7}$$

where $\otimes$ denotes element-wise multiplication. For the shared encoder, starting from the second layer, once the multi-modal features are fused through cross attention, the output is further combined with the previous level output through a $Conv_{3\times3}$.

Remarks. Our fusion design differs from concurrent works [69,191,199,210] in several aspects: (**A**) We leverage both spatial and channel attention to aggregate multi-modal features, while [69,199] only focus on channels; (**B**) Different from ASTA [191], our calibration is bi-directional (RGB to depth and depth to RGB), while ASTA is asymmetric which only leverages depth cues to improve RGB features. Hence, it does not tackle depth noise; (**C**) SPnet [210] also adopts the symmetric fusion strategies. Our work differs from SPnet in that we fully explore the attention modules for feature fusion, while SPnet is built upon simple convolutions to combine features; (**D**) The fusion scheme can also be implemented by the CBAM [161]. However, vanilla CBAM is modality-specific and cannot explore its relevance in cross-domain features. The ablation study in Section 3.5 Tab. 3.9 shows the gain with the cross interaction.

3.3.3 Decoder Aggregation with Efficient Multi-Input Fusion Module

To aggregate the learned features from both RGB and depth decoders into the shared decoder, a simple concatenation may not be adaptive enough due to the tripled number of descriptors. Thus, we propose an efficient multi-input fusion strategy. Specifically, as shown in Fig. 3.6, after the simple concatenation between different inputs (RGB f_R, depth f_D, and previous-level shared features f_h), we adopt the vanilla ECA [121] module (termed GECA while G stands for global pooling) to explore the inter-dependencies of different features. Thus, the most responded features are adaptively selected to form the shared decoder. A residual addition is adapted to reinforce the contribution of the previous level features. We have the overall process:

$$f_{shared} = GECA(Conv_{3\times3}(Concat(f_R, f_D, f_h))) + f_h. \tag{3.8}$$

The shared decoded features are then fed into our cross dual-attention scheme to realize the skip-connection between the shared encoder-decoder.

Remarks. Our encoder fusion (CDA) and decoder fusion (EMI) are technically different. We observe that the spatial cues gradually lose during encoding and become limited for decoders. This motivates us to apply both spatial and channel attention for the encoder fusion, while only using channel attention for the decoder fusion.

3.3.4 Optimization

To take full advantage of the hierarchical information, we supervise multi-level outputs for both RGB, depth, and shared/fused branches. For outputs from each level, the predicted map is upsampled to form the same resolution mask as the ground truth. We adopt BCE loss $\mathcal{L}^{BCE}$ for pixel restriction and IoU loss $\mathcal{L}^{IoU}$ for global restriction [122, 159, 199]. Therefore, we have the loss $\mathcal{L}_i$ for the i^{th} level output:

$$\mathcal{L}_i = \mathcal{L}_i^{BCE} + \mathcal{L}_i^{IoU}. \tag{3.9}$$

In total, we have five-level outputs (after each RFB in Fig. 3.2). Thus, by combining the loss from each branch (R for RGB, D for depth, and S for shared branches), the overall multi-level loss function $\mathcal{L}_{ml}$ becomes:

$$\mathcal{L}_{ml} = \sum_{i=1}^{5} \lambda_i(\mathcal{L}_i(R) + \mathcal{L}_i(D) + \mathcal{L}_i(S)), \tag{3.10}$$

where λ_i is the weight of the different-level loss. To correlate with the network hierarchies, we follow [19, 199] and set the weight λ as $\{1, 0.8, 0.6, 0.4, 0.2\}$.

We expect the multi-level loss to measure the difference between the generated mask and ground truth at various layers, and to force the network to learn hierarchical features that capture long- and short-range spatial relationships between pixels. The gain by adopting the multi-level loss can be found in ablation study Section 3.5 Tab. 3.8.

3.4 Experiments

3.4.1 Benchmark Datasets

To verify the effectiveness of our approach, we firstly train with the conventional training dataset following the protocol presented in [38, 69, 100, 199, 210] with 2,195 samples: 1,485 samples from the NJU2K-train [72] and 700 samples from the NLPR-train [112]. For testing, experiments are conducted on five classical benchmark RGB-D datasets. DES [22] : includes 135 images of indoor scenes captured by a Kinect camera. NLPR-test [112]: contains 300 natural images captured by a Kinect under different illumination conditions. NJU2K-test [72]: contains 500 stereo image pairs from different sources such as the Internet, 3D movies, and photographs taken by a Fuji W3 stereo camera, where several depth maps are estimated through an optical flow method [138]. STERE [106]: includes 1,000 stereoscopic images downloaded from the Internet where the depth map is estimated using the SIFT flow method [90]. SIP [37]: contains 929 images with humans in the scene, and images are acquired by a mobile device. We further evaluate our model on a newly published dataset COME15K [190] where the depth is estimated through a modified optical flow algorithm [151]. In this case, our model is trained with provided 8,025 training samples and tested on the "Difficult" set with 3,000 images.

3.4.2 Experimental Settings

Our model is implemented based on Pytorch and trained with a V100 GPU. Our backbone is initialized with the pre-trained weights obtained from ImageNet. For the depth stream, we modify the first convolution to start from one channel. The input RGB-D resolution is fixed to 352×352. We choose the Adam algorithm as our optimizer. We initialize the learning rate to be $1e{-}4$ which is further divided by 10 every 60 epochs. The total training time takes around 6 hours for 100 epochs. During training, we adopt random flipping, rotating, and border clipping for data augmentation. During inference, the prediction maps from the shared branch are the final outputs (middle branch of Fig. 3.2).

We evaluate our performance with four generally-recognized metrics: F-measure is a region-based similarity metric that takes into account both Precision (P) and Recall

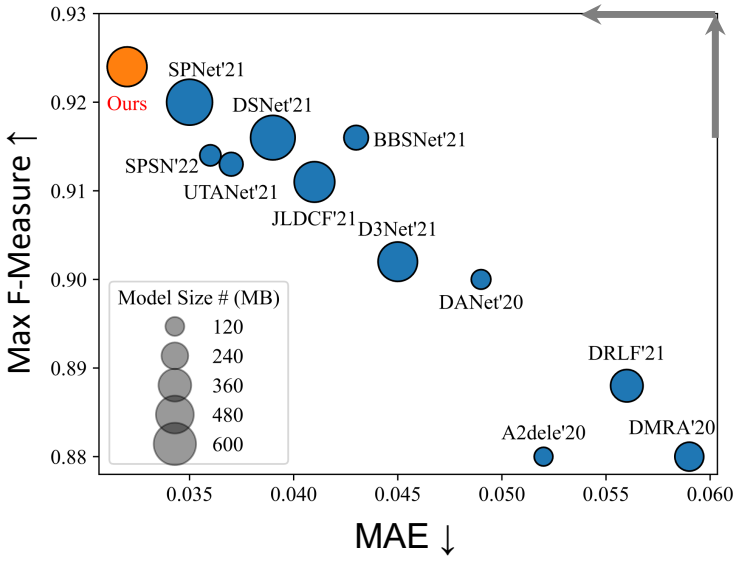


FIGURE 3.7: Average **Max F-Measure, MAE, and Model Size** of different methods on benchmark datasets. The circle size denotes the model size. Note that better models are shown in the upper left corner (i.e., with a larger F-measure and smaller MAE). Methods with smaller size perform inferior, making our method both efficient and accurate.

(R). Mathematically, we have : $F_\beta = \frac{(1+\beta^2)\cdot P\cdot R}{\beta^2\cdot P+R}$. The value of β^2 is set to be 0.3 as suggested in [1] to emphasize the precision. In this paper, we report the **maximum F-measure** (F_β) score across the binary maps of different thresholds. **Mean Absolute Error** (M) measures the approximation degree between the saliency map and ground-truth map at the pixel level. **S-measure** (S_m) [35] evaluates the similarities between object-aware (S_o) and region-aware (S_r) structures of the saliency map compared to the ground truth. Mathematically, we have: $S_m = \alpha \cdot S_o + (1-\alpha)\cdot S_r$, where α is set to be 0.5. **E-measure** (E_m) evaluates both image-level statistics and local pixel-matching information. Mathematically, we have: $E_m = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\phi_{FM}(i,j)$, where $\phi_{FM}(i,j)$ stands for the enhanced-alignment matrix as presented in [36]. To make a fair comparison, we use the same protocol as [210] to evaluate the officially released saliency maps for each SOTA method.

3.4.3 Comparison with SOTA RGB-D models

Quantitative Comparison: We provide in Figure 3.7 an overview of the average performance on conventional benchmark datasets, i.e., DES [22], NLPR [112], NJU2K [72], STERE [106], and SIP [37]. The detailed quantitative performances can be found in Tab. 3.1. We also present in Tab. 3.2 the quantitative comparison on the newly published challenging COME15K [190] dataset. All saliency maps are directly provided by authors or computed by authorized codes.

TABLE 3.1: Quantitative comparison with SOTA models. ↑ (↓) denotes that the higher (lower) is better. (**Bold**: best, Underline: second best).

Dataset	Size	DES				NLPR				NJU2K				STERE				SIP			
Metric	Mb	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$
Performance of RGB-D Models with VGG Backbones																					
$DMRA_{19}$ [113]	278	.030	.907	.900	.934	.031	.888	.899	.940	.051	.896	.886	.920	.047	.895	.886	.930	.086	.852	.806	.847
$A2dele_{20}$ [115]	116	.029	.897	.886	.917	.029	.895	.899	.943	.051	.890	.871	.914	.044	.892	.879	.926	.070	.856	.829	.887
$ATSA_{20}$ [191]	131	.022	.931	.917	.954	.027	.907	.909	.947	.046	.905	.885	.928	.038	.912	.896	**.940**	.063	.884	.849	.895
$CMMS_{20}$ [76]	546	.018	.934	.934	.958	.028	.914	.919	.946	.044	.905	.900	.929	.045	.899	.894	.925	.058	.893	.872	.901
$DANet_{20}$ [203]	128	.029	.916	.904	.932	.047	.904	.897	.926	.045	.910	.899	.927	.048	.895	.892	.919	.054	.900	.888	.912
$CMWNet_{20}$ [78]	327	.022	.939	.934	.959	.029	.913	.917	.941	.046	.913	.903	.925	.043	.911	.905	.930	.062	.889	.867	.901
$HDFNet_{20}$ [109]	308	.021	.932	.926	.962	.023	.926	.923	.957	.039	.922	.908	.939	.042	.910	.900	.933	.048	**.909**	.886	.924
$PGAR_{20}$ [14]	62	.032	.894	.886	.906	.027	.912	.917	.941	.042	.918	.909	.932	.045	.902	.894	.919	.072	.852	.838	.875
SSF_{20} [192]	126	.026	.912	.904	.930	.027	.912	.915	.947	.043	.911	.899	.929	.065	.859	.837	.882	.091	.810	.799	.855
$CASGNN_{20}$ [102]	160	.027	.917	.893	.926	.025	.914	.919	.953	.036	.927	.910	.944	**.038**	.913	.899	**.940**	-	-	-	-
$D3Net_{21}$ [37]	518	.031	.909	.897	.923	.030	.907	.912	.942	.049	.910	.900	.928	.039	.911	.902	**.940**	.063	.886	.866	.897
$CDINet_{21}$ [187]	217	.020	.943	**.937**	.962	.024	.923	.927	.953	**.030**	.928	**.918**	.945	.040	.912	**.913**	.937	.054	.904	.875	.908
$UCNet_{21}$ [188]	120	.018	.936	.934	**.970**	.025	.915	.920	.953	.043	.908	.897	.932	.039	.908	.902	.938	.051	.896	.875	.915
$DRLF_{21}$ [157]	351	.030	.909	.895	.918	.032	.904	.903	.929	.055	.896	.886	.914	.050	.897	.888	.916	.071	.869	.850	.882
$HAINet_{21}$ [77]	228	.018	**.945**	.935	.967	.024	.920	.924	.956	.037	.924	.911	.940	.040	**.917**	.907	.938	.052	.907	.879	.917
$BIANet_{21}$ [197]	189	.020	.939	.931	.955	.025	.921	.925	.954	.039	.928	.915	.939	.043	.910	.903	.932	.052	.904	.883	.916
$DCMF_{22}$ [149]	78	.022	.934	.932	.956	.029	.913	.922	.940	.041	.911	.902	.935	.043	.916	.910	.928	-	-	-	-
Ours (VGG16)	269	**.017**	.944	.929	.968	**.021**	**.927**	**.928**	**.962**	.034	**.930**	**.918**	**.947**	.039	.915	.902	.939	**.045**	**.909**	**.889**	**.927**
Performance of RGB-D Models with ResNet Backbones																					
$JLDCF_{21}$ [45]	548	.020	.934	.931	.961	.022	.925	.925	.955	.041	.912	.902	.936	.040	.913	.903	.934	.049	.903	.880	.918
$RD3D_{21}$ [12]	179	.019	.941	.935	.965	.022	.927	.930	.959	.036	.923	.916	.941	.037	.917	.911	.939	.048	.906	.885	.918
$BIANet_{21}$ [197]	244	.020	.939	.930	.958	.023	.924	.926	.956	.036	.929	.917	.942	.039	.912	.905	.935	.047	.904	.887	.920
$CoNet_{20}$ [70]	162	.024	.920	.914	.944	.027	.903	.911	.943	.046	.902	.896	.926	.037	.909	.905	.941	.058	.887	.860	.911
$DASNet_{20}$ [199]	141	.024	.926	.905	.932	.021	**.929**	.929	.960	.042	.911	.902	.935	.037	.915	.910	.939	.051	.900	.877	.918
$BBSNet_{21}$ [186]	200	.021	.942	.934	.955	.023	.927	.930	.953	.035	.931	.920	.941	.041	.919	.908	.931	.055	.902	.879	.910
DCF_{21} [69]	435	.024	.910	.905	.941	.022	.918	.924	.958	.036	.922	.912	.946	.039	.911	.902	.940	.052	.899	.876	.916
$DSA2F_{21}$ [141]	-	.021	.896	.920	.962	.024	.897	.918	.950	.039	.901	.903	.923	.036	.898	.904	.933	-	-	-	-
$DSNet_{21}$ [160]	661	.021	.939	.928	.956	.024	.925	.926	.951	.034	.929	.921	.946	.036	.922	**.914**	.941	.052	.899	.876	.910
$UTANet_{21}$ [204]	186	.026	.921	.900	.932	**.020**	.928	**.932**	**.964**	.037	.915	.902	.945	.033	.921	.910	.948	.048	.897	.873	.925
$C2DFNet_{22}$ [193]	198	.020	.937	.922	.948	.021	.926	.928	.956	-	-	-	-	.038	.911	.902	.938	.053	.894	.782	.911
$MVSalNet_{22}$ [208]	-	.019	.942	.937	**.973**	.022	.931	.930	.960	.036	.923	.912	.944	.036	.921	.913	.944	-	-	-	-
$SPSN_{22}$ [74]	149	.017	.942	.937	**.973**	.023	.917	.923	.956	.032	.927	.918	.949	.035	.909	.906	.941	.043	.910	.891	**.932**
Ours (ResNet50)	523	**.015**	**.947**	**.939**	**.973**	.022	.927	.925	.957	**.030**	**.937**	**.924**	**.952**	**.033**	**.926**	**.914**	**.948**	**.043**	**.915**	**.893**	.930
Performance of RGB-D Models with Res2Net Backbones																					
$BIANet_{21}$ [197]	244	.017	.948	.942	.972	.022	.926	.928	.957	.034	.932	.923	.945	.038	.916	.908	.935	.046	.908	.889	.922
$SPNet_{21}$ [210]	702	.014	.950	.945	**.980**	**.021**	.925	.927	.959	**.028**	.935	.925	**.954**	.037	.915	.907	.944	**.043**	.916	**.894**	**.930**
Ours (Res2Net50)	525	**.013**	**.952**	**.946**	**.980**	**.021**	**.929**	**.930**	**.961**	.029	**.939**	**.926**	**.954**	**.035**	**.921**	**.911**	**.946**	**.043**	**.919**	.892	.927

TABLE 3.2: Quantitative comparison on the challenging COME15K. The performance is evaluated with *Difficult* test set [190].

	JLDCF	*A2dele*	*DMRA*	*CoNet*	*BBSnet*	*SPnet*	*CMINet*	**Ours**
$M \downarrow$	.075	.092	.137	.113	.071	.065	.064	**.062**
$E_m \uparrow$	.870	.838	.775	.813	.876	.888	**.893**	**.893**

TABLE 3.3: Quantitative comparison with different **fusion designs**. We replace our fusion module with four SOTA fusion modules and retrain the new networks under the same training setting. We use the Mean Absolute Error (M), max F-measure (F_m), S-measure (S_m), and max E-measure (E_m) as evaluation metrics. (**Bold**: best.)

Dataset	Size	NLPR		NJU2K		STERE		SIP	
Metric	Mb	$F_\beta \uparrow$	$E_m \uparrow$	$F_\beta \uparrow$	$E_m \uparrow$	$F_\beta \uparrow$	$E_m \uparrow$	$F_\beta \uparrow$	$E_m \uparrow$
Res2Net50 + Ours	525	**.929**	**.961**	**.939**	**.954**	.921	**.946**	**.919**	.927
Res2Net50 + BBS [38]	509	.922	.953	.918	.939	.890	.909	.916	.917
Res2Net50 + CDI [187]	531	.926	.958	.927	.946	**.922**	.945	.907	.920
Res2Net50 + DCF [69]	347	.927	.958	.933	.948	.916	.939	.911	.923
Res2Net50 + SP [210]	737	.925	.959	.935	**.954**	.915	.944	.916	**.930**

Under the consideration of a fair comparison, we conduct experiments with different backbones such as VGG16 [133], ResNet50 [59], and Res2Net50 [47]. It can be seen that our HiDAnet with each backbone achieves comparable and superior performance compared to the SOTA models with the same backbone. Specifically, our HiDANet with VGG16 backbones achieves significantly better performance on NLPR and SIP datasets, while being very competitive on the model size with 269 MB and around 6 FPS. Our HiDAnet with ResNet50 backbones further sets new SOTA records on DES, NLPR, and NJU2K datasets with 523 MB and around 12 FPS. We also follow the SOTA SPNet and replace our backbone with Res2Net50. It can be seen that our method performs favorably compared to SPNet with only 525 MB compared to that of SPNet with 702 MB. Our FPS is around 11. We also exhibit in Fig. 3.9 the PR curves with several latest published models to further demonstrate the superior performance of our model.

Finally, in addition to the difference in the backbone, we observe that existing works adopt different architectures, i.e., design of decoder, supervision, training settings, etc. Under the consideration of fair comparison and to purely analyze the effectiveness of encoder fusion design, we re-implement several fusion alternatives under the same architecture (Res2Net50 + fusion). Specifically, we choose the same backbone (Res2Net50), the same decoder (the SOTA [210]), loss (multi-scale supervision), and the same training settings as ours. The only difference between one model to another is in the fusion module. The quantitative comparison can be found in Table 3.3. It can be seen that by replacing our fusion with other methods, the empirical results significantly drop. This validates the superior effectiveness of our granularity and CDA in leveraging RGB-D cues compared to other alternatives.

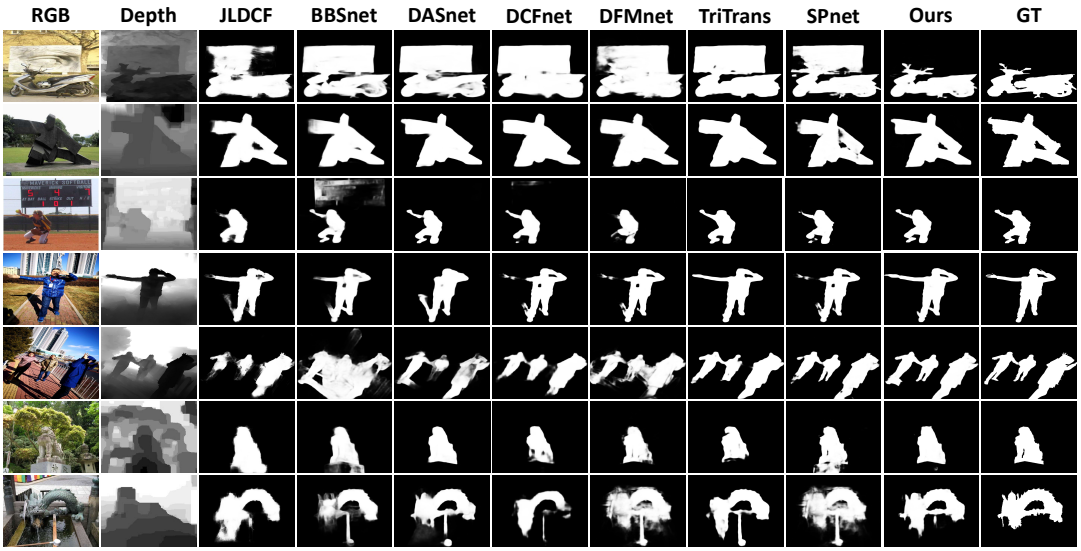


FIGURE 3.8: Visual comparison between our HiDAnet and SOTA methods in various challenging cases.

Qualitative Comparison: Fig. 3.8 illustrates generated saliency maps of different methods on challenging cases: cluttered background and foreground with a similar appearance ($1^{st} - 2^{nd}$ rows), human in scene ($3^{rd} - 5^{th}$ rows), and low contrast on the depth map ($6^{th} - 7^{th}$ rows). Compared to the SOTA models, our HiDAnet yields results closer to the ground-truth masks. For the motorbike in the 1^{st} row, our model can selectively remove the background region (board). For the sculpture in the 2^{nd} row, our network pays local attention to the foreground and thus the hollow part can be detailed. We can also accurately extract the human with large deformations in the $3^{rd} - 5^{th}$ rows.

Robustness against Depth Noise: Tab. 3.4 reports the robustness analysis on the depth quality. To make a fair comparison, we conduct experiments and compare with the SOTA SPnet [210] and CMINet [190] under the same inferior condition with a simulated Gaussian noise on depth. We further evaluate the performances on the simulated noisy testing dataset. The noise level is defined by the conventional metrics $RMSE$ and $\delta 1$. While $RMSE$ and $\delta 1$ are 0, we report the performance tested with the vanilla dataset (without noise). **Drop** Δ denotes the performance degradation by % under the simulated depth noise.

Note that CMINet designs a multi-scale mutual information minimization during the encoding stage and lately merge multi-modal features at the semantic level, yielding an unsatisfactory performance while dealing with noisy datasets (drop 2.0% S_m and 2.3% E_m for noisy DES). Differently, both SPnet and ours fuse features at each stage, leading to superior robustness against the noise. Compared to SPnet, it can be seen that our

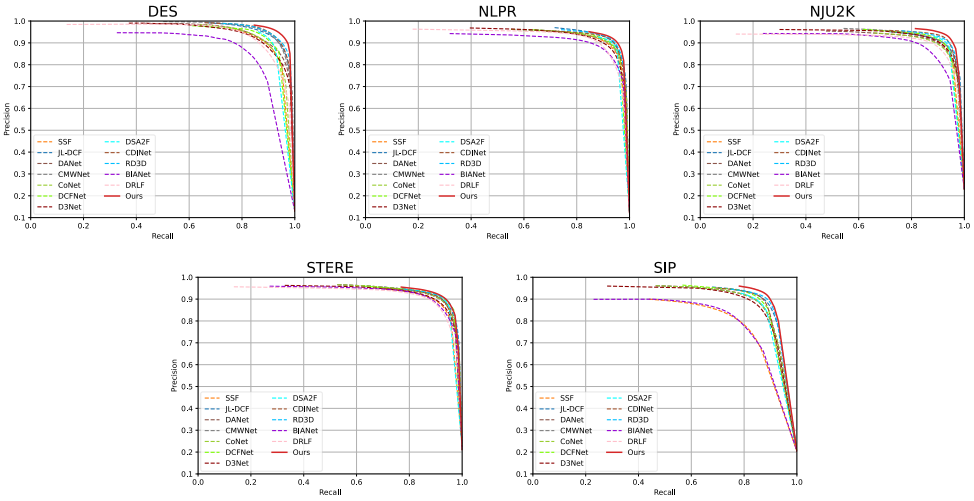


FIGURE 3.9: Comparison on **PR curves**. Our HiDANet achieves better performance compared to the 12 listed SOTA methods across different datasets.

performance is more stable, which can be attributed to our granularity attention and fusion designs. The gain of each component can be found in Tab. 3.8.

3.5 Ablation Study

Comparison with Vanilla Channel Attention. We propose granularity-aware attention (GBA) referring to geometric priors, which differs from the traditional channel attention on the pooling strategies. Formally, let $z \in \mathbb{R}^C$ be the squeezed spatial information from feature $x \in \mathbb{R}^{H\times W\times C}$. Accordingly, we can obtain three variations of average pooling:

$$(I)\ z = \frac{\sum\sum x(.)}{H\times W};\quad (II)\ z = \frac{\sum\sum x(.)\cdot m_i()}{H\times W};\quad (III)\ z = \frac{\sum\sum x(.)\cdot m_i(.)}{\sum\sum m_i(.)} \tag{3.11}$$

where (I) denotes the vanilla global average pooling, (II) is the global pooling with local region $m_i(.)$, and (III) is our proposed GBA module that applies local pooling with local region $m_i(.)$. Note that when depth data is constant, i.e., all the pixels belong to the same granularity, our local average becomes the global average pooling and our model is equivalent to the conventional channel attention [64, 121]. To verify our effectiveness, we conduct experiments by replacing our local pooling with the aforementioned poolings. Empirical results in Tab. 3.5 show that compared to (I), (II) can better leverage local awareness which spatially constrains attention around the local region. However, with a large $H \times W$, the attention activation is limited. Hence, we further propose a local pooling to automatically adjust the weight (III). Our superior performance validates the effectiveness of our local design.

TABLE 3.4: Experiments under **inferior conditions** with simulated depth noises (*RMSE*, $\delta 1$). While *RMSE*, $\delta 1$ are 0, it represents the result without simulated noises. **Drop** Δ denotes the absolute performance difference. Our HiDAnet leads to a more stable performance compared to the SOTA methods with a lower Δ under different inferior conditions, proving that our model is more robust against depth noises. We use the Mean Absolute Error (M), max F-measure (F_m), S-measure (S_m), and max E-measure (E_m) as evaluation metrics. (**Bold**: best.)

Dataset	DES						NLPR						NJU2K					
Metric	*RMSE*	$\delta 1$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	*RMSE*	$\delta 1$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	*RMSE*	$\delta 1$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$
CMINet$_{21}$	0	0	.016	.944	.940	.975	0	0	.020	.931	.932	.959	0	0	.028	.940	.929	.954
CMINet$_{21}$	.261	.270	.022	.925	.920	.952	.259	.342	.021	.929	.932	.960	.236	.413	.032	.934	.922	.948
Drop Δ(%)	-	-	.6	1.9	2.0	2.3	-	-	.1	0.2	**0**	**.1**	-	-	0.4	0.6	.7	.6
SPNet$_{21}$	0	0	.014	.950	.945	.980	0	0	.021	.925	.927	.959	0	0	.028	.935	.925	.954
SPNet$_{21}$	.261	.270	.017	.944	.935	.972	.259	.342	.020	.922	.924	.956	.236	.413	.033	.931	.920	.946
Drop Δ(%)	-	-	.3	.6	1	.8	-	-	.1	.3	.3	.3	-	-	.5	**.4**	.5	.8
Ours	0	0	.013	.952	.946	.980	0	0	.021	.929	.930	.961	0	0	.029	.939	.926	.954
Ours	.261	.270	.015	.948	.943	.980	.259	.342	.021	.930	.930	.962	.236	.413	.029	.935	.925	.953
Drop Δ(%)	-	-	**.2**	**.4**	**.3**	**0**	-	-	**0**	**.1**	**0**	**.1**	-	-	**0**	**.4**	**.1**	**.1**

TABLE 3.5: Ablation study on attention designs with different average pooling methods.

#	*Description*	DES		NLPR		NJU2K		STERE	
		$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$
I	Global Attention + Global Pool	.019	.940	**.020**	**.929**	.030	.936	.037	.918
II	Local Attention + Global Pool	.015	.947	.021	.927	.032	.928	.038	.915
III	Local Attention + Local Pool (ours)	**.013**	**.952**	.021	**.929**	**.029**	**.939**	**.035**	**.921**

Why GDA in both streams: We analyze in Tab. 3.6 the contribution of GBA for both RGB and Depth feature modelings: (A) We remove the GBA from our network, denoted as RGB+D; (B) GBA is only applied in the RGB stream, denoted as RGB(G) + D; (C) GBA is applied in both streams, denoted as RGB(G) + D(G). We observe that the performance augments by gradually inserting GBA into the encoders. This shows that GBA can be considered as depth-aware attention for the RGB stream and as a self-enhancement module for the Depth stream to produce regions with favorable objectness.

Number of Otsu Regions for GBA: Our fine-grained details are determined by the number of Otsu regions as shown in Figure 3.10. The two first columns represent the paired RGB-D inputs. On the 3^{rd}, 5^{th}, and 7^{th} columns we list the Otsu regions with different numbers of multi granularities, respectively. On the 4^{th}, 6^{th}, and 8^{th} columns we list the generated masks with different numbers of thresholds $T = 1, 2, 3$, respectively.

By comparing the 3^{rd} and 5^{th} columns, it can be seen that a small number of Otsu threshold $T = 1$ cannot get the full benefit from the geometric priors. For example, the building in the 1^{st} row cannot be perfectly distinguished from the background;

Table 3.6: Ablation of GBA module. Experiments by gradually adding GBA module on RGB and Depth streams. RGB(G)/D(G) denotes the case when granularity attention is applied to RGB/Depth branch.

Dataset	DES				NLPR				NJU2K				STERE				SIP			
Metric	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$
(A) RGB + D	.015	.949	.940	.972	.022	.925	.927	.960	.030	.932	.923	.952	.037	.913	.901	.936	.046	.914	.889	.923
(B) RGB(G) + D	.014	.951	.943	**.980**	**.021**	.927	.926	.960	.030	.936	.923	.953	.036	.916	.907	.945	**.043**	**.919**	**.894**	**.928**
(C) RGB(G) + D(G)	**.013**	**.952**	**.946**	**.980**	**.021**	**.929**	**.930**	**.961**	**.029**	**.939**	**.926**	**.954**	**.035**	**.921**	**.911**	**.946**	**.043**	**.919**	.892	.927

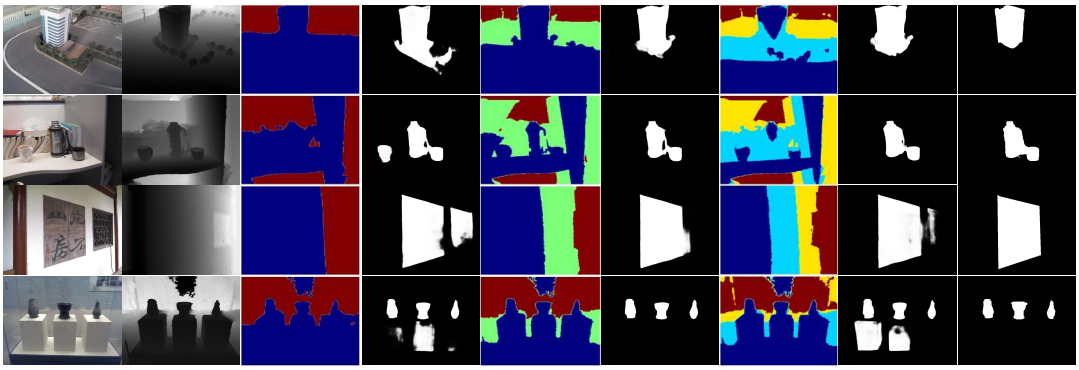


FIGURE 3.10: Qualitative comparison with **different numbers of Otsu** thresholds ($T = 1, 2, 3$) for our granularity-based attention. With the threshold T, we divide the depth map into $T + 1$ regions with different colors. Each region shares the same granularity of geometric information. With one threshold $T = 1$, the local regions are coarse and cannot get the full benefit from the geometric priors. This results in unsatisfactory salient masks (4^{th} column). With two thresholds $T = 2$, the depth map is better discretized with more fine-grained details, yielding salient masks closer to the ground truth (6^{th} column). With three thresholds $T = 3$, the depth map is over-discretized, resulting in sub-optimal salient masks (8^{th} column). Our plain HiDAnet is built upon $T = 2$.

the cups in the 2^{nd} row are mixed with the table and a part of the wall. The unsatisfactory thresholding on the depth histogram leads to sub-optimal performance of granularity-based attention that the discriminatory power cannot be fully exploited. While augmenting the number of thresholds to $T = 2$, we observe from the 5^{th} column that the scene can be better discretized. The fine-grained details contribute to the clearer boundary generation as shown in the 6^{th} column. We further augment the number of thresholds to $T = 3$ and observe the over-discretization, leading to the misunderstanding on the depth map. Thus, it results in lower quality salient masks as shown in the 8^{th} column.

Thus, we perform the experiments with different numbers of thresholds T. Tab. 3.7 shows that the best overall performance is achieved with $T = 2$ thresholds, thus with $n = 3$ regions. It can be considered as a scene discretization into three parts: close, middle, and far regions. Our plain HiDAnet is with $T = 2$ thresholds and achieves the best performance. We also discover that the sensitivity to thresholding varies from one dataset to another, especially the NLPR dataset which is not highly sensitive to the granularity. This is mainly due to the fact that NLPR contains objects residing in the background. In such circumstances, the target object has the mixed depth response as the background, leading to less-noticeable granularity as shown in the last two rows of Figure 3.10. In more common and popular cases (DES, NJU2K, STERE,

and SIP), our fine-grained details achieve significant improvement compared to our baseline with conventional attention as shown in Tab. 3.7.

Ablation study on Key Components: Tab. 3.8 presents a thorough ablation study for each key component. We observe that by gradually adding proposed modules, our network leads to better performance. We also conduct experiments by replacing our proposed modules with several SOTA counterparts. Specifically, we compare our Granularity-Based Attention with the DEDA module proposed in [203]. Both our GBA and DEDA belong to the mask-guided attention modules. Specifically, DEDA leverages the depth map to dynamically learn the masked-guided attention map which is supervised by the ground truth. The learned attention map refers to the contrast to guide RGB learning. Differently, our mask is statically computed by the Otsu threshold by maximizing inter-class variance. The computed local regions refer to the fine-grained details which are further integrated with semantics cues. Empirically, by comparing ($\#6 - \#8$), our GBA performs favorably against DEDA, showing that our method can better leverage the depth cues to distinguish objects with different camera distances. We also replace our encoder fusion (CDA) with the concurrent DCF [69] built upon channel attention. The main difference is that DCF is based on channel attention, while our CDA additionally leverages the spatial attention for better localization. By comparing ($\#7 - \#8$), we can observe that while CDA is replaced by the DCF, the performance drops significantly. This validates the effectiveness of our CDA with both channel and spatial attention.

Design of Cross Dual Attention: We verify in Tab. 3.9 the design of our encoder fusion by removing or replacing each component: ($C1$) Features are simply fused through addition; ($C2$) Features are fused through concatenation-convolution (**CC**); ($C3$) Features are firstly self-enhanced with vanilla CBAM before the addition fusion. ($C4$) Features are firstly self-enhanced and later fused through **CC**. ($C5$) We explore the attention in a cross manner and fuse features with addition. We can observe the gain of attention modules by comparing ($C1 - C3 - C5$), the improvement from cross-domain interaction by comparing ($C3 - C5$), and the contribution of **CC** by comparing ($C5 - Ours$). These results validate the effectiveness of our proposed encoder fusion scheme.

Design of Efficient Multi-Input Fusion: We also verify the design of our decoder fusion in Tab. 3.9: ($E1$) Features are fused with **CC**. ($E2$) Features are concatenated and fed into the ECA model before the convolution. ($E3$) Features are fused with CC and then fed into the ECA. ($E4$) Based on the configuration $E2$, we further add a residual addition. By comparing ($E2 - E3$) and ($E4 - Ours$), we can observe that the ECA module performs better with a reduced channel size. The comparison

TABLE 3.7: Ablation study on the Ostu number. We present the quantitative comparison with different Otsu thresholds. Our plain HiDAnet is with $T = 2$ thresholds. $T = 2$ achieves the best performance with a reasonable FPS.

Dataset		DES				NLPR				NJU2K				STERE				SIP			
Metric	$FPS\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$
$T = 0$	**13.3**	.019	.941	.927	.955	.020	**.929**	**.931**	**.961**	.031	.936	.924	.952	.037	.919	.908	.943	.046	.915	.888	.924
$T = 1$	12.6	.015	.951	**.948**	.979	.023	.927	.927	.960	**.029**	.933	.924	.953	.035	.918	.908	.944	.044	.918	**.894**	**.927**
$T = 2$	11.3	**.013**	**.952**	.946	**.980**	.021	**.929**	.930	**.961**	**.029**	**.939**	**.926**	**.954**	**.035**	**.921**	**.911**	**.946**	**.043**	**.919**	.892	**.927**
$T = 3$	10.5	.015	.949	.942	.979	**.020**	.929	.928	**.961**	.031	.929	.920	.949	.036	.914	.900	.940	.044	.916	.891	.925

TABLE 3.8: Ablation study on key components of HiDAnet. We partially remove key components or replace the fusion designs with simple addition.

#	Baseline	*GBA*	*CDA*	*Skip*	*EMI*	$\mathcal{L}_{ml}$	*DEDA* [203]	*DCF* [69]	DES		STERE	
									$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$
1	✓								.018	.941	.038	.917
2	✓	✓							.016	.944	.037	.917
3	✓	✓	✓						.016	.946	.036	.919
4	✓	✓	✓	✓					.015	.947	.036	**.923**
5	✓	✓	✓	✓	✓				.014	.949	**.034**	.921
6	✓			✓	✓	✓	✓		.016	.946	.041	.914
7	✓	✓		✓	✓	✓		✓	.017	.946	.037	.918
8	✓	✓	✓	✓	✓	✓			**.013**	**.952**	.035	.921

TABLE 3.9: Ablation study on encoder fusion and decoder fusion designs. CC stands for Concatenation-convolution. S stands for Self-interaction. R stands for residual connection.

Dataset	# Descrip.	C1 Add	C2 CC	C3 S+C1	C4 S+C2	C5 Cross+C1	E1 CC	E2 Middle	E3 Later	E4 E2+ R	**Ours**
DES	$M\downarrow$	.017	.016	.015	.014	.015	.015	.016	.015	.014	**.013**
	$F_\beta\uparrow$	.945	.946	.948	.949	.947	.947	.945	.949	.950	**.952**
STERE	$M\downarrow$	.039	.039	.036	.037	.036	.038	.038	.037	.036	**.035**
	$F_\beta\uparrow$	.915	.916	.918	.917	.919	.914	.915	.916	.920	**.921**

between ($E2-E4$) validates the effectiveness of residual addition which propagates the hierarchical features.

3.6 Conclusion

In this chapter, we propose an end-to-end HiDAnet for RGB-D saliency detection. Different from previous networks, we fully leverage fine-grained details and merge them with semantic cues through local channel attention. Extensive evaluations on challenging RGB-D benchmarks indicate that our HiDAnet improves saliency detection in several challenging scenarios where the SOTA approaches fail, notably in cases where multiple objects with similar appearances but at distinct camera distances (granularity). In addition to the channel axis, the spatial direction also plays an important role in CNN. Therefore, in the following chapter, we will discuss how to leverage the depth cues to better design non-local depth-adapted spatial attention.

Chapter 4

Depth As Offset - A Novel Spatial Attention For CNN

In the previous chapter, we have discussed the depth-wise channel attention to improve the discriminability of CNN with respect to geometric priors. In addition to the channel axis, spatial information is also important which helps to precisely localize the object on the image. Hence, in this chapter, we present how can we leverage the depth cues to design a local and deformable depth-adapted spatial attention. We validate our approach on both semantic segmentation and saliency detection tasks. It is worth noting that few existing methods explicitly leverage the contribution of depth cues to adjust the sampling position on RGB images. Therefore, we propose a novel framework to incorporate the depth information in the RGB convolutional neural network (CNN), termed Z-ACN (Depth-Adapted CNN)r. Specifically, our Z-ACN generates a 2D depth-adapted offset which is fully constrained by low-level features to guide the feature extraction on RGB images. With the generated offset, we introduce two intuitive and effective operations to replace basic CNN operators: depth-adapted convolution and depth-adapted average pooling. Extensive experiments on semantic segmentation and saliency detection tasks demonstrate the effectiveness of our approach.

4.1 Application in Semantic Segmentation

4.1.1 Introduction

As one of the fundamental tasks in computer vision, semantic segmentation aims to understand the pixel-wise label from an input image of a generic target scene. Recent advances in deep neural networks, as well as the GPU, have set new state-of-the-art (SOTA) performance in semantic segmentation. Despite significant progress in the last decade, semantic segmentation based on RGB input remains challenging in many challenging scenarios, i.e., low-contrast light, object occlusion, and separating objects sharing a similar visual appearance.

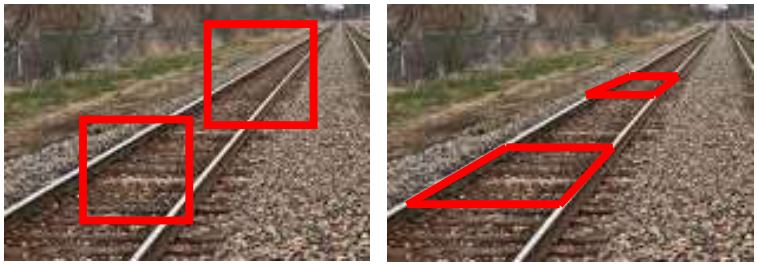


FIGURE 4.1: A sketch of Depth-Adapted Sampling position. We explicitly leverage the depth priors to compute a locally-deformable sampling position, yielding a simple but efficient manner to introduce a global-local attention into CNN.

Recent developments in RGB-D sensors make RGB-D inputs accessible at a low cost, motivating research interests in designing various fusion strategies to merge multi-modal features. A number of works have demonstrated the benefit of spatial cues to improve the accuracy of semantic segmentation, affirming the effectiveness of learning from complementary modalities. In the literature, two main designs have been widely exploited: single-stream design and two-stream design. The single-stream often realizes an early fusion where RGB and depth images are simply concatenated from the input side. Different from conventional RGB networks with 3-channel, several works merge multi-modal inputs at the channel axis to form a 4-channel input (RGB-D) or 6-channel input (RGB-HHA where HHA is encoded from depth referring to disparity, height above ground, and norm angle). However, these networks directly extract features from early-mixed modalities that cannot fully explore the correlation between RGB and depth images. The two-stream strategy adopts parallel encoders that extract multi-modal features separately and further fuse them at different semantic levels. Nevertheless, compared to single-stream networks, two-stream designs inevitably increase the computational cost. Furthermore, the fusion mechanism is often pre-defined that cannot adapt to different scenarios without handcraft adjusting.

In this chapter, we explore differently the relationship between RGB and depth by explicitly leveraging the perspective effect. Recent non-local attention [28, 64, 98, 147, 156, 161] works in vision tasks have proved their effectiveness in modeling contextualized awareness. Several recent works [24, 99, 108] further figure out that aggregating both global and local attention can lead to better performance since neighboring pixels tend to have high similarity and correlation. Sharing the same idea, we seek to improve CNN with local but contextualized awareness which is fully constrained by the geometry. As shown in Figure 4.1(a), the conventional convolution is designed to have a regular and fixed structure on the image plane. With additional priors on

camera parameters and depth cues, we show that the sliding windows can adapt to the geometry, e.g., the vanishing effect as shown in Figure 4.1(b).

Inspired by this observation, we develop a Depth-Adapted Convolutional Network, denoted Z-ACN. Z stands for the Z-axis of the camera coordinates representing the depth information. Specifically, we propose a depth-adapted offset that can be integrated into basic functions of CNN, i.e, convolution and pooling, and introduce two new operators: depth-adapted convolution and depth-adapted average pooling.

Our proposed depth-adapted convolution replaces conventional neighboring pixels with geometrically similar ones. Concretely, we reshape the receptive field to cover pixels sharing the same 3D plane with the center of the kernel, yielding a simple but efficient manner to articulate both photometric and geometric information. The second introduced operator is depth-adapted average pooling. Sharing the same idea as the depth-adapted convolution, we re-define the notion of neighboring pixels for average pooling such that the geometrical relations will be considered while computing the mean of the local region of the feature map. Both operators break the limits on the conventional definition of neighboring pixels, forcing the network to pay attention to a larger and more malleable receptive field.

Depth-adapted operations are based on the intuition that pixels with the same geometrical character should be more likely to share the same semantic label. One common example is the vanishing effect, as illustrated in Figure 4.1. We assume that pixels on the same 3D plane tend to share the same class. This 3D plane and depth variance have a high correlation. As shown in Figure 4.1, we display the projection of the 3D plane of the rail on the image plane as the adapted sampling position. The depth-adapted field should be more correlated to the real scene compared to the conventional neighboring field. Essentially, our method uses depth to transform planes into a canonical pose relative to the camera, such that the extracted feature maps are also in a canonical reference frame and thus invariant to scale changes and out-of-plane rotation. The main advantages of such operations are summarized as follows:

- We propose a novel depth-adapted convolutional network termed Z-ACN, that can integrate the geometric constraint into the conventional receptive field, hence improving the convolution with depth-aware contextualized attention.
- Our grid adaptation is processed by the non-learning method that does not introduce extra learning parameters compared to conventional counterparts.
- Experiments on both indoor and outdoor RGB-D semantic segmentation benchmarks demonstrate that our method can perform favorably over the baseline performance with large margins and set the new state-of-the-art performance.

4.1.2 Related Work

4.1.2.1 3D Representation

We have discussed in previous sections the success of PointNet in dealing with 3D data. Since then, different 3D CNN methods are trying to adapt to the irregularity of the point cloud. [82] integrates an x-transformation to leverage the spatially-local correlation of point cloud [146] introduces a spatially deformable convolution based on kernel points to study the local geometry. [96] learns the mapping from geometry relations to high-level relations between points to get a shape awareness. [95] defines convolution as an SLP (Single-Layer Perceptron) with a nonlinear activator.

Besides, a number of efforts have been made to reduce the model complexity. [145] adapts CRF (Conditional Random Fields) to reduce the model parameters. Multi-view methods [16, 48, 75, 117] reform 3D CNN to become the combination of 2D CNNs. [16] profits from Lidar to get bird-view and front-view information in addition to a traditional RGB image. [48] uses depth image to generate the 3D volumetric representation after which projections on X, Y, and Z planes are learned respectively by 2D CNNs. 3D CNN achieves better results than RGB CNN but requires further development on problems such as memory cost, data resolution, and computing time.

4.1.2.2 2D RGB-D Fusion

Through the development of RGB-D fusion models, it can be seen that there exist a number works that aim to guide the feature extraction with enhanced depth awareness. D-CNN [154] enhances the network with a depth similarity term which re-weight the standard convolution with the depth-related local context. Since then, various works have been developed on the forms of weight functions. [18] extends the idea of [154] to dilated convolution. [175, 176] develop 2.5 D convolutions with a more generalized weight function. [23] projects 3D convolution on 2D images to form a depth-aware multi-scale 2D convolution. [177] uses depth information to define local neighborhoods by introducing a learned Gaussian kernel. Sharing the same idea of re-weighting the convolution, ShapeConv [7] integrates the channel attention into the convolution function and forms a more generalized convolution that is not limited to RGB-D context.

It can be seen that contextualized awareness has played a vital role in RGB-D fusion. For two-stream designs, multi-modal features are often firstly fed into attention module before the data fusion: [85] with ConvLSTM modules, ACNet [65] with channel attention [64], [177] with a learned Gaussian convolution kernel, and [15] with a modified CBAM [161]. For single-stream design, the contextualized awareness is directly integrated into the basic convolution function to re-calibrate the filter weight: [154]

with depth similarity, [176] with a malleable depth-aware function, and [7] with channel attention. Despite the popularity of attention modules in previous works, the capability of modeling long-range dependencies is still limited due to the fixed shape of the convolutional receptive field, i.e., within the 8 neighboring pixels for a conventional 3×3 convolution. In contrast, we propose a depth-adapted sampling position to explicitly leverage both global and local awareness in a simple yet efficient manner. By designing a geometry-constrained offset, we aim to break the conventional receptive field to adapt to the perspective effect, yielding an effective depth-guided 2D CNN to improve the RGB understanding.

4.1.2.3 Non-local Adaptive Model

In previous chapters, we have briefly reviewed several non-local networks. Despite the demonstrated promising results, we observe that the contextualized awareness are or learned through gradient descent, e.g., the positional encoding in CPVT and the offset in Deformable works, or learned through a pre-defined large receptive field, e.g., global attention in transformer and large kernel size in ConvNext. In the case of multi-modal feature learning, we seek to compute the global awareness from the additional prior. This perspective has been widely studied in the field of spherical images where the global attention is computed according to the distortion priors [25, 27, 41, 143]. Inspired by these works, we propose to compute the non-local awareness from the depth priors, making the convolution geometry-aware for RGB-D semantic segmentation. A concurrent work SConv [11] learns the offset from depth image. Our approach resembles the SConv in that both methods belong to depth-adapted convolution frameworks. However, one main difference is that our offset is purely defined by the geometric without requiring any gradient descent, while SConv applies convolutional layers to learn the offset from latent space. Our approach explicitly leverages the scale changes along the Z-axis of camera coordinates and out-of-the-plane rotation. Instead of adding extra learning parameters, we show that a simple and intuitive local deformation can contribute to semantic segmentation with minimal cost.

4.1.3 Depth-Adapted Convolutional Network

In this section, two depth-adapted operations are presented: depth-adapted convolution and depth-adapted average pooling. Figure 4.2 shows the information propagation in our network.

First, we take a 2D conventional regular and fixed area on the depth map, which corresponds to a conventional receptive field, e.g., 3×3 convolution. We back-project the pixels to the 3D scene to get the 3D position in the camera coordinate. Second,

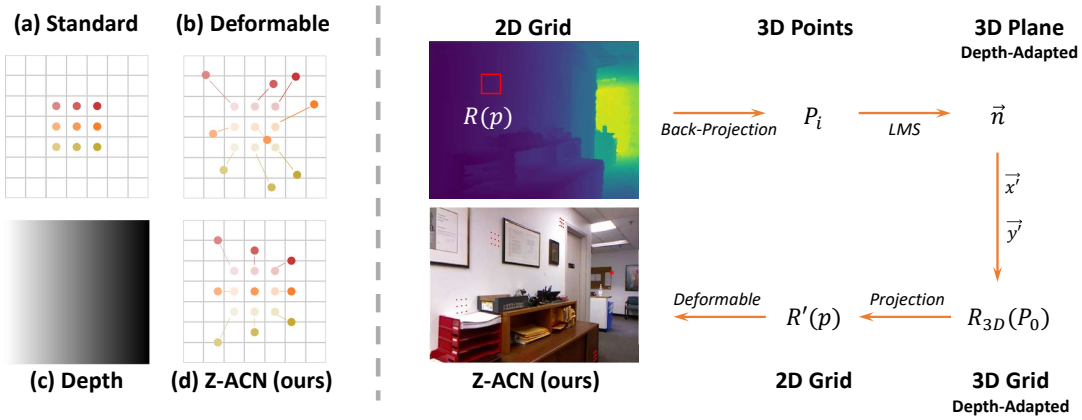


FIGURE 4.2: On the left we show the example of a 3 × 3 kernel: a) shows a standard 2D convolution with dilation equal to 1. b) shows the offset computed from deformable convolution [28]. c) is the available depth data. The represented figure shows a linear change with the depth value. From left to right, the scene becomes deeper. d) illustrates offset computed by Z-ACN which is adapted to depth. On the right, we illustrate the overview of our approach. **LMS** stands for Least Mean Square algorithm. $(\vec{x'}, \vec{y'})$ are the 3D unit axis. Firstly, pixels within the 2D receptive field are back-projected to 3D space to form a point cloud, based on which a 3D plane is computed with normal $\vec{n}$. Secondly, a new 3 × 3 grid on the 3D space is created with the help of 3D axis $(\vec{x'}, \vec{y'})$ which are perpendicular to the normal $\vec{n}$. Finally, the 3D grid is projected to the image plane, forming our depth-adapted sampling position. Zoom in for more details on the depth-guided sampling position on the RGB image.

we compute a depth-aware plane that passes through the real-world position of the kernel center and fits the best to all 3D points. Third, we create a 3D regular grid on this plane with an adapted orientation to fit the geometry. Last, we project this 3D grid on the image plane to form a 2D depth-adapted sampling grid.

Our model requires 2 inputs: input feature map and depth map (ground truth or estimated). The feature map is denoted as $\mathbf{x} \in \mathbf{R}^{c_i \times h \times w}$, where c_i is the number of input feature channel, h and w are the height and weight of the input feature map. The depth map is denoted as $\mathbf{D} \in \mathbf{R}^{h \times w}$. $\mathbf{D}$ is used to adapt the spatial sampling locations by computing the offset, denoted as $\Delta p \in \mathbf{R}^{c_{off} \times h_1 \times w_1}$, where h_1 and w_1 are the height and weight of the output feature map and $c_{off} = 2 \times N \times N$ for a $N \times N$ filter. Different from Deformable ConvNet, our offset does not require gradient during back-propagation. The output feature map is denoted as $\mathbf{y} \in \mathbf{R}^{c_o \times h_1 \times w_1}$, where c_o is the number of output feature channel.

4.1.3.1 Depth-Adapted Convolution

A standard image convolution is formulated as:

$$\mathbf{y}(p) = \sum_{p_n \in \mathbf{R(p)}} \mathbf{w}(p_n) \cdot \mathbf{x}(p + p_n), \tag{4.1}$$

where $\mathbf{w}$ is the weight matrix. $\mathbf{R(p)}$ is the grid for point p. Physically it represents a local neighborhood on input feature map, which conventionally has regular shape with certain dilation Δd, such that :

$$\mathbf{R(p)} = a\vec{u} + b\vec{v} \tag{4.2}$$

where $(\vec{u}, \vec{v})$ is the pixel coordinate system of input feature map and $(a, b) \in (\Delta d \cdot \{-1, 0, 1\})^2$.

To exploit the 3D planarity, depth-adapted convolution simply adds an adapted deformation term Δp to adjust the spatial sampling locations :

$$\mathbf{y}(p) = \sum_{p_n \in \mathbf{R(p)}} \mathbf{w}(p_n) \cdot \mathbf{x}(p + p_n + \Delta p_n) \tag{4.3}$$

The convolution may be operated on the irregular positions $p_n + \Delta p_n$ as the offset Δp_n may be fractional. To address the issue, we use the bilinear interpolation which is the same as that proposed in [28]. In the following subsections, we will present how to process this offset from traditional computer vision algorithms.

4.1.3.2 3D Planarity

To compute the offset, firstly we assume that the camera fits the pinhole model. Therefore, with the camera parameters, we can back-project 2D pixels within the conventional field $\mathbf{R(p)}$ into camera coordinates, forming the 3D point cloud $P_i = (X_i, Y_i, Z_i)$. An analysis of the intrinsic parameters is presented in Section 4.1.5. Let $p = (u_0, v_0)$ be the center of 2D receptive field and $P_0 = (X_0, Y_0, Z_0)$ the associated back-projection on 3D space. The plane π passing through P_0 and fitting the best to all P_i can be extracted by applying the least square method:

$$\overrightarrow{n} = \arg \min_{\substack{\overrightarrow{n}=(n_1,n_2,n_3) \\ ||\overrightarrow{n}||=1}} \sum_i ||\overrightarrow{n} \cdot \overrightarrow{P_0 P_i}||^2 \tag{4.4}$$

where $\overrightarrow{n} = (n_1, n_2, n_3)$ is an approximation of the normal of the plane π.

Basing on the plane π, we build a new planar and regular grid, denoted as $R_{3D}(P_0)$, which is centered on P_0. The regular shape is defined by an orthonormal basis $(\vec{x'}, \vec{y'})$ on the plane π. We fix $\vec{x'}$ as horizontal ($\vec{x'} = (\alpha, 0, \beta)$). As $\vec{x'}$ is on the plane π defined by its normal $\vec{n} = (n_1, n_2, n_3)$ and $(\vec{x'}, \vec{y'})$ are the orthonormal basis, we have :

$$\vec{x'} \cdot \vec{n} = 0; \quad ||\vec{x'}||^2 = 1; \quad ||\vec{n}||^2 = 1; \quad \vec{n} \times \vec{x'} = \vec{y'}. \tag{4.5}$$

Analytically, we can obtain $(\vec{x'}, \vec{y'})$ as follow:

$$\vec{x'} = \begin{bmatrix} \frac{n_3}{\sqrt{1-n_2^2}} \\ 0 \\ -\frac{n_1}{\sqrt{1-n_2^2}} \end{bmatrix}, \quad \vec{y'} = \begin{bmatrix} -\frac{n_1 n_2}{\sqrt{1-n_2^2}} \\ \sqrt{1-n_2^2} \\ -\frac{n_2 n_3}{\sqrt{1-n_2^2}} \end{bmatrix} \tag{4.6}$$

To conclude, $R_{3D}(P_0)$ is defined as :

$$R_{3D}(P_0) = a\vec{x'} + b\vec{y'} \tag{4.7}$$

with $(a, b) \in (-k_u, 0, k_u) \times (-k_v, 0, k_v)$ where (k_u, k_v) are scale factors.

A conventional 2D convolution on the image plane can be considered as realizing a planar convolution on a fronto-parallel plane on the 3D camera basis. While the depth value is constant, our depth-adapted plane $\overrightarrow{n}$ becomes the same as the fronto-parallel plane. Otherwise, our plane $\overrightarrow{n}$ can better explore the perspective effect compared to the counterpart, yielding a depth adapted sampling position $R_{3D}(P_0)$ in the camera basis.

4.1.3.3 Scale Factor

The scale factors are designed to be constant such that the 3D receptive field of each point from the feature map has the same size. In such a way, with the variance of depth, due to the perspective effect, the projected 2D receptive field on the image plane will have different sizes. The value of scale factors can be empirically set in different tasks. In our application, we want the adapted convolution performs the same as a conventional 2D convolution on a particular point $p(u_0, v_0)$ whose associated plane in Eq. 4.4 is fronto-parallel $\{Z|Z = Z_0\}$. By taking into account the dilation Δd and the camera focal length (f_u, f_v), we have:

$$\begin{aligned} k_u &= \Delta d \times \frac{Z_0}{f_u} \\ k_v &= \Delta d \times \frac{Z_0}{f_v}. \end{aligned} \tag{4.8}$$

4.1.3.4 Depth-Adapted Sampling Position

To form the depth-adapted sampling position, we denote **R'(p)** as the projection of $R_{3D}(P_0)$ on the image plane :

$$\begin{aligned} \mathbf{y}(p) &= \sum_{p_n \in \mathbf{R'(p)}} \mathbf{w}(p) \cdot \mathbf{x}(p + p_n) \\ &= \sum_{p_n \in \mathbf{R(p)}} \mathbf{w}(p) \cdot \mathbf{x}(p + p_n + \Delta p_n). \end{aligned} \tag{4.9}$$

Different from the conventional grid **R(p)**, the newly computed **R'(p)** breaks the regular size and shape structure with the additional offset. In such a way, the geometry information is incorporated in RGB CNN.

4.1.3.5 Depth-Adapted Average Pooling

A standard average pooling is defined as :

$$\mathbf{y}(p) = \frac{1}{|\mathbf{R(p)}|} \sum_{p_n \in \mathbf{R(p)}} \mathbf{x}(p + p_n). \tag{4.10}$$

This treats every pixel equally regardless of its associated geometry information, e.g. whether they belong to the same plane or not. To address this issue, similar to depth-adapted convolution, we add an extra offset to adjust the pooling field to the geometry. We force pixels sharing the same plane to contribute more to the corresponding output. For each pixel location p, the depth-adapted average pooling operation becomes :

$$\mathbf{y}(p) = \frac{1}{|\mathbf{R(p)}|} \sum_{p_n \in \mathbf{R(p)}} \mathbf{x}(p + p_n + \Delta p_n). \tag{4.11}$$

4.1.3.6 Understanding Depth-Adapted operations

In Figure 4.2 we show several examples of depth-adapted sampling positions of given input neurons (the center) on an RGB image. We seek to profit from the depth cues to articulate both photometric and geometric information for RGB CNN. Our method integrates the geometry into the convolution by adjusting the 2D sampling grid. This pattern is integrated into Eq. 4.3. In the case of conventional CNN, the shape of the grid is fixed as regular, which has difficulty adapting to the perspective effect. With the proposed Z-ACN, we can better leverage the geometric constraint in the sampling position. As shown in Figure 4.2, the receptive field for a closer input neuron in the 3D space is larger than that of a geometrically farther neuron. Sampling positions on the same plane also have different shapes that are adapted to the camera-projection effect. These patterns improve 2D CNN's performance with contextualized awareness without complicating the network with extra learning parameters.

4.1.4 Experiments

4.1.4.1 Experimental setup

Dataset and metrics. We evaluate the effectiveness of our approach on both indoor and outdoor RGB-D semantic segmentation benchmarks, including NYUv2 dataset [132], SUN RGBD dataset [135] and KITTI dataset [50]. For the NYUv2 dataset, it contains 1,449 RGB-D images which are split into 795 training images and

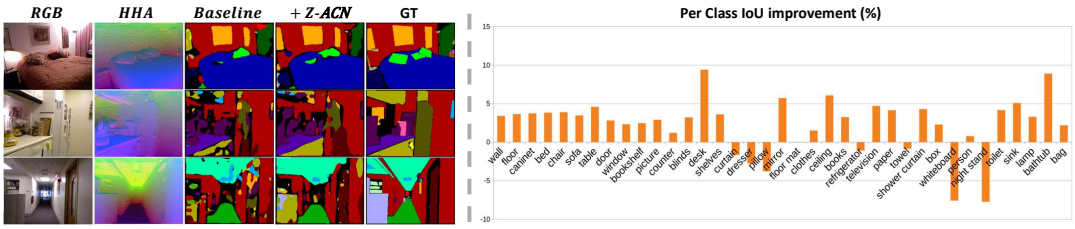


FIGURE 4.3: On the left we illustrate the qualitative comparison on the NYUv2 dataset. The two first columns are the input RGB and HHA, respectively. *Baseline* represents the semantic map obtained with early fused RGB-HHA input. + Z-ACN stands for the results obtained by inserting our depth-adapted sampling position into the baseline. It can be seen that by explicitly leveraging non-local attention, our method reasons about semantic maps closer to the ground truth (GT). The black regions in "GT" are the ignoring category. On the right we illustrate the per-class improvement above the baseline. We improve 29/37 classes with 5.2% mean IoU increment.

654 testing images. For SUN-RGBD, it contains 37 categories of objects and consists of 10,335 RGB-D images which are split into 5,285 training images and 5,050 testing images. For the KITTI dataset, we use the semantic segmentation annotation provided in [179], which contains 70 training and 37 testing images from different scenes, with high-quality pixel annotations in 11 categories. The performance is evaluated with common metrics, i.e., Pixel Accuracy (PixelAcc), Mean Accuracy (mAcc.), Mean Region Intersection Over Union (mIoU), and Frequency Weighted Intersection Over Union (f.w.IoU).

Implementation details. Our approach requires paired RGB-D images as input. The depth map is first used to generate the geometry-aware offset which is further integrated into the network. As HHA encoding, the offset generation can be also realized during pre-processing since our method does not require gradient descent. We follow the same learning settings for both our proposed network and the baseline counterpart. Experiments are realized with 2 Nvidia V100 GPUs under the PyTorch framework. During inference, we apply a single-scale inference strategy.

Comparison protocol. We evaluate the generalization capability of our approach with different backbones, including old-fashioned VGG-16 encoders and popular ResNet encoders. We seek to demonstrate that our approach can constantly improve the baseline performance. To purely analyze the gain by applying our approach, we only replace the vanilla convolution and average pooling with our proposed depth-adapted operators.

4.1.4.2 With VGG-16 backbone

Comparison with D-CNN [154]: D-CNN is the pioneering work that integrates depth into the basic operations of convolutional networks. The depth is used to

TABLE 4.1: Comparison with the concurrent D-CNN [154] which uses depth to re-design the convolutional weight. Both models are trained from scratch with the same training settings. Our method achieves better performance under different datasets, show that the depth priors are better exploit with our Z-ACN.

Dataset	NYUv2			SUN-RGBD		
Method	RGB	D-CNN	**Z-ACN**	RGB	D-CNN	**Z-ACN**
PixelAcc (%)	50.1	60.3	**73.5**	66.6	72.4	**78.4**
mIoU (%)	15.9	27.8	**28.4**	22.8	29.7	**30.5**

compute a similarity term to re-calibrate convolutional weight. We build our approach upon the DeepLab architecture, which is the same as D-CNN [154]. Note that both the D-CNN and our approach belong to the depth-aware convolution framework. Unlike the D-CNN model, we update the depth information to break the limitation of a fixed structure, which can better leverage long-range dependencies while D-CNN seeks to refine the sampling position within the conventional receptive field. To evaluate our superior design, we follow the same training settings as D-CNN and conduct experiments on both NYUv2 and SUN-RGBD datasets. Since conventional backbones are pre-trained with RGB input, i.e., ImageNet [29], which is not designed for RGB-D tasks. Hence, we follow D-CNN and train our model from scratch. We refer authors to [154] for more details on the training strategies.

The quantitative comparison can be found in Table 4.1. We only extract features from RGB input images. The baseline model is with VGG-16 encoder under Deeplab [9] architecture. D-CNN stands for the performance obtained by adding depth-aware re-calibration on both convolution and pooling. Z-ACN is the result obtained with our proposed convolution and pooling where we explicitly integrate the contextualized awareness in the basic operators. Our method can achieve superior performance over the counterpart, validating the effectiveness of our depth guided sampling position which can better model geometric priors compared to D-CNN.

Comparison with pre-trained methods: We also evaluate our method with pre-trained weights, i.e., we initialize the weight with the pre-trained models and further fine-tune it on NYUv2 datasets. We build our approach upon the DeepLab architecture, which is the same as D-CNN [154]. We report in Table 4.2 the performances of different methods. It can be seen that our method performs favorably over other methods. Compared to [119] which adapts a 3D CNN, our model remains a 2D CNN that requires less computational cost but achieves superior performance. Compared to our baseline, i.e., vanilla Deeplab, our Z-ACN enables significant improvements by encoding the depth information into the network. As D-CNN, our approach can also work well with the early fused RGB-HHA input, yielding a further improvement in

TABLE 4.2: Quantitative comparison with VGG-16 based methods on NYUv2 dataset. Our method significantly boosts the performance over the baseline and sets a new record on VGG-16-based approaches.

Model	Learned features	mIoU (%)
SurfConv [23]	RGB + HHA	31.0
Eigen et al. [34]	RGB + HHA	34.1
3DGNN [119]	RGB	39.9
Std2p [61]	RGB + HHA	40.1
D-CNN [154]	RGB	41.0
CFN [88]	RGB + HHA	41.7
D-CNN [154]	RGB + HHA	43.9
Baseline	RGB + HHA	40.4
Z-ACN (Ours)	RGB	42.5
Z-ACN (Ours)	**RGB + HHA**	**45.6**

the performance. The quantitative results validate that our operators are more effective in merging multi-modal features compared to the counterpart and set the new state-of-the-art performance with the VGG-16 encoder.

The qualitative comparison can be found in Figure 4.3 which shows the improvement of our approach over the baseline. The two first columns show the input RGB image and input HHA map. *Baseline* is the result obtained with early fused RGB-HHA input. + Z-ACN denotes that we further apply our approach over the baseline. It can be seen that our approach can favorably improve scene understanding over the counterparts by explicitly leveraging the depth cues, yielding more accurate semantic maps.

4.1.4.3 With ResNet backbones

Plug in SOTA ESAnet [128]: The current SOTA CNN performance on RGB-D semantic segmentation is achieved with ESAnet. To evaluate the generalization properties of our approach, we plug our Z-ACN into ESAnet, aiming to further improve the performance with additional depth-awareness. Compared to VGG encoders, ResNet encoders are deeper with more convolutions. Hence, replacing all convolutions with depth-adapted convolutions will yield more computational cost. As suggested in previous work [3, 131], the geometric cues play a more vital role in the first convolutional layers. Therefore, to find the best trade-off between the computational cost and the performance, we simply add a 3×3 depth-adapted convolution before the RGB encoder. This operation can be regarded as am early fusion to merge RGB and depth images at the stemming layer.

TABLE 4.3: Quantitative comparison with the baseline ESAnet on NYUv2 dataset. By simply adding an depth-adapted convolution, our method performs favorably over the baseline with different backbones, demonstrating the generalization capability or our Z-ACN.

Backbone	Setting	mIoU (%)	Improvement Δ (%)
ResNet-18	ESAnet	46.28	0.74
	Ours	**47.02**	
ResNet-34	ESAnet	48.13	1.02
	Ours	**49.15**	
ResNet-50	ESAnet	49.02	1.03
	Ours	**50.05**	
ResNet-101	ESAnet	49.44	1.76
	Ours	**51.24**	

The gain by further adding our Z-ACN can be found in Table 4.3. With our depth-adapted operator, the new model performs favorably over the ESAnet baseline under different backbones, demonstrating the generalization capability of our approach which can easily be embedded into any existing backbones. Furthermore, since both our approach and the counterpart shares the same architecture, the improvement is purely attributed to our depth-awareness, validating the effectiveness of our geometry-guided sampling position.

Comparison with RGB-D attention convolutions: To evaluate our Z-ACN, we compare our approach with two recent RGB-D attention convolutions, ShapeConv [7] and SConv [11]. ShapeConv decomposes the features within the receptive field into a base component and the remaining which are then calibrated with two additional learning weights before the convolution. The base component is computed by the mean function to squeeze the spatial resolution, which can be regarded as the additional channel attention for convolution. Different from ShapeConv which is not specially dedicated to RGB-D tasks, we explicitly leverage the depth prior to deform the convolutional sampling position, yielding a simple but efficient manner to integrate the spatial attention into convolution. Meanwhile, the concurrent SConv proposes a learning strategy to infer a depth-aware offset from latent space. However, for the same scene, the learned offset may vary under different settings such as different training strategies or backbones. As shown in Figure 4.4, while the backbone changes, SConv yields different sampling positions. Intuitively, the depth-aware offset should be only dependent on the geometry and independent of the learning factors. Different from SConv, our offset is computed without any learning parameters, making our depth-awareness constant under different environments. Further, we show through Figure 4.4 that our computed receptive field can favorably describe the perspective effect over the counterpart. Besides, we report in Table 4.4 the model size for each method. Similar to ShapeConv, our method does not add additional parameters

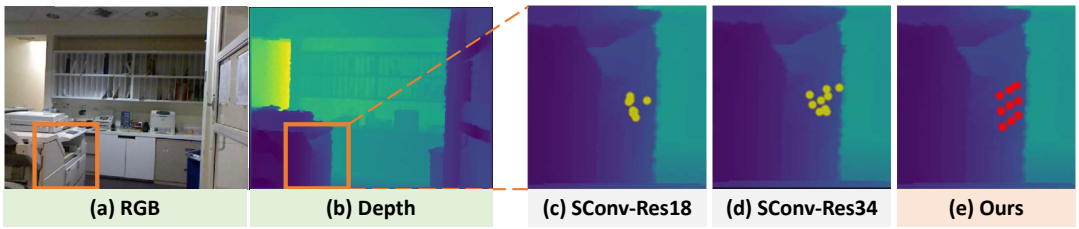


FIGURE 4.4: Visual comparison with concurrent depth-aware offset SConv. (a-b) are the input RGB and Depth. (c-d) illustrates the learned sampling position from SConv [11] for different ResNet backbones. (e) is the receptive field computed by our approach. SConv adopts a learning diagram to generate the receptive field, resulting in different shapes for different backbones. However, our method explicitly leverages the geometric constraint for the perspective effect. Our whole process is realized without learning parameters, making the depth-adapted sampling position independent from the neural network.

TABLE 4.4: Model size with different attention convolutions. We choose ResNet-18 as the backbone. Similar to ShapeConv, our method do not add extra learning parameters on top of the baseline. Different from SConv, we compute the offset in a non learning manner, yielding a efficient manner to explicitly leverage the depth attention in 2D CNN.

ResNet-18	ESAnet [128]	+ SConv [11]	+ ShapeConv [7]	+ Ours
Size (Mb)	304	+1	+0	+0

above the baseline and is more efficient compared to SConv which requires additional learning costs.

Table 4.5 illustrates the quantitative comparison with other attention convolutions. Under the consideration of a fair comparison, we embed all the operators into the ESAnet baseline and retrain them under the same settings. It can be seen that our Z-ACN outperforms the concurrent approaches with a large margin under all backbones. This highlights the effectiveness of our depth-constraint attention compared to channel attention (ShapeConv) and learned depth attention (SConv).

Comparison with SOTA performance: We compare the performance of our Z-ACN with other state-of-the-art models. The quantitative results can be found in Table 4.6. Our Z-ACN sets the new state-of-the-art performance in NYUv2 datasets. Compared to ShapeConv when both methods use ResNet-101 as the backbone and adopt single-scale inference, our approach achieves 3.8% mIoU improvement.

Outdoor scene: We also evaluate our approach to the outdoor scene, e.g., KITTI [50]. The vanilla KITTI dataset provides RGB and lidar input. We take the dataset presented in [179] which provides a dense depth map. We validate all methods on the held-out testing set due to the smaller size and lack of a proposed validation split.

TABLE 4.5: Quantitative comparison with other attention convolution methods on NYUv2 dataset. All methods are implemented on the ESAnet baseline and trained under the same settings. Our approach achieves better mIoU compared to concurrent works under different backbones, validating the effective of our geometry-constrained sampling position.

Backbone	Setting	PixelAcc	mAcc	mIoU	f.w.IoU
ResNet-18	+SConv	74.19	60.01	46.93	60.26
	+ShapeConv	74.11	59.37	46.38	60.61
	+Ours	**74.35**	**59.82**	**47.02**	**60.73**
ResNet-34	+SConv	74.95	61.08	47.99	61.77
	+ShapeConv	74.68	61.07	47.70	61.20
	+Ours	**75.78**	**62.81**	**49.15**	**62.64**
ResNet-50	+SConv	76.13	62.36	49.04	63.00
	+ShapeConv	**76.17**	62.45	49.58	**63.02**
	+Ours	75.88	**63.55**	**50.05**	62.99
ResNet-101	+SConv	76.49	63.65	50.43	63.67
	+ShapeConv	76.45	63.28	50.10	63.46
	+Ours	**77.00**	**64.26**	**51.24**	**64.32**

We adopt the same modified ResNet-18 as presented in [23] as our backbone with skip-connected fully convolutional architecture [101]. The conventional convolution is replaced by our proposed operator.

Our model is compared with 3D representation such as PointNet [116], Conv3D [136, 145] and 2D representation such as DeformCNN [28] and SurfConv [23]. Conv3D [136, 145] and PointNet [116] use the hole-filled dense depth map provided by the dataset to create 3D input. For PointNet, the source code is used to use RGB plus gravity-aligned point cloud (pcl). The recommended configuration [116] is used to randomly sample points. The sample number is set to be 25k. For Conv3D, the SSCNet architecture [145] is used and is trained with flipped - TSDF and RGB. The resolution is reduced to $240 \times 144 \times 240$ voxel grid. For DeformCNN, RGB images and HHA images are chosen as input for a fair comparison. For SurfConv, we compare with their best performance, which requires a resampling on the input image to be adapted to the 8 levels of depth. For all the above-mentioned models, we follow the same configuration and learning settings as discussed in [23].

The quantitative result is reported in Table 4.7 that all methods are trained from scratch following [23]. While dealing with an outdoor scene, 3D methods such as point cloud suffer from computational costs compared to 2D CNN which extracts features from images. It is also the case for Conv3D [136, 145] since voxelizing the whole 3D space is time-consuming. Compared to these 3D methods, our model remains 2D CNN but achieves a better result. DeformCNN [28] takes into RGB + HHA as input and learns offsets to deform the sampling position. Nevertheless, the offset is

TABLE 4.6: Performance comparison with SOTA methods on NYUv2 dataset. ⋆ denotes the multi-scale strategy. Our method is tested with single-scale inference strategy and sets the new state-of-the-art performance among ResNet based models.

Method	Backbone	PixelAcc	mAcc	mIoU	f.w.IoU
ACNet [65]	ResNet-50	-	-	48.3	-
2.5*D* [175]	ResNet-101	75.9	-	49.1	-
ShapeConv [7]	ResNet-101	74.5	59.5	47.4	60.8
⋆*CFN* [88]	ResNet-152	-	-	47.7	-
⋆3*DGNN* [120]	ResNet-101	-	55.7	43.1	-
⋆*RDFNet* [110]	ResNet-152	76.0	62.8	50.1	-
⋆*ShapeConv* [7]	ResNet-101	75.5	60.7	49.0	61.7
⋆*Malleable* [176]	ResNet-101	76.9	-	50.9	-
⋆*SGNet* [11]	ResNet-101	76.8	63.1	51.1	-
⋆*CANet* [207]	ResNet-101	76.6	63.8	51.2	-
Z-ACN (Ours)	ResNet-101	**77.0**	**64.3**	**51.2**	**64.3**

TABLE 4.7: Comparison on KITTI test set. Our methods achieve better performance compared to 3D approaches and the concurrent SurfConv. It is worth noting that with single RGB input, our depth-adapted sampling position enables significant improvement over our baseline, validating the effectiveness of depth-guided non-local attention. Models are trained from scratch.

Model	Learned features	Acc (%)	mIoU (%)
PointNet [116]	RGB + pcl	55.1	9.4
Conv3D [136, 145]	RGB + voxel	64.5	17.5
DeformCNN [28]	RGB + HHA	79.2	34.2
SurfConv-8 [23]	RGB + HHA	79.4	35.1
Baseline	RGB	79.3	31.3
Z-ACN (Ours)	RGB	79.7	33.5
Z-ACN (Ours)	**RGB + HHA**	**80.1**	**35.8**

learned from the input feature maps which do not explicitly leverage the geometric constraints. In contrast, our model computes the offset from low-level constraint, i.e., 1-channel depth, with traditional algorithms and does not require gradient descent. The result in Table 4.7 shows that our model performs favorably over DeformCNN without extra learning parameters, validating the effectiveness of our depth-adapted sampling position. SurfConv is a concurrent work that incorporation 3D information into 2D CNN. However, it requires additional pre-processing on the input data such that depth-guided image resampling. Instead, we encode the depth into the CNN via the bias of offset. Compared to the concurrent method, our approach achieves large performance gains.

We present in Table 4.8 the quantitative comparison over the baseline with weight initialization. $Baseline_1$ and $Baseline_2$ represent the result obtained with RGB input

TABLE 4.8: Quantitative comparison on KITTI test set. Networks are trained from pre-trained models.

KITTI	*Baseline*$_1$	+ Z-ACN	*Baseline*$_2$	+ **Z-ACN**
mAcc (%)	48.3	**49.5**	51.8	**55.1**
mIoU (%)	39.1	**40.6**	41.6	**45.3**

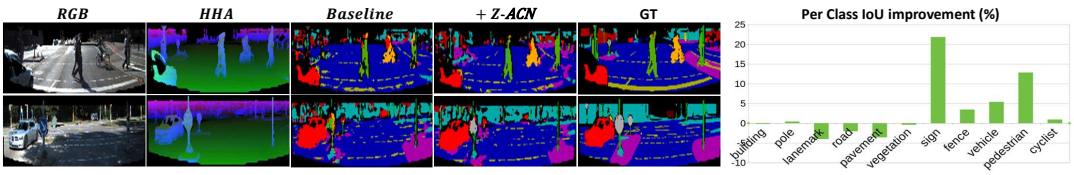


FIGURE 4.5: On the left we illustrate the qualitative comparison on the NYUv2 dataset. The two first columns are the input RGB and HHA, respectively. *Baseline* represents the semantic map obtained with early fused RGB-HHA input. + Z-ACN stands for the results obtained by inserting our depth-adapted sampling position into the baseline. It can be seen that our approach can also improve the baseline performance in outdoor scenes. The black regions in "GT" are the ignoring category. On the right, we illustrate the per-class improvement above the baseline. We improve 7/11 objects with 3.7% mean IoU increment. Segmentation results on the KITTI test dataset. GT stands for ground truth. The black regions in "GT" are the ignoring category.

and early fused RGB-HHA input, respectively. + Z-ACN stands for the results obtained by inserting our depth-adapted offset into the baseline. It can be seen that our methods can significantly enable gains over the baseline performance with improved depth-awareness. We illustrate in Figure 4.5 the per-class IoU improvement with RGB input. Compared to the baseline, our approach enables improvement on 7/11 objects, especially "salient" objects in the urban scene such as the vehicle, cyclists, and pedestrians. However, we also observe that our approach achieves lower performance in detecting lanemark. This is because the lanemark is co-planar as the road that the confusing geometrical information may add noises for our depth-adapted model.

The qualitative comparison over the baseline is shown in Figure 4.5. The two first columns show the input RGB image and input HHA map. *Baseline* is the result obtained with early fused RGB-HHA input. + Z-ACN denotes that we further replace the baseline convolution with our approach. By explicitly leveraging the geometry, our approach constrains the network to pay more attention to boundaries and reason about semantic maps with higher accuracy. We observe that objects like the vehicle, pedestrian, and cyclist are better segmented, as well as the sign. Recognizably, these objects do not share the same depth compared to the background (road or pavement). Hence, our adapted sampling position contributes to improving the discriminability of these salient objects.

TABLE 4.9: Empirical analysis on the influence of the intrinsic parameters. All methods are trained from pre-trained model under the same setting.

NYUv2 (%)	mAcc	mIoU
Baseline	51.9	40.4
Z-ACN (k_r)	53.4	41.6
Z-ACN (GT)	**55.2**	**42.5**

4.1.5 Additional Studies and Discussions

In this section, we conduct additional studies on the NYUv2 dataset to validate the efficiency, robustness, and flexibility of our operators. We choose VGG-16 with Deeplab as the baseline. The features are extracted from RGB images. The depth map is used to guide the sampling position.

4.1.5.1 Intrinsic Parameters

Our model requires the intrinsic parameters to back-project the pixels to the 3D scene and project the depth-adapted 3D planar grid to the image plane. This pattern is integrated into Eq. 4.4 and Eq. 4.8. Demanding camera parameters as priors can be a strong assumption that limits the application. Therefore, we evaluate the performance with a randomly set camera matrix.

The intrinsic parameters include the principal point and the focal length. However, most models resize the input image shape, which results in difficulties in using the official principal point value. Hence, we assume that the principal point is the same as the center of the input image and chose a random value for focal length. We set $(f_u, f_v) = (100, 100)$ for NYUv2 dataset, where the official value is around $(519, 519)$. We retrain the new model under the same training setting. The quantitative result is reported in Table 4.9. We denote k_r, the result obtained with randomly chosen intrinsic parameters, and GT, the result obtained with official values.

It can be seen that with an arbitrary value for intrinsic parameters, our model can still achieve favorable performance compared to the baseline. Compared to the result obtained with GT intrinsic value, the loss is only 0.2% for mAcc and 0.9% for mIoU. The result validates that our model can get rid of the assumption of the input intrinsic parameters under the condition that they are logically chosen.

4.1.5.2 Ablation Study

To further verify the functionality of both depth-adapted convolution and depth-adapted average pooling, the following experiments are conducted.

TABLE 4.10: Results of using depth-adapted operators in different layers. Experiments are conducted on NYUv2 test set. *i* stands for the number of convolution layers.

	Configuration	mIoU (%)
Result from scratch		
a)	Baseline	24.0
b)	Z-Conv5_1	27.6
c)	Z-Convi_1	29.7
d)	**Z-Convi_1 + Z-AvgPool**	**30.4**
Result from pre-trained		
a)	Baseline	40.4
b)	Z-Conv5_1	42.2
c)	Z-Conv5_2	41.7
d)	Z-Conv5_3	41.7
e)	**Z-Conv5_1 + Z-AvgPool**	**42.5**

- For results trained from scratch: we analyze a) baseline performance, b) a deep layer convolution replaced by Z-ACN, c) first convolutions from all layers replaced by Z-ACN, d) CNN replaced by Z-ACN including the average pooling.
- For results trained from pre-trained weight: we analyze a) baseline performance, b) the first convolution from a deep layer replaced by Z-ACN, c) the second convolution from a deep layer replaced by Z-ACN, d) the third convolution from a deep layer replaced by Z-ACN, e) CNN replaced by Z-ACN including the average pooling.

Experimental results are reported in Table 4.10. While learning from scratch, our operators can effectively extract features with geometric relationships and improve the segmentation performance. By comparing (a) and (b), we only replace deep convolution with our approach, i.e., the first convolution of layer 5 of VGG-16, we achieve a 3.6% gain on mIoU. (c) illustrates the result with the first convolution of all layers replaced by our approach. Our Z-ACN enables a 5.7% gain compared to the baseline (a). Finally, by introducing the depth-adapted average pooling (d), we observe that the performance can be further promoted, validating the effectiveness of our depth-adapted pooling method.

While learning from the pre-trained model, we firstly want to argue that the existing weight may not be fair nor suitable for our adapted convolution. The existing weight is learned with a fixed size and shape structure, while our adapted convolution breaks this limitation. The most suitable pre-trained weight for our operator might require training our depth-adapted model on ImageNet, which is impossible since the depth information is not available on this dataset.

Nevertheless, we still show that our approach can benefit from the conventional pre-trained weights. By fine-tuning the weights, Table 4.10 illustrates that replacing the first convolution from a deep layer contributes the most to the performance by 1.8% over the baseline. By introducing the depth-adapted average pooling, the performance can be further promoted.

4.2 Application in Saliency Detection

4.2.1 Introduction

In the last decade, RGB-based deep learning models for salient object detection (SOD) [30, 91, 172, 194, 202] achieved significant success thanks to the advances of GPU and CNN. Given an input image, the goal of SOD is to compute the pixel-wise location of the prominent objects that visually attract human attention the most. However, RGB SOD models focus more on photometric information instead of geometry. This is due to the fixed shape and size kernel design of CNN that is not invariant to scale changes and to 3D rotations. By the lack of geometric information on the input side, it is inevitable for RGB models to add additional learning modules in the network to attend to salient objects, resulting in model complexity and computational cost.

Recent RGBD-based SOD has motivated research interest thanks to the accessibility of cross-modal information from the input side. State-of-the-art RGBD models [44, 109, 115, 203] achieve superior performance over the RGB baseline, affirming the effectiveness of learning from two modalities. Most architectures adapt fusion-wise models, such as early fusion [203] where the depth map is fed as the fourth channel to RGB image, or multi-scale and late fusion [109] where two-stream networks are adopted. However, early fusion contains more low-level features than semantic ones. Multi-scale or late fusion inevitably requires more learning parameters. As shown in Figure 4.6, the size of RGBD models is often larger than that of RGB networks.

We explore differently the relationship between depth map and RGB image. Taking human beings as an example, to distinguish salient objects from the 3D world, the input is the visual appearance through human eyes. With the color information and thanks to the depth estimation capability, humans further discover geometric information. This prior guides the understanding of RGB images. It should be the same case for intelligent machines.

To this end, we propose a novel Modality-Guided Subnetwork (MGSnet) which adaptively transforms convolutions by fusing information from one modality to another (e.g., depth to RGB or RGB to depth). Our network matches perfectly both RGB and RGB-D data and dynamically estimates depth if not available by simply applying an off-the-shelf depth prediction model. We design a subnetwork mechanism alongside

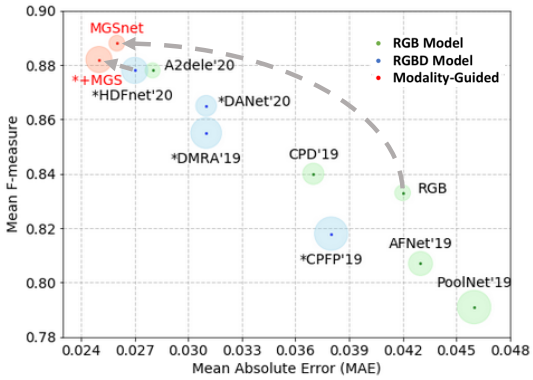


FIGURE 4.6: Comparison with SOTA saliency model. We report the performance analysis on NLPR dataset [112]. Note that better models are shown in the upper left corner (i.e., with a larger mean F-measure and smaller MAE). The circle size denotes the model size. Our proposed MGSnet for RGB SOD achieves the best performance with the lightest model size. The MGS design can also be embedded to the state-of-the-art RGBD model HDFnet [109] to enable further progress (denoted as $* + MGS$).

the master streaming pipeline. The subnetwork can be treated like a light residual-addition branch as the ResNet [59]. It takes one modality map as the master input, e.g. RGB, and enhances its robustness by deforming the convolution kernel with the supervision of the complementary modal prior, e.g. depth, and vice versa.

In summary, the main contributions of this chapter are listed as follows :

- By exploiting the nature of CNN sampling position, we propose a novel cross-modal fusion design (MGS) for salient object detection, where we use a subsidiary modality, i.e., RGB/depth, to guide the main modality streaming, i.e., depth/RGB.
- For RGB-only input, we suggest using an off-the-shelf depth prediction model to mimick the multi-modality input. Our MGSnet enables dramatical performance gain on benchmark datasets and achieves state-of-the-art performance among RGB SOD models.
- The proposed MGS can also be embedded in RGBD two-stream network with the advantage of cross-modality cues while being lightweight.

4.2.2 Related Work

RGB SOD: In the past decade, the development of GPU and CNN contributes to the advances of RGB SOD. One core problem is understanding the geometric information from the image. Fully Convolutional Network (FCN) [101] is a pioneering work in leveraging spatial information in CNN. Most recent researches dominating RGB SOD are FCN-based, such as [194] which designs a single stream encoder-decoder

system, [79] which adopts a multi-scale network on input, and most currently [30, 91, 172, 202] which fuse multi-level feature map. Some branch designs also have achieved impressive results such as C2S-Net [81] which bridges contour knowledge for SOD. By inserting additional transformation parameters in networks, it contributes to the model performance. Nevertheless, the inference time and computational cost become more significant.

RGBD SOD: The complementary depth map may provide extra clues on the geometry. How to efficiently joint RGB and depth modality is the key challenge for RGBD SOD. One possible solution is to treat the depth map as an additional channel and adapt a single-stream system as shown in DANet [203]. It further designs a verification process with a depth-enhanced dual attention module. An alternative is to realize multi-stream networks followed by a feature fusion mechanism. PDNet [215] designs a depth-enhanced stream to extract geometric features and further fuses with the RGB features. D3net [37] adopts separate networks to respectively extract features from RGB, depth map, and RGBD four-channel input. A late fusion is further realized. HDFnet [109] adopts two streaming networks for both RGB image and depth map. These features are further fused to generate region-aware dynamic filters. JL-DCF [44] proposes joint learning from cross-modal information through a Siamese network. Generally, RGBD networks achieve superior performance compared to RGB as shown in Figure 4.6. However, these methods rely on the quality and accessibility of the depth map. A high-quality depth map requires expensive depth sensors and is still sparse compared to an RGB image as suggested in [37, 44]. To this end, DCF [69] proposes to calibrate the raw depth to improve the quality. Nevertheless, the high computational cost due to the two-streaming network requires more development.

Some recent researches [70, 115, 199] propose to learn from RGBD images and tests on RGB. This design enables an RGB CNN to achieve a comparable result with RGBD SOD during testing. Different from it, we propose to firstly discover the hidden geometric modality behind RGB images by simply using an off-the-shelf depth prediction method. With the estimated depth, we further propose a Modality-Guided Subnetwork mechanism to enhance the master RGB network understanding of the contour problem. Our proposed MGSnet achieves state-of-the-art performance with real-time inference speed compared to other RGB models. It can also be embedded in RGBD two-stream models to enable further progress with raw depth.

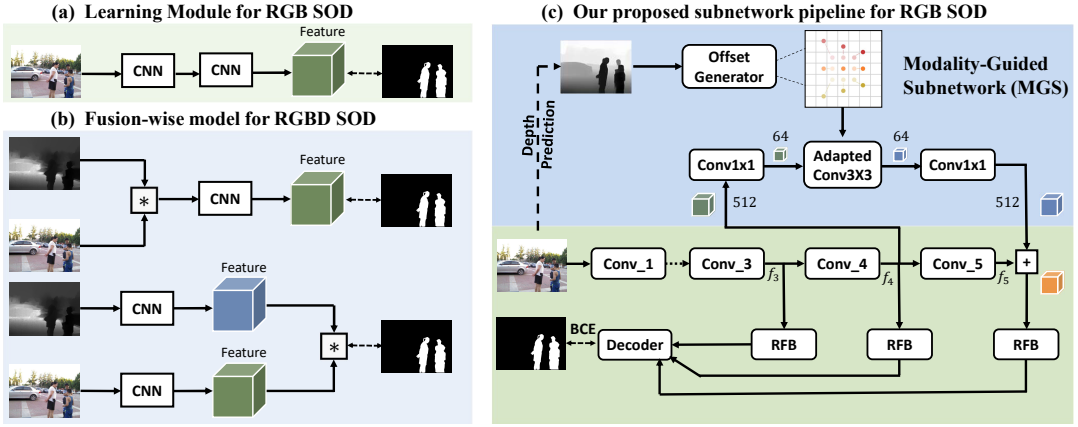


FIGURE 4.7: Overview of our MGSnet. (a) Conventional RGB models [91, 172, 194] insert additional modules to learn geometry-invariant features. (b) RGBD models [44, 109, 203] adopt fusion-wise design to learn both photometric and geometric information. (c) Our proposed MGSnet which takes only RGB image for both training and testing. We use depth prior to guide sampling position on RGB feature map through a subnetwork design to compensate the master streaming.

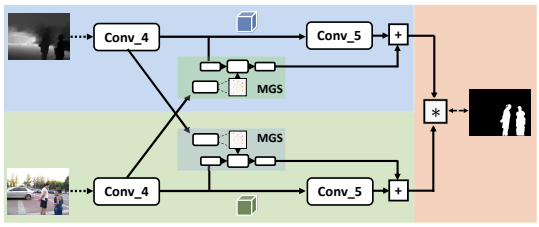


FIGURE 4.8: Illustration of embedded MGS on a RGBD two-streaming network.

4.2.3 Modality-Guided Subnetwork

4.2.3.1 Overview

In Figure 4.7 (c), our network only takes RGB as input that then estimates the pseudo-depth. Our MGSnet only takes the pseudo-depth to deform the RGB streaming. In other words, only the RGB modality is fed through Conv_4.

Note that our model is not limited by the nature of the modality. It can be a depth-guided RGB convolution as well as an RGB-guided depth convolution. Figure 4.8 presents our model embedded on an RGBD two-streaming network and Figure 4.9 illustrates the idea of modality-guided sampling position. We learn the offset from both semantic RGB and depth features to create a cross supervision mechanism.

For simplicity, we present in the following section a depth-guided subnetwork for RGB features. It contains three parts: a master RGB streaming network, an off-the-shelf prediction model to estimate a pseudo-depth map if not available, and a depth-guided subnetwork design. For simplicity, VGG-16 [133] architecture is adopted as our basic convolutional network to extract RGB features for its wide application in SOD. We

use RFB [94] on the steamer layers (f_3, f_4, f_5) which contains high level features for SOD as suggested in [44, 109, 115]. We further embed our subnetwork to enhance the edge understanding of the encoder output. We take the same decoder as proposed in [115] and a simple binary cross-entropy (BCE) as the loss.

4.2.3.2 Depth-guided Subnetwork

To proceed with the geometric prior, the depth map D and the RGB feature map (output of $Conv\_4$) are fed together to our model. We use $f_4 \in \mathbb{R}^{b\times 512\times h\times w}$ to denote the input RGB feature. The depth prior and RGB feature maps are articulated through an adaptive convolution to compute depth-aware RGB feature maps as output. The last is added to the master RGB stream to form the final feature map.

The subnetwork contains three convolutions of different filter sizes: $1\times 1, 3\times 3$, and 1×1. It shares the same architecture of plain baseline of ResNet [59] that the 1×1 layers are used for reducing (512→64) and then increasing dimensions (64→512), allowing the 3×3 layer with smaller input/output dimensions. We denote $\mathcal{D}$ and $\mathcal{U}$ for the first and the last 1×1 convolution, which stands for down-sample and up-sample, respectively. This design can significantly reduce the learning parameters, which contributes to the lightweight design of our subnetwork. Different from ResNet that uses the three layers as a bottleneck, we use them as the residual-addition branch which serves as complementary information to the plain network.

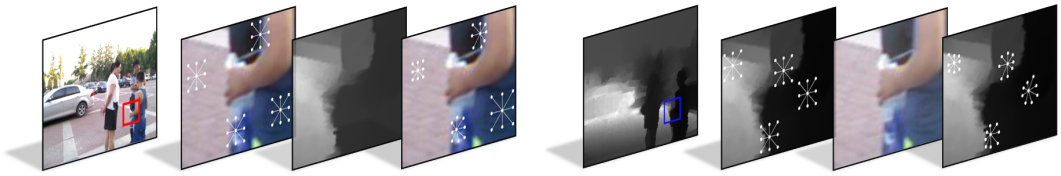
Similar to previous section 4.1.3, we replace the conventional 3×3 convolution by deformable convolution (DeformConv) [28], where the kernels are generated with different sampling distributions which is adapted to depth modality. Mathematically, we have:

$$\mathbf{y}(p) = \sum_{p_n\in\mathbf{R(p)}} \mathbf{w}(p_n)\cdot\mathbf{x}(p+p_n+\Delta p_n) \tag{4.12}$$

The convolution may be operated on the irregular positions $p_n + \Delta p_n$ as the offset Δp_n may be fractional. To address the issue, we use the bilinear interpolation which is the same as that proposed in [28]. The adapted convolution is denoted as $\mathcal{A}$.

Thanks to the depth input of the subnetwork, the scale and geometric transformation of objects on the RGB feature map can be directly analyzed with the adapted offsets. This process is expressed as:

$$\Delta p_n = \eta(D) \tag{4.13}$$



(A) Depth-guided sampling position. (B) RGB-guided sampling position.

FIGURE 4.9: Visual understanding of MGSnet. A pair of RGB and depth images from an RGBD dataset are illustrated on the left. While extracting features through two streaming networks, the cross-modal information beyond the fixed sampling position is not utilized (second left). Our proposed modality-guided sampling position breaks the limit of fixed-local configurations. The new sampling position incorporates supporting modality into the basic function of CNN on the main modality: the fixed sampling position is replaced by relevant neighbors defined by the supporting modality without limitation (right).

We present two types of offset generators according to different plain networks. More details are discussed in the following section. The newly defined sampling position becomes depth-aware and helps to better articulate the RGB feature and geometric information. Finally, the output of MGS is added to the master stream, which serves as complementary depth-aware guidance on RGB features.

The entire process to compute the modality-guided feature f_M can be formulated as follows:

$$\begin{aligned} f_M &= MGS(f_4, D) \\ &= \mathcal{U}(\mathcal{A}(\mathcal{D}(f_4), \eta(D))) \end{aligned} \tag{4.14}$$

The output of RGB encoder can be formulated as :

$$out = f_5 + \lambda f_M \tag{4.15}$$

where λ is the weight parameter.

4.2.3.3 Offset generator

We use another modal prior to deform the main stream convolution. When the offset exceeds the input size, the output will be computed as if the zeros padding is applied. For RGB input, the pseudo-depth is used to deform the RGB sampling position. The offset is generated through Z-ACN [166, 169] or previous section 4.1.3. It firstly back-projects the 2D conventional grid to form a 3D point cloud according to the depth. Based on the point cloud, it extracts a depth-aware 3D plan and further creates a depth-aware 3D regular grid. Then it projects the 3D regular grid to the image plan to form the deformable sampling position. More details can be found in Z-ACN [166, 169] paper. Different to DeformConv [28] that learns offset from the RGB

feature map to deform RGB sampling position, Z-ACN computes offset according to low-level geometric constraint (one-channel depth) and does not require gradient descent, thus perfectly matches our light-weight subnetwork design. The computed offset allows the RGB convolution to be scale and rotation independent. We verify through experiments the superior performance of our model in the ablation study.

For RGBD input, current Sconv [11] suggests learning the RGB offset from a semantic depth feature map. We share the same motivation as Sconv. However, Sconv firstly projects the depth into a high-dimensional feature space and secondly learns a depth-aware offset and mask. Unlike Sconv, we learn the offset from the encoder or high-level features to avoid the additional projection. In other words, in our case, the offset generator η is realized through a simple 3×3 convolution to minimize the computational cost. Furthermore, we adapt to different modalities as input, i.e., it learns offset from both RGB and depth, while Sconv only learns from depth.

4.2.3.4 Understand adaptive sampling position

Our model aims to compensate for the single modality streaming. As shown in Figure 4.9, while extracting features from RGB images, the conventional sampling position is limited by the lack of capability to include geometry due to the fixed shape. We propose to use the depth prior to accurately locate the sampling position. For RGB input without depth prior, we suggest mimicking the depth map by using a monocular depth estimation model. Some pseudo-depth images may be inaccurate due to the domain gap between SOD and monocular depth estimation. In such a case, the offset will converge to 0 so that the deformation becomes minimal and local. The contribution of the depth-aware RGB feature is further regularized by the weight parameter λ of Eq. 4.15. In Fig. 4.10, we show that our method is robust to non-optical depth through several examples.

While extracting features from raw depth, conventional sampling positions may produce sub-optimal results due to some inaccurate measurements. The raw depth maps for SOD are obtained by camera measurements such as Kinect and Light Field cameras, or estimated by classic computer vision algorithms as [90, 138]. Thus, the raw depth images may contain noise and ambiguity. We can visualize several low-quality samples on the third row of Figure 4.10. To this end, we propose to use the RGB image to deform the depth sampling position. In such a case, the RGB-guided sampling position can make up for the measurement error on geometry.

4.2.4 Experiments

4.2.4.1 Benchmark Dataset

To verify the effectiveness of our method, we conduct experiments on seven following benchmark RGBD datasets. DES [22] : includes 135 images about indoor scenes captured by Kinect camera. LFSD [80]: contains 100 images collected on the light field with an embedded depth map and human-labeled ground truths. NLPR [112]: contains 1000 natural images captured by Kinect under different illumination conditions. NJUD [72]: contains 1,985 stereo image pairs from different sources such as the Internet, 3D movies, and photographs taken by a Fuji W3 stereo camera and with estimated depth by using optical flow method [138]. SSD [216]: contains 80 images picked up from stereo movies with estimated depth from flow map [138]. STEREO [106]: includes 1000 stereoscopic images downloaded from the Internet where the depth map is estimated by using SIFT flow method [90]. DUT-RGBD [114]: contains 1200 images captured by Lytro camera in real-life scenes.

4.2.4.2 Experimental Settings

Our model is implemented basing on the Pytorch toolbox and trained with a GTX 3090Ti GPU. We adopt several generally-recognized metrics for quantitative evaluation: F-measure is a region-based similarity metric that takes into account both Precision (Pre) and Recall (Rec). Mathematically, we have : $F_\beta = \frac{(1+\beta^2)\cdot Pre\cdot Rec}{\beta^2\cdot Pre+Rec}$. The value of β^2 is set to be 0.3 as suggested in [1] to emphasize the precision. In this chapter, we report the **maximum F-measure** (F_β) score across the binary maps of different thresholds, the **mean F-measure** (F_β^{mean}) score across an adaptive threshold and the **weighted F-measure** (F_β^w) which focuses more on the weighted precision and weighted recall. **Mean Absolute Error** (MAE) studies the approximation degree between the saliency map and ground-truth map on the pixel level. **S-measure** (S_m) evaluates the similarities between object-aware (S_o) and region-aware (S_r) structure between the saliency map and ground-truth map. Mathematically, we have: $S_m = \alpha \cdot S_o + (1-\alpha)\cdot S_r$, where α is set to be 0.5. **E-measure** (E_m) studies both image level statistics and local pixel matching information. Mathematically, we have: $E_m = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\phi_{FM}(i,j)$, where $\phi_{FM}(i,j)$ stands for the enhanced-alignment matrix as presented in [36].

4.2.4.3 Performance Comparison with RGB Input

We firstly compare with RGB models, including R3Net [30], PoolNet [91], CPD [172], AFnet [40]. All saliency maps are directly provided by authors or computed by authorized codes. For fair comparisons, we adopt the same training set as suggested in [115], which contains 1485 samples from NJUD, 700 samples from NLPR, and 800

TABLE 4.11: Quantitative comparisons of with RGB input. The off-the-shelf depth estimation is realized with MiDaS [125] which presents 86Mb model size. ↑ & ↓ denote larger and smaller is better, respectively. (**red**: best, **blue**: second best).

Dataset		DES			NLPR			NJUD			STEREO			DUT-RGBD		
Metric	*Size* ↓	*MAE* ↓	F_β^{mean} ↑	F_β^w ↑	*MAE* ↓	F_β^{mean} ↑	F_β^w ↑	*MAE* ↓	F_β^{mean} ↑	F_β^w ↑	*MAE* ↓	F_β^{mean} ↑	F_β^w ↑	*MAE* ↓	F_β^{mean} ↑	F_β^w ↑
RGB input																
R^3Net_{18}	225	.066	.728	.693	.101	.649	.611	.092	.775	.736	.084	.800	.752	.113	.781	.709
$PoolNet_{19}$	279	.031	.852	.814	.046	.791	.771	.057	.850	.816	.045	.877	.849	.049	.871	.836
CPD_{19}	112	**.028**	.860	**.841**	.037	.840	.829	.059	.853	.821	.046	.880	.851	.055	.872	.835
$AFNet_{19}$	144	.034	.840	.816	.043	.807	.796	.056	.857	.832	.046	.876	.850	.064	.851	.817
$EGNet_{19}$	412	.035	.831	.797	.047	.800	.774	.060	.846	.808	.049	.876	.835	.059	.866	.805
+ Pseudo Depth (86 Mb extra model size)																
$HDFnet_{20}$	177 (+86)	.070	.721	.664	.062	.758	.741	.124	.716	.656	.106	.743	.684	-	-	-
$CoNet_{20}$	171 (+86)	.037	.820	.808	.049	.744	.835	.068	.827	.795	.050	.848	.825	.045	.865	.847
Ours	**62** (+86)	**.028**	**.871**	**.837**	**.025**	**.888**	**.874**	**.047**	**.882**	**.856**	**.041**	**.881**	**.857**	**.037**	**.906**	**.889**

samples from the DUT-RGBD dataset. The remaining images of all listed datasets are used for testing. The quantitative comparison is presented in Table 4.11. Our model is trained with 50 epochs with 256×256 input image size.

For the RGB model, we can conclude from Table 4.11 that the improvement on the saliency map is attributed to different learning modules, which results in high computational cost (size). Different from traditional RGB models which do not exploit the depth information, we propose to take full advantage of the pseudo-geometry estimated with an existing monocular depth estimation method.

We re-train two RGB-D SOD network (HDFnet [109], CoNet [70]) with the additional estimated pseudo-depth. We observe a significant performance gap between the recent RGB-D models and the previous RGB models. The main reason is the quality of depth estimation: the domain gap between the depth estimation dataset and the SOD dataset leads to some failure depth maps. This can be noticed in the poor performance of HDFnet that extracts features from both RGB and depth images. CoNet, however, is more robust to the depth quality since the depth map is only used to supervise the feature extraction on RGB images. Our model shares the same motivation as CoNet to use depth prior to guide SOD but in a completely different manner. In our model, we directly learn a geometric-aware offset from the depth map to the sampling position on the RGB image. Our model achieves consistent superior performance compared with other models.

4.2.4.4 Performance Comparison with RGB-D Input

We also compare with state-of-the-art RGBD models with raw depth input in the Table 4.12, including CoNet [70], A2dele [115], DANet [203], cmMS [76], HDFnet [109], and DSA2F [141]. For fair comparisons, all saliency maps and the FPS are directly provided by authors or computed by authorized codes. Note that the FPS depends on the GPU for inference. Thus, only the FPS of HDFnet is tested on the same GPU as ours.

While depth is only used as supervision during training and only RGB image is required during testing, our model surpasses existing efficient A2dele significantly on performance with only an + around 5Mb model size. Compared to CoNet, the model size is minimized by 63% and achieves a comparable result. As presented in Figure 4.9, our proposed module can take advantage of cross-modality cues while being lightweight. Thus, we further incorporate with the HDFnet [109] to show the performance gain by integrating our approach. It achieves the state-of-the-art (SOTA) performance on VGG16 based models ($HDF + Ours$). To better demonstrate the superiority of the proposed method, we also use a larger backbone (VGG19) to compare

TABLE 4.12: Quantitative comparisons of with recent RGBD models. ↑ & ↓ denote larger and smaller is better, respectively. MGS can also be embedded to the HDFnet [109] to enable further progress. The scores/numbers better than ours are underlined (extracting RGB feature, extracting RGBD feature with VGG16, and extracting RGBD feature with VGG19 models are labeled separately).

		Extract RGB feature			Extract RGBD feature						
		$CoNet_{20}$	$A2dele_{20}$	*Ours*	$DANet_{20}$	$cmMS_{20}$	$HDFnet_{20}$	*+Ours*	$DSA2F_{21}$	$HDFnet_{20}$	*+Ours*
	Backbone	Resnet101	– VGG16 –		– VGG16 –				– VGG19 –		
	Size ↓	167	57	**62**	102	430	177	**178**	-	220	**221**
	FPS ↑	-	120	**150**	32		62	**58**	-	-	-
DES	*MAE* ↓	.027	.029	**.028**	.023	-	.030	**.019**	.021	.017	**.017**
DES	F_β^{mean} ↑	.862	.870	**.871**	.887	-	.843	**.920**	.896	.918	**.923**
DES	S_m ↑	.910	.881	**.882**	.904	-	.899	**.935**	.920	.937	**.937**
DES	E_m ↑	.945	.918	**.922**	.967	-	.944	**.979**	.962	.976	**.979**
NLPR	*MAE* ↓	.031	.031	**.025**	.028	.027	.027	**.025**	.024	.027	**.025**
NLPR	F_β^{mean} ↑	.848	.871	**.888**	.871	.869	.878	**.885**	.891	.883	**.882**
NLPR	S_m ↑	.908	.889	**.908**	.915	.899	.898	**.918**	.918	.915	**.918**
NLPR	E_m ↑	.934	.937	**.952**	.949	.945	.948	**.954**	.950	.951	**.951**
NJUD	*MAE* ↓	.047	.052	**.047**	.045	.044	.039	**.037**	.039	.038	**.035**
NJUD	F_β^{mean} ↑	.872	.873	**.882**	.871	.886	.887	**.893**	.898	.887	**.898**
NJUD	S_m ↑	.895	.867	**.879**	.899	.900	.907	**.911**	.903	.911	**.912**
NJUD	E_m ↑	.911	.914	**.928**	.922	.914	.931	**.935**	.923	.932	**.942**
STEREO	*MAE* ↓	.037	.044	**.041**	.047	.043	.042	**.039**	.039	.040	**.039**
STEREO	F_β^{mean} ↑	.885	.875	**.881**	.858	.879	.864	**.864**	.893	.875	**.878**
STEREO	S_m ↑	.908	.878	**.887**	.901	.895	.900	**.904**	.897	.903	**.902**
STEREO	E_m ↑	.928	.929	**.936**	.914	.922	.929	**.937**	.933	.934	**.938**

with the plain version HDFnet and the SOTA method DSA2F. Note that DSA2F uses neural architecture search to automate the model architecture while ours is hand-designed. Our model enables significant gains on the plain version with minimal cost (+ around 1 Mb on model size) and achieves comparable results with the DSA2F.

4.2.4.5 Qualitative Evaluation

We present the qualitative result with some challenging cases in Figure 4.10: low density (1^{st} columns), similar visual appearance between foreground and background ($2^{nd} - 5^{th}$ columns), small objects (6^{th} columns), far objects ($7^{th} - 9^{th}$ columns), human in scene (10^{th} columns), and similar and low contrast on depth map ($11^{th} - 13^{th}$ columns). It can be seen that our MGSnet yields the results closer to the ground truth mask in various challenging scenarios, especially for the last three columns with low-quality depth clues. Different from two-stream networks that tend to treat sub-optimal depth equally as RGB input, MGSnet extracts features from RGB images while the depth map serves only as complementary guidance, thus becoming robust to depth bias. By analyzing the response on HDFnet (sixth row) and HDFnet with embedded MGS (seventh row), we observe that our approach enables the plain network better discrimination of salient objects from the background.

4.2.5 Ablation Study

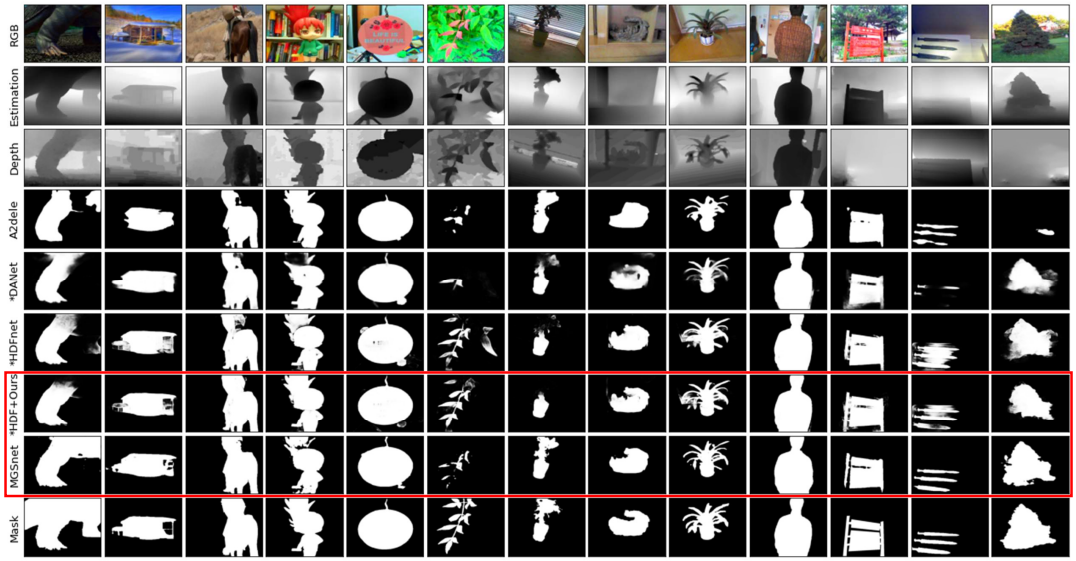


FIGURE 4.10: [
Visual comparison between the proposed MGSnet and the state-of-the-art RGB/RGBD methods.]Visual comparison between the proposed MGSnet and the state-of-the-art RGB/RGBD methods. $*$ denotes that the ground truth depth is used during testing. We also embed MGS on the HDFnet [109] to enable further improvement, denoted as $*HDF + Ours$.

	Dataset	DES		LFSD		NLPR		NJUD		SSD		STEREO	
#	Metric	$MAE \downarrow$	$F_{max} \uparrow$	$MAE \downarrow$	$F_{max} \uparrow$	$MAE \downarrow$	$F_{max} \uparrow$	$MAE \downarrow$	$F_{max} \uparrow$	$MAE \downarrow$	$F_{max} \uparrow$	$MAE \downarrow$	$F_{max} \uparrow$
1	RGB Baseline	.036	.880	.104	.835	.037	.883	.059	.877	.082	.785	.054	.882
2	RGB + Self Deform	.042	.860	.110	.798	.031	.885	.052	.879	.062	.814	.046	.882
3	RGB pseudo D Fusion	.032	.888	.093	.819	.029	.893	.061	.863	.077	.776	.049	.878
4	RGB + Depth-Deform	**.028**	**.899**	**.078**	**.849**	**.025**	**.905**	**.047**	**.897**	**.063**	**.801**	**.041**	**.898**
5	RGB-D Baseline (B)	.020	.933	.089	.856	.028	.921	.039	.922	.054	.867	.042	.911
6	B + Self Deform	.019	.936	.089	.853	.026	.916	.038	.925	.053	.878	.044	.895
7	B + Cross-Modal Deform	**.019**	**.936**	**.079**	**.871**	**.025**	**.921**	**.037**	**.925**	**.049**	**.867**	**.039**	**.917**

TABLE 4.13: Ablation study of modality-guided sampling position

Effect of Modality-Guided Sampling Position: Our modality-guided sampling position aims to incorporate multi-modal information through the basic function of CNN - the sampling position of convolution. This pattern is integrated in Eq. 4.12 and Eq. 4.13. To verify the effectiveness of the proposed modality-guided sampling position, a series of experiments with different learning strategies are realized.

(1) - (4) are experiments on RGB model: (1) RGB Baseline. (2) Self-guided deformable sampling position. We learn the offset from the RGB feature map. (3) RGB pseudo-depth early fusion. We form a four-channel input with pseudo depth. (4) Depth-guided deformable position. We compute an offset from pseudo-depth using Z-ACN to guide RGB streaming. (5) - (7) are experiments on RGBD model: (5) Baseline. We use the same architecture as HDFnet. (6) Self-guided deformable sampling position. The offset applied to RGB streaming is learned from the RGB feature. Idem for depth streaming. (7) Cross modality-guided deformable position. We learn an offset from depth to guide RGB streaming, and vice versa.

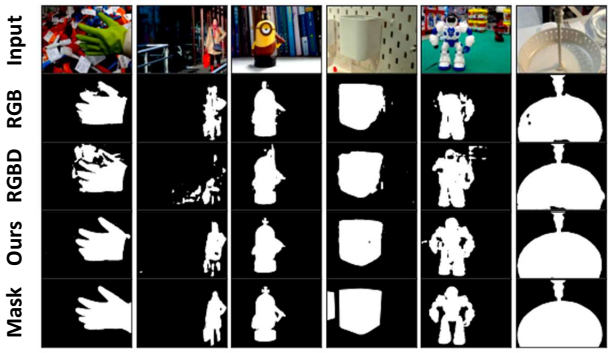


FIGURE 4.11: Visual analysis of embedded depth with MGSnet.

Table 4.13 (1) and (3) compare the performance of the baseline RGB three-channel input and mimicked RGBD four-channel input with pseudo-depth, respectively. The mimicked multi-modality early fusion achieves better performance, indicating that the pseudo-depth provides additional semantic. However, by comparing (3) and (4), we observe that the proposed depth-guided deformable sampling position can better use the complementary information to supervise RGB streaming, compared with early fusion. By comparing (2) and (4), we show that the depth-guided deformable position is more accurate on saliency compared to that of the self/RGB-guided. This verifies the assumption that depth cues can help the RGB model to better distinguish the foreground and background. Note that in (4) we only extract features from RGB images. The additional awareness of the geometry is only treated as a 2D offset to better locate the sampling position. This new integration design contributes to the model performance with minimal cost. For better understanding, the qualitative result presented in Figure 4.11 shows that our approach provides more accurate saliency maps

with better contrast. On the RGBD model (5-7), we also observe the superior performance with the cross-modality deformable sampling position achieves as it directly compensates for the single modal streaming.

Performance with different depth qualities: We also conduct an experiment to show the impact of depth quality. We choose the HDFnet [109] as the baseline and further embed it with our method. We present the average metric on all testing datasets in Table 4.14 with pseudo-depth (estimated) and raw depth from the RGBD dataset. Results obtained with pseudo-depth are denoted with *.

AvgMetric	*HDFnet∗*	*+Ours∗*	*HDFnet*	*+Ours*
$MAE \downarrow$	.1053	**.0758**	.0405	**.0375**
$F_\beta \uparrow$	.8410	**.8599**	.9121	**.9166**
$F_\beta^{mean} \uparrow$	.7326	**.7868**	.8730	**.8831**
$F_\beta^w \uparrow$	.6789	**.7488**	.8569	**.8672**
$S_m \uparrow$	.8010	**.8390**	.9013	**.9053**
$E_m \uparrow$	.8359	**.8797**	.9312	**.9377**

TABLE 4.14: Performance variation with different depth qualities. (*) denotes results obtained with pseudo-depth.

It shows that the quality of depth has an important influence on performance. Features extracted from raw depth describe better the salient object and were in line with our expectations. However, in both cases, our MGS can significantly enable progress compared to the plain networks. For pseudo-depth, the contribution of our MGS is more significant, which can be explained by the effectiveness of our RGB-guided sampling position for depth streaming. It can efficiently help to alleviate depth errors.

4.3 CONCLUSIONS

In this chapter, we discuss a novel 2D CNN to include geometric information in RGB CNN. We firstly validate our approach in semantic segmentation tasks. Different from previous works that integrates the channel or spatial attention into convolution through learning methods, our network fully explores the geometric constraint in a statistic manner, making the depth-awarenss independent to the learning settings. We introduce two basic depth-adapted operators that can be easily integrated into the existing CNN model. Extensive studies generalization property of our methods which perform favorably over the baseline and other convolutions. Experiments on challenging RGB-D datasets demonstrate that our approach performs well over the state-of-the-art methods by large margins.

Furthermore, we test the idea of depth-guided convolution on RGB-D saliency tasks. Since our method can only generates offset from depth cues, we inspire from concurrent method and propose a learnable offset generate to enable bi-directional guidance

(Depth to RGB and RGB to depth). Extensive experiments against RGB baselines demonstrate the performance gains of the proposed module, and the addition of the proposed module to existing RGB-D models further improved results.

In this chapter, we present how to explore depth as a form of spatial attention to better guide convolutional sampling position. Recent development on NLP, especially on transformer attention has shown great advantages in modeling contextualized awareness. Regular convolution and deformable convolution can also be regarded as a special case of transformer: the query is the center pixel and the key are the neighboring pixels within conventional sampling position or deformable sampling position. Despite the plausible results achieved by convolutional networks, the limited capability of modeling the contextualized features is the main performance bottleneck for both backbones and feature fusion modules. Therefore, in the following chapter, we present the transformer attention to fuse RGB-D cues.

Chapter 5

Transformer Fusion for RGB-D Semantic Segmentation

Fusing geometric cues with visual appearance is an imperative theme for RGB-D indoor semantic segmentation. Existing methods commonly adopt convolutional modules to aggregate multi-modal features, paying little attention to explicitly leveraging the long-range dependencies in feature fusion. Therefore, it is challenging for existing methods to accurately segment objects with large-scale variations. In this chapter, we propose a novel transformer-based fusion scheme, named TransD-Fusion, to better model contextualized awareness. Specifically, TransD-Fusion consists of a self-refinement module, a calibration scheme with cross-interaction, and a depth-guided fusion. The objective is to firstly improve modality-specific features with self- and cross-attention, and then explore the geometric cues to better segment objects sharing a similar visual appearance. Additionally, our transformer fusion benefits from a semantic-aware position encoding which spatially constrains the attention to neighboring pixels. Extensive experiments on RGB-D benchmarks demonstrate that the proposed method performs well over the state-of-the-art methods by large margins.

5.1 Introduction

Recent developments in depth sensors provide geometric information at a low cost. Since the depth information along with images can naturally contribute to scene understanding, RGB-D semantic segmentation has drawn increasing attention [154, 158, 166, 206].

When merging the depth cues and images, three typical challenges arise: (1) Multi-modal fusion. RGB input contains rich information on visual changes, while depth images are sensitive to occluded boundaries. How to extract, preserve, and fuse these modality-specific features is as yet an open issue for RGB-D semantic segmentation. (2) Noisy response in each modality. On the one hand, the similar visual appearance between neighboring objects can adversely affect the model discriminability. On the

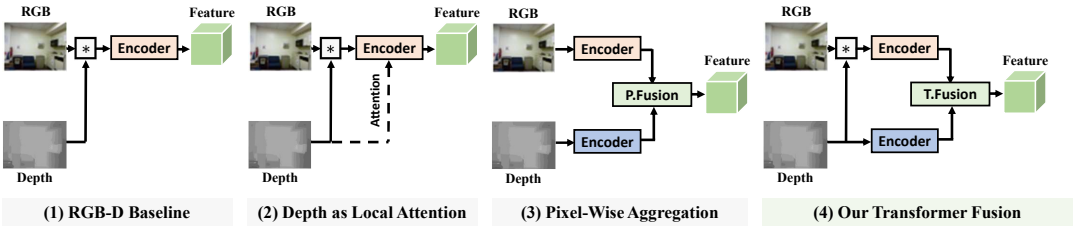


FIGURE 5.1: **Comparison of different RGB-D fusion strategies.** (1) Conventional RGB-D early fusion schemes. (2) Previous attempts to improve the RGB-D learning with local depth awareness [154, 166]. (3) Pipeline of most existing two-stream networks with pixel-wise feature fusion strategies [15, 65]. **P.** stands for **Pixel-Wise Correlation**. (4) Our transformer fusion which explores contextualized geometric cues to better deal with objects sharing the similar visual appearance. **T.** stands for **Transformer Fusion**.

other hand, the depth quality may be influenced by environmental factors during acquisition, such as object distances, as discussed in previous works [15, 37, 69]. (3) Feature alignment. As shown in Fig. 3.1(3), current fusion approaches assume that the sensor calibration is precise and different modalities are accurately aligned at the pixel level, which is not always the case in practice. Despite the recent advances [11, 15, 65, 154], we observe that most existing works are still based on pixel-wise fusion, whose limited awareness of contextualized cues causes the main performance bottleneck.

Recently, transformer has shown its capability in modeling long-range dependencies in various vision tasks [17, 32, 98, 218]. Compared to convolution, transformer is built upon global attention with inter key-query correlation. We observe that by extending the inter key-query correlation to cross-modal key-query correlation, transformer attention suggests a natural way to aggregate RGB-D features. Inspired by this observation, we propose to firstly extract both mixed RGB-D and modality-specific depth features. Then we leverage the depth cues to retrieve geometric information from mixed RGB-D features. As shown in Fig. 3.1(4), the key idea is to leverage contextualized transformer attention to improve the early fusion with enhanced awareness of depth cues. As such, we can better deal with objects sharing a similar visual appearance but at different camera distances or with occlusion, which is challenging for indoor semantic segmentation.

Specifically, our transformer fusion with geometric cues, termed TransD-Fusion, consists of three parts: a self-enhancement module, a bi-directional cross-calibration module, and a depth-guided query design. The enhancement module is realized through the vanilla transformer self-attention. The bi-directional calibration module aims to refine each modality with complementary information: for the depth image, we expect to suppress unsatisfactory responses due to measurement bias; while for the RGB

image, we expect to strengthen the edge awareness on neighboring objects with a similar visual appearance. Finally, the depth-guided query strategy ensures effectively segmenting objects with strengthened discriminability.

To enable position-awareness and leverage locality into our TransD-Fusion, we propose a semantic-aware position encoding generator (S-PE) built upon convolutions. It takes a modality-specific sequence as input and generates a category-aware position encoding. We expect our encoding to spatially constrain the attention around the neighboring area to better segment objects. Moreover, our positional embedding can be learned from hierarchical features, yielding a simple yet efficient encoding for RGB-D fusion. Finally, to tackle the limitations of CNN-based backbones, we implement our TransD-Fusion on Swin-Transformer [98] to better model contextualized dependencies. In brief, our contributions are summarized as follows:

- We propose a novel transformer-based multi-modal fusion to replace the existing pixel-wise fusion modules for RGB-D semantic segmentation.
- We design a semantic-aware position encoding (S-PE) scheme to improve our transformer fusion. The S-PE is dynamically generated from a modality-specific sequence of tokens by a convolutional layer, yielding a spatial constraint on neighboring features for accurate segmentation.
- Our proposed network performs favorably over the state-of-the-art methods on large-scale benchmark datasets by large margins.

5.2 Related Work

5.2.1 RGB-D Semantic Segmentation

How to deal with the complementary depth is a key research topic for RGB-D semantic segmentation. At an early stage, [55] proposes to explore the geometric cues by transforming the depth map into an HHA image. Afterward, researchers take RGB-HHA as input and design various fusion strategies. Several preliminary works [55, 57, 150] fuse the RGB-D images from the input side, treating depth/HHA as additional channels. D-CNN [154] further proposes a depth-aware re-calibration weight to strengthen the discriminatory power during feature modeling. Since then, networks with early-fused RGB-HHA have shown great advances with different forms of weight functions [18, 175, 176]. However, the proposed depth-aware operations are sensitive to depth noise, which might be the performance bottleneck while dealing with unsatisfactory geometry.

To address this issue, several works propose to re-calibrate feature representation with the attention modules. ACNet [65] adopts self-enhancement module with the channel attention [64]. Sharing the same idea, ShapeConv [7] directly integrates the

channel attention into the convolution function. SA gate [15] further leverage spatial attention [161] to calibrate each modality. Another group of works proposes to enhance feature representation with long-range attention. [85] introduces ConvLSTM models in RGB-D fusion to better model contextualized cues. VCD [177] introduces a learned Gaussian convolution kernel to improve spatial-context awareness. Several works [11, 166] integrate depth cues with the deformable convolution [28] to create a more malleable receptive field. Despite the popularity of non-local attention in RGB-D semantic segmentation [11, 85, 166, 177], the capability of modelling long-range dependencies is still limited due to convolution-based feature extraction and fusion. Furthermore, one basic assumption for existing approaches is that the RGB and depth maps are perfectly aligned at the pixel level, which is not always the case in practice due to sensor calibration errors. To tackle these dilemmas, we propose a transformer-based aggregation scheme to explicitly leverage contextualized awareness in multi-modal feature fusion.

5.2.2 Transformer Fusion

There are extensive surveys [56, 73, 144] of transformer applied in vision tasks. ViT and its successors [32, 98, 123] explore the transformer on feature modeling. DERT and its successors [8, 46, 218] adopt transformer on the detection head. Several works on video object tracking [17, 152, 180] adopt transformer to analyze the correlation between search image and template image. Another work on saliency detection [93] adopts transformer as a dimension regulator to convert the sequence of tokens from the encoder space to the decoder space. Different from previous works, our model aims to explore multi-modal cues for feature aggregation. We make full use of attention modules to explicitly preserve, calibrate, and fuse multi-modal information.

By design, attention modules cannot capture order awareness of input tokens. Hence, various researches on position encoding (PE) have been conducted to address this issue. In the literature, two main groups of solutions are proposed: absolute PE and relative PE. Absolute PE generates a unique encoding vector for each position, e.g., 2D sinusoidal embeddings [49, 147], while relative PE proposes to focus on the relative distance of the elements [4, 129, 183]. In vision tasks, previous studies [32, 60, 98, 162, 218] have shown that the relative position enables better performance on the image classification task, while the absolute encoding is more suitable for object detection where the pixel position plays a vital role in segmenting and locating objects. CPVT [24] proposes a conditional PE to leverage the local awareness through a single 2D convolution to improve ViT. However, extending such an idea to RGB-D feature fusion at the semantic level is non-trivial due to the limited feature resolution. In

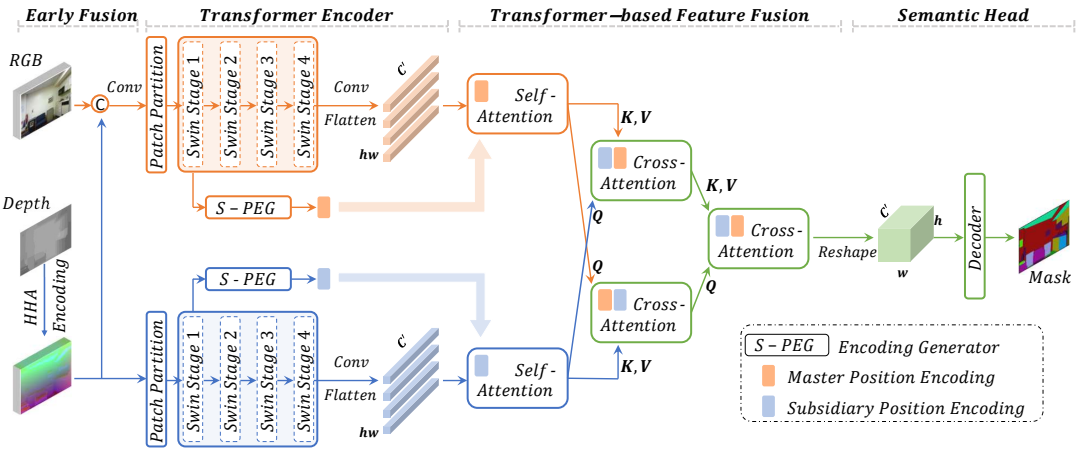


FIGURE 5.2: Overview of the proposed network for RGB-D semantic segmentation. Our TransD-Fusion leverages the transformer attention to aggregate multi-modal features. The self-attention aims to refine modality-specific features, while the cross-attention makes full use of cross-domain cues to firstly calibrate and then combine multi-modal information. The transformer fusion benefits from dynamically generated position encodings to constrain the attention around category-aware neighboring pixels.

contrast, we propose a modality-dependant and semantic-aware PE to improve our transformer fusion with a better position and category awareness.

5.3 Our Approach: TransD-Fusion

5.3.1 Overview

Fig.5.2 presents the overall framework of our network which is composed of a master network, a subsidiary network, and our proposed transformer feature fusion (TransD-Fusion). The master network is an encoder-decoder pipeline with early-fused RGB-HHA images. The encoder stage takes the transformer backbone to extract features from concatenated RGB-HHA input, while the decoder stage takes the classical convolutional head to output the semantic map. The subsidiary network takes HHA images as input. It processes depth features and aims to enhance the master network with geometric cues via our TransD-Fusion. Details are presented in the following sections.

5.3.2 Master-Subsidiary Network

Early fusion has been widely exploited in RGB-D semantic segmentation [18, 154, 175, 176]. It promotes the geometric constraint in the visual appearance from the input side. Nevertheless, the inflexibility of further analysis of multi-modal features at the semantic level severely limits the model performance. To address this issue, we design a master network with early-fused input and a subsidiary stream to enable high-level manipulation with transformer fusion.

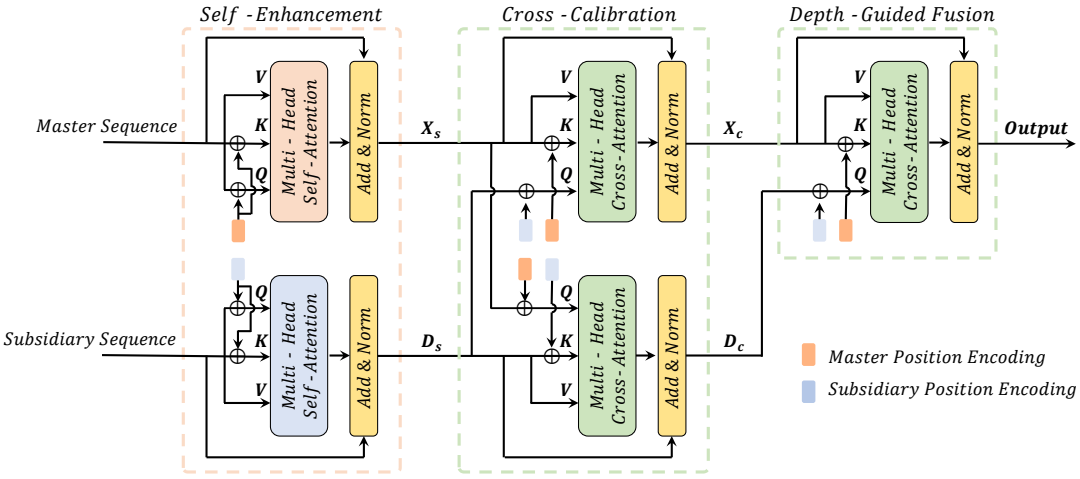


FIGURE 5.3: Our proposed feature enhancement, calibration, and fusion scheme with transformer attention. Best viewed in color.

Given the RGB image $I \in \mathbb{R}^{3\times H\times W}$ and the geometric feature HHA map $D \in \mathbb{R}^{3\times H\times W}$, we can obtain the *master* feature $X \in \mathbb{R}^{3\times H\times W}$:

$$X = Conv_{1\times 1}([I, D]), \tag{5.1}$$

where [] denotes the concatenation along the channel dimension. In such a way, the master feature contains both photometric and geometric information and feats the input shape of the transformer backbone.

To extract multi-modal features, X and D are firstly fed into the patch partition to form two sequences of token separately, and then fed into the Swin-Transformer [98] encoders. A Swin-Transformer layer contains window-based multi-head self-attention (W-MSA), shifted window partitioning configurations (SW-MSA), and a point-wise multi-layer perceptron (MLP) with layer norm (LN). For the i^{th} layer, $i \in \{1, ..., L\}$, it takes the sequence z_{i-1} as input, and outputs the new sequence z_{i+1}:

$$\begin{aligned}
\hat{z}_i &= W\text{-}MSA(LN(z_{i-1})) + z_{i-1};\\
z_i &= MLP(LN(\hat{z}_i)) + \hat{z}_i;\\
\hat{z}_{i+1} &= SW\text{-}MSA(LN(z_i)) + z_i;\\
z_{i+1} &= MLP(LN(\hat{z}_{i+1})) + \hat{z}_{i+1}.
\end{aligned} \tag{5.2}$$

Compared to CNN backbones [59, 133], transformer encoders [98, 147, 198] can better model long-range features. Furthermore, we particularly build upon Swin-Transformer [98] with window attention which reduces the computational complexity. We refer readers to the original paper [98] for more details.

5.3.3 Transformer feature fusion

Given two sequences of tokens $f_X \in \mathbb{R}^{c\times h\times w}$ and $f_D \in \mathbb{R}^{c\times h\times w}$ from different streams, we firstly apply convolutions to f_X and f_D and and output two new feature maps. We expect to strengthen the local awareness and/or to reduce the channel size from c to c'. These two new feature maps are further flattened in spatial dimension, obtaining $f_x \in \mathbb{R}^{c'\times hw}$ and $f_d \in \mathbb{R}^{c'\times hw}$. These flattened features are the inputs of our transformer fusion.

As shown in Fig. 5.3, we propose a three-stage fusion scheme. Firstly, the modality-specific features are enhanced through self-attention. Secondly, a bi-directional calibration is applied with cross-attention. Finally, we initialize a geometry-guided query scheme to accurately segment objects. The attention module is equipped with learnable position encoding to enable both local and semantic awareness. In the following paragraphs, we introduce the details of each component. The benefit of each component can be found in the ablation study Section 5.4.2.3 Table 5.6.

5.3.3.1 Multi-Head Attention in Transformer.

The attention mechanism is the key component of our TransD-Fusion. Given an input sequence of tokens, it is firstly flattened to a 1D vector and generates three intermediate representations: queries Q, keys K, and values V. The attention is formulated as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{5.3}$$

where d_k is the scaling factor. [147] shows that multi-head attention with h heads can further contribute to the model performance by paying diverse attention to features from different positions. The multi-head attention is formulated as follows:

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, ..., head_h)W^O \\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{5.4}$$

where W^O, W_i^Q, W_i^K, W_i^V are the projection matrices.

5.3.3.2 Self-Enhancement.

While (Q, K, V) are from the same input modality, the attention module becomes multi-head self-attention which can be considered as a self-enhancement. It analyzes long-range dependencies and explores contextual information to further improve the modality-specific features. Taking flattened global feature f_x as an example, the self-enhanced global feature X_s can be formulated as:

$$X_s = f_x + MultiHead(Q_x + P_x, V_x + P_x, K_x), \tag{5.5}$$

where (Q_x, K_x, V_x) are the associated intermediate representations and P_x is the associated position encoding. Similarly, we can obtain the self-enhanced geometric feature D_s with the associated position encoding P_d.

5.3.3.3 Cross-Calibration.

The objective of cross-calibration is to reduce the ambiguity in a single modality, e.g., the limited awareness of the geometric cues in visual appearance and measurement bias in geometric features. Different from previous dual attention [15, 161], our cross-calibration is based on transformer attention. We take the queries from one input feature, e.g., Q_{D_s}, to compute the correlation with the keys from the other modality, e.g., K_{X_s}. Formally, we have:

$$\begin{aligned} X_c &= X_s + MultiHead(Q_{D_s} + P_d, K_{X_s} + P_x, V_{X_s}), \\ D_c &= D_s + MultiHead(Q_{X_s} + P_d, K_{D_s} + P_x, V_{D_s}), \end{aligned} \tag{5.6}$$

where (X_s, D_s) are the outputs of the self-enhancement module, $(Q_{X_s}, K_{X_s}, V_{X_s})$ are the associated intermediate representations for master feature X_s, and $(Q_{D_s}, K_{D_s}, V_{D_s})$ for subsidiary feature D_s. We use the same position encodings (P_x, P_d) as in previous self-enhancement module.

5.3.3.4 Depth-Guided Fusion.

To combine master and subsidiary streams, similar to cross-calibration, we use the geometry stream to initialize the query strategy. The difference compared to cross-calibration is that the depth-guided fusion here is non-symmetrical version. We have:

$$Output = X_c + MultiHead(Q_{D_c} + P_d, K_{X_c} + P_x, V_{X_c}) \tag{5.7}$$

where (X_c, D_c) are the outputs of the cross-calibration module, in which the same position encodings (P_x, P_d) are used. The depth-guided fusion module contributes to deal with objects with similar appearance.

5.3.4 Semantic-Aware Position Encoding

We propose a novel position encoding to equip with our transformer attention. Specifically, for each modality, we dynamically generate the position encoding from a lower-dimensional feature map with a larger resolution to make full benefits of spatial information, i.e., the output of the first stage of the encoder.

As illustrated in Fig. 5.4, given the two sequences with higher resolution, we first project the input sequence into a high-dimensional feature space through semantic

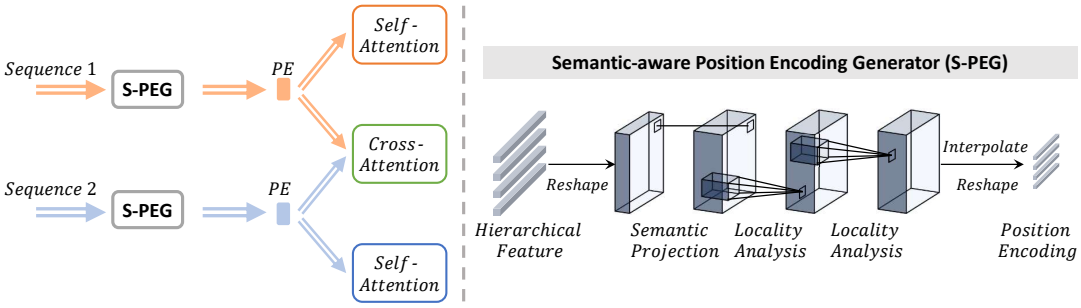


FIGURE 5.4: Our proposed semantic-aware position encoding (S-PE). Left: position encoding flows. Right: illustration of encoding generator. Best viewed in color.

projector $\mathcal{P}$. Here we use the term "semantic" since the embedding dimension is significantly higher than the input dimension. Therefore, we assume that the projection can allow the feature map to contain more semantic cues. Then, we utilize two convolutional modules $\mathcal{F}$ to strengthen the local awareness of the input sequence. Each module consists of 3×3 convolution, batch normalization, and ReLU activation.

5.3.5 Architecture

We follow [43, 207] and apply our transformer fusion on the highest-dimensional features where the resolution is minimized. To generate the output semantic map, we adopt the classical DeeplabV3+ [9] architecture. The whole training process is supervised by the conventional cross-entropy.

In our model, we adopt early fusion together with late fusion. The objective is to fully leverage the depth cues at both the geometric level and semantic level. The idea of using HHA cues to guide RGB-D learning has been widely used in previous RGB-D works, such as DCNN [154], 2.5D [175], Malleable [176], DACN [166], etc. The main difference is that previous works compute local attention (depth weight/offset) from the depth and embed them in convolution, while we explicitly leverage the contextualized awareness to better deal with feature misalignment.

Our fusion strategy substantially differs from the recent fusion works. Specifically, CCFFNet [163] adopts spatial and channel attention on features, while our work is fully based on contextualized attention with tokens. Compared to DeepFusion [83], our cross-modal interaction is bi-directional, while DeepFusion is single-directional (Lidar to camera). Finally, compared to CPVT [24], our positional embedding can better leverage both hierarchical and semantic cues, yielding a simple yet efficient encoding for RGB-D fusion as shown in ablation study.

TABLE 5.1: Performance comparison on RGB-D benchmark datasets.

Source	Method	Backbone	PixelAcc	mAcc	mIoU	f.w.IoU
Comparison on NYUv2 datasets						
ECCV'20	*Malleable* [176]	ResNet-101	76.9	-	50.9	-
ECCV'20	*SAGate* [15]	ResNet-50	77.9	-	52.4	-
SPL'21	*RTLNet* [185]	ResNet-50	77.2	-	53.1	-
TIP'21	*SGNet* [11]	ResNet-101	76.8	63.1	51.1	-
ICRA'21	*ESAnet* [128]	ResNet-34	-	-	51.6	-
CVPR'21	*InverseForm* [6]	ResNet-101	78.1	-	53.1	-
ICCV'21	*ShapeConv* [7]	ResNext-101	76.4	63.5	51.3	63.0
PR'22	*CANet* [206]	ResNet-101	77.1	64.6	51.5	-
TMM'22	*PGDENet* [212]	ResNet-34	78.1	66.7	53.7	-
TMM'22	*TET* [196]	ResNet-50	77.3	59.7	51.8	-
CVPR'22	*Omnivore* [51]	Swin-B	-	-	54.0	-
CVPR'22	*TokenFusion* [158]	SegFormer	**79.0**	66.9	54.2	-
TransD-Fusion (Ours)		Swin-B	78.5	**69.4**	**55.5**	**66.3**
Comparison on SUN-RGBD datasets						
ECCV'18	*DCNN* [154]	VGG-16	-	53.5	42.0	-
ICIP'19	*2.5D* [175]	ResNet-101	82.4	-	48.2	-
ACCV'20	*CANet* [207]	ResNet-101	81.9	-	47.7	-
SPL'21	*RTLNet* [185]	ResNet-50	81.3	-	45.7	-
ICRA'21	*ESAnet* [128]	ResNet-50	-	-	48.3	-
TIP'21	*SGNet* [11]	ResNet-101	82.0	60.7	48.6	-
ICCV'21	*ShapeConv* [7]	ResNet-101	82.2	59.2	48.6	71.3
TETCI'22	*RFNet* [211]	ResNet-34	87.3	59.0	50.7	-
JSTSP'22	*FRNet* [213]	ResNet-34	**87.4**	62.2	51.8	-
TransD-Fusion (Ours)		Swin-B	83.2	**64.1**	**51.9**	**72.8**
Comparison on SID datasets						
TPAMI'17	*Deeplab* [9]	VGG-16	64.3	46.7	35.5	48.5
ECCV'18	*DCNN* [154]	VGG-16	65.4	55.5	39.5	49.9
ArXiv'19	*MMAFNet* [42]	ResNet-152	76.5	62.3	52.9	-
ICCV'21	*ShapeConv* [7]	ResNet-101	**82.7**	70.0	60.6	71.2
TransD-Fusion (Ours)		Swin-B	**82.7**	**72.0**	**62.2**	**71.5**

5.4 Experiments

We evaluate our model on three benchmark RGB-D datasets, i.e., NYUv2 [132], SUN-RGBD [135], and Stanford 2D-3D-Semantic Indoor Dataset (SID) [2]. We analyze the performance with common metrics, i.e., Pixel Accuracy (PixelAcc), Mean Accuracy (mAcc.), Mean Region Intersection Over Union (mIoU), and Frequency Weighted Intersection Over Union (f.w.IoU). Let s_i be the number of pixels with the ground truth class i. n_{ij} denotes the number of pixels with ground truth class i and but predicted as class j. n_c denotes the number of total classes, and $s = \sum_i s_i$ is the number of all pixels. Mathematically, the metrics are defined by:

- Pixel Acc: $PixelAcc = \sum_i \frac{n_{ii}}{s}$
- mean Acc: $mAcc = \frac{1}{n_c} \sum_i \frac{n_{ii}}{s}$
- mean Intersection over Union: $mIoU = \frac{1}{n_c} \sum_i \frac{n_{ii}}{s_i + \sum_j n_{ji} - n_{ii}}$
- Frequency Weighted Intersection over Union: $f.w.IoU = \frac{1}{s} \sum_i s_i \frac{n_{ii}}{s_i + \sum_j n_{ji} - n_{ii}}$

We follow conventional train-test protocols for RGB-D benchmarks experiments [7,15, 206]. On NYUv2 with 40 categories, we follow the widely-used split with 795 images used for training and the rest 654 images are for testing among the 40 classes. On SUN-RGBD with 37 categories, we follow the widely-used split with 5,285 images for training and the rest 5,050 images for testing. On SID with 13 categories, we train our model on areas 1, 2, 3, 4, and 6 and Area 5 is for testing. During training, we resize the images to a random ratio between 0.5 and 2.0 and explore left-right flipped images. We choose the standard SGD optimizer with momentum to train our model following the "poly" learning rate policy. The initial learning rate is set to 0.007, the momentum is fixed to 0.9, and the weight decay is set to 0.0001. For inference, we evaluate our model with multi-scale testing strategies, i.e., {0.5, 0.75, 1.0, 1.25, 1.5, 1.75}. Similar to previous works [7,15,55,154], we take RGB and HHA images as input. The HHA maps are generated according to [55] during pre-processing. To make a fair comparison, our transformer backbone is initialized with the weights pre-trained on ImageNet-1K [29] as CNN backbones.

5.4.1 Comparison with the State-of-the-Art Models

5.4.1.1 Quantitative Comparison.

Table 5.1 illustrates the quantitative comparison on NYUv2. We observe that the models with transformer encoders [51,158] outperform CNN approaches. Our TransD-Fusion even surpasses transformer counterparts on mIoU and sets a new state-of-the-art record, i.e., 55.5% with 1.7 FPS. We also report the performance of the SUN-RGBD dataset and SID dataset. Our TransD-Fusion (Swin-B) outperforms the concurrent

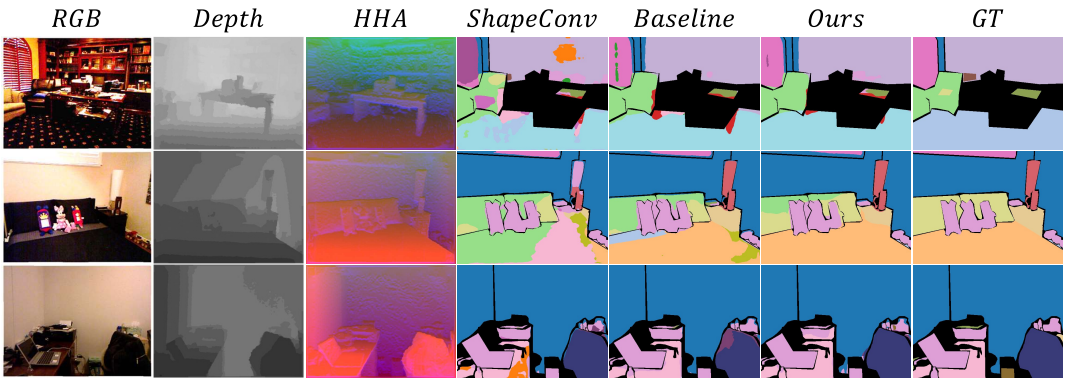


FIGURE 5.5: Qualitative comparison. We compare our TransD-Fusion with SOTA CNN model and with vanilla transformer backbone. The black regions in semantic maps indicate the ignored category.

ShapeConv [7] which is also built upon DeepLabV3+ with a large margin: 1.4% ↑ mIoU on SUN-RGBD and 1.6% ↑ mIoU on SID. The leading performances on indoor benchmarks validate our effectiveness.

5.4.1.2 Qualitative Comparison.

Fig. 5.5 illustrates semantic maps generated by SOTA CNN model ShapeConv [7], transformer baseline (with DeeplabV3+ [9]), and our TransD-Fusion. Compared to ShapeConv, we observe that transformer models can better generate contextualized features and yield results closer to the ground truth. Compared to the transformer baseline, TransD-Fusion can further explore geometric cues to distinguish objects sharing similar visual appearances, leading to a more accurate semantic segmentation.

5.4.2 Ablation Studies

5.4.2.1 Robustness against Alignment Bias.

We analyze the robustness of different fusion approaches against sensor misalignment, i.e., RGB and Depth maps are not accurately aligned at the pixel level. Specifically, we simulate a calibration error on NYUv2 by additionally cropping 20 pixels from the RGB input and obtaining a misaligned dataset. We retrain our TransD-Fusion (Swin-B) and the SOTA CNN model ShapeConv with early-fused input. To make a fair comparison, we additionally build two late-fusion baseline networks with the Swin-B backbone. The features are combined with attention modules such as SA gate [15] (denoted as Swin + **SA**), or with simple pixel-wise concatenation and convolution (denoted as Swin + **Conv**).

The performances under the inferior condition are presented in Fig. 5.6 and in Tab. 5.2. Since **SA** and **Conv** are built upon the pixel-wise correlation between different modalities at the semantic level, their performances significantly drop when the features are no more accurately aligned. We observe 1.8% mIoU degradation on Swin

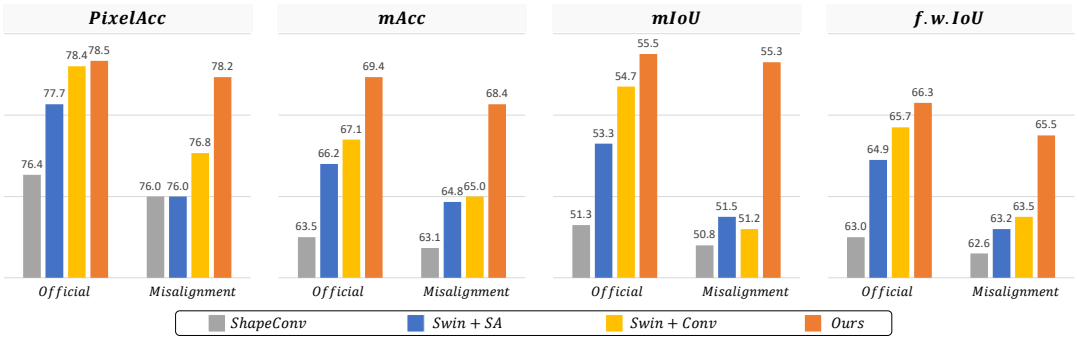


FIGURE 5.6: Robustness analysis on the simulated misaligned NYUv2 dataset. Our TransD-Fusion leads to a more stable performance compared to SOTA fusion approaches.

TABLE 5.2: **Robustness analysis** on the simulated misaligned NYUv2 dataset. Our TransD-Fusion leads to a more stable and superior performance.

Method	Crop (pixel)	PixelAcc	mAcc	mIoU	f.w.IoU
ShapeConv	40	74.7	62.5	49.2	61.1
Swin + CC	40	76.1	64.1	50.5	62.8
Swin + SA	40	75.7	63.1	50.7	62.2
TransD-Fusion	40	**78.1**	**69.1**	**55.1**	**65.7**
ShapeConv	60	74.6	60.7	48.2	60.8
Swin + CC	60	74.8	63.1	48.8	61.4
Swin + SA	60	75.3	63.7	49.7	61.9
TransD-Fusion	60	**77.9**	**68.8**	**54.8**	**65.5**

+ **SA** and 3.5% mIoU degradation on Swin + **Conv**, respectively. In contrast, our TransD-Fusion only drops 0.2% on mIoU. The stable performance against misalignment can be attributed to our fusion design which is built upon the contextualized correlation, yielding a more soft and robust fusion scheme for RGB-D semantic segmentation.

5.4.2.2 Generalization Capability.

Our TransD-Fusion can be used as a plug-in module. To demonstrate its generalization properties, we conduct experiments with several widely used semantic segmentation architectures, such as Segmenter [137], PSPnet [198], and DeeplabV3 [10] or DeeplabV3+ [9]. We use Swin-B as the backbone for all architectures and report the performances on the NYUv2 dataset in Table 5.3. "Baseline" presents the result obtained with RGB-HHA input through the Swin backbone under the corresponding architecture. "Ours" presents the result obtained by further applying TransD-Fusion between backbone and decoder. "+↑" shows the performance gain with our approach. We observe that TransD-Fusion consistently enables progress over the baseline performance in each architecture, demonstrating the flexibility and effectiveness of our method.

TABLE 5.3: Generalization capability. We report the performance comparison with different architectures on NYUv2 dataset.

Architecture	Segmenter [137]			PSPnet [198]			DeeplabV3 [10]			DeeplabV3+ [9]		
	Baseline	Ours	+↑	Baseline	Ours	+↑	Baseline	Ours	+↑	Baseline	Ours	+↑
PixelAcc	69.6	70.6	**1.0**	75.0	75.6	**0.6**	76.5	77.3	**0.8**	78.0	78.5	**0.5**
mAcc	55.3	56.9	**1.6**	61.4	63.4	**2.0**	64.1	64.4	**0.3**	66.1	69.4	**3.3**
mIoU	42.5	44.3	**1.8**	49.2	50.6	**1.4**	52.3	53.5	**1.2**	53.8	55.5	**1.7**
f.w.IoU	54.5	55.7	**1.2**	60.7	62.1	**1.4**	62.7	66.5	**3.8**	65.1	66.3	**1.2**

TABLE 5.4: Ablation study on our fusion design. We report the performance comparison with other fusion alternatives on NYUv2 dataset.

# Descrip.	F1 (Early)	F2 (Add)	F3 (Conv)	F4 (SA)	F5 (TransT)	**Ours**
PixelAcc	78.0	77.4	77.0	77.7	76.5	**78.5**
mAcc	66.1	64.6	63.9	66.2	63.2	**69.4**
mIoU	53.8	52.8	52.2	53.3	51.4	**55.5**
f.w.IoU	65.1	64.3	63.6	66.2	62.9	**66.3**

5.4.2.3 Comparison with Previous Fusion schemes.

To verify whether our transformer fusion with contextualized awareness is efficient, we conduct experiments by replacing TransD-Fusion with other approaches. To make a fair comparison, all the experiments use the Swin-B backbone with DeeplabV3+ architecture. We have "$F1$" denoting the early fusion for RGB-HHA input. "$F2$"-"$F5$" adopt the conventional two-streaming design with different late fusion designs: "$F2$" with pixel-wise addition; "$F3$" with concatenation-convolution; "$F4$" with SA gate [15]; "$F5$" with TransT [17]. The quantitative results can be found in Table 5.4. We observe that our TransD-Fusion enables a better result compared to other fusion methods.

Note that SA gate [15] and TransT [17] are initially applied with CNN backbones and are re-employed with our transformer backbone. Our TransD-Fusion differs from these two designs in several aspects: **(A)** Compared to SA gate, our work is based on transformer attention, while SA gate adopts conventional dual attention [161]. Our superior performance ("$Ours$" >"$F4$") shows that we can better model long-range dependencies to effectively aggregate multi-modal features. **(B)** Both TransT and our TransD-Fusion belong to transformer fusion frameworks. However, TransT is initially designed to compute the correlation between two RGB images, hence focusing on shared features between two inputs. Extending TransT to RGB-D fusion is not trivial since there exist both common and different information in these two modalities. Empirically, as shown in Table 5.4, TransT ("$F5$") leads to significantly dropped performance which is even lower than simple fusion designs such as addition and convolution ("$F2$"-"$F3$"). Different from TransT, we design a depth-guided query strategy to deal with objects that share a similar visual appearance. Furthermore, we leverage a category-aware position embedding to equip with our attention, while vanilla TransT uses an absolute encoding which is not suitable for multi-modal fusion.

5.4.2.4 Comparison with other position encodings (PEs).

Prior works adopt different PEs that focus on order awareness to improve feature extraction. The PE in our TransD-Fusion plays a more vital role since it should

TABLE 5.5: Ablation study on positional encoding. We replace our positional encoding with other alternatives and report the performance comparison on NYUv2 dataset.

# Descrip.	P1 (w/o)	P2 (Abs)	P3 (Relative)	P4 (L4)	P5 (L3)	P6 (L2)	P7 (CPVT)	**Ours**
PixelAcc	77.8	78.1	**78.5**	78.3	78.3	78.4	78.4	**78.5**
mAcc	68.2	67.9	67.5	66.6	67.4	68.8	68.3	**69.4**
mIoU	53.9	54.2	54.9	54.2	54.3	54.9	54.8	**55.5**
f.w.IoU	65.1	65.6	65.7	65.5	65.8	66.2	66.0	**66.3**

be locality-aware for better segmentation and be category-dependent for multi-modal fusion. To validate the superiority of our proposed PE, we conduct experiments by removing or replacing our encoding with other approaches and report the performance in Table 5.5. We have: "*P1*" without PE; "*P2*" with absolute PE; "*P3*" with relative PE. Since our PE can be learned from a hierarchical feature with higher resolution to fully excavate the spatial cues, we also conduct experiments to analyze the influence of feature resolution. We denote: "*P4*" for PE learned from the output of Layer 4; "*P5*" learned from Layer 3 output; "*P6*" learned from Layer 2 output. We replace our PE with the concurrent CPVT [24] by re-implementing it in our TransD-Fusion, denoted as "*P7*". Under consideration of a fair comparison, we apply CPVT to learn features from Layer 1 output as our S-PE.

Empirical results in Table 5.5 show that there exists significant degradation on mIoU after removing or replacing our S-PE with conventional PEs. This validates the effectiveness of our S-PE that can better constrain the transformer attention for multi-modal fusion. We also observe that the spatial dimension plays an imperial role for our S-PE. When the spatial resolution decreases, i.e., from Layer 1 output to Layer 4 output, the performances with our S-PE drop as well. Compared to the concurrent CPVT, our superior performance demonstrates that we can better leverage locality awareness.

5.4.2.5 Key Components Analysis of TransD-Fusion.

In this section, we conduct studies to verify the importance of the key components of TransD-Fusion: Master stream (master), Subsidiary stream (sub), Self-Enhancement (SE), Cross-Calibration (CC), and Depth-Guided Fusion (DGF). All the experiments are built upon the Swin-B backbone and we report the associated model size for each module. We remove partially or entirely the key components. To make a fair comparison, we additionally conduct experiments with conventional fusion strategies such as element-wise addition (**Add**), concatenation-convolution (**Conv**), and the concurrent SA module [15] under the same architecture. Note that the SA module is initially applied for middle fusion. Under the consideration of a fair comparison we

TABLE 5.6: **Key components analysis** on NYUv2 dataset.

#	*Master*	*Sub*	*Add*	*Conv*	*SA − M*	*SE* 42 Mb	*CC* 37 Mb	*DGF* 5Mb	Metric	
									mAcc	mIoU
1	✓								66.1	53.8
2	✓					✓			67.0	53.9
3	✓	✓	✓						61.4	51.2
4	✓	✓		✓					67.1	54.6
5	✓	✓						✓	68.5	55.1
6	✓	✓				✓		✓	68.6	55.2
7	✓	✓			✓			✓	66.5	54.3
Ours	✓	✓				✓	✓	✓	**69.4**	**55.5**

adopt the same middle fusion design to merge RGB-D features at each scale. This is denoted as SA-M in Table 5.6.

We observe from Table 5.6 that after removing the cross-calibration module, the performance drops since the modality-specific features can no more benefit from complementary cues. Without self-enhancement, the performance further degrades. While further replacing the depth-guided fusion strategy with pixel-wise fusion module, we can observe a significantly drop, i.e., 3.9% ↓ on mIoU with **Add** and 0.5% ↓ on mIoU with **Conv**. These results validate the necessity of leveraging the long-range dependencies for feature fusion. Finally, by comparing lines #5-#6, we observe that the SE plays a minimal role. Therefore we try to replace our SE with the SA module [15]. However, the performance significantly drops, which shows the importance of our self-attention that fully leverages and preserves modality-specific features with contextualized cues.

5.5 Conclusion

In this chapter, we propose a novel RGB-D fusion scheme for semantic segmentation. Different from previous fusion designs built upon pixel-wise correlation, our network fully explores the transformer attention to aggregate multi-modal features with contextualized cues. Additionally, we design a novel position encoding generator to better leverage the locality awareness into our transformer fusion. Extensive ablation studies verify the robustness against misalignment and the generalization property of our TransD-Fusion. The comparison with previous works on fusion design and position encoding further validates the effectiveness of our proposed approach. Experiments on challenging RGB-D benchmarks demonstrate that our TransD-Fusion performs well over the state-of-the-art methods by large margins.

Despite the fact that there exist a number of various fusion methods from pixel-wise aggregation to cross-modal contextualized attention, it is still unclear which should layer apply the fusion method. In the literature, early, middle, and late fusion designs

have been widely explored. However, existing fusion architectures are designed in a handcrafted manner, which is agnostic of input data. For example, researchers have shown that stemming layers focus more on low-level geometric features, while deeper layers focus more on semantic cues. Since a depth map is a sort of low-level geometric input, it is trivial and intuitive that while the depth is good, it should play a more important role at the stemming stage. While the depth quality is unsatisfactory due to the measurement bias, it should play a more important role at a deeper stage. From this perspective, we discuss in the following chapter our proposed robust fusion design which can learn the trade-off between the early and late fusion with respect to the depth quality.

Chapter 6

Robust RGB-D Fusion for Saliency Detection

Efficiently exploiting multi-modal inputs for accurate RGB-D saliency detection is a topic of high interest. Most existing works leverage cross-modal interactions to fuse the two streams of RGB-D for intermediate features' enhancement. In this process, a practical aspect of the low quality of the available depths is not considered. In this chapter, we aim for RGB-D saliency detection that is robust to the low-quality depths which primarily appear in two forms: inaccuracy due to noise and the misalignment to RGB. To this end, we propose a robust RGB-D fusion method that benefits from (1) layer-wise, and (2) trident spatial, attention mechanisms. On the one hand, layer-wise attention (LWA) learns the trade-off between early and late fusion of RGB and depth features, depending upon the depth accuracy. On the other hand, the trident spatial attention (TSA) aggregates the features from a wider spatial context to address the depth misalignment problem. The proposed LWA and TSA mechanisms allow us to efficiently exploit the multi-model inputs for saliency detection, while being robust against low-quality depths. Our experiments on five benchmark datasets demonstrate that the proposed fusion method performs consistently better than the state-of-the-art fusion alternatives. The source code will be made publicly available.

6.1 Introduction

Saliency detection aims to segment image contents that visually attract human attention the most. Existing RGB-based saliency detection methods [91, 172, 194, 199] achieve promising results in generic settings. However, in cluttered and visually similar backgrounds, they often fail to perform accurate detection. Therefore, many recent works [37, 70, 115, 199] exploit image depths as additional geometric cues, in the form of RGB-D inputs, to improve the saliency detection performance in difficult scenarios.

Given accurate and well-aligned depths, existing RGB-D methods perform well even in difficult scenarios. Unfortunately, this is not often the case in practice. Sometimes,

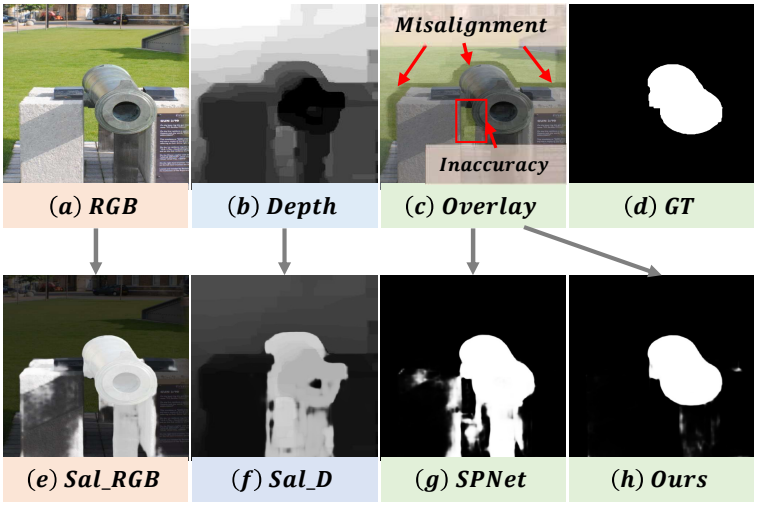


FIGURE 6.1: Motivation of layer-wise attention. (a) and (b) are the paired RGB-D inputs. (e) and (f) are the associated saliency maps generated from the single-modal input which are sub-optimal. (c) is the overlay between RGB and depth image. It can be seen that there exists inaccurate measurement from the depth map and misalignment between both modalities, which is the main performance bottleneck for existing models. To address this issue, we propose a robust RGB-D fusion to explicitly model the depth noise for saliency detection. Compared to the state-of-the-art method SPNet [210] (g), our model favorably yields results (h) closer to the ground-truth mask (d).

only low-quality depths can be acquired, depending upon the scene and the source of depths. For example, depths from multi-view stereo cameras are often noisy [21, 178] and asynchronous depth cameras are spatially misaligned [111], as shown in Figure 6.1. Other environmental factors such as object distance, texture, or even lighting conditions during the acquisition can also degrade the depth quality [37, 69, 168, 195]. Therefore, a method that can still exploit the geometric cues, while being robust to the depth quality discrepancy is highly desired.

We observe that most existing methods perform unsatisfactorily on datasets with low-quality depths. This is primarily because of the commonly used fusion technique [38, 100, 109, 164, 187, 210] that merges the parallel streams of RGB and depth with equal importance, while being agnostic to misalignment. Less accurate depths are evidently expected to play a smaller role than their counterpart. On the other hand, the possibility of misalignment between RGB and depth needs to be considered during the fusion process.

In this work, we propose a robust RGB-D fusion method that addresses the aforementioned problems of inaccurate and misaligned depths. The proposed method uses a layer-wise attention (LWA) mechanism to enable the depth quality aware fusion of RGB and depth features. Our LWA attention learns the trade-off between early and

late fusions, depending upon the provided depth quality. More precisely, LWA encourages the early fusion of the depth features for high-quality depth inputs, and vice versa. Such fusion avoids the negative influence of the spurious depths, while being opportunistic when high-quality depths are provided. In other words, the good-quality depth should play an important role in early layers thanks to its rich and exploitable low-level geometric cues, while low-quality depth should be more activated at semantic levels.

To address the problem of misaligned depths, we introduce the trident spatial attention (TSA) that aggregates features from a wider spatial context. The introduced TSA is used to replace the vanilla spatial attention, enabling the aligned aggregation of the misaligned features. In particular, our TSA requires only minor additional parameters and computation, while being sufficient to address the problem of misalignment. Note that the misalignment problem often exists only locally therefore the global context (at the cost of additional computation) may not be necessary. Such an example is shown in Figure 6.1(c). We improve the vanilla spatial attention with different scales of receptive fields, yielding a simple yet efficient manner to replace the pixel-wise correspondence with region-wise correlation. Finally, the new spatial attention is adaptively merged with channel attention to form our hybrid fusion module.

In summary, our major contributions are listed below:

- We study the problem of RGB-D fusion in a real-world setting, highlighting two major issues, inaccurate and misaligned depths, for accurate saliency detection.
- We introduce a novel layer-wise attention (LWA) to automatically adjust the depth contribution through different layers and to learn the best trade-off between early and late fusion with respect to the depth quality.
- We design a trident spatial attention (TSA) to better leverage the misaligned depth information by means of aggregating the features from a wider spatial context.
- Extensive comparisons on five benchmark datasets validate that our fusion performs consistently better than state-of-the-art alternatives, while being very efficient.

6.2 Related Work

There are extensive surveys of salient object detection [5, 153, 209] and on attention modules [73, 144] in the literature. In the following, we briefly review related works.

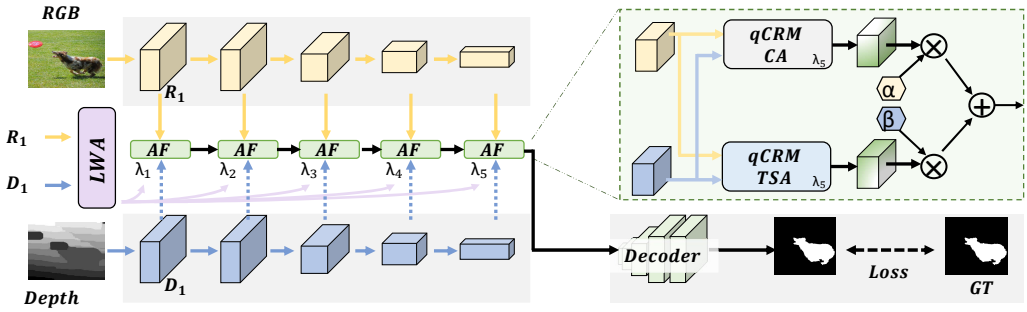


FIGURE 6.2: Architecture. Our proposed network consists of a layer-wise attention (LWA, see Section 6.3.1) and an adaptive Attention Fusion (AF, see Section 6.3.2). LWA aims to find the best trade-off for early and late fusion depending on the depth quality, while AF leverage cross-modal cues to compute the shared representation with channel attention and improved spatial attention (TSA). CRM is from [69].

6.2.1 RGB-D Fusion for Saliency Detection

In the literature, we can divide current models into two types of architectures: single-stream and multi-stream schemes. The main difference is in the number of encoders. Single-stream networks are commonly lighter compared to multi-stream works. In [44, 203], the authors proposed the concatenation of RGB-D images from the input side and then feed them into a single encoder-decoder architecture. From another perspective, [115] introduces a depth distiller to enable cross-modal knowledge distillation, leading to a lightweight inference strategy with RGB-only input. Other works [70, 141, 200] propose to directly integrate low-level geometric cues in the RGB stream to strengthen the RGB features. Despite the proven result with a lightweight model, single-stream models fail to explicitly model cross-modal correlation in complex scenarios, which is the main performance bottleneck.

Recently, multi-stream architectures have drawn increasing research interests. A number of works [38, 39, 100, 109, 168, 190, 210] propose to explicitly model RGB and depth cues through two parallel encoders and then aggregate multi-modal features through multi-scale fusion schemes, leading to better performance compared to their counterpart. In the literature, we can group existing works into three categories based on the fusion schemes: 1) depth-guided fusion, 2) discrepant fusion, and 3) multi-scale fusion. Depth enhanced fusion models [38, 109, 195] often adopts an asymmetric fusion scheme that the depth features are fused into RGB features at each level to improve the boundary awareness. However, these models are sensitive to depth noise and the performance is significantly degraded when depth maps are under inferior conditions. Other works [69, 109, 168, 187, 192] propose to merge multi-modal cues through a discrepant design. In [187], the authors adopt different fusion designs for low-level and high-level features, i.e., RGB to calibrate depth in earlier layers and depth to calibrate RGB in deeper layers. [69, 192] only fuse features at semantic levels, i.e.,

outputs from the last three layers. Different from discrepant and asymmetric designs, a number of works [39, 190, 199, 210] realize bi-directional cross-modal interaction at each scale of the neural network. This fusion design, also known as middle fusion, has shown plausible performance in saliency benchmarks. Nevertheless, we observe that most existing works treat RGB and depth equally to form the shared features, paying little attention to explicitly model the measurement bias and alignment issue. [195] has introduced a weighting strategy to deal with the measurement bias. However, their weighting scheme assumes the perfect alignment between multi-modal features. Different from previous works, we estimate the depth quality index by leveraging contextualized awareness. We show through empirical comparison that our approach can better model the depth quality to adjust the contribution.

6.2.2 Attention for Cross-Modal Interaction

Self-attention modules [43, 98, 121, 147, 156, 161] have been proven to be efficient for visual tasks. Inspired by their success, a number of RGB-D saliency works [38, 69, 92, 187, 199] leverage self-attention as an augmentation to better preserve, calibrate, and fuse multi-modal features. [69, 199] explicitly leverages the attention along the channel direction to calibrate each modality. [92] introduces a mutual and non-local strategy to learn the spatial cues from one modality and apply it to the other. Several recent works [39, 93, 100] further explore the long-range dependencies with transformer attention [147].

Despite the popularity of contextualized attention, we observe that these modules often require a significant computational cost. Therefore, fusion with transformer attention is often realized with a small resolution feature map, i.e., at deeper layers of encoders [39, 92, 100]. To benefit from the spatial cues at each stage, a number of works [38, 39, 187] adopt the hybrid models with vanilla spatial and channel attention from [161] to aggregate features at each stage. However, vanilla spatial attention is agnostic of feature misalignment. Moreover, these hybrids treat spatial and channel attention equally, failing to be adjusted with respect to the network depth. Different from previous works, we propose a simple yet efficient trident spatial attention that can better model contextualized awareness compared to its counterpart. Furthermore, we integrate our spatial attention with channel attention in a parallel scheme, yielding a more robust fusion strategy with adaptive weights.

6.3 Method

Figure 6.2 presents the overall framework of our network. We first extract RGB and depth features through parallel encoders. Then, these features are gradually merged through our proposed fusion module with respect to the depth noise. Specifically,

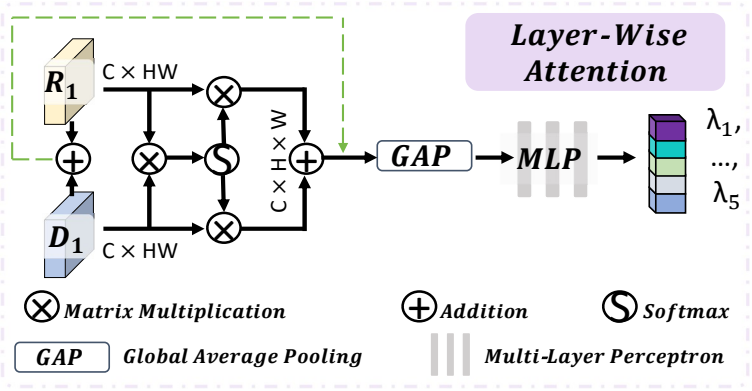


FIGURE 6.3: Layer-Wise Attention (LWA). It takes paired RGB and depth low-level features as input, i.e., features from first layer R_1 and D_1, and outputs confidence values λ_i to adjust the depth contribution for the i^{th} stage fusion. Specifically, we first leverage non-local attention to enable bi-directional interaction. Then, the cross-calibrated features are merged together and fed into a MLP to model the depth contribution. The dashed shortcut (**green**) stands for the residual addition for reducing gradient vanishment.

to tackle the inaccurate measurement bias, we propose a layer-wise attention (LWA) to control the depth contribution. To deal with feature misalignment, we propose a hybrid attention fusion (AF) module with a trident spatial attention and an adaptively merged channel attention. Details of each component are presented in the following sections.

6.3.1 Layer-Wise Attention

We observe that there exist several depths with unsatisfactory quality as shown in Figure 6.1. Inspired by this observation, we propose a depth quality indicator that aims to explicitly model the depth contribution. Our intuition is that while dealing with low-quality depth at early layers, the network should have a higher confidence value on the RGB feature instead of equally average the multi-modal cues.

As depicted in Figure 6.3, our layer-wise attention takes outputs from the first encoder layer as input, i.e., $R_1 \in \mathbb{R}^{C\times H\times W}$ and $D_1 \in \mathbb{R}^{C\times H\times W}$. We argue that these features contain more heterogeneous and modality-specific cues compared to semantic-level features which are homogenized. With R_1 and D_1, we first compute the similarity between the two modalities. Instead of directly realizing the pixel-wise multiplication, we leverage the contextualized awareness to avoid the feature misalignment and focus on the measurement bias. Specifically, R_1 and D_1 are firstly fed into $Conv_{1\times1}$ and flattened to form $R_1' \in \mathbb{R}^{C\times HW}$ and $D_1' \in \mathbb{R}^{C\times HW}$. These new features are then fed into the matrix multiplication as shown in Eq. 6.1. To normalize the obtained attention map, we further apply the softmax function to adjust the weight. Further,

the normalized weight map is multiplied to flattened R_1 and D_1 to improve the cross-modal awareness. Finally, the retrieved RGB and depth attention maps are merged together through addition. Formally, the similarity matrix can be formulated as:

$$Attention(R_1', D_1') = softmax(\frac{R_1' D_1'^T}{\sqrt{c}})(R_1' + D_1'). \tag{6.1}$$

Similar to self-attention works [147, 156], we add a skip connection with early fused RGB-D features to stabilize the training procedure. Once we obtain the similarity matrix, we seek to explicitly quantify the depth measurement bias. Specifically, we first extract the feature vector with the help of global average pooling (GAP) and then feed it into a multi-level perceptron (MLP) to estimate the confidence values. We particularly estimate distinct values to explicitly guide feature fusion at different scales. The adaptive weight $\lambda \in \mathbb{R}^5$ can be formulated as:

$$\lambda = MLP(GAP(Attention(R_1', D_1'))). \tag{6.2}$$

Finally, let R_i and D_i be the encoded RGB-D features from the i^{th} layer. Instead of equally averaging both feature maps by $R_i + D_i$ which is agnostic of input depth quality, our proposed fusion by $R_i + \lambda_i D_i$ can better merge multi-modal features with context awareness.

At first glance, our attention map is similar to non-local attention [156] which has been applied in S2MA [92] or to transformer attention [147] which has been applied in TriTrans [100]. However, our method differs from previous works in two aspects, i.e., the purpose and the model size. Compared to S2MA which uses non-local attention for cross-modal calibration, our work aims to analyze the similarity between multi-modal features and assign a confidence value to the depth cues. Compared to TriTrans which adopts multi-head transformer attention to fuse features at the deepest layer, our design is significantly lighter with only one head and is applied to low-level features with higher resolution. The concurrent work DFMnet [195] adopts Dice similarity coefficient [104] to analyze the depth quality. However, it simply multiplies RGB and depth features with the pixel-wise association, paying little attention to explicitly model measurement bias and the misalignment in a separate manner.

6.3.2 Adaptive Attention Fusion

Existing methods [38, 39, 69, 187, 199] often adopt attention modules, i.e., spatial attention (SA) and channel attention (CA), to enable cross-modal interaction, with few methods pay attention to inherent feature misalignment. While by design CA is more robust to this issue due to the squeezed spatial resolution, the vanilla SA has more difficulties dealing with this inferior condition since it assumes a perfect alignment

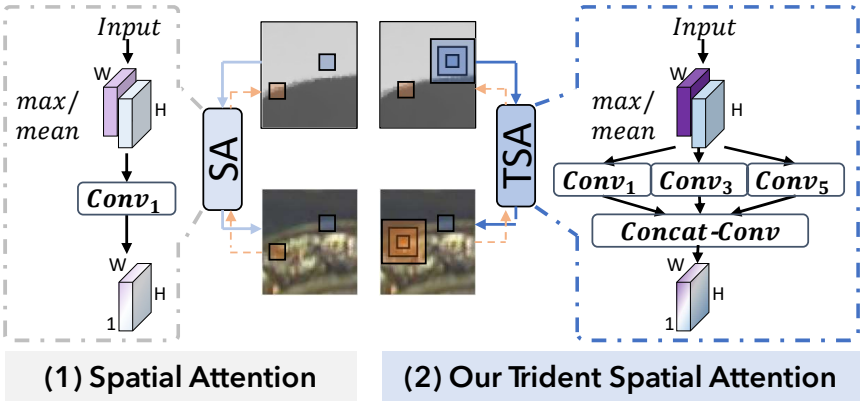


FIGURE 6.4: Motivation of attention fusion. (1) Vanilla spatial attention [38, 161, 187] which is not suitable for cross-modal interaction due to feature misalignment. (2) We propose a trident spatial attention (TSA) with dilated receptive field to better leverage contextualized awareness. Better to zoom in.

between different modalities. To address this dilemma, we propose to improve the current SA with enlarged global awareness, yielding a simple yet efficient manner to replace the pixel-wise alignment with region-wise correlation. Furthermore, current works simply apply CA and SA one by another [38, 39] or equally average them to form the output [187]. These works are agnostic to the network depth that SA and CA still contribute equally at each stage. Previous work [108] has shown that layers with different depths will pay attention to different contexts. Therefore, we seek to introduce an adaptive fusion strategy with learnable weights to automatically adjust the contribution of each attention at different levels.

Formally, let an input feature map $f \in \mathbb{R}^{C\times H\times W}$. The vanilla SA firstly squeeze the channel dimension with average and max pooling across the channel, denoted as $CAP(\cdot)$ and $CMP(\cdot)$, respectively, to obtain the spatial map $f' \in \mathbb{R}^{2\times H\times W}$. Then, from f' SA learns a 2-D weight map $SA \in \mathbb{R}^{1\times H\times W}$:

$$\begin{aligned} f' &= Concat(CAP(f), CMP(f)); \\ SA(f) &= \sigma(Conv_1(f'))), \end{aligned} \tag{6.3}$$

where $\sigma(\cdot)$ is the Sigmoid activation, $Conv_1$ stands for the convolution with dilation 1. To improve global awareness, we replace the current convolution with trident branches where each branch focuses on learning features with different scales, as shown in Figure 6.4. Our proposed trident spatial attention can be formulated as:

$$\begin{aligned} TSA(f) = \sigma(&Concat(Conv_1(f') \\ &Conv_3(f') \\ &Conv_5(f')). \end{aligned} \tag{6.4}$$

where $Conv_1$, $Conv_3$, $Conv_5$ stand for convolutions with different dilation values.

To attentively aggregate multi-modal features, we follow the pipeline of the Cross-Reference Module (CRM) as suggested in DCF [69]. Formally, let R_i and D_i the paired RGB-D input for the fusion module, we first compute the modality-specific channel CA_r and CA_d, as well as the shared channel attention CA_f as follow:

$$\begin{aligned} CA_r &= CA(R_i); \quad CA_d = CA(D_i); \\ CA_f &= norm(max(CA_r, CA_d)); \end{aligned} \tag{6.5}$$

The vanilla CRM benefits from channel attention to realize the self- and cross-calibration before the feature fusion. We have:

$$\begin{aligned} CRM(R_i, D_i) = Concat(&CA_f \otimes CA_r \otimes R_i; \\ &CA_f \otimes CA_d \otimes D_i); \end{aligned} \tag{6.6}$$

We refer readers to the original paper [69] for more details on the cross-modal interaction. In our application, we replace the final concatenation with adaptive addition with respect to depth quality and form our $qCRM^{CA}$ as follow:

$$\begin{aligned} qCRM^{CA}(R, D) = &CA_f \otimes CA_r \otimes R_i + \\ &\lambda \cdot CA_f \otimes CA_d \otimes D_i; \end{aligned} \tag{6.7}$$

Moreover, we additionally design another branch where the CA is replaced by our proposed TSA. This new branch is termed as $qCRM^{TSA}$. We further learn two scalar values α and β to adaptively weight CRM^{TSA} with the original branch CRM with channel attention. Finally, our adaptive fusion (AF) can be formulated as:

$$AF(R, D) = \alpha \cdot qCRM^{CA}(R_i, D_i) + \beta \cdot qCRM^{TSA}(R_i, D_i) \tag{6.8}$$

6.3.3 Architecture

In this chapter, we propose a novel fusion design that can be easily adapted to any existing architecture. To compete with the state-of-the-art performance, we choose Res2Net [47] as our backbone to extract features. Our decoder is the same as SP-Net [210]. Specifically, it consists of five-level RFB blocks [94]. Each block is skipped and connected with the fused encoded features. However, different from SPNet with a triple decoder to explicitly both modality-specific and shared features, we only maintain one decoder to decode our efficiently fused features. Our network is supervised by conventional IoU and BCE losses.

TABLE 6.1: Quantitative comparison with different fusion designs. We replace our fusion module with five SOTA fusion modules and retrain the new networks with the same training setting. ↑ (↓) denotes that the higher (lower) is better.

Dataset	Size↓	DES				NLPR				NJU2K				STERE				SIP			
Metric	(Δ Mb)	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$
BBS [38]	460(+95)	.015	**.946**	**.941**	.976	.023	.920	.923	.953	.035	.924	.915	.944	.040	.913	.902	.935	.053	.904	.877	.912
DFM [195]	495(+130)	.015	**.946**	**.941**	.974	.022	.919	.926	.956	.034	.922	.917	.943	.041	.909	.902	.933	.049	.909	.885	.919
CDI [187]	520(+155)	.023	.919	.914	.950	.024	.918	.921	.952	.035	.927	.915	.944	.036	.918	.910	.941	.055	.900	.870	.911
DCF [69]	336(**-29**)	.015	.944	.938	.976	**.020**	.927	**.931**	.960	.030	.930	.924	.949	.038	.913	.904	.937	.044	.913	.891	.928
SPNet [210]	593(+228)	.016	.944	.936	.973	.022	.924	.925	.956	.032	.928	.919	.945	.038	.913	.904	.938	.048	.907	.884	.921
MobSal [164]	723(+358)	**.015**	.945	.940	.976	.024	.924	.923	.955	.033	.926	.915	.945	.038	.913	.902	.937	**.042**	.915	.892	.930
Ours	365	**.015**	**.946**	**.941**	**.977**	**.020**	**.932**	**.931**	**.962**	**.029**	**.936**	**.926**	**.951**	**.035**	**.921**	**.911**	**.944**	**.042**	**.916**	**.893**	**.931**

6.4 Experimental Validation

6.4.1 Datasets, Metrics and Training Settings

We follow previous works [38, 69, 168, 210] and train our model on the conventional training set which contains 1,485 samples from the NJU2K-train [72] and 700 samples from the NLPR-train [112]. For testing benchmarks, we observe that the depth quality within each dataset varies, which is mainly due to acquisition methods. Specifically, DES [22] contains 135 images of indoor scenes captured by a Kinect camera. SIP [37] provides a human dataset that contains 929 images captured by a mobile device. Therefore, these two datasets can be considered moderate with less noisy depths.

However, the remaining NLPR-test [112], NJU2K-test [72] and STERE [106] datasets are more challenging. NLPR-test [112] contains 300 natural images which are captured by a Kinect sensor. However, the images are obtained under different illumination conditions. NJU2K-test [72] contains 500 stereo image pairs from different sources such as the Internet and 3D movies. A number of depth maps are estimated through the optical flow method [138]. STERE [106] contains 1,000 stereoscopic images where the depths are estimated with SIFT flow method [90]. Due to the measurement or estimation error, these datasets contain more noisy depths. Therefore, to purely analyze the performance under different conditions, we additionally report the average metric (AvgMetric) for datasets with good quality depths and for datasets with more challenging depths.

To quantify the performance of our methods, we compute conventional saliency metrics such as Mean Absolute Error, F-measure, S-measure, and E-measure. Specifically, **Mean Absolute Error (M)** measures the pixel-level similarity between the estimated saliency map and the ground-truth map. For F-measure, we report the **maximum F-measure** (F) score across the binary maps of different thresholds. **S-measure (S)** and **E-measure (E)** are more specialized metrics for saliency detection. The prior (S) was firstly introduced in [35] to evaluate the similarities between object-aware (S_o) and region-aware (S_r) structures of the saliency map compared to the ground truth. The latter (E_m) is introduced in [36] to evaluate both image-level statistics and local pixel matching information. We refer readers to the original paper for more details.

Our method is based on the Pytorch framework and is learned with a V100 GPU. The encoder is initialized with the pre-trained weights. For the 1-channel depth input, we replace the first convolution of backbone to feet with the depth size. The learning rate is initialized to $1e{-}4$ which is further divided by 10 every 60 epochs. We fix and resize the input RGB-D resolution to 352×352. During training, we adopt random

flipping, rotating, and border clipping for data augmentation. The total training time takes around 5 hours with batch size 10 and epoch 100.

6.4.2 Comparison with SOTA fusion alternatives

We observe that existing works adopt different architectures, i.e., choice of backbones, design of decoder, supervision, training settings, etc. For example, light models [164, 195] always choose MobileNet [127] to extract features. Several works [109,115,203] are based on VGG [133] encoders, while another group of models [39,195] takes ResNet [59] as encoders. Recent works [100, 210] are based on more powerful backbones such as Res2Net [47] and ViT [32]. The choice of backbone will undoubtedly impact the final performance. Furthermore, the design of the decoder varies from one work to another. Several works are based on DenseASPP [182], while others are based on RFB [94]. Under the consideration of a fair comparison, we re-implement six SOTA fusion works under the same architecture. Specifically, we choose the same backbone, same decoder, loss, and same training settings as ours. The only difference between one model to another is in the fusion module. We refer readers to previous sections for more experimental details. Note that several fusion designs [69, 164] were initially applied only to certain layers. To fairly and purely analyze the fusion performance, we implement all the fusion modules at each scale as ours.

Table 6.1 illustrates the quantitative comparison. We also report the model size of each embedded fusion module. $\Delta Size$ stands for the difference in model size compared to ours. It can be seen that our fusion strategy yields significantly better results compared to our counterparts. Compared to the lightest DCF fusion which only applies channel attention during feature fusion, we add additional spatial attention, yielding a slightly heavier model size (+29 Mb) but favorably improving the performance. Elsewise, our model size is significantly lighter compared to other counterparts, validating the effectiveness of our proposed fusion module.

6.4.3 Quantitative Comparison

Table 6.2 illustrates the quantitative comparison. For challenging datasets (NLPR, NJU2K, and STERE), our method performs favorably over the existing methods and sets a new state-of-the-art (SOTA) record, validating the superior robustness of our approach against depth bias. We further illustrate in Figure 6.5 the trade-off between model size and SOTA performances. Compared to the current SOTA TriTrans [100], our model is significantly smaller with only one-third of the model size but with better performance on S-measure. For other datasets with less depth noise (DES and SIP), we also achieve competitive performance with almost halved the model size compared to the current SOTA SPNet [210]. Note that both SPNet and ours adopt

TABLE 6.2: Quantitative comparison with state-of-the-art models. ↑ (↓) denotes that the higher (lower) is better. The best and second best are highlighted in **bold** and underline, respectively. We further report the average metric (AvgMetric) for datasets with more challenging depths and with less less noisy depths.

Dataset	Size↓	Benchmarks with challenging depth																Benchmarks with less noisy depth											
		NLPR				NJU2K				STERE				**AvgMetric**				DES				SIP				**AvgMetric**			
Metric	(Mb)	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$	$M\downarrow$	$F\uparrow$	$S\uparrow$	$E\uparrow$
$CPFP_{19}$ [201]	278	.036	.867	.888	.932	.053	.877	.878	.923	.051	.874	.879	.925	.049	.873	.880	.925	.038	.846	.872	.923	.064	.851	.850	.903	.060	.850	.852	.905
$DMRA_{19}$ [113]	238	.031	.879	.899	.947	.051	.886	.886	.927	.047	.886	.886	.938	.045	.884	.888	.936	.030	.888	.900	.943	.085	.821	.806	.875	.078	.829	.817	.883
$A2dele_{20}$ [115]	**116**	.029	.882	.898	.944	.051	.874	.871	.916	.044	.879	.878	.928	.043	.878	.879	.927	.029	.872	.886	.920	.070	.833	.828	.889	.060	.850	.852	.905
$JLDCF_{20}$ [44]	548	.022	.916	.925	.962	.043	.903	.903	.944	.042	.901	.905	.946	.038	.904	.907	.948	.022	.919	.929	.968	.051	.885	.879	.923	.047	.889	.885	.928
$CMMS_{20}$ [76]	546	.027	.896	.915	.949	.044	.897	.900	.936	.043	.893	.895	.939	.040	.894	.899	.939	.018	.930	.937	.976	.058	.877	.872	.911	.052	.883	.880	.918
$CoNet_{20}$ [70]	162	.031	.887	.908	.945	.046	.893	.895	.937	.040	.905	.908	**.949**	.040	.898	.904	.945	.028	.896	.909	.945	.063	.867	.858	.913	.058	.870	.864	.917
$DANet_{20}$ [203]	128	.028	.916	.915	.953	.045	.910	.899	.935	.043	.892	.901	.937	.041	.901	.902	.939	.023	.928	.924	.968	.054	.892	.875	.918	.050	.896	.881	.924
$DASNet_{20}$ [199]	-	.021	.929	.929	-	.042	.911	.902	-	.037	.915	.910	-	.035	.916	.910	-	.023	.928	.908	-	-	-	-	-	-	-	-	-
$HDFNet_{20}$ [109]	308	.031	.839	.898	.942	.051	.847	.885	.920	.039	.863	.906	.937	.041	.854	.898	.933	.030	.843	.899	.944	.050	.904	.878	.920	.047	.896	.880	.923
$BBSNet_{20}$ [38]	200	.023	.918	.930	.961	.035	.920	.921	.949	.041	.903	.908	.942	.036	.910	.915	.947	.021	.927	.933	.966	-	-	-	-	-	-	-	-
DCF_{21} [69]	435	.021	.891	-	.957	.035	.902	-	.924	.039	.885	-	.927	.034	.890	-	.931	-	-	-	-	.051	.875	-	.920	-	-	-	-
$D3Net_{21}$ [37]	518	.030	.897	.912	.953	.041	.900	.900	.950	.046	.891	.899	.938	.041	.894	.901	.943	.031	.885	.898	.946	.063	.861	.860	.909	.058	.864	.864	.913
$DSA2F_{21}$ [141]	-	.024	.897	.918	.950	.039	.901	.903	.923	.036	.898	.904	.933	.034	.898	.906	.933	.021	.896	.920	.962	-	-	-	-	-	-	-	-
$TriTrans_{21}$ [100]	927	.020	-	.928	.960	.030	-	.920	.925	**.033**	-	.908	.927	.030	-	.914	.931	.014	-	.943	.981	.043	-	.886	.924	.039	-	.893	.931
$CDINet_{21}$ [187]	217	.024	.916	.927	-	.035	.922	.919	-	.041	.903	.906	-	.036	.910	.913	-	-	-	-	-	-	-	-	-	-	-	-	-
$SPNet_{21}$ [210]	702	.021	.925	.927	.959	**.028**	.935	.925	**.954**	.037	.915	.907	.944	.031	.922	.915	**.949**	**.014**	**.950**	**.945**	**.980**	.043	**.916**	**.894**	.930	.039	**.920**	**.900**	**.936**
RFNet (ours)	364	**.020**	**.932**	**.931**	**.962**	.029	**.936**	**.926**	.951	.035	**.921**	**.911**	.944	**.030**	**.927**	**.918**	.948	.015	.946	.941	.977	**.042**	**.916**	.893	**.931**	**.038**	.919	.899	**.936**

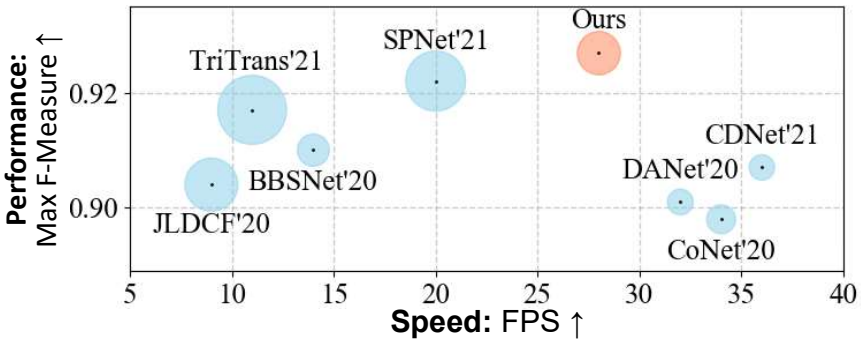


FIGURE 6.5: Average **Performance, Speed, and Model Size** of different methods on challenging datasets (NLPR, NJUK, STERE). The circle size denotes the model size. Note that better models are shown in the upper right corner (i.e., with a larger F-measure and larger FPS). Our method finds the best trade-off of the three measures. Methods with higher speed perform inferior, making our method both efficient and accurate.

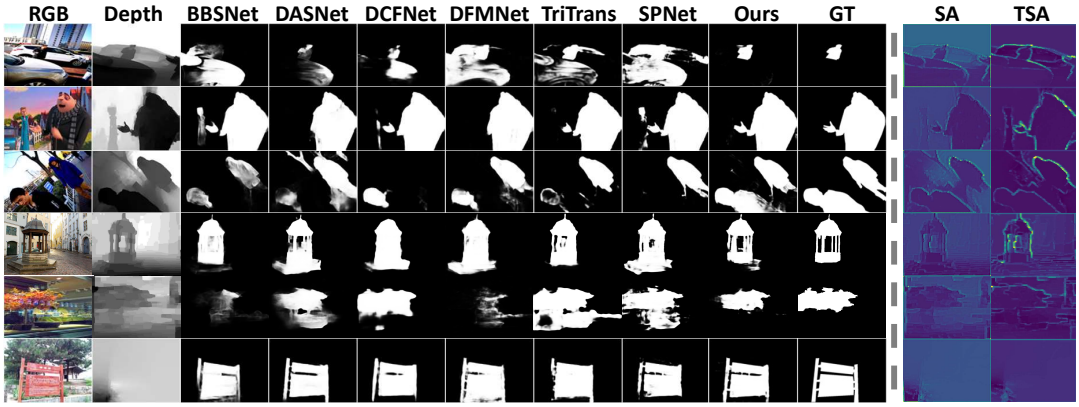


FIGURE 6.6: Qualitative comparison. We also illustrate the depth features enhanced by vanilla SA and by our proposed TSA, respectively. Our work yields more boundary activation compared to the counterpart. Better to zoom in.

Res2Net50 [47] as the backbone. Thus, our performance can be contributed to our proposed fusion solely.

6.4.4 Qualitative Comparison

Figure 6.6 presents the generated saliency maps of different methods on challenging cases. Specifically, the $1^{st} - 3^{rd}$ rows show the cases with a single human in the scene with the depth captured by a mobile camera (1^{st} row) or estimated by algorithms ($2^{nd} - 3^{rd}$ rows). $4^{th} - 6^{th}$ rows show the cases when there are multiple humans in the scene. The associated depths are captured by a mobile camera. $7^{th} - 8^{th}$ illustrates the cases with clustered foreground-background. 9^{th} row shows the case with low-quality depth. It can be seen that our methods consistently reason about saliency masks closer to the ground truth.

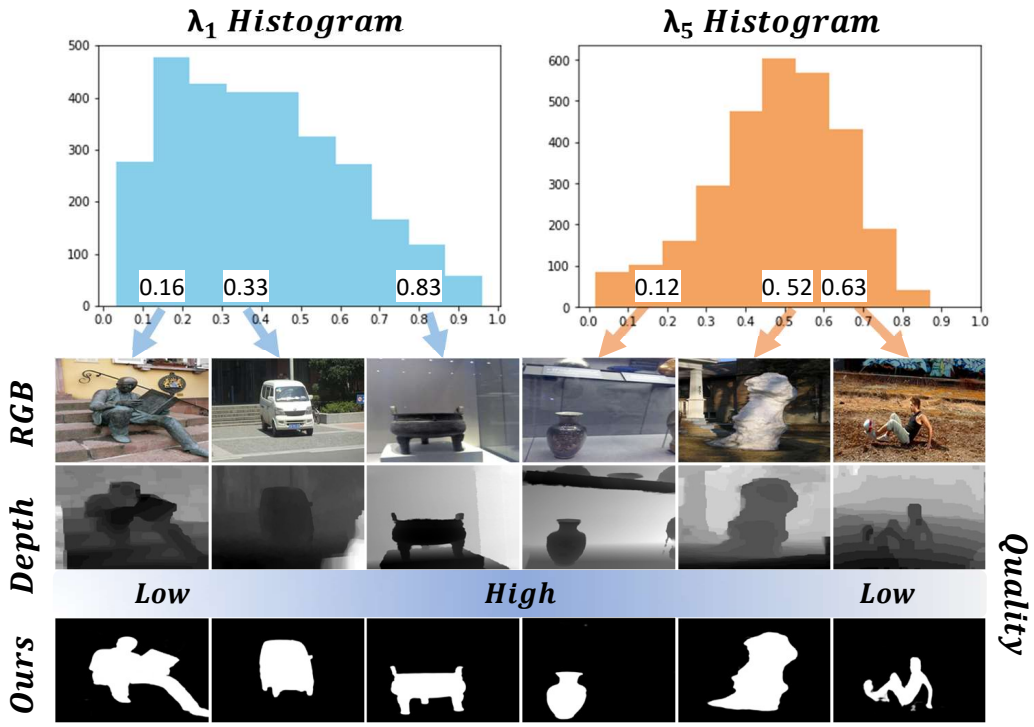


FIGURE 6.7: Trade-off between early and late fusion. Our layer-wise attention can adaptively model the depth contribution during feature fusion. While are with low quality, we assign less weight for early fusion since the noisy geometric cues are difficult to be exploited. Meanwhile, we assign more weight for late fusion to leverage the multi-modal semantic cues for feature fusion.

We further illustrate the comparison between depth feature maps enhanced with our proposed spatial attention (TSA) and with the counterpart (TA). For the cases with multi-objects at different camera distances, i.e., $1^{st}-6^{th}$ rows, we can visualize that our attention can better segment object regions. This can be contributed to our trident branches with different scales. Furthermore, our attention yields more activation on the boundary, facilitating the network to better leverage geometric for saliency detection.

Finally, we illustrate in Figure 6.7 the histogram for our layer-wise attention. We particularly choose λ_1 and λ_5 to facilitate the understanding of the trade-off between early and late fusion. We can observe that while depths are of low quality, our LWA assigns more weights for late fusion (with low λ_1 value and high λ_5 value). While depths are of good quality, our LWA assigns more weights for early fusion (with high λ_1 value and low λ_5 value). This observation is consistent with previous studies [39, 69, 109, 187] with discrepant fusion. We hope our analysis of layer-wise attention can inspire future adaptive fusion works.

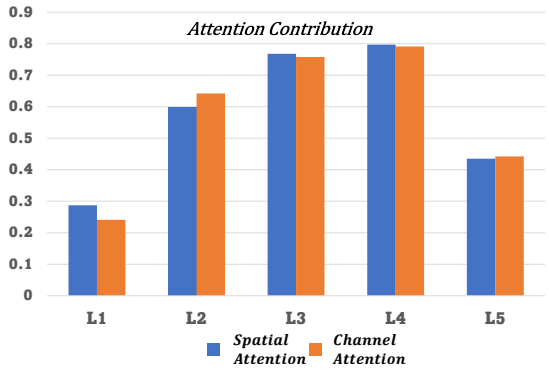


FIGURE 6.8: Attention contribution during feature fusion. $L_1, ...L_5$ stands for the different layers. We realize attention fusion through $\alpha \cdot CA + \beta \cdot TSA$.

6.4.5 Distribution of Spatial and Channel Attention

Since we propose an adaptive weighting strategy to merge our spatial attention (TSA) and channel attention (CA), we illustrate in Figure 6.8 the distribution of weights of each attention at different stages of the network. We can observe that TSA and CA contribute differently with respect to the network depth. At layer 1 (L_1), the network assigns more weight to TSA. This can be explained by the significant spatial resolution of the features. For deeper layers, it can be seen that SA and CA tend to play a similar role at each stage to enhance the feature modeling with equal importance. However, we show that the contributions from different layers to form the final output are different. Specifically, it can be seen that attention from the third layer (L_3) and the fourth layer (L_4) contribute more compared to the first two layers ($L_1 - L_2$) and the last layer (L_5). The different contribution with respect to the network depth is also consistent with previous works [108] and to our layer-wise attention that shallow and deep layers play different roles for feature fusion.

6.4.6 Ablation Study

In this section, we conduct an ablation study to validate the effectiveness of each proposed component. The quantitative result of each combination can be found in Table 6.3. To analyze the effectiveness of our trident spatial attention (TSA), we replace ours with vanilla spatial attention [161] and observe a dropped performance. This is mainly due to the limited receptive field of vanilla attention that assumes a local correlation between different features. In contrast, our TSA can significantly improve performance by leveraging contextualized awareness. The boosted performance on the aforementioned datasets validates the design of our TSA.

We also conduct experiments by replacing our LWA with another depth quality module presented in DFM [195]. While the LWA is replaced, the performance is significantly degraded. The difference between DFM and ours is in the manner to compute the

TABLE 6.3: Ablation study on key components. B stands for the baseline performance where RGB-D features are merged through simple addition without any form of attention.

B	CRM				DFM	LWA	Size	Overall Metric			
	CA	TSA	SA	α, β	([195])	(Ours)	Mb↓	$M \downarrow$	$F \uparrow$	$S \uparrow$	$E \uparrow$
✓							**305**	.039	.915	.904	.935
✓	✓						336	.035	.918	.907	.940
✓	✓	✓					364	.035	.923	.910	.943
✓	✓		✓				363	.035	.920	.908	.941
✓	✓	✓		✓			364	.034	**.924**	.910	.943
✓	✓	✓		✓	✓		364	.035	.921	.908	.941
✓	✓	✓		✓		✓	365	**.033**	**.924**	**.911**	**.944**

similarity matrix. Specifically, DFM assumes a perfect alignment between multi-modalities and realizes a pixel-wise matrix multiplication, while we leverage the non-local attention with flattened vectors to compute the similarity.

6.5 Conclusion

In this chapter, we propose a novel fusion architecture for RGB-D saliency detection. Different from previous works, we improve the robustness against inaccurate and misaligned depth inputs. Specifically, we proposed a layer-wise attention to explicitly leverage the depth quality by learning the best trade-off between early and late fusion. Furthermore, we improved the vanilla spatial attention to a broader context, yielding a simple yet efficient mechanism to address the depth misalignment problem. Extensive comparisons on benchmark datasets validate the effectiveness and robustness of our approach compared to the state-of-the-art alternatives. Our method also sets new records on challenging datasets with smaller model sizes. The method developed in this chapter can potentially be used for other tasks, such as semantic segmentation and object detection, in a similar setting of RGB-D inputs in a robust manner.

Chapter 7

Conclusion and Perspectives

7.1 Conclusion

In this thesis, we are interested in fusing RGB-D information for a more effective and robust scene understanding. The objective is therefore to propose novel and more adapted fusion modules to improve the RGB baseline performance. We are particularly interested in designing fusion methods with different forms of attention, which have recently drawn great research interest and set a new state-of-the-art performance. The main strength of attention modules is that they can leverage the most informative features from the input. These cues can be from the channel and spatial dimension, as well as contextualized correlation and layer-wise fusion architecture. The computed attention can therefore better guide the RGB-D fusion to alleviate local noise and improve the feature representation.

Our first contribution in this thesis is to merge the depth information, i.e., the granularity with the semantic cues, i.e., channel attention. We show that by creating a depth-wise channel attention, the deep neural network can pay better attention to local regions. Therefore, the feature discriminability can be naturally enhanced by locally constrained attention. We observe that the proposed attention module correlates well with the network hierarchy, which learns different scales of information during the encoder stage. Hence, by integrating our proposed depth-wise channel attention into the feature extraction, we can extract features with better awareness of the geometric constraint.

The second contribution is on the spatial attention with depth-awareness. We observe that the concurrent learning methods cannot fully leverage the low-level constraint. Without explicit supervision, the learned attention is not consistent on the same pixel but with different chosen backbones. Differently, we propose to compute statically the offsets and do not require any learning parameters, yielding a more consistent attention fully dependent on the geometry. We show that the static spatial attention performs significantly better than the dynamic counterpart on semantic segmentation. However, one limitation of the static model is that it cannot be extended to all modalities, such

as computing a RGB attention to improve the depth learning. Therefore, we follow the learning strategy to perform a bi-directional supervision pipeline to improve both RGB and depth stream for saliency detection.

The third contribution is to explore the contextualized attention for RGB-D fusion. Conventional methods focus more on fusing multi-modal features, with few methods paying attention to feature alignment. Therefore, while the RGB and depth images are not perfectly aligned due to calibration bias, previous methods built upon pixel-wise correlation fails to perform well under inferior condition. To tackle this issue, we propose a transformer-based RGB-D fusion design that can better leverage global attention and hence become more robust to feature misalignment. We dig into the basic operators of the transformer module: we leverage the self-attention for feature self-enhancement and make full use of the cross-attention for cross-modal calibration and feature fusion. We also propose a local and context-aware positional encoding to constrain the global attention into local regions to boost the segmentation accuracy.

Last but not least, we propose a layer-wise attention for an adaptive and robust RGB-D fusion. Previous methods often adopt pre-defined and fixed fusion architecture such as early, middle, and late fusion, which cannot be adapted to the data context. For example, a depth image with good quality can directly provide rich low-level geometric cues that correlate well with the stemming layers. Therefore, it is more intuitive and straightforward to apply early fusion for such a case. While the quality of depth image is unsatisfactory that geometric cues cannot be easily extracted at an early stage, it becomes necessary to apply an encoder to extract the desired feature and fuse RGB-D features at the semantic level, i.e., late fusion. Inspired by this intuition, we propose a context-aware layer attention that learns the trade-off between early and late fusion. Hence, our model can automatically define which layer/stage is more adequate to fuse multi-modal features, yielding a simple yet robust manner to control the depth contribution with respect to the image quality.

7.2 Perspective

From my point of view, we are nowadays in the transition stage where people have the choice between a CNN model and a transformer model. CNN is well known for its sliding receptive field which can pay attention to local pixels. The transformer, from another perspective, can better model the long-range dependencies while requiring more computational cost compared to a fixed-size local convolutional window. Each design has its intrinsic pros and cons. Hence, it is yet unclear to the vision society which model can consistently lead to better performance. A recent work [108] proposes to combine both CNN and transformer and shows the superior performance of a hybrid

backbone. Therefore, one possible future direction for RGB-D fusion is to leverage both local and global attention to merge multi-modal features.

Another perspective is on the availability of depth images. One inherent shortage of RGB-D fusion is the requirement of multi-modal data from the input side, which is hard to achieve in practice. One possible alternative is to estimate a pseudo-depth to mimick the RGB-D input. Nevertheless, the pseudo-depth can be with sub-optimal quality due to the domain gap. Despite the recent tentative [134] which re-calibrates the pseudo-depth attention according to the trustful regions, these regions are still computed by a pre-trained saliency model and there are no theoretical supports that the salient regions contain good depth estimations. Furthermore, in this thesis we focus more on the spatial correlation between different modalities, paying little attention to temporal consistency which is important for video tasks. How to ensure, enhance, and benefit from the consistency thanks to the depth priors (ground truth or estimated) can be another future research direction.

Finally, the objective of perception is to promote machines with sensing capability and further realize complex tasks. In the industrial context, an algorithm should be able to provide good performance while being embedded in a low-cost system. This criterion inspires us to design a lightweight RGB-D fusion module in the future, i.e., proposing methods to minimize the redundancy between multi-modal features. A recent work [190] has shown that by minimizing the mutual information, the network can learn more complementary and informative cues from RGB-D inputs. However, this work does not focus on the lightweight model, which leaves large room for an efficient fusion module.

Appendix A

Academic Experience

A.0.1 Publication

The works developped durnig this thesis yield in the following accepted publications:

- Robust RGB-D Fusion for Saliency Detection [170], 3DV, 2022
- Modality-Guided Subnetwork for Salient Object Detection [168], 3DV, 2021
- Depth-Adapted CNN for RGB-D cameras [166], ACCV (Oral), 2020

We also have several accepted or under-reviewing submissions developed during this thesis:

- RGB-Event Fusion for Moving Object Detection in Autonomous Driving, ICRA 2023 [214]
- Depth-Adapted CNN for RGB-D Semantic Segmentation [169] (Extension of previous conference paper)
- RGB-D Salient Object Detection via Hierarchical Depth Awareness [165]
- Transformer Fusion for Indoor RGB-D Semantic Segmentation [173]

A.0.2 Reviewer

During the thesis, I am glad and fortunate to contribute to the computer vision society as a reviewer. I help my supervisors and collaborators to review top-level conference and journal papers such as NeurIPS, ECCV, AAAI, IROS, ICRA, CVMJ, and TIP. I also serve as a reviewer for 3DV, BMVC, ACCV, and RA-L.

A.0.3 Teaching assistant and Master Thesis Supervision

During the PhD, I serve as teaching assistant in IUT Le Creusot, University of Bourgogne Franche Comte. I supervise students from different years, from bachelor's to master. I realized more than 100 hours on algorithms, programming, image processing, computer vision, and artificial intelligence.

I have also got the chance to help and advise some master students:

- Shriarulmozhivarman GOBICHETTIPALAYAM (Master Thesis at the University of Bourgogne Franche Comte – 2022): "Robust RGB-D Fusion for Saliency Detection".
- Hugo LEBLOND (Research Intern at the University of Bourgogne Franche Comte – 2022): "RGB-D Nerf for Deformable Objects".

Bibliography

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[2] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[3] Charles-Olivier Artizzu, Haozhou Zhang, Guillaume Allibert, and Cédric Demonceaux. Omniflownet: a perspective neural network adaptation for optical flow estimation in omnidirectional images. In *25th International Conference on Pattern Recognition (ICPR)*, 2021.

[4] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2019.

[5] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational Visual Media (CVM)*, 5(2):117–150, 2019.

[6] Shubhankar Borse, Ying Wang, Yizhe Zhang, and Fatih Porikli. Inverseform: A loss function for structured boundary-aware segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 40(4):834–848, 2017.

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[11] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. Spatial information guided convolution for real-time rgbd semantic segmentation. *IEEE Transactions on Image Processing (TIP)*, 30:2313–2324, 2021.

[12] Qian Chen, Ze Liu, Yi Zhang, Keren Fu, Qijun Zhao, and Hongwei Du. Rgb-d salient object detection via 3d convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[13] Qian Chen, Zhenxi Zhang, Yanye Lu, Keren Fu, and Qijun Zhao. 3-d convolutional neural networks for rgb-d salient object detection and beyond. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2022.

[14] Shuhan Chen and Yun Fu. Progressively guided alternate refinement network for rgb-d salient object detection. In *European Conference on Computer Vision (ECCV)*. Springer, 2020.

[15] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[16] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[18] Yunlu Chen, Thomas Mensink, and Efstratios Gavves. 3D neighborhood convolution: Learning depth-aware features for RGB-D and RGB semantic segmentation. In *International Conference on 3D Vision (3DV)*, 2019.

[19] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[20] Xiaolong Cheng, Xuan Zheng, Jialun Pei, He Tang, Zehua Lyu, and Chuanbo Chen. Depth-induced gap-reducing network for rgb-d salient object detection: An interaction, guidance and refinement approach. *IEEE Transactions on Multimedia (TMM)*, 2022.

[21] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[22] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *Proceedings of International Conference on Internet Multimedia Computing and Service (ICIMCS)*, 2014.

[23] Hang Chu, Wei-Chiu Ma, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Surfconv: Bridging 3d and 2d convolution for RGBD images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[24] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021.

[25] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. 2018.

[26] Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. Review of visual saliency detection with comprehensive information. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 29(10):2941–2959, 2018.

[27] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[28] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017.

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 2009.

[30] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. R3net: Recurrent residual refinement network for saliency detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[31] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[32] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold,

Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations, (ICLR)*, 2021.

[33] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[34] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2015.

[35] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[36] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[37] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems (TNNLS)*, 32(5):2075–2089, 2021.

[38] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[39] Xian Fang, Jinshao Zhu, Xiuli Shao, and Hongpeng Wang. GroupTransNet: Group transformer network for RGB-D salient object detection. *arXiv preprint arXiv:2203.10785*, 2022.

[40] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] C. Fernandez-Labrador, J. M. Facil, A. Perez-Yus, C. Demonceaux, J. Civera, and J. J. Guerrero. Corners for layout: End-to-end layout recovery from 360 images. *IEEE Robotics and Automation Letters*, 5(2):1255–1262, 2020.

[42] Fahimeh Fooladgar and Shohreh Kasaei. Multi-modal attention-based fusion model for semantic segmentation of RGB-Depth images. *arXiv preprint arXiv:1912.11691*, 2019.

[43] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[44] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[45] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, Qijun Zhao, Jianbing Shen, and Ce Zhu. Siamese network for rgb-d salient object detection and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[46] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of DETR with spatially modulated co-attention. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[47] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(2):652–662, 2021.

[48] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[49] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning (ICML)*, 2017.

[50] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[51] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[52] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2014.

[53] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. Recent advances in convolutional neural networks. *Pattern recognition (PR)*, 77:354–377, 2018.

[54] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, pages 1–38, 2022.

[55] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *Proceedings of European conference on computer vision (ECCV)*, 2014.

[56] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.

[57] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In *Proceedings of Asian conference on computer vision (ACCV)*, 2016.

[58] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[60] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*, 2021.

[61] Yang He, Wei-Chen Chiu, Margret Keuper, and Mario Fritz. STD2P: RGBD semantic segmentation using spatio-temporal data-driven pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[62] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[63] Armin Hornung, Kai M Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. Octomap: An efficient probabilistic 3d mapping framework based on octrees. *Autonomous robots*, 34(3):189–206, 2013.

[64] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[65] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. ACNet: Attention based network to exploit complementary features for RGB-D semantic segmentation. In *IEEE International Conference on Image Processing (ICIP)*, 2019.

[66] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition Workshops (CVPRW*, 2020.

[67] Nianchang Huang, Yongjiang Luo, Qiang Zhang, and Jungong Han. Discriminative unimodal feature selection and fusion for rgb-d salient object detection. *Pattern Recognition (PR)*, 122:108359, 2022.

[68] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[69] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, and Li Cheng. Calibrated RGB-D salient object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[70] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate RGB-D salient object detection via collaborative learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[71] Wen-Da Jin, Jun Xu, Qi Han, Yi Zhang, and Ming-Ming Cheng. Cdnet: Complementary depth network for rgb-d salient object detection. *IEEE Transactions on Image Processing (TIP)*, 30:3376–3390, 2021.

[72] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *2014 IEEE International Conference on Image Processing (ICIP)*, 2014.

[73] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021.

[74] Minhyeok Lee, Chaewon Park, Suhwan Cho, and Sangyoun Lee. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In *European Conference on Computer Vision (ECCV)*. Springer, 2022.

[75] Bo Li, Tianlei Zhang, and Tian Xia. Vehicle detection from 3d lidar using fully convolutional network. *arXiv:1608.07916*, 2016.

[76] Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy. RGB-D salient object detection with cross-modality modulation and selection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[77] Gongyang Li, Zhi Liu, Minyu Chen, Zhen Bai, Weisi Lin, and Haibin Ling. Hierarchical alternate interaction network for rgb-d salient object detection. *IEEE Transactions on Image Processing (TIP)*, 30:3528–3542, 2021.

[78] Gongyang Li, Zhi Liu, Linwei Ye, Yang Wang, and Haibin Ling. Cross-modal weighting network for rgb-d salient object detection. In *European Conference on Computer Vision (ECCV)*. Springer, 2020.

[79] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[80] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[81] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[82] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Advances in Neural Information Processing Systems*, pages 820–830, 2018.

[83] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V. Le, Alan Yuille, and Mingxing Tan. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[84] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021.

[85] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin. LSTM-CF: Unifying context modeling and fusion with LSTMs for RGB-D scene labeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[86] Zhengfa Liang, Yulan Guo, Yiliu Feng, Wei Chen, Linbo Qiao, Li Zhou, Jianfeng Zhang, and Hengzhu Liu. Stereo matching using multi-level cost volume and multi-scale feature constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):300–315, 2019.

[87] Ping-Sung Liao, Tse-Sheng Chen, and P. C. Chung. A fast algorithm for multilevel thresholding. *Journal of Information Science and Engineering (JISE)*, 17:713–727, 2001.

[88] Di Lin, Guangyong Chen, Daniel Cohen-Or, Pheng-Ann Heng, and Hui Huang. Cascaded feature network for semantic segmentation of RGB-D images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[89] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.

[90] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(5):978–994, 2011.

[91] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[92] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for rgb-d saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[93] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[94] Songtao Liu, Di Huang, and Yunhong Wang. Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[95] Yongcheng Liu, Bin Fan, Gaofeng Meng, Jiwen Lu, Shiming Xiang, and Chunhong Pan. Densepoint: Learning densely contextual representation for efficient point cloud processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[96] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[97] Yun Liu, Guolei Sun, Yu Qiu, Le Zhang, Ajad Chhatkuli, and Luc Van Gool. Transformer in convolutional neural networks. *arXiv preprint arXiv:2106.03180*, 2021.

[98] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[99] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

[100] Zhengyi Liu, Wang Yuan, Zhengzheng Tu, Yun Xiao, and Bin Tang. TriTransNet: RGB-D salient object detection with a triplet transformer embedding network. *Poceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, 2021.

[101] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[102] Ao Luo, Xin Li, Fan Yang, Zhicheng Jiao, Hong Cheng, and Siwei Lyu. Cascade graph neural networks for rgb-d salient object detection. In *European Conference on Computer Vision (ECCV)*. Springer, 2020.

[103] Daniel Maturana and Sebastian Scherer. 3d convolutional neural networks for landing zone detection from lidar. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 3471–3478. IEEE, 2015.

[104] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, 2016.

[105] Will Nash, Tom Drummond, and Nick Birbilis. A review of deep learning in the study of materials degradation. *npj Materials Degradation*, 2(1):1–12, 2018.

[106] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[107] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics (TSMC)*, 9(1):62–66, 1979.

[108] Xuran Pan, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, and Gao Huang. On the integration of self-attention and convolution. *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[109] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[110] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017.

[111] Danda Pani Paudel, Cédric Demonceaux, Adlane Habed, and Pascal Vasseur. 2D-3D synchronous asynchronous camera fusion for visual odometry. *Autonomous Robots*, 43(1):21–35, 2019.

[112] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. RGBD salient object detection: a benchmark and algorithms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[113] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[114] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[115] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[116] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[117] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[118] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5099–5108, 2017.

[119] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3D graph neural networks for RGBD semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[120] Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3D graph neural networks for RGBD semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[121] Wang Qilong, Wu Banggu, Zhu Pengfei, Li Peihua, Zuo Wangmeng, and Hu Qinghua. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *The IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[122] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[123] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning, (ICML)*, 2021.

[124] Aman Raj, Daniel Maturana, and Sebastian Scherer. Multi-scale convolutional architecture for semantic segmentation. *Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RITR-15-21*, 2015.

[125] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[126] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, 2015.

[127] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[128] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient RGB-D semantic segmentation for indoor scene analysis. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[129] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[130] Xiaoke Shen. A survey of object classification and detection based on 2d/3d data. *arXiv:1905.12683*, 2019.

[131] Hao Shi, Yifan Zhou, Kailun Yang, Yaozu Ye, Xiaoting Yin, Zhe Yin, Shi Meng, and Kaiwei Wang. Panoflow: Learning optical flow for panoramic images. *arXiv preprint arXiv:2202.13388*, 2022.

[132] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

[133] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[134] Hwanjun Song, Eunyoung Kim, Varun Jampan, Deqing Sun, Jae-Gil Lee, and Ming-Hsuan Yang. Exploiting scene depth for object detection with multimodal transformers. In *32nd British Machine Vision Conference (BMVC)*, 2021.

[135] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[136] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[137] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[138] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[139] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[140] Guolei Sun, Yun Liu, Hao Tang, Ajad Chhatkuli, Le Zhang, and Luc Van Gool. Mining relations among cross-frame affinities for video semantic segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[141] Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi Li. Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[142] Hao Tang, Song Bai, and Nicu Sebe. Dual attention gans for semantic image synthesis. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 2020.

[143] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[144] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *arXiv preprint arXiv:2009.06732*, 2020.

[145] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *2017 international conference on 3D vision (3DV)*, pages 537–547. IEEE, 2017.

[146] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[147] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[148] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[149] Fengyun Wang, Jinshan Pan, Shoukun Xu, and Jinhui Tang. Learning discriminative cross-modality features for rgb-d saliency detection. *IEEE Transactions on Image Processing (TIP)*, 31:1285–1297, 2022.

[150] Jinghua Wang, Zhenhua Wang, Dacheng Tao, Simon See, and Gang Wang. Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[151] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Kaihao Zhang, Pan Ji, and Hongdong Li. Displacement-invariant matching cost learning for accurate optical flow estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[152] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[153] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–1, 2021.

[154] Weiyue Wang and Ulrich Neumann. Depth-aware CNN for RGB-D segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[155] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(1):20–33, 2017.

[156] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[157] Xuehao Wang, Shuai Li, Chenglizhao Chen, Yuming Fang, Aimin Hao, and Hong Qin. Data-level recombination and lightweight fusion scheme for rgb-d salient object detection. *IEEE Transactions on Image Processing (TIP)*, 30:458–471, 2020.

[158] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[159] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: Fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[160] Hongfa Wen, Chenggang Yan, Xiaofei Zhou, Runmin Cong, Yaoqi Sun, Bolun Zheng, Jiyong Zhang, Yongjun Bao, and Guiguang Ding. Dynamic selective network for rgb-d salient object detection. *IEEE Transactions on Image Processing (TIP)*, 30:9179–9192, 2021.

[161] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[162] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[163] Wei Wu, Tao Chu, and Qiong Liu. Complementarity-aware cross-modal feature fusion network for rgb-t semantic segmentation. *Pattern Recognition*, 131:108881, 2022.

[164] Yu-Huan Wu, Yun Liu, Jun Xu, Jia-Wang Bian, Yu-Chao Gu, and Ming-Ming Cheng. Mobilesal: Extremely efficient rgb-d salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[165] Zongwei Wu, Guillaume Allibert, Fabrice Meriaudeau, Chao Ma, and Cédric Demonceaux. Hidanet: Rgb-d salient object detection via hierarchical depth awareness. In *arXiv preprint arXiv:2301.07405*, 2023.

[166] Zongwei Wu, Guillaume Allibert, Christophe Stolz, and Cédric Demonceaux. Depth-adapted CNN for RGB-D cameras. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020.

[167] Zongwei Wu, Guillaume Allibert, Christophe Stolz, Chao Ma, and Cédric Demonceaux. Modality-guided subnetwork for salient object detection. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021.

[168] Zongwei Wu, Guillaume Allibert, Christophe Stolz, Chao Ma, and Cédric Demonceaux. Modality-guided subnetwork for salient object detection. In *International Conference on 3D Vision (3DV)*, 2021.

[169] Zongwei Wu, Guillaume Allibert, Christophe Stolz, Chao Ma, and Cédric Demonceaux. Depth-adapted CNNs for RGB-D semantic segmentation. *arXiv preprint arXiv:2206.03939*, 2022.

[170] Zongwei Wu, Shriarulmozhivarman Gobichettipalayam, Brahim Tamadazte, Guillaume Allibert, Danda Pani Paudel, and Cédric Demonceaux. Robust rgb-d fusion for saliency detection. In *International Conference on 3D Vision (3DV)*, 2022.

[171] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015.

[172] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[173] Zongwei WU, Zhuyun ZHOU, Guillaume Allibert, Christophe Stolz, Cédric Demonceaux, and Chao Ma. Transformer fusion for indoor rgb-d semantic segmentation. *Available at SSRN 4251286*, 2022.

[174] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[175] Yajie Xing, Jingbo Wang, Xiaokang Chen, and Gang Zeng. 2.5 D convolution for RGB-D semantic segmentation. In *IEEE International Conference on Image Processing (ICIP)*, 2019.

[176] Yajie Xing, Jingbo Wang, and Gang Zeng. Malleable 2.5 D convolution: Learning receptive fields along the depth-axis for RGB-D scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[177] Zhitong Xiong, Yuan Yuan, Nianhui Guo, and Qi Wang. Variational context-deformable convnets for indoor scene parsing. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[178] Hongbin Xu, Zhipeng Zhou, Yali Wang, Wenxiong Kang, Baigui Sun, Hao Li, and Yu Qiao. Digging into uncertainty in self-supervised multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[179] Philippe Xu, Franck Davoine, Jean-Baptiste Bordes, Huijing Zhao, and Thierry Denœux. Multimodal information fusion for urban scene understanding. *Machine Vision and Applications*, 2016.

[180] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[181] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[182] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[183] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems (NIPS)*, 2019.

[184] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122*, 2015.

[185] Yuchun Yue, Wujie Zhou, Jingsheng Lei, and Lu Yu. Rtlnet: Recursive triple-path learning network for scene parsing of rgb-d images. *IEEE Signal Processing Letters*, 29:429–433, 2021.

[186] Yingjie Zhai, Deng-Ping Fan, Jufeng Yang, Ali Borji, Ling Shao, Junwei Han, and Liang Wang. Bifurcated backbone strategy for rgb-d salient object detection. *IEEE Transactions on Image Processing (TIP)*, 30:8727–8742, 2021.

[187] Chen Zhang, Runmin Cong, Qinwei Lin, Lin Ma, Feng Li, Yao Zhao, and Sam Kwong. Cross-modality discrepant interaction network for RGB-D salient object detection. In *Poceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, 2021.

[188] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Saleh, Sadegh Aliakbarian, and Nick Barnes. Uncertainty inspired rgb-d saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[189] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[190] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. RGB-D saliency detection via cascaded mutual information minimization. In *Poceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[191] Miao Zhang, Sun Xiao Fei, Jie Liu, Shuang Xu, Yongri Piao, and Huchuan Lu. Asymmetric two-stream architecture for accurate rgb-d saliency detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[192] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for RGB-D saliency detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[193] Miao Zhang, Shunyu Yao, Beiqi Hu, Yongri Piao, and Wei Ji. C2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection. *IEEE Transactions on Multimedia (TMM)*, 2022.

[194] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[195] Wenbo Zhang, Ge-Peng Ji, Zhuo Wang, Keren Fu, and Qijun Zhao. Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In *Poceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, 2021.

[196] Xiaoya Zhang, Shumin Zhang, Zhen Cui, Zechao Li, Jin Xie, and Jian Yang. Tube-embedded transformer for pixel prediction. *IEEE Transactions on Multimedia (TMM)*, 2022.

[197] Zhao Zhang, Zheng Lin, Jun Xu, Wen-Da Jin, Shao-Ping Lu, and Deng-Ping Fan. Bilateral attention network for rgb-d salient object detection. *IEEE Transactions on Image Processing (TIP)*, 30:1949–1961, 2021.

[198] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[199] Jiawei Zhao, Yifan Zhao, Jia Li, and Xiaowu Chen. Is depth really necessary for salient object detection? In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 2020.

[200] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for RGBD salient object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[201] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for rgbd salient object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[202] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. EGNet: Edge Guidance Network for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[203] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[204] Yifan Zhao, Jiawei Zhao, Jia Li, and Xiaowu Chen. Rgb-d salient object detection with ubiquitous target awareness. *IEEE Transactions on Image Processing (TIP)*, 30:7717–7731, 2021.

[205] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 30(11):3212–3232, 2019.

[206] Hao Zhou, Lu Qi, Hai Huang, Xu Yang, Zhaoliang Wan, and Xianglong Wen. CANet: Co-attention network for RGB-D semantic segmentation. *Pattern Recognition (PR)*, 124:108468, 2022.

[207] Hao Zhou, Lu Qi, Zhaoliang Wan, Hai Huang, and Xu Yang. RGB-D co-attention network for semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020.

[208] Jiayuan Zhou, Lijun Wang, Huchuan Lu, Kaining Huang, Xinchu Shi, and Bocong Liu. Mvsalnet: Multi-view augmentation for rgb-d salient object detection. In *European Conference on Computer Vision (ECCV)*. Springer, 2022.

[209] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Rgb-d salient object detection: A survey. *Computational Visual Media (CVM)*, pages 1–33, 2021.

[210] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving RGB-D saliency detection. In *Poceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[211] Wujie Zhou, Sijia Lv, Jingsheng Lei, Ting Luo, and Lu Yu. Rfnet: Reverse fusion network with attention mechanism for rgb-d indoor scene understanding. *IEEE Transactions on Emerging Topics in Computational Intelligence (Trans. Emerg. Topics Comput.)*, 2022.

[212] Wujie Zhou, Enquan Yang, Jingsheng Lei, Jian Wan, and Lu Yu. Pgdenet: Progressive guided fusion and depth enhancement network for rgb-d indoor scene parsing. *IEEE Transactions on Multimedia (TMM)*, 2022.

[213] Wujie Zhou, Enquan Yang, Jingsheng Lei, and Lu Yu. Frnet: Feature reconstruction network for rgb-d indoor scene parsing. *IEEE Journal of Selected Topics in Signal Processing (J. Sel. Top. Signal Process.)*, 2022.

[214] Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, Fan Yang, Cédric Demonceaux, and Dominique Ginhac. Rgb-event fusion for moving object detection in autonomous driving. *arXiv preprint arXiv:2209.08323*, 2022.

[215] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. PDNet: Prior-model guided depth-enhanced network for salient object detection. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2019.

[216] Chunbiao Zhu and Ge Li. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[217] Jinchao Zhu, Xiaoyu Zhang, Xian Fang, Muhammad Rameez Ur Rahman, Feng Dong, Yuehua Li, Siyu Yan, and Panlong Tan. Boosting rgb-d salient object detection with adaptively cooperative dynamic fusion network. *Knowledge-Based Systems*, page 109205, 2022.

[218] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021.