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 [10]. 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. [11]

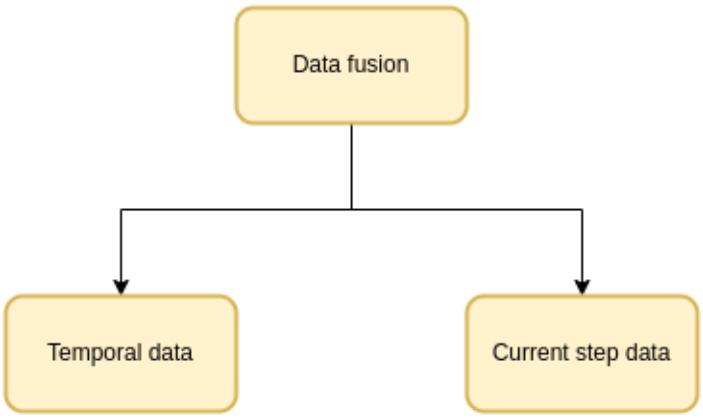

Figure 1.1: Data fusion categories based on timestamp

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 [12], video-based depth estimation [13], and segmentation [14].

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 [15]. 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 [16]. The temporal arrangements of events are captured, thereby incorporating the cause and effect phenomenon [16]. Temporal fusion can be commonly observed in human activity detection [17], context-aware mobile phones [19], and online batch process monitoring [18].

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 [19] [20] [21] [22] and video object detection [23] [24] [25] 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% [26]. 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 [27]. Multi-camera video surveillance fuses the spatiotemporal frames from different sources to reliably find the motion trajectories [28]. 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 [29].

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 [30], 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 [31] 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 [32]. For specialized tasks, different algorithms are developed [33]. Work by Shervin surveyed the various state-of-the-art segmentation algorithms [34]. 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 [35] 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 [36] 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 [37], [38], [39]. 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 [34].

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 [40]. 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 [41].

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 [42]. 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 [43].

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

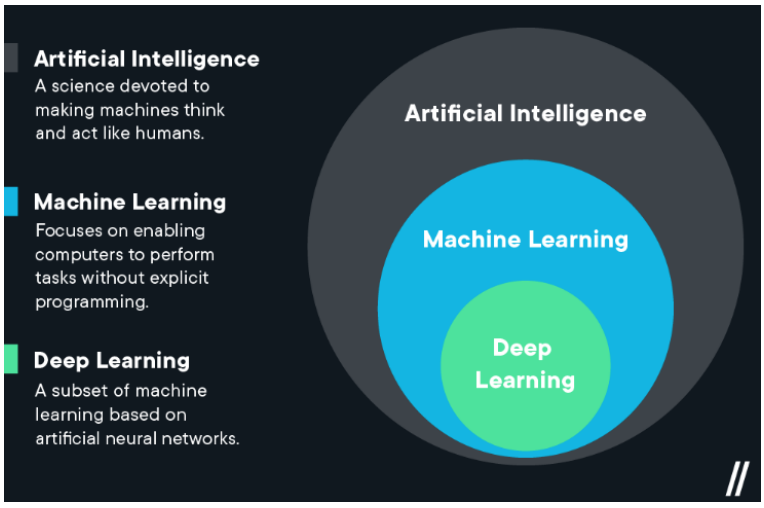

Figure 2.1: Deep learning in the artificial intelligence domain. Courtesy of [1]

Classical machine learning system uses the raw input and domain expert carefully represent the data as a feature vector from which the data is fed to the models to learn the patterns and classify into appropriate classes [44]. Deep learning is a representation learning that takes raw data and find the patterns in the data with different levels of representation in the multiple layers [44]. Deep learning can learn any complex representation of the data. For example, a image is represented as pixels and are fed to the neural network, at each layer of the network different feature are learned. In the first layer, higher level features such as edges at a specific orientation and location is determined. In the second layer motifs are learned and so on. The important aspect of the deep learning is that the features are not designed by the field expert rather than learned from the data with a specific set of learning procedures [44].

Many current state of the art learning models uses the deep learning approach to learn the complex function from data. Currently deep learning method can be found in image recognition [45], speech technologies [46], discovery of drug molecule [47] , understanding the particle accelerator data [48], DNA sequencing [49], ,and natural language processing [50].

Computer vision is the field of computer science that deals with replicating the functionalities of the human visual system. Traditionally computer vision solved the vision problem by finding the hand crafted features. However, the performance of the classical apporach is outperformed by the advancement of the deep learning based methods. Hand crafted feature descriptors such as Speeded Up Robust Features (SURF), Hough Transforms are used as feature vectors for the classical machine learning methods for learning [51]. Deep learning methods automatically learns the patterns from the data. Computer vision solves wide variety of problems in the perception domain. Latest approaches helps to solve the detection [52], [53], classification [54], image synthesis and segmentation tasks [55].

Temporal data are the time varying information and can be commonly observed in financial portfolio management, accounting, medical records, inventory management, data from airline, hotel, train industries contains time component with it [56]. Video data are constructed from combining time variant frames and is a common example of a temporal data. Temporal fusion deals with combining of the past information into the current step computation with a aim of improved performance.

In general approach segmentation is done frame by frame or by skipping in between frames and computing the segmentation on the nth frame. Temporal fusion can be applied in these settings to perform improved segmentation task by combining the past rich information in the current step.

2.2 Temporal Fusion

Temporal fusion can be defined as the process of fusing the temporal data onto the current step with a aim of improving the performance of the model. Temporal data can be observed in many fields such as social media, healthcare, accounting, agriculture, transportation, physics, crime data, traffic dynamics and climate science [57]. Temporal data can be encountered with different data types, Common data types are video, audio, tabular data, and sensors data. Forecasting is a common application of temporal fusion. Multi horizon forecasting is a important problem in the domain of time series. Multi horizon forecast allow the user to optimize the process across the entire path. A novel Temporal fusion Transformer (TFT) [58] is a attention based DNN architecture for forecasting by fusing the important past features into

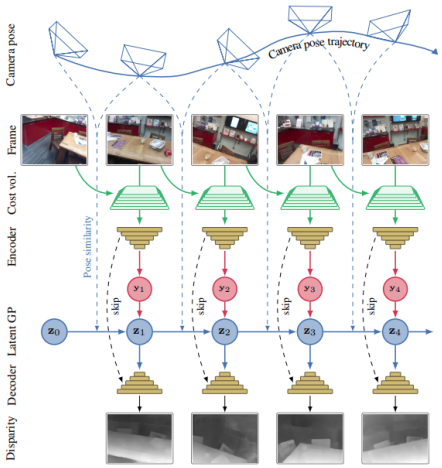

Figure 2.2: Mulit view stereo architecture for depth estimation. Courtesy of [2]

the current step. Temporal fusion play a major role in the improved video action recognition. Temporal fusion helps in two ways, firstly by understanding the temporal data the accuracy of the recognition for the dynamic action is improved, secondly removing the redundant temporal data saves the computation overhead. A temporal fusion network known as the AdaFuse, fuses the current and past features with a goal of improved accuracy and efficiency [59]. A temporal non parametric fusion aims to fuse the temporal pose data to the computation of the depth map thereby improving accuracy and efficiency [2]. The architecture of the multi view stereo can be depicted in the Fig 2.2. An online multi view depth prediction approach where the depth estimated in the previous step is fused onto the current step in a sensible manner. The network is named as DeepVideoMVS and it is based on the encoder decoder architecture. A ConvLSTM is placed at the latent space to fuse the information from the previous step. The proposed approach outperformed all the existing state of the art multi view stereo method evaluated on the standard metrics [60]. A Multiple Fusion Adaptation (MFA) method improves the segmentation accuracy on a unlabeled datasets. Three fusion approach was proposed under MFA, cross model fusion, temporal fusion and novel online-offline pseudo labels. The MFA produced improved semantic segmentation result of 58.2% and 62.5% on GTA5-to-Cityscapes and SYNTHIA-to-Cityscapes respectively [61].

2.3 Semantic Segmentation

Humans can perceive the surrounding environment and make sense of it with high accuracy. Due to the advancement of computer vision these capabilities are transferred to the machines, performing even better than humans. Today, we have computer vision models that can detect objects, find shapes, track the object movement and perform action based on the data. Computer vision is most commonly used in the autonomous driving cars, aerial mapping, surveillance applications, virtual reality and augmented reality and so on. One of the common problem in computer vision is labeling the each pixels of the image to a particular categories. Also known as the segmentation. Mathematically image segmentation can be defined as

If I is set of all image pixels of a image, then segmentation generate unique regions $S_1, S_2, S_3, S_4, ...S_n$ such that combining all these regions will return I.

Image segmentation can be classified into three categories Semantic segmentation, Instance segmentation and Panoptic segmentation. Semantic segmentation finds the shape, size and form of the objects in addition to their location. Instance segmentation finds one more parameter of number of unique object present in the image. Panoptic segmentation is the combination of the semantic and instance segmentation. The difference between all the types of semantic segmentation can be observed in the Fig 2.3.

Figure 2.3: Semantic and Instance segmentation example. Courtesy of [3]

2.3.1 Classical Semantic Segmentation

Most commonly used traditional segmentation techniques are threshold based technique [62], histogram-based bundling, region-growing [63], k-means clustering, watersheds, active contours, graph cuts, conditional and Markov random fields [64], sparsity based methods [65]. However, in the recent years deep learning (DL) yielded a new generation of image segmentation models with state of the art performance.

2.3.2 Deep Learning based Semantic Segmentation

Deep learning based segmentation network can be classified into following categories [4]

- Fully convolutional networks
- Convolutional models with graphical models

- Encoder-decoder based models
- Multi-scale and pyramid network based models
- R-CNN based models (for instance segmentation)
- Dilated convolutional models and DeepLab family
- Recurrent neural network based models
- Attention-based models
- Generative models and adversarial training
- Convolutional models with active contour models

Deep learning based computer vision model most commonly use the convolutional neural network [66], recurrent neural network (RNNs), and Long short term memory (LSTM), encoder-decoder [6] and generative adversarial networks (GANs) based networks [4]. The master thesis work is concentrated on the encoder-decoder based deep learning models. Encoder-Decoder based network are a two stage network that learns to map from input point to the output point. In the encoder stage the input data is compressed into a latent space representation $z = f(x)$ and decoder decompress the latent space representation to the output $a = g(z)$ [67]. Latent representation of the input data in compressed form. It can be commonly observed in image-to-image translation problem as well as in sequence-to-sequence models in NLP. A reconstruction loss $L(y, \hat{y})$ is defined at the output that measure the differences between the ground truth output y and corresponding reconstruction $\hat{y}$. Autoencoders are the special version of the encoder-decoder models that have similar input and output.

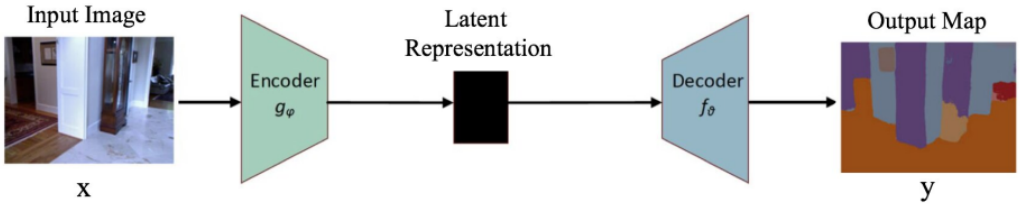

Figure 2.4: Simple encoder-decoder architecture. Courtesy of [4]

Most of the segmentation network are encoder-decoder based architecture. A novel semantic segmentation network was proposed by Noh et al [5]. The network is based on the deconvolution. The encoder network is based on the VGG 16-layer network and the decoder network takes the latent space encoding and outputs the pixel wise class probabilities. The segmentation mask and pixel-wise class labels are predicted by the deconvolutional and unpooling layers. The network generated a accuracy of 72.5 % on the PASCAL VOC 2012 dataset.

Badrinarayanan et al proposed a convolutional encoder-decoder architecture for image segmentation called as SegNet [6]. The architecture of the SegNet described in the Figure 2.6

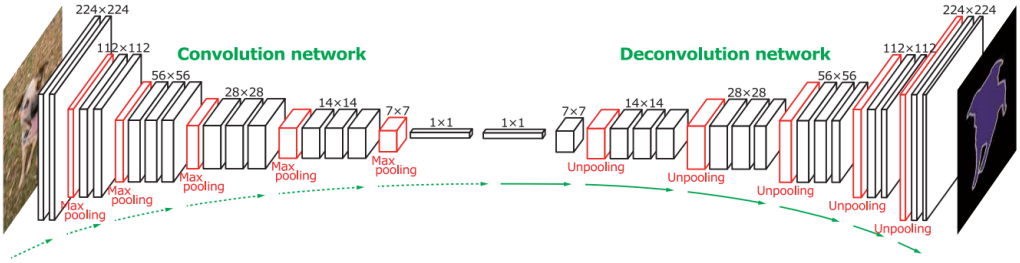

Figure 2.5: Simple encoder-decoder architecture. Courtesy of [5]

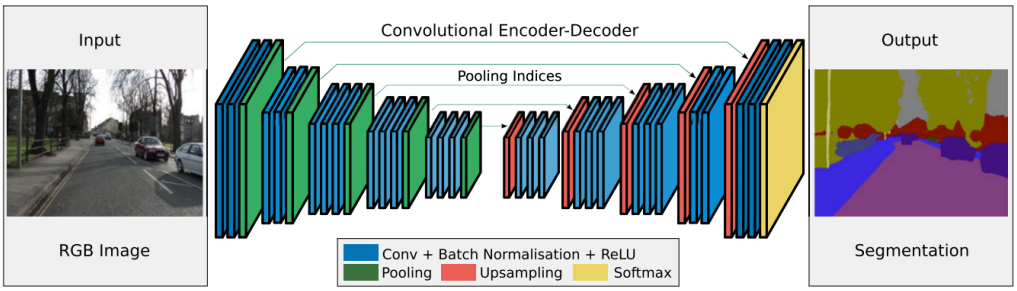

Figure 2.6: SegNet architecture. Courtesy of [6]

The encoder part of the SegNet consist of 13 convolutional layers in the VGG16 network, and followed by the pixel wise classification layer. Decoder upsample the low resolution feature maps in a unique fashion. Non-linear upsampling is performed by using the pooling indices computed in the max-pooling step of the encoder. This process of reusing the encoder output helps to eliminate the need for learning to up-sample. Dense feature maps are generated by convolving with the trainable filters. To account for the uncertainty involved with the encoder-decoder network, scene segmentation is proposed [68]. HRNet [69] is the recently developed high resolution network by connecting the high to low resolution convolutions streams in parallel and exchanging information between different resolutions. HRNet maintain high resolution representation through the encoding process. Many recent architecture use HRNet as the backbone. Other encoder decoder segmentation models are Stacked Deconvolutional Network [70], Linknet [71], W-net [72].

Many segmentation models are developed for medical application and among those U-Net [7] and V-Net [73] are the famous architecture. These architecture are now used outside of the medical domain.

Ronneberger et al [7] proposed a segmentation model to perform semantic segmentation on medical microscopy images. The architecture of the U-Net is described in Fig 2.7. The context is captured by the contracting part and localization of the target area is identified by the expanding decoder path. The network heavily dependent on the annotated images efficiently. The encoder part has a 3x3 convolutions features extractor, similar to the FCN-like architecture. The decoder part increases the dimensions and reducing the number of feature maps. The feature map from the encoder is mapped to the upscaled decoder to retain the pattern information. A 1x1 convolution at the output process the feature maps to

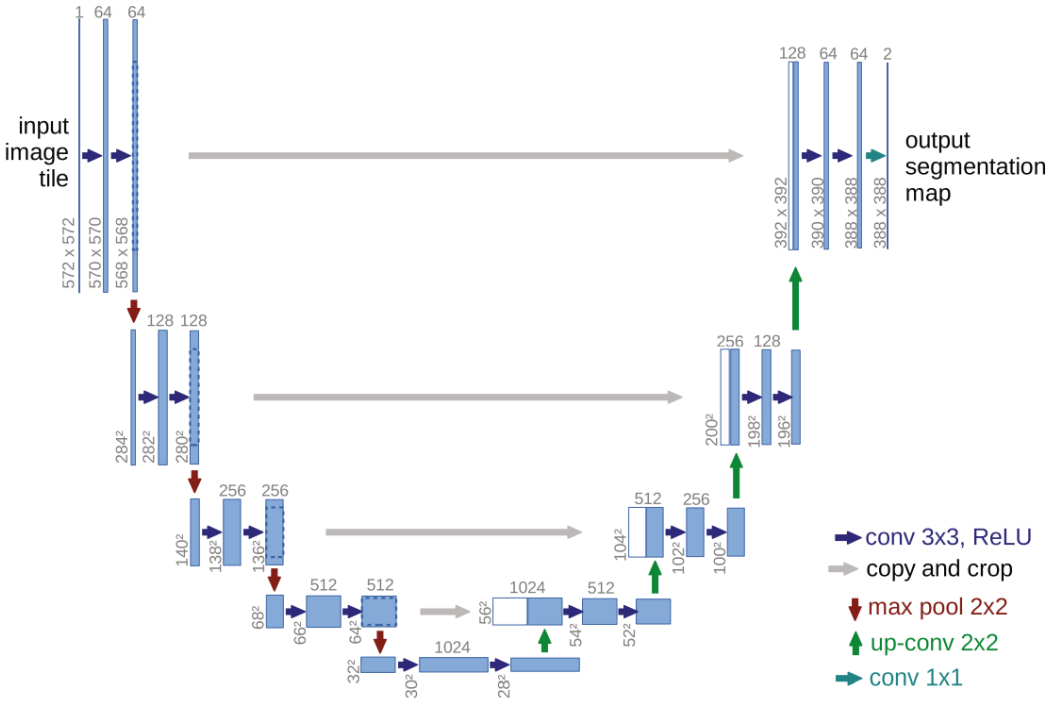

Figure 2.7: Unet architecture. Courtesy of [7]

generate segmentation output by categorizing each pixels of the input image to a particular class. Original U-Net was trained on the electronic microscopic images and outperformed by a large margin on the ISBI challenge. The network is fast and produce result on 512x512 image in less than a second on the modern GPU [7], [4]. In a sequence data the information from the previous frames can be utilized to segment the current frame with a aim of improved performance in comparison to the segmentation without the temporal fusion.

2.4 Temporal Fusion in Semantic Segmentation

Semantic segmentation of sequence data aims to assign pixel-wise semantic labels to the video frames. It is a important task in the visual understanding [74]. Strong representation of the feature map are important for the segmentation task. One of the common approach in the video segmentation is to perform the image segmentation to each frame independently. However the temporal information of the dynamic scenes are not captured by this approach. A common solution to the problem is to apply semantic segmentation to the each and every frame and add additional layer on top to capture the temporal data to extract the better features [75], [76], [77]. However, such approach doesn't help to improve the performance as feature needs to be computed at each and every frame. So a good approach is to apply the segmentation at key frames and reuse the already computed features for the other frames [78], [79]. A new highly efficient and low accuracy neural network based model is developed for semantic video

segmentation called as Temporally Distributed Network (TDNet) [8]. In TDNet feature extraction is distributed evenly across the sequential frames to eliminate the re-computation and then these features are combined together using the Attention Propagation Module (APM) to get the strong features for accurate segmentation [8]. The pictorial representation of the same is described in the Fig 2.8.

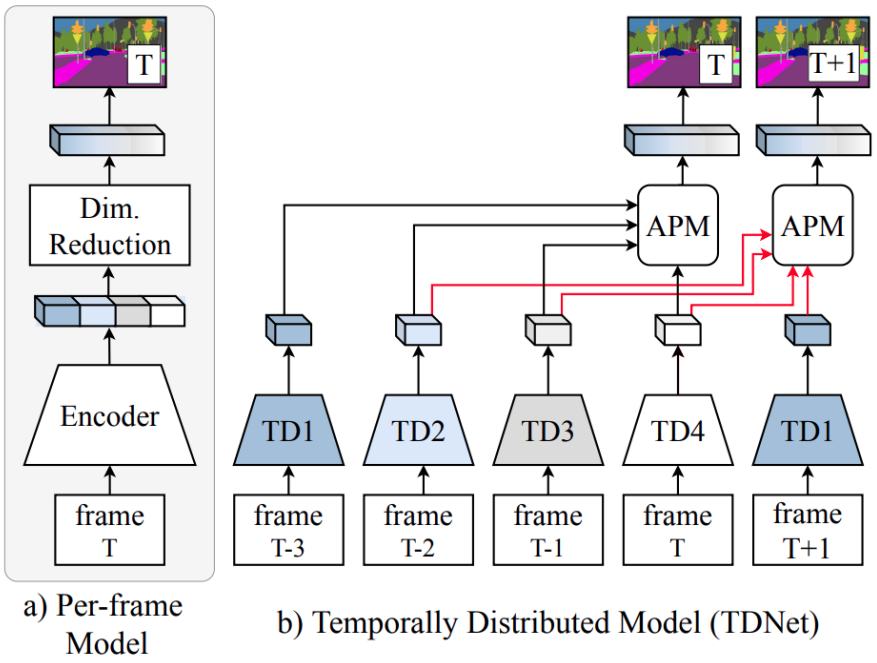

Figure 2.8: TDNet. Courtesy of [8]

2.5 Limitations of Previous Work

The perception system of the modern ADAS uses segmentation to understand the surrounding environment by capturing the surrounding environment with the help of modern cameras. The high FPS data collected by the camera are in continuous sequence where the every frame is related to its previous frames. In general setting segmentation is performed on these frames to understand the object and their boundaries, number of objects, types of objects present in the frame. A work by Hou et al [2] integrate the camera pose data onto the computation of the depth maps, however similar strategy is not studied for a segmentation task. Also study of temporal data fusion in the latent space using LSTM is not studied in any of the previous work. This thesis aims to study the impact of temporal pose data onto the computation of the semantic segmentation and taking the previous frame data onto the current frame semantic segmentation task using LSTM network is studied.

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

Evaluation and experimental result chapter contains the metrics used to evaluate the conducted experiment and the discussion on the research questions. Results of research question is answered in the subsequent sections with listing of the result in a table. Three research questions answered in the section are the results of state of the art temporal fusion techniques, How are the results in comparison to each other and finally cross transfer the temporal fusion techniques from depth estimation to the semantic segmentation. The model is trained and evaluated on three datasets Scannet [80], Virtual kitti [81] and VIODE [82] datasets.

4.1 Evaluation Metric

The proposed model needs to be validated to understand the impact of the newly trained model. To validate the model different evaluation metrics are proposed. The description of the proposed model is listed below.

4.1.1 Pixel Accuracy

Pixel accuracy is commonly defined as percent of pixels in a image correctly classified into a particular class. Accuracy is defined as below

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

Per class mean pixel accuracy (mPA)

Per class mean pixel accuracy is the average of pixel accuracies of all the classes.

Pixel accuracy (PA) and Mean pixel accuracy (mPA) can also be defined as below [83]

$$\textbf{PA} = \frac{\sum_{j=1}^{k} n_{jj}}{\sum_{j=1}^{k} t_j}$$

$$\textbf{mPA} = \frac{1}{k} \frac{\sum_{j=1}^{k} n_{jj}}{\sum_{j=1}^{k} t_j}$$

where n_{jj} is the total number of pixels both classified and labeled as class j.

4.1.2 IoU

Intersection over Union (IoU) also known as the Jaccard index is a method to quantify the overlapping between the target mask and the predicted output. In other words, it is number of pixels common between the target and prediction masks by the total number of pixels exist between both the masks [84]. pictorial representation of the same is presented in Fig 4.1

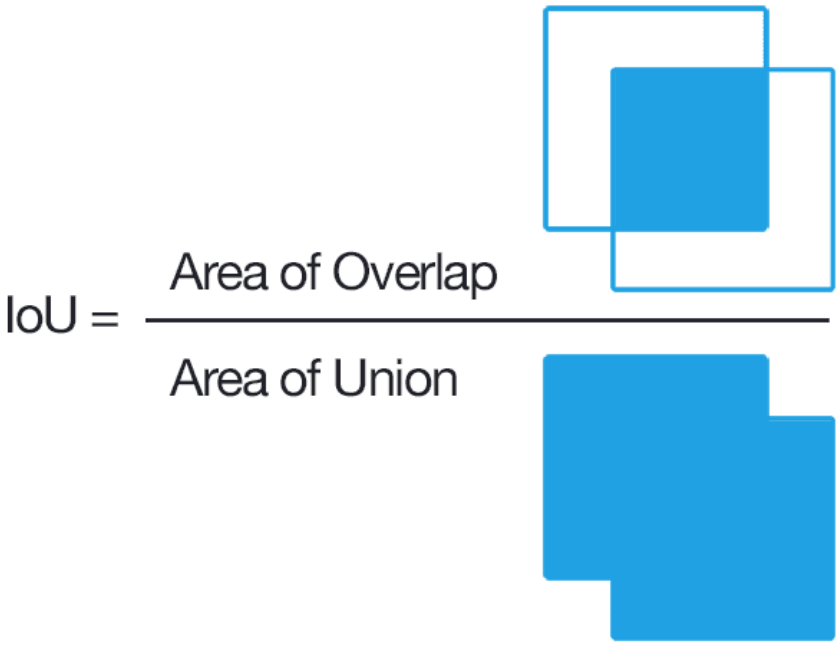

Figure 4.1: IoU. Courtesy of [9]

IoU is calculated for each class separately and then averaged over all the classes to provide mean IoU.

$$\textbf{IoU} = \frac{\sum_{j=1}^{k} n_{jj}}{\sum_{j=1}^{k}(n_{ij} + n_{ji} + n_{jj})}, i \neq j$$

where n_{ij} is the pixels which are labeled as class i but classified as class j and n_{ji} is the total number of pixels labeled as class j, but classified as class i. [83]

$$\textbf{mIoU} = \frac{1}{k} \sum_{j=1}^{k} \frac{n_{ij}}{n_{ij} + n_{ji} + n_{jj}}, i \neq j$$

Frequency weighted IoU (FwIoU). It is a metric derived from the mIoU which weighs each class importance depending on appearance frequency using t_j [83]

$$\textbf{FwIoU} = \frac{1}{\sum_{j=1}^{k} t_j} \sum_{j=1}^{k} t_j \frac{n_{jj}}{n_{ij} + n_{ji} + n_{jj}}, i \neq j$$

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

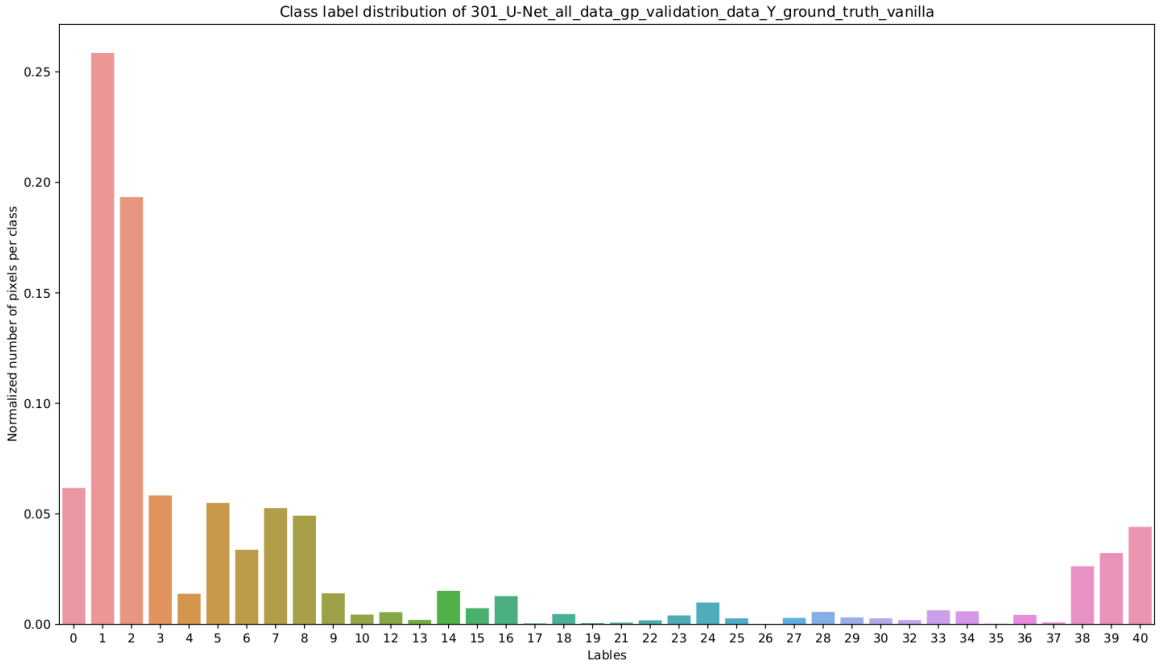

Figure 4.2: Per class pixel distribution of the ground truth pixel class label

Metric	Value
Pixel Accuracy	0.5096
Pixel Mean accuracy	0.1907
meanIOU	0.1102
IoU	[1.8345e-01, 5.6047e-01, 6.0881e-01, 1.6476e-01, 3.8241e-01, 1.1697e-01, 2.3122e-02, 8.9410e-02, 2.5848e-01, 1.6661e-01, 2.0180e-03, 2.0504e-01, 4.5985e-02, nan, 4.2167e-02, 1.1143e-01, 2.3969e-01, 1.2545e-02, 1.1324e-01, 2.7658e-03, 0.0000e+00, 7.0415e-02, 7.0606e-02, 0.0000e+00, 0.0000e+00, 1.2208e-01, 1.4680e-02, 1.1862e-04, 2.2112e-04, 6.5610e-03, 4.3742e-03, nan, 7.6680e-02, 3.5784e-02, 1.1516e-01, 5.6912e-02, 2.9310e-04, 3.7764e-02, 2.2634e-02, 1.1269e-01, 2.2094e-01]
FwIoU	0.3531

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

Distance and covariance matrix

Ordered set of images

Results of the experiment

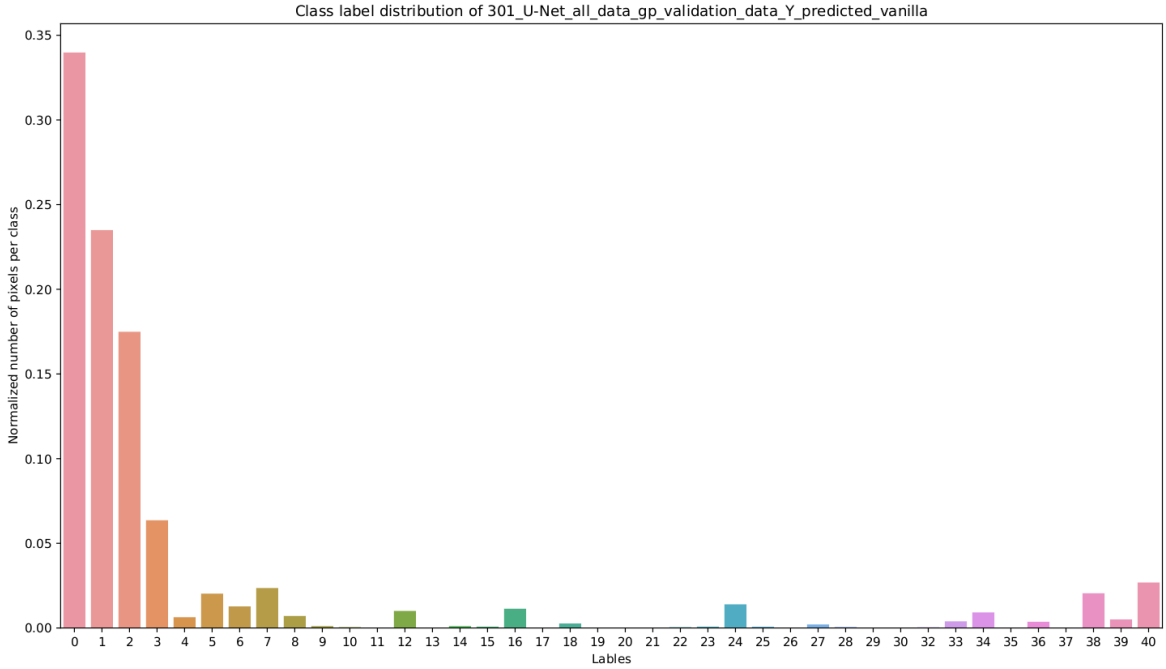

Figure 4.3: Per class pixel distribution of the predicted pixel class label

Metric	Value
Pixel Accuracy	0.5184
Pixel Mean accuracy	0.1679
meanIOU	0.1161
IoU	[1.7087e-01, 5.1271e-01, 5.9998e-01, 2.1256e-01, 4.1160e-01, 1.5834e-01, 3.8634e-02, 2.3669e-01, 1.6056e-01, 1.1568e-01, 7.9677e-02, 1.0454e-02, 2.4003e-02, 0.0000e+00, 1.2199e-01, 4.3193e-02, 3.3956e-01, 6.6473e-02, 1.4712e-01, 2.8003e-03, 6.0475e-05, 2.6127e-01, 5.7962e-02, 0.0000e+00, 3.4611e-04, 1.6519e-02, 0.0000e+00, 4.3417e-04, 4.4221e-02, 6.6478e-03, 1.2108e-02, nan, 5.3272e-02, 5.8480e-02, 2.2352e-01, 4.2175e-02, 8.4644e-02, 1.1630e-04, 4.9106e-02, 1.0338e-01, 1.7687e-01]
FwIoU	0.3497

4.4.3 Experiment1.3: U-Net temporally fused lstm model

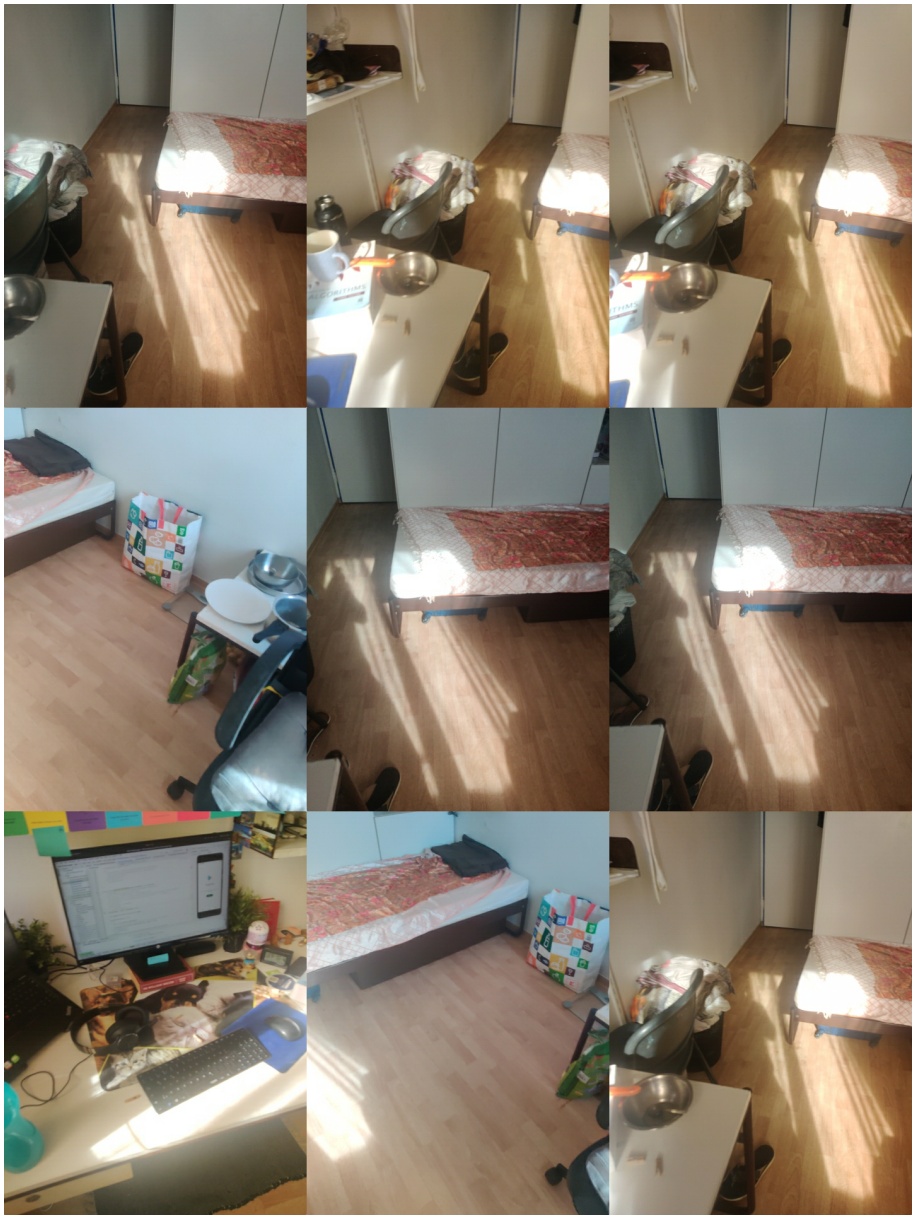

Figure 4.4: Ordered set of images

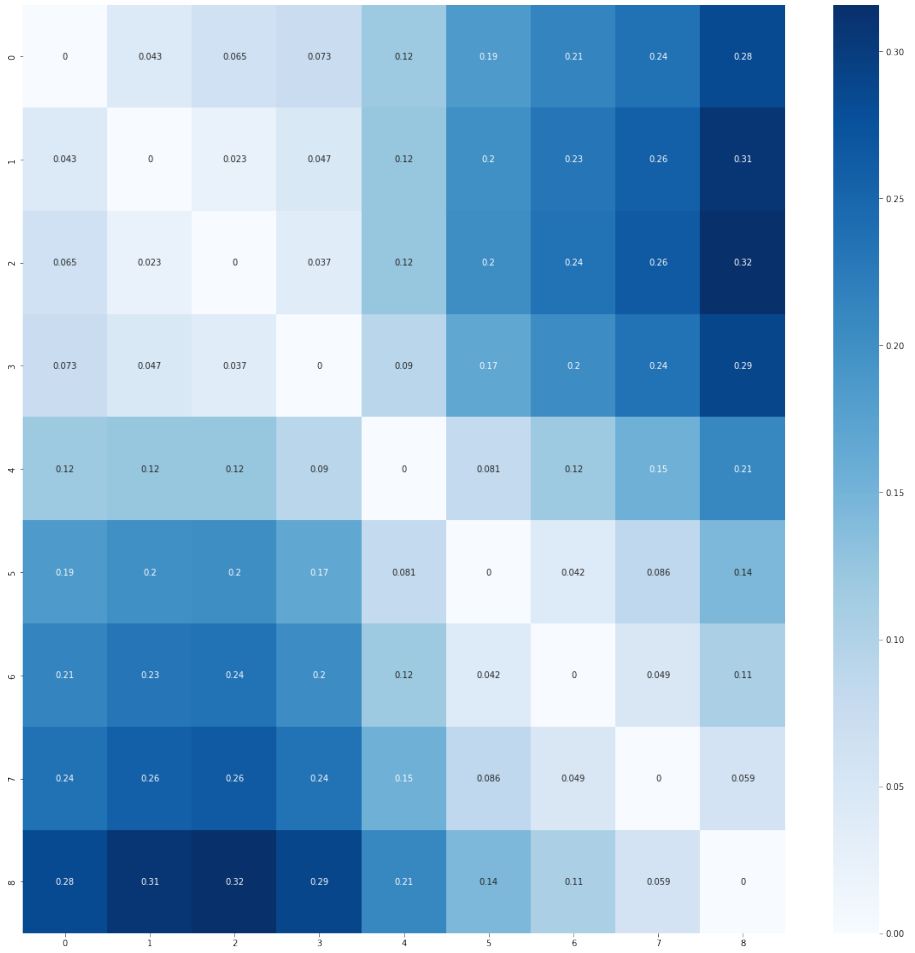

Figure 4.5: Distance matrix depicted as heatmap

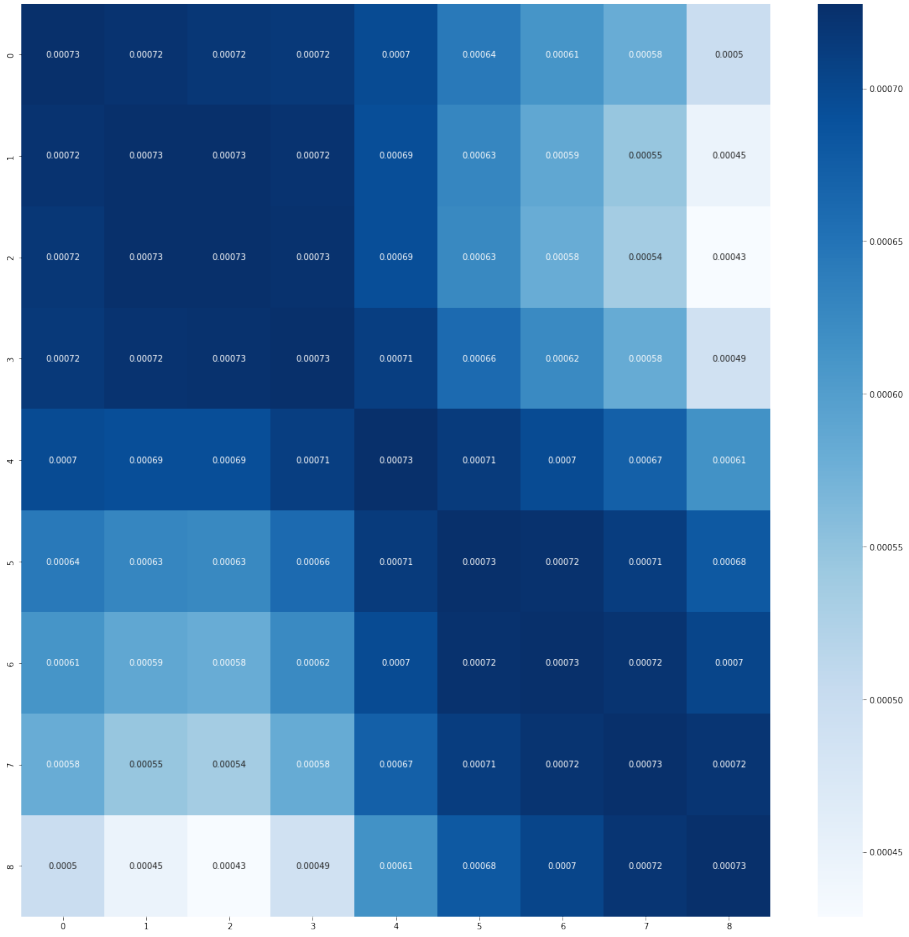

Figure 4.6: Kernel matrix depicted as heatmap

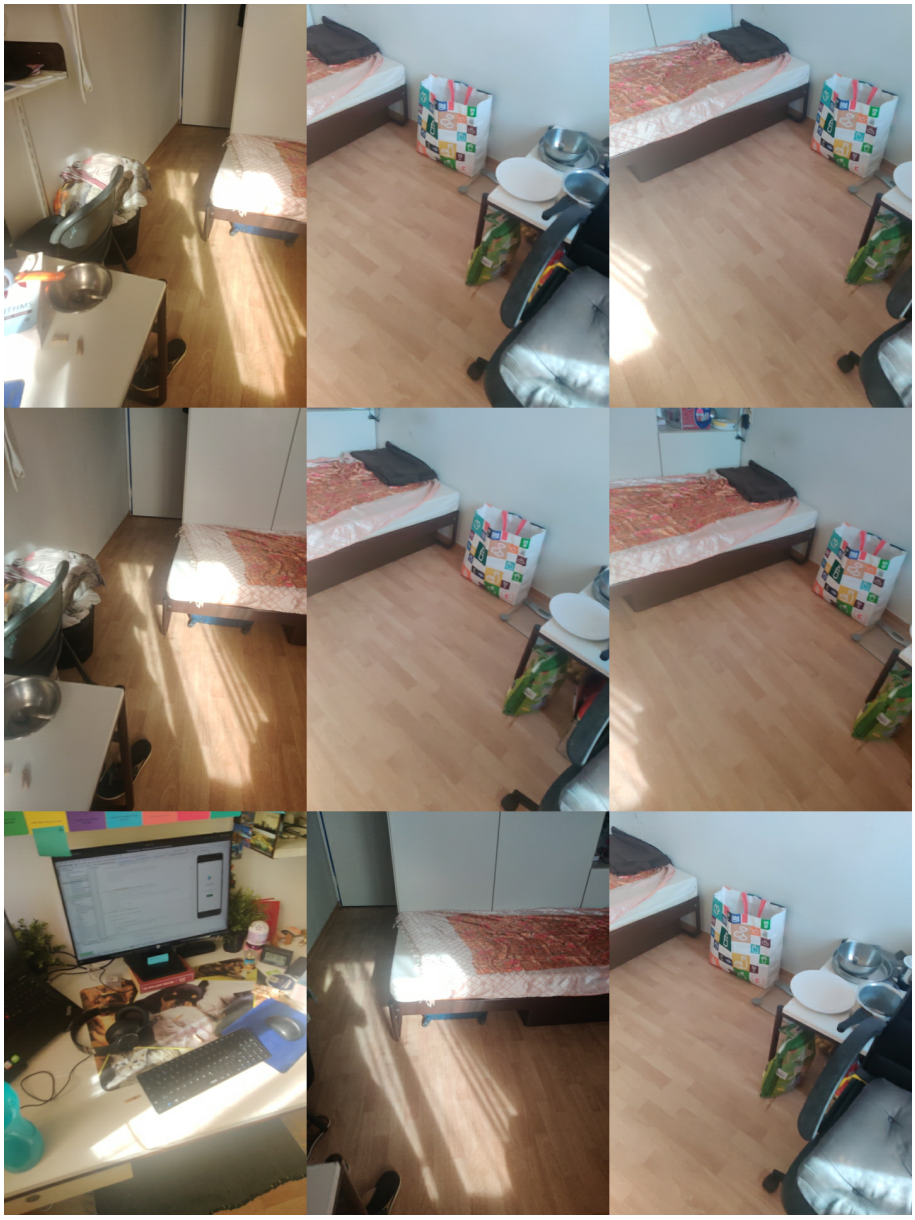

Figure 4.7: Ordered set of images

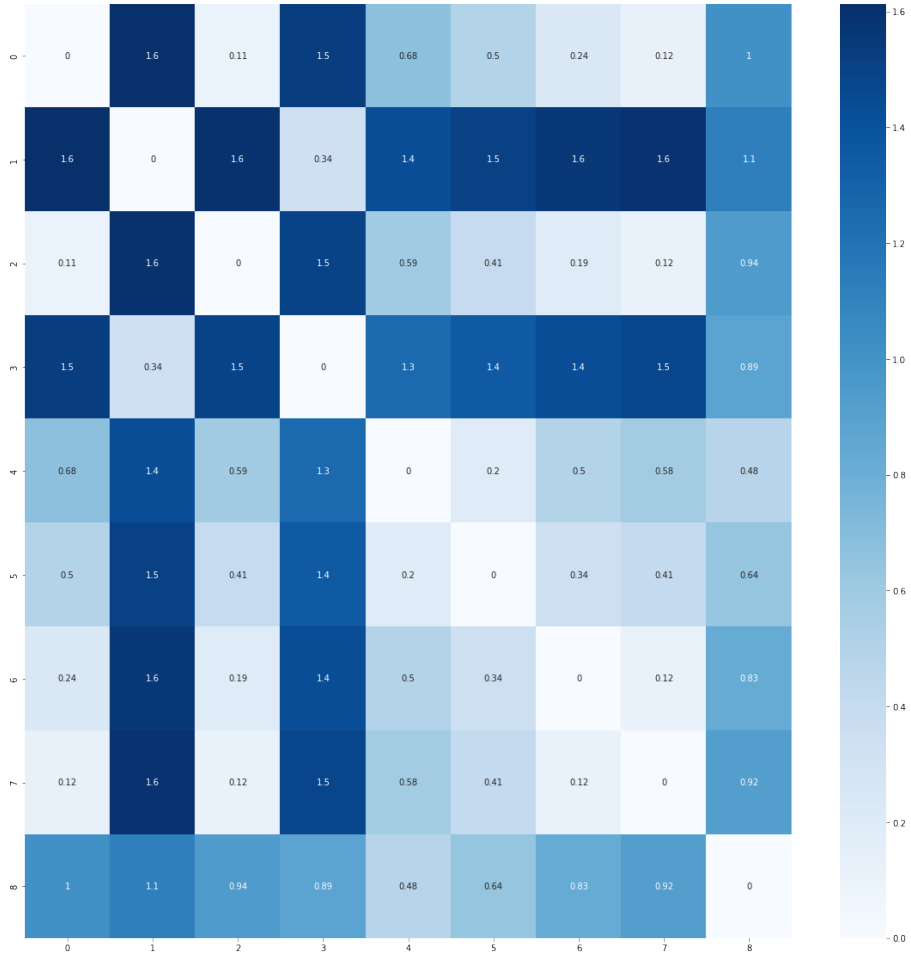

Figure 4.8: Distance matrix depicted as heatmap

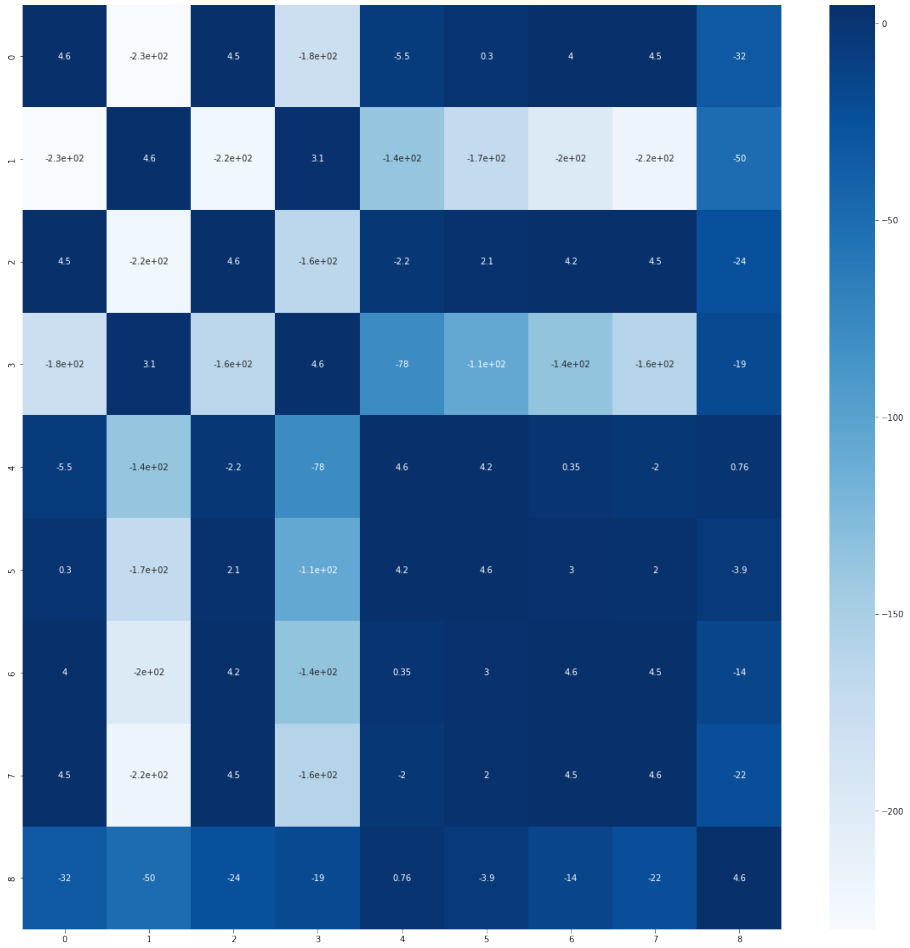

Figure 4.9: Kernel matrix depicted as heatmap

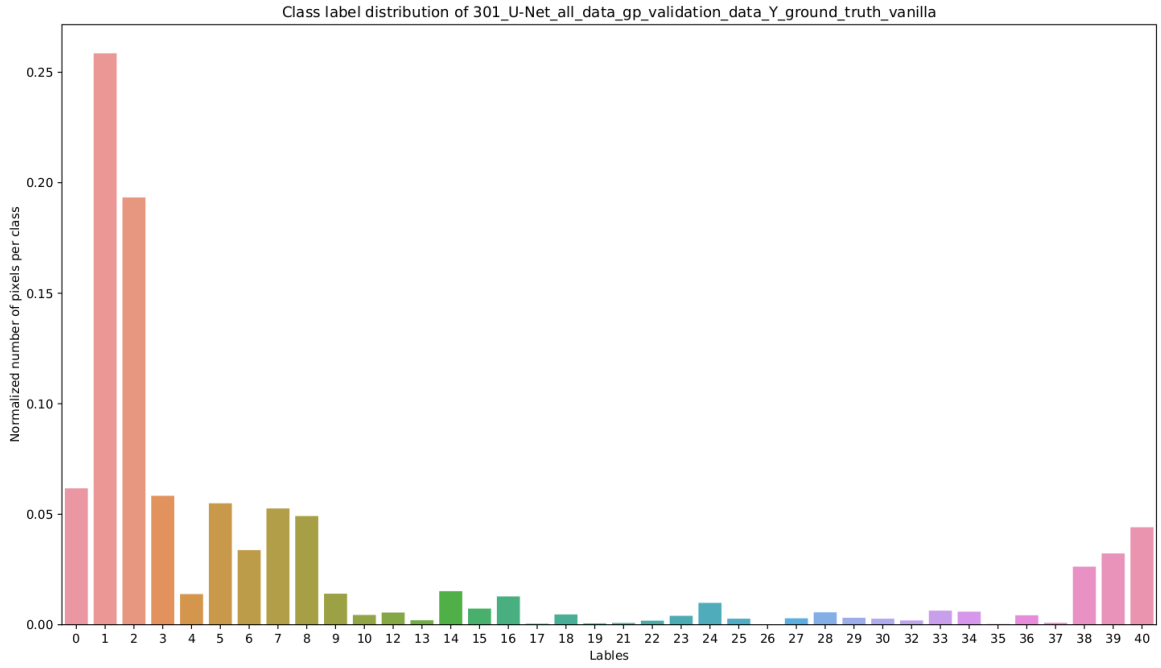

Figure 4.10: Per class pixel distribution of the ground truth pixel class label for gp model

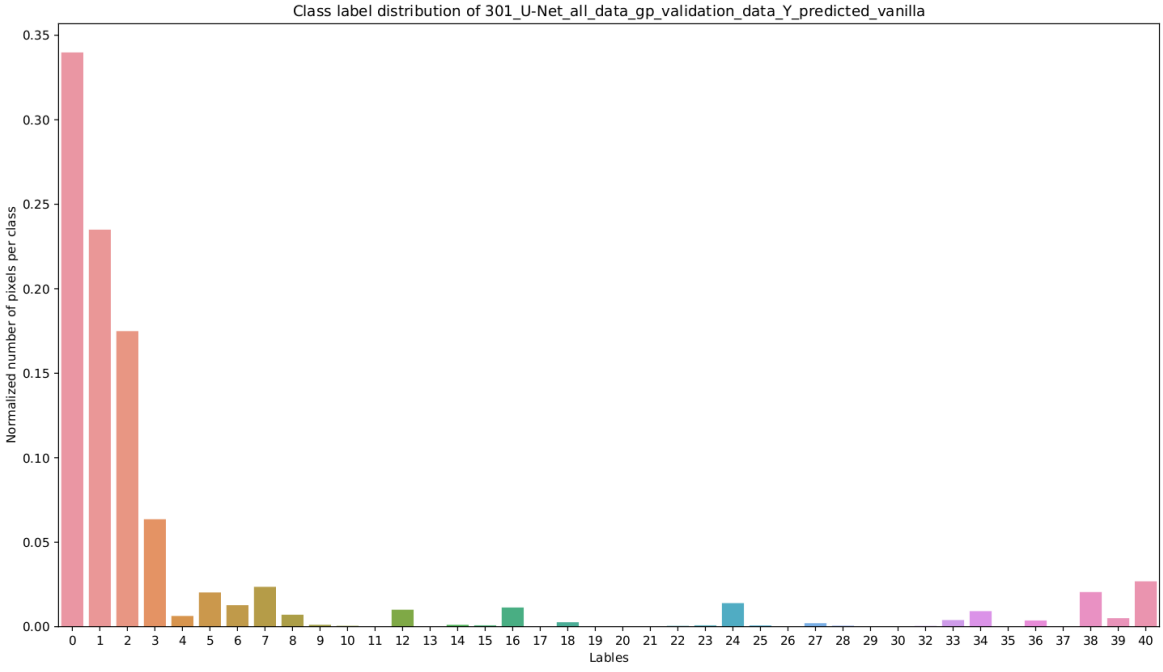

Figure 4.11: Per class pixel distribution of the predicted pixel class label for gp model

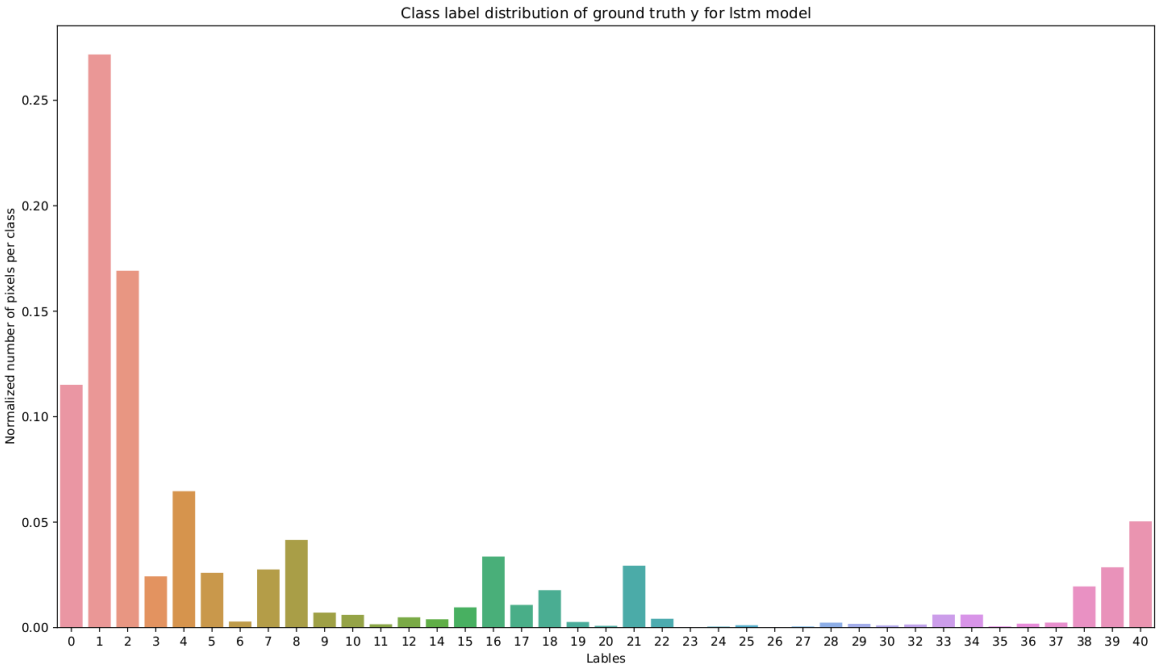

Figure 4.12: Per class pixel distribution of the ground truth pixel class label for lstm model

Metric	Value
Pixel Accuracy	0.5601
Pixel Mean accuracy	0.2254
meanIOU	0.145
IoU	[0.2664, 0.6187, 0.6561, 0.0875, 0.4232, 0.1989, 0.0567, 0.1796, 0.2832, 0.1905, 0.0258, 0.2313, 0.0013, 0.0000, 0.0681, 0.1169, 0.2300, 0.0685, 0.1384, 0.0182, 0.0000, 0.0600, 0.1695, 0.0515, 0.0000, 0.3135, 0.0000, 0.0000, 0.0069, 0.0335, 0.1894, nan, 0.0804, 0.1759, 0.1104, 0.1517, 0.0053, 0.0996, 0.1119, 0.1184, 0.2613]
FwIoU	0.3996

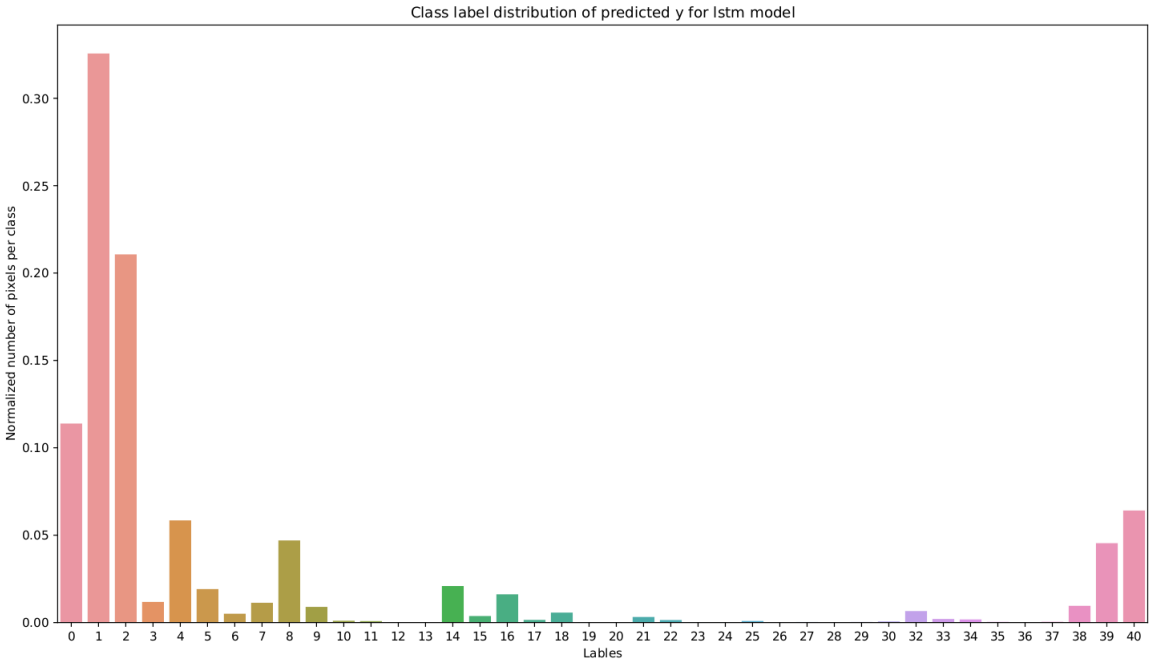

Figure 4.13: Per class pixel distribution of the predicted pixel class label for lstm model

4.4.4 Temporal fusion on a continuous sequence data

4.4.5 RQ3.1: Which fusion method is good for the scannet data?

4.4.6 RQ3.2: Which fusion method is good for the virtual kitti data?

4.4.7 RQ3.3: Which fusion method is good for the VIODE data?

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?, 2021.

[2] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2660, 2019.

[3] Your complete guide to image segmentation. `https://www.telusinternational.com/articles/guide-to-image-segmentation#:~:text=Different%20types%20of%20image%20segmentation%20tasks,types%20of%20image%20segmentation%20tasks.` Accessed: 2022-08-22.

[4] 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.

[5] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[6] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[7] 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.

[8] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8818–8827, 2020.

[9] Adrian Rosebrock. Intersection over union (iou) for object detection, 2016.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[45] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2012.

[46] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[47] Lauv Patel, Tripti Shukla, Xiuzhen Huang, David W Ussery, and Shanzhi Wang. Machine learning methods in drug discovery. *Molecules*, 25(22):5277, 2020.

[48] T Ciodaro, D Deva, JM De Seixas, and D Damazio. Online particle detection with neural networks based on topological calorimetry information. In *Journal of physics: conference series*, volume 368, page 012030. IOP Publishing, 2012.

[49] Jinny X Zhang, Boyan Yordanov, Alexander Gaunt, Michael X Wang, Peng Dai, Yuan-Jyue Chen, Kerou Zhang, John Z Fang, Neil Dalchau, Jiaming Li, et al. A deep learning model for predicting next-generation sequencing depth from dna sequence. *Nature communications*, 12(1):1–10, 2021.

[50] Julia Hirschberg and Christopher D Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

[51] Niall O'Mahony, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh. Deep learning vs. traditional computer vision. In *Science and information conference*, pages 128–144. Springer, 2019.

[52] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016.

[53] Zhongchun Han and Anfeng Xu. Ecological evolution path of smart education platform based on deep learning and image detection. *Microprocessors and Microsystems*, 80:103343, 2021.

[54] Shrey Srivastava, Amit Vishvas Divekar, Chandu Anilkumar, Ishika Naik, Ved Kulkarni, and V Pattabiraman. Comparative analysis of deep learning image detection algorithms. *Journal of Big Data*, 8(1):1–27, 2021.

[55] 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.

[56] Christian S Jensen and Richard T Snodgrass. Temporal data management. *IEEE Transactions on knowledge and data engineering*, 11(1):36–44, 1999.

[57] Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys (CSUR)*, 51(4):1–41, 2018.

[58] 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.

[59] Yue Meng, Rameswar Panda, Chung-Ching Lin, Prasanna Sattigeri, Leonid Karlinsky, Kate Saenko, Aude Oliva, and Rogerio Feris. Adafuse: Adaptive temporal fusion network for efficient action recognition. *arXiv preprint arXiv:2102.05775*, 2021.

[60] 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.

[61] Kai Zhang, Yifan Sun, Rui Wang, Haichang Li, and Xiaohui Hu. Multiple fusion adaptation: A strong framework for unsupervised semantic segmentation adaptation. *arXiv preprint arXiv:2112.00295*, 2021.

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

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

[64] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2001.

[65] J-L Starck, Michael Elad, and David L Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE transactions on image processing*, 14(10):1570–1582, 2005.

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

[67] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[68] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[69] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

[70] Jun Fu, Jing Liu, Yuhang Wang, Jin Zhou, Changyong Wang, and Hanqing Lu. Stacked deconvolutional network for semantic segmentation. *IEEE Transactions on Image Processing*, 2019.

[71] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4233–4241, 2018.

[72] Xide Xia and Brian Kulis. W-net: A deep model for fully unsupervised image segmentation. *arXiv preprint arXiv:1711.08506*, 2017.

[73] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[74] Xiaojie Jin, Xin Li, Huaxin Xiao, Xiaohui Shen, Zhe Lin, Jimei Yang, Yunpeng Chen, Jian Dong, Luoqi Liu, Zequn Jie, et al. Video scene parsing with predictive feature learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5580–5588, 2017.

[75] Raghudeep Gadde, Varun Jampani, and Peter V Gehler. Semantic video cnns through representation warping. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4453–4462, 2017.

[76] Xiaojie Jin, Xin Li, Huaxin Xiao, Xiaohui Shen, Zhe Lin, Jimei Yang, Yunpeng Chen, Jian Dong, Luoqi Liu, Zequn Jie, et al. Video scene parsing with predictive feature learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5580–5588, 2017.

[77] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6819–6828, 2018.

[78] Samvit Jain, Xin Wang, and Joseph E Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8866–8875, 2019.

[79] Behrooz Mahasseni, Sinisa Todorovic, and Alan Fern. Budget-aware deep semantic video segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1038, 2017.

[80] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[81] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020.

[82] Koji Minoda, Fabian Schilling, Valentin Wüest, Dario Floreano, and Takehisa Yairi. VIODE: A simulated dataset to address the challenges of visual-inertial odometry in dynamic environments. *IEEE Robotics and Automation Letters*, 6(2):1343–1350, 2021.

[83] Irem Ulku and Erdem Akagündüz. A survey on deep learning-based architectures for semantic segmentation on 2d images. *Applied Artificial Intelligence*, pages 1–45, 2022.

[84] JEREMY JORDAN. Evaluating image segmentation models, 2018.

[85] 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.