Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

Master's Thesis

Multi-View Temporal Fusion in Semantic Segmentation

Manoj Kolpe Lingappa

Submitted to Hochschule Bonn-Rhein-Sieg,
Department of Computer Science
in partial fullfilment of the requirements for the degree
of Master of Science in Autonomous Systems

Supervised by

Prof. Dr. Nico Hochgeschwender
Prof. Dr. Sebastian Houben
M.Sc. Deebul Sivarajan Nair

Month 20XX

I, the undersigned below, declare that this work has not previously been submitted to this or any other university and that it is, unless otherwise stated, entirely my own work.

Date	Manoj Kolpe Lingappa

Abstract

Your abstract

Acknowledgements

Thanks to ....

Contents

List of Figures

List of Tables

1

Introduction

1.1 Motivation

Any task to make a prediction by combining data from different sources uses data fusion. Data fusion combines the information from multiple sources to achieve improved performance and inferences. According to Hall and Llinas [1] data fusion can be defined as "data fusion techniques combine data from multiple sensors and related information from associated databases to achieve improved accuracy and more specific inferences than could be achieved by the use of a single sensor alone." The living organisms fuse information from various sources and past data to make an informed decision [2]. Data fusion aims to reduce the prediction error probability and improve the model's reliability. Data from multiple sources can be fused at different levels, such as raw data, features, or decision levels. The data sources can be from various fields or different data types. Data fusion is described in different contexts and in other application areas. The most common areas include decision fusion and multisensor data fusion. [3]

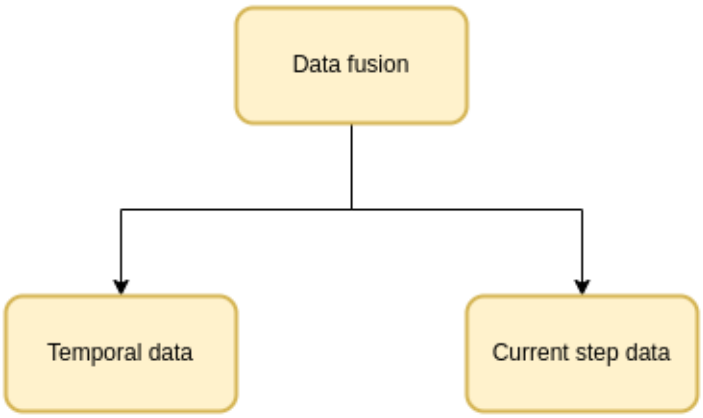

Figure 1.1: Data fusion categories based on timestamp

Data fusion is divided into temporal and current data based on the timestamp factor. Temporal data are the data collected from the former steps. In the current data fusion approach, the data are extracted

from the current step and are fused for improved prediction.Information fusion is applied in different fields such as time series prediction [4], video-based depth estimation [5], and segmentation [6].

Understanding surrounding regions and decision making of a human is based on the signals obtained from different sensors. However, with the knowledge of the past helps to better recognize the nearby activity or to make a educated choice. Thereby fusing the information from different sources and the former data adds to achieve a improved outcome. Temporal fusion is a process of fusing the information to the current step to make the prediction better at each timestamp. Common temporal data types include weather data, frames in a video sequence, and different sensor data. Semantic segmentation take advantage of the temporal fusion to make a better decision.

1.1.1 Temporal fusion

In a general setting the previous data is not utilized to make a current prediction, resulting in information loss. The rich features from the past can be utilized in the current step, thereby making a robust and efficient model prediction. Temporal fusion is one of the dimensions of data fusion, and different data sources collected over a period of time are fused for improvement in the prediction [7]. Classical multi-data fusion finds application in automatic target tracking, autonomous vehicle detection, surveillance systems, robotics, wearable devices, and manufacturing monitoring. In all of the mentioned areas, the data is collected over a period of time and contains important features. These temporal features are combined together to make an efficient prediction.

The temporal fusion can model the behavioral aspect of the collected data rather than just the current timestep data. The temporal fusion extracts the relationship between contextual and temporal proximity [8]. The temporal arrangements of events are captured, thereby incorporating the cause and effect phenomenon [8]. Temporal fusion can be commonly observed in human activity detection [9], context-aware mobile phones [19], and online batch process monitoring [10].

A 3D object detection approach on two popular datasets, KITTI[4] and nuScenes[5] take single-time step LIDAR data for prediction, resulting in the loss of valuable forecast and features computed during the previous step [3]. Using the rich features present in the successive frames to accommodate the past data at a time is extensively studied in neural network-based action recognition [11] [12] [13] [14] and video object detection [15] [16] [17] methods. The fusion of features from the previous step to the current step to improve the 3D object detection is studied in the Temp-Frustum Net architecture [3]. A Temporal Fusion Module (TFM) is proposed to combine the object-specific features. The temporal fusion method improved the average result by 6% [18]. Depth map using dynamic MRFs fuse Time-of-Flight (TOF) and passive stereo to get an enhanced depth map. The depth map estimation is extended to the temporal domain resulting in accurate depth maps [19]. Multi-camera video surveillance fuses the spatiotemporal frames from different sources to reliably find the motion trajectories [20]. A moving object is detected and segmented with the unmanned aerial vehicle (UAV) data by stacking the consecutive frames containing objects of interest resulting in constant object position and moving background, thereby improving the segmentation efficiency [21].

1.1.2 Semantic segmentation

Segmentation of images is an essential task of the visual understanding systems. It involves dividing the image into multiple segments. Image segmentation can be framed as classifying the individual pixels into a particular class or semantic labels. Segmentation can be classified as semantic segmentation, instance segmentation, and panoptic segmentation. Segmentation of images finds a broad range of application areas [22], such as medical for boundary extraction and tissue volume estimation, autonomous systems for detecting a boundary for path planning, and surveillance to track objects. Semantic segmentation is not only about the data but the problem segmentation addresses. For example, in a pedestrian detection system, pixels belonging to a person are categorized into a single class; however, for action recognition, the different parts of the body are classified into other classes. Instance segmentation [23] solves the problem of counting unique objects present in an image and is a common task in image retrieval tasks. Many traditional techniques have been developed to solve the segmentation problem [24]. For specialized tasks, different algorithms are developed [25]. Work by Shervin surveyed the various state-of-the-art segmentation algorithms [26]. However, many algorithms are developed, but few works are proposed for multi-view semantic segmentation. From the previous work, it is evident that temporal fusion improves the model's performance. This work aims to study the impact of temporal fusion of information in the latent space and cross-transfer the technology to the semantic segmentation task.

1.2 Challenges and Difficulties

Semantic segmentation benefited from the advancement of deep learning methods. Building a temporal fusion for semantic segmentation model is a challenging task due to the presence of high number of variables involved and different choices of fusion architectures. Common challenges involved are the

- Datasets
- Fusion architecture
- Computation cost
- Application areas

1.2.1 Dataset

In many application areas the deep learning model can be trained from scratch given that we have large datasets. However, for new domain there are not enough datasets available to train the model, in such cases transfer learning can be applied. In the transfer learning approach a model is trained on some data and the part of trained model weights are used for building a new application areas architecture. Many deep learning based models are trained on the ImageNet datasets and take the pretrained encoder weights which capture the features needed to do the segmentation thereby reducing the dependency on the requirement of large datasets. Image augmentation is the another approach to increase the number of data points. Data augmentation helps to create more data by applying transformation to the existing

small datasets so that variety of the input data is generated from the existing small datasets. Some of the common transformation on the input images are translation, reflection, rotation, warping, scaling, color space shifting, projecting onto the principle component. It helps to faster convergence and reduce the over-fitting probability, and improving the generalization capability of the model. For some task data augmentation showed improvement in the performance of the model. For temporal fusion there is a need of 2D datasets along with the pose of the camera. Pose information cane be fused at the latent space to improve the prediction efficiency of the model. There is a need for different kinds of datasets such as the still images, navigation datasets, Unmanned aerial vehicle (UAV) datasets to validate the model in different environment, helps to evaluate the model performance.

1.2.2 Fusion architecture

With the advancement of the deep learning more and more segmentation models are developed with improved efficiency and with variety of fusion architecture. Fusing of features are commonly used in the segmentation task. Adaption of fusion features in the increased depth deep learning model showed significant improvement in the prediction. U-net [27] model effectively use the already learned features by fusing the information from the encoder to the decoder. To tackle the decrease of initial image resolution at the output a RefineNet [28] network was proposed. Deeper layers captures the high-level semantic features is refined by fusing the fine-grained features from the earlier convolutions. Dense connection is employed in the many of the recent neural network architecture [29], [30], [31]. Choosing the appropriate fusion architecture depends on the factor of problem at hand available resources to solve the problem.

1.2.3 Computation cost

Many state of the art segmentation network requires high computation cost during training as well during the inference time. So the recent research is focused on decreasing the computation cost and also keeping the accuracy of the model high. To deploy the model in the low computational mobile devices simpler models needs to be developed that fits in the computation cost of the device. This can be done by compressing the model or using the knowledge distillation techniques to build the low computational model [26].

1.2.4 Real time inference for various application areas

Most of the recent top performing semantic segmentation models are based on the fully convolutional network [32]. For real time application or at the frame rate of the camera needs to have reasonable accuracy and prediction speed. Real time prediction is extremely critical in the autonomous driving and medical fields. However, most the fully convolutional network is not upto the mark with respect to the maximum requirements defined by application areas. Models with the dilated convolution improved the performance of the model however the benchmark can be still improved. ICNet takes multiple input sizes to capture objects of varying sizes to tackle the real time deployment [33].

1.3 Use cases

Semantic segmentation finds application in many areas of the computer vision. Some of them are listed below,

1.3.1 Autonomous driving and Robotics

Important components of the autonomous driving systems are the object recognition, object localization and segmentation. Semantic segmentation classify each pixels of the image into a particular class thereby identifying different classes such as street, traffic sign, trees, cars, sky, pedestrians or sidewalks. It is critical to classify each pixel with high accuracy due to the safety concerns. The rich information captured in last step can be used in the current step calculation to make a better prediction at the current computational step. With the development of the robotics system to perform a complex task the interaction with the environment also increased. So, there is a need to develop a robust system to understand the knowledge about the workspace.

1.3.2 Weed mapping using Unmanned Aerial Vehicle (UAV)

Mapping of the fields are essential for weed control and spraying applications. The presence of the weed can be mapped by unmanned aerial vehicle remote sensing technology. The targeted spraying onto the weed area helps to curb the weed growth by inspecting the weed map obtained from the UAV. The entire process involves real-time image processing hardware that integrates the map visualization, flight control, image collection [34]. To build a weed map semantic segmentation can be employed with reasonable performance and real time capability.

1.3.3 Real-Time Hand Gesture Recognition

Hand Gesture Recognition (HGR) is an essential component in human-computer interactions. With the advancement of vision-based HGR systems, HGR is widely used in the automotive sector, consumer electronics, home automation, etc. Important feature of the HGR is the real time performance. HGR should perform without any lag to control the location of the cursor. HGR is based on the semantic segmentation method to locate the position of the hand, therefore an efficient real time segmentation network needs to be developed [35].

1.4 Problem Statement and Contribution

Research question answered and contribution in the thesis work is listed below

1.4.1 Research question

RQ1 What are the works on state-of-the-art temporal fusion?

RQ2 How are the results from RQ1 compared with each other to perform temporal fusion?

RQ2.1 What are the results in comparison with different error metrics?

RQ3 How to cross-transfer the depth estimation temporal fusion technique to semantic segmentation?

RQ3.1 How do different loss criteria impact semantic segmentation perfor- mance?

RQ3.2 What is the semantic segmentation performance for different Gaussian kernels?

1.4.2 Contribution

- Literature review on the temporal fusion in the context of depth estimation and semantic segmentation
- Analysis of the state-of-the-art temporal fusion architectures
- Create a baseline of temporal fusion with sequence images
- Compare performances of state-of-the-art temporal fusion techniques with different error metrics
- Cross transfer the temporal fusion architecture to the segmentation task
- Performance of multi-view temporal fusion with different Gaussian kernels

1.5 Report outline

Theoretical background of deep learning, semantic segmentation, temporal fusion and their limitations is discussed in the Chapter 2. Datasets, preprocessing steps, experimental designs, training procedures and hardware configuration used to training and inferences are listed down in the Chapter 3. Evaluation of the temporal fusion architecture with different experimental settings, metrics and research questions are discussed in the Chapter 4. Deployment of the model in the android is described in the android deployment Chapter 5. Finally contribution of the thesis work, lesson learned and future work is explained in the conclusion chapter 6.

2

State of the Art

Introduction to the modern deep learning and their impact onto the various vision tasks are described in the Deep Learning section. Information fusion in the temporal domain to fuse information is explained in Temporal Fusion. State of the art segmentation of the input images, in particular semantic segmentation task is illustrated in the Semantic Segmentation section. State of the art segmentation in the classical era and in modern deep learning play crucial role in the temporally fused semantic segmentation. However, there is very little work of fusing the camera pose onto the segmentation task in temporal fashion. More details are discussed in the Semantic Segmentation. Finally chapter 2 is ended with the discussion on the limitations of the previous work with respect to the temporal fusion.

2.1 Deep Learning

Deep learning is a sub field of machine learning that aims to learn the features present in the data by utilizing the hierarchical architectures. The area deep learning falls in the artificial intelligence is depicted in the picture 2.1

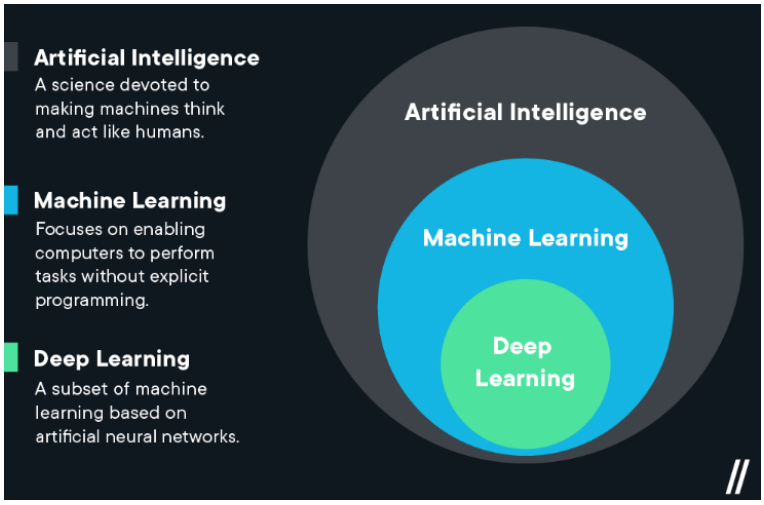

Figure 2.1: Deep learning in the artificial intelligence domain. Courtesy of [1]

2.2 Temporal Fusion

2.3 Segmentation

2.4 Semantic Segmentation

2.4.1 Classical Semantic Segmentation

2.4.2 Deep Learning based Semantic Segmentation

2.5 Temporal Fusion in Semantic Segmentation

Use as many sections as you need in your related work to group content into logical groups
Don't forget to correctly cite your sources [**?**].

2.6 Limitations of previous work

3

Methodology

Semantic segmentation can be evaluated using the
How you are planning to test/compare/evaluate your research. Criteria used.

3.1 Dataset

3.1.1 ScanNet

3.1.2 Virtual KITTI 2

3.1.3 VIODE

3.2 Data Collection and Preprocessing

3.3 Experimental Design

3.3.1 U-Net Vanilla model

3.3.2 U-Net with Temporal Fusion

3.3.3 W-Net Vanilla model

3.3.4 W-Net with Temporal Fusion

3.4 Training and Evaluation Pipeline

3.5 Training Procedure

3.6 Hardware Configuration

4

Evaluation and Experimental Result

Implementation and measurements.

4.1 Evaluation Metric

4.1.1 Pixel Accuracy

4.1.2 Precision

4.1.3 Recall

4.1.4 ROC and AUC

4.1.5 IOU

4.2 RQ1: What are the works on state-of-the-art temporal fusion?

4.2.1 Experiment1.1: U-Net and W-Net model with single sequence data

4.2.2 Experiment1.2: U-Net and W-Net model with two sequence data

4.2.3 Experiment1.3: U-Net and W-Net model with three sequence data

4.2.4 Experiment1.4: U-Net and W-Net model with four sequence data

4.2.5 Experiment1.5: U-Net and W-Net model with all sequence data

4.3 RQ2: How are the results from RQ1 compared with each other to perform temporal fusion?

4.3.1 Experiment1.1: U-Net and W-Net model with single sequence data

4.3.2 Experiment1.2: U-Net and W-Net model with two sequence data

4.3.3 Experiment1.3: U-Net and W-Net model with three sequence data

4.3.4 Experiment1.4: U-Net and W-Net model with four sequence data

4.3.5 Experiment1.5: U-Net and W-Net model with all sequence data

4.4 RQ3: How to cross-transfer the temporal fusion technique to semantic segmentation?

4.4.1 Experiment1.1: U-Net vanilla model

4.4.2 Experiment1.2: U-Net temporally fused gp model

5

Android Deployment

5.1 Framework

Describe results and analyse them

5.2 Pipeline

5.3 Deployment and Results

6

Conclusions

6.1 Contributions

6.2 Lessons learned

6.3 Future work

A

Design Details

Your first appendix

B

Parameters

Your second chapter appendix

References

[1] Michael Middleton. Deep Learning vs. Machine Learning — What's the Difference?, year = 2021, url = https://flatironschool.com/blog/deep-learning-vs-machine-learning/, urldate = 2022-07-28.

[2] Danilo P Mandic, Dragan Obradovic, Anthony Kuh, Tülay Adali, Udo Trutschell, Martin Golz, Philippe De Wilde, Javier Barria, Anthony Constantinides, and Jonathon Chambers. Data fusion for modern engineering applications: An overview. In *International Conference on Artificial Neural Networks*, pages 715–721. Springer, 2005.

[3] Federico Castanedo. A review of data fusion techniques. *The scientific world journal*, 2013, 2013.

[4] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

[5] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15324–15333, 2021.

[6] Minghan Li, Shuai Li, Lida Li, and Lei Zhang. Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11215–11224, 2021.

[7] Ko Ming Hsiao, Geoff West, Svetha Venkatesh, and Mohan Kumar. Temporal data fusion in multi-sensor systems using dynamic time warping. In *SENSORFUSION 2005: Workshop on Information Fusion and Dissemination in Wireless Sensor Networks*, pages 1–9. IEEE, 2005.

[8] Ko Ming Hsiao, Geoff West, Svetha Venkatesh, and Mohan Kumar. Temporal data fusion in multi-sensor systems using dynamic time warping. In *SENSORFUSION 2005: Workshop on Information Fusion and Dissemination in Wireless Sensor Networks*, pages 1–9. IEEE, 2005.

[9] Andreas Krause, Daniel P Siewiorek, Asim Smailagic, and Jonny Farringdon. Unsupervised, dynamic identification of physiological and activity context in wearable computing. In *ISWC*, volume 3, page 88, 2003.

[10] Jong-Min Lee, ChangKyoo Yoo, and In-Beum Lee. On-line batch process monitoring using a consecutively updated multiway principal component analysis model. *Computers & Chemical Engineering*, 27(12):1903–1912, 2003.

[11] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465*, 2016.

[12] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2896–2907, 2017.

[13] Guanghan Ning, Zhi Zhang, Chen Huang, Xiaobo Ren, Haohong Wang, Canhui Cai, and Zhihai He. Spatially supervised recurrent convolutional neural networks for visual object tracking. In *2017 IEEE international symposium on circuits and systems (ISCAS)*, pages 1–4. IEEE, 2017.

[14] Zhichao Lu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Retinatrack: Online single stage joint detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14668–14678, 2020.

[15] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019.

[16] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[17] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[18] Emeç Erçelik, Ekim Yurtsever, and Alois Knoll. Temp-frustum net: 3d object detection with temporal fusion. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1095–1101. IEEE, 2021.

[19] Jiejie Zhu, Liang Wang, Jizhou Gao, and Ruigang Yang. Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):899–909, 2009.

[20] Gang Wu, Yi Wu, Long Jiao, Yuan-Fang Wang, and Edward Y Chang. Multi-camera spatio-temporal fusion and biased sequence-data learning for security surveillance. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 528–538, 2003.

[21] Michael Teutsch and Wolfgang Krüger. Spatio-temporal fusion of object segmentation approaches for moving distant targets. In *2012 15th International Conference on Information Fusion*, pages 1988–1995. IEEE, 2012.

[22] David Forsyth and Jean Ponce. *Computer vision: A modern approach.* Prentice hall, 2011.

[23] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3150–3158, 2016.

[24] King-Sun Fu and JK Mui. A survey on image segmentation. *Pattern recognition*, 13(1):3–16, 1981.

[25] W Ladys law Skarbek and Andreas Koschan. Colour image segmentation a survey. *IEEE Transactions on circuits and systems for Video Technol-ogy*, 14(7):1–80, 1994.

[26] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[27] 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*, pages 234–241. Springer, 2015.

[28] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[29] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 11–19, 2017.

[30] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, and Kurt Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

[31] 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*, pages 3684–3692, 2018.

[32] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[33] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 405–420, 2018.

[34] Jizhong Deng, Zhaoji Zhong, Huasheng Huang, Yubin Lan, Yuxing Han, and Yali Zhang. Lightweight semantic segmentation network for real-time weed mapping using unmanned aerial vehicles. *Applied Sciences*, 10(20):7132, 2020.

[35] Chen-Chiung Hsieh, Dung-Hua Liou, and David Lee. A real time hand gesture recognition system using motion history image. In *2010 2nd international conference on signal processing systems*, volume 2, pages V2–394. IEEE, 2010.

[36] Ko Ming Hsiao, Geoff West, Svetha Venkatesh, and Mohan Kumar. Temporal data fusion in multi-sensor systems using dynamic time warping. In *SENSORFUSION 2005: Workshop on Information Fusion and Dissemination in Wireless Sensor Networks*, pages 1–9. IEEE, 2005.