Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it
Bonn-Aachen
International Center for
Information Technology

Master Thesis Proposal

Multi-View Temporal Fusion in Semantic Segmentation

Manoj Kolpe Lingappa

Supervised by

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

July 2022

1 Introduction

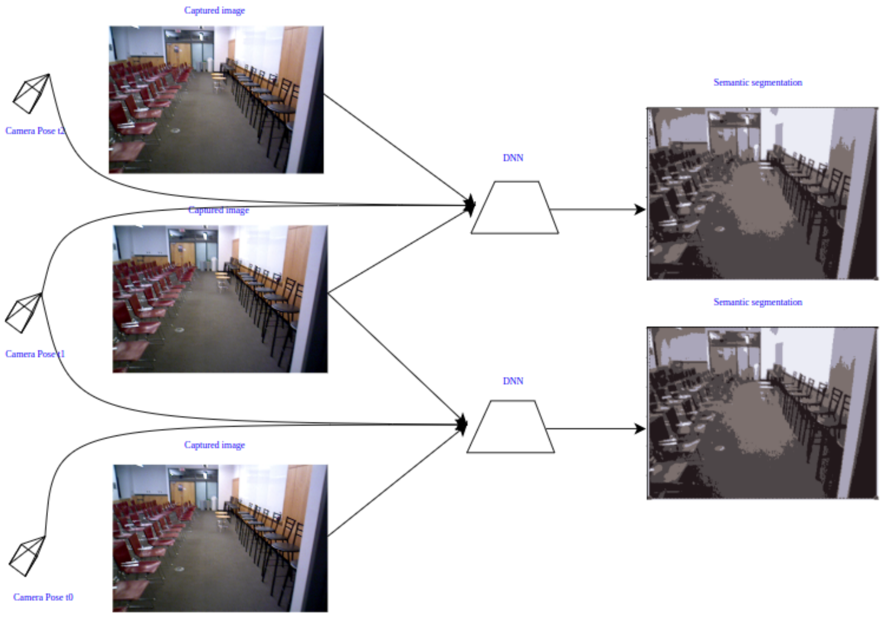

Figure 1: Semantic segmentation from a pair of images and pose

A common temporal fusion problem is predicting variables of interest at different future timestamps by integrating information at each step from past or other modalities. Fusing information at every step helps optimize the future target prediction and the temporal fusion of information observed in classical and deep learning-based architecture. An image-pose pair encoder-decoder-based architecture for disparity estimation successfully fuses information in the latent space with the help of the Gaussian process at the latent space [16]. To extract overlapping information an end-to-end video-based person re-identification architecture was proposed which fuses features in a spatial-temporal fashion. In this work, both the spatial and temporal features of the input image sequences are calculated by collecting the feature vector $f^{(t)}$ from the convolutional neural network (CNN) that connects the temporal pooling layer. The aggregated features combined into a single feature vector averaged over the entire sequence [5]. Information can be fused at the single-stage output of a multi-stage encoder-decoder network. An online multi-view feature prediction network on the posed video frame, fusing richer scene geometry calculated at the last step, is fused into the current step efficiently [12]. Geometric information from the output of the previous encoder-decoder network is fused to compute a cost volume and fed as input to the second stage encoder-decoder architecture. Also, the output of the second stage network is fed back to the convolutional long short-term memory (ConvLSTM) network in the latent space by warping [12].

Computer vision aims to understand the surrounding environment using various mathematical modeling techniques. Moreover, semantic segmentation is the key problem in computer vision that targets complete scene understanding. The semantic segmentation segments the input images based on the semantic information and categorizes each pixel into a class from the list of labels [28]. Semantic segmentation can be applied to 2D images, volumetric data, or videos. Number of field take advantage of understanding the surrounding environment such as human-machine interaction [31], computational photography [45], image search engines [40], augmented reality [22], autonomous driving [39], biomedical imaging [3], fashion [44]. Traditional semantic segmentation is performed with conditional random fields [32], clustering [8]. With the development of deep learning architecture, semantic segmentation problem are solved with CNN's [30], [6], [15]. A segmentation network can be thought of as an encoder-decoder network, usually with a pre-trained classification network on the encoder side followed by semantically projecting the learned discriminative feature onto the feature space to get a dense classification.

A general semantic segmentation framework involves independently applying a deep image segmentation model to every image in a sequence. However, these approaches result in information loss by not considering the correlation among the consecutive frames. The data loss problem tackled by taking the information from the previous step is fusing at the current step in a timely manner resulting in an efficient semantic segmentation. Estimation of semantic segmentation from unconstrained camera images is a challenging task. State-of-the-art temporal semantic segmentation based on deep learning. A temporally distributed network for semantic segmentation finds the correlation between the frames and the attention propagation module, effectively combining the distributed feature groups [18]. Efficient semantic segmentation can be performed by fusing current, and the previous camera poses information onto the latent space in an encoder-decoder-based architecture.

This thesis aims to study the state-of-the-art temporal fusion techniques and compare the results from these findings. Furthermore, cross-transfer the temporal fusion technique to the segmentation task by fusing information from the previous step to the current step to extract the correlation between the frames and optimize the prediction at every time stamp.

2 Problem Statement

Multi-view temporal fusion in the context of images is a field that targets fusing the temporal information at every step to extract the temporal correlation between the consecutive frames to optimize the prediction at each time stamp. Different fields use temporal fusion to optimize the prediction, and one such field is the segmentation task. Semantic segmentation aims to label each pixel in an image, which is an ill-posed problem. Over the past years, many algorithms and architectures have proposed finding the scenes' semantic segmentation. Conventional semantic segmentation algorithms compute semantic maps for the individual frame. They do not consider the geometric correlation between the consecutive frames, thereby losing crucial overlapping information in the successive images. Images can be obtained from a stereo camera or a monocular camera. Estimating semantic segmentation from the unconstrained monocular camera image is challenging. However, there are potential benefits to using monocular camera images. Firstly, it reduced cost and flexibility in the movement of the monocular camera, thereby gathering rich information about the surrounding environment. Secondly, multiple varying point images can fuse all the information for robust and stable semantic segmentation in a temporal fashion. Most of the state-of-the-art semantic segmentation architecture is computationally heavy. Therefore an excellent lightweight architecture with acceptable performance needs to be developed. This work concentrates on the literature survey of state-of-the-art temporal fusion techniques, comparing results from these findings, and cross-transfer of temporal fusion techniques to the segmentation task. Finally, end the thesis with the deployment of the model on a low computational android device.

3 Related Work

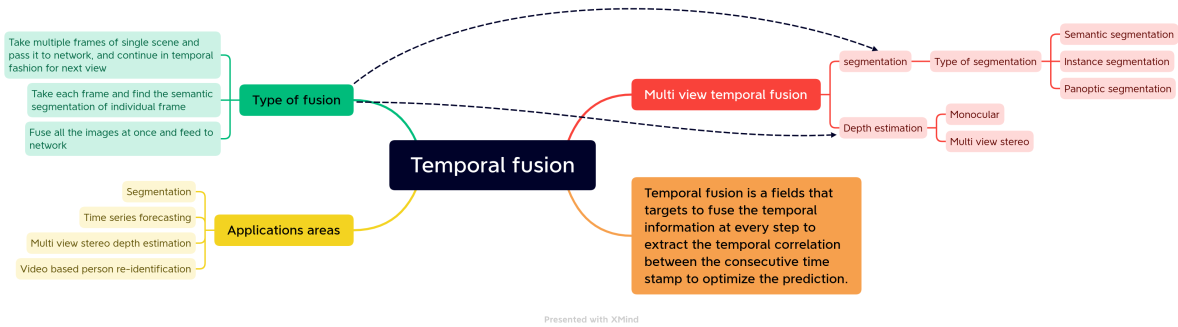

Figure 2: Temporal fusion mind map

Temporal fusion is a technique to fuse the gathered information in the time axis, making the model robust and reliable. The rich correlation between the consecutive data sources is gathered and combined in a temporal fusion fashion. Data collected over time fused in several different ways. For example, combine the collected data over time and feed it to the detection module [41]. Fusion can be done at each time stamp; we can take the last step data and pass it onto the prediction module to make the current step prediction efficient [12]. Alternate variant would be fusing at a different location of the prediction model [17], [16], [12]. In stereo vision, the stereo pair is passed onto the pair of ResNet-50, followed by many convolutional layers. The tokens are afterward reassembled in different locations with variable resolution and fused progressively [37]. Number of field take advantage of the fusion technique, such as multi-camera video surveillance [42], myoelectric interfaces to decode the Electromyogram (EMG) time-series data [21], time series forecasting [26], multi-view stereo depth estimation [12], video-based person re-identification [20], [5], segmentation [24], [19], [34].

Multi-view temporal fusion deals explicitly with the fusion of image information taken at different time stamps. Combining the data from consecutive frames makes the following prediction step robust and reliable. Fan Zhu et al. work solves the multi-view action recognition problem with a local segment of silhouettes similarity voting scheme followed by a multi-sensor fusion method. This approach gives impressive results by fusing the spatial multi-sensor inputs [46]. A multi-view summarization work by Yanwei Fu et al. consumes a large number of video files and finds the critical information quickly from these data by representing the multi-view video structure with a Spatio-temporal shot graph since it carries the information of the shots and at the same time find the correlation among the shots [13]. A real-time texture montage for dynamic multi-view reconstruction uses the dilated depth discontinuities and majority voting from Holoportation to suppress the ghosting effect while blending textures [11]. The renderer computes the view-dependent weights of each texture using the rendering camera's parameters. On the other hand, the texture weights field can change quickly if there is an abrupt change in viewpoint. Montage4D uses the concept of temporal texture weights to solve this problem, allowing texture weights to change seamlessly across time [11]. An online-based multi-view depth prediction approach on the posed video streams propagates the scene geometry information computed in the previous step onto the current step in a geometrically feasible way. A lightweight encoder-decoder architecture with the cost volume computed from the pair of images with ConvLSTM at the latent space to accommodate the past information in the current step [12]. Minghan Li et al. proposed an effective one-stage video segmentation architecture spatially calibrated with STMask temporal fusion in the pipeline. The temporal correlation between the video frames is captured with STMask to find the instance mask for the current frame from the adjacent frame. Thereby solving the motion blur and partial occlusion problem [25].

Segmentation is one of the sub-problems of the computer vision domain to classify the pixels or segments of an image into a particular class. Image segmentation is a super-set of image classification by detecting the objects and specifying the location of objects present in the image. Traditional segmentation methods used region growing and snake algorithms to compare the pixel values to get the segment map. The advent of deep learning-based architecture in computer vision pushed the accuracy and performance of the segmentation task. Image segmentation can be broadly classified into the semantic, instance, and Panoptic. Semantic segmentation is the task of classifying pixels in an image into a semantically meaningful class [27]. Instance segmentation is a method that classifies the pixel based on instances rather than classes. Panoptic segmentation is a task that can be expressed as a combination of semantic and instance segmentation. In this, each segment class is identified.

Different fields use semantic segmentation, such as unmanned aerial vehicles (UAV), Autonomous robots, medical imaging and diagnosis, and facial recognition. Classical semantic segmentation is based on the decision trees [35] or Markov random fields [43]. Ciresan et al. [7] suggested a CNN-based semantic segmentation architecture on a biological image. Long et al. developed an efficient fully convolutional network (FCN) segmentation network that generated dense prediction of images of any size much faster than any previously developed method. The success of FCN paved the way for the development of efficient segmentation architecture. In the context of medical imaging, a state-of-the-art encoder-decoder-based semantic architecture was developed named U-Net [33]. The spatial dimension is reduced on the encoder side to find what information to capture, and the same is up-sampled using the decoder to find the information. Following years a similar architecture to U-Net known as SegNet [2] was introduced. Furthermore, it does not transfer feature maps from the encoder to the decoder entirely. Instead, transfer only the max-pooling indices [4]. Yi Zhu et al. work on improving the semantic segmentation via joint image-label propagation to scale up the training set. Given a sequence of video frames with labels for only a subset of frames, the model's ability to segment exploited to predict the unlabelled frames by label propagation and joint label propagation [47]. A video segmentation technique that makes use of temporal coherence in video and reuses single-stage image segmentation using a NetWarp module is proposed by Raghudeep Gadde et al. [14]. The NetWarp network efficiently transforms the image CNNs into video CNNs. There was much work related to semantic segmentation; however, the problem of temporal fusion of camera pose with semantic segmentation is rarely addressed. NYU-Depth v2 [29] is a video sequence from various indoor scenes recorded by the RGB and depth camera. The dataset contains a densely labeled pair of RGB and depth images. The accelerometer's roll, yaw, pitch, and tilt angle are captured during each frame. The intrinsic camera parameter of the camera is provided. Annotation tool [38] can be used to create a small dataset for the evaluation purpose. Transfer learning of semantic segmentation models can also be performed to start from a baseline with a focus on real-time application. This thesis work aims to temporally incorporate the camera's pose to learn the overlapping geometric information from the consecutive image frames to improve semantic segmentation prediction efficiently.

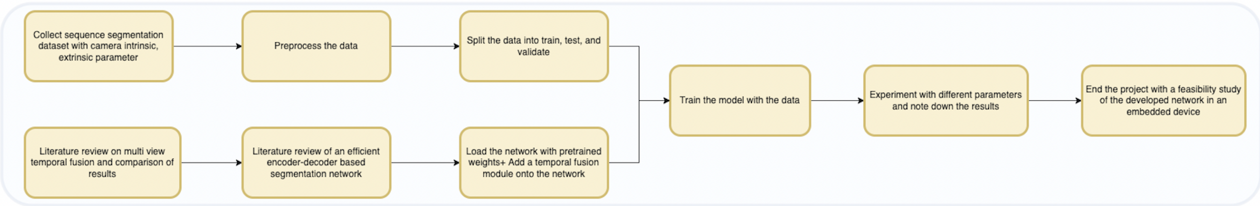

Figure 3: Pipeline of master thesis project

4 Research questions

Three research questions were defined for the master thesis.

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 temporal fusion technique to the other tasks, such as object detection or semantic segmentation?

RQ3.1 How do different loss criteria impact semantic segmentation performance?

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

5 Project Plan

The following sections explain work packages, milestones, project schedules, and deliverables.

5.1 Work Packages

The bare minimum will include the following packages:

WP1 Literature Search

This section aims to search for references to papers related to temporal fusion.

T1.1 Literature review

This task aims to collect literature on temporal fusion and conceptual understanding of the fusion of previous frame information into the current stage prediction to improve the prediction task.

WP2 Data aggregation and preprocessing

This section explains the data collection and data preprocessing.

T2.1 Data collection

This section targets to collect data from multiple sources and the nature of the data examined and analyzed using visualization tools or statistical methods. An analysis to ensure data is diverse, unbiased, and abundant in nature. Publicly available data related to the segmentation task is collected. Some of the datasets related to the segmentation task is listed below,

- Cityscapes dataset [9]
- ScanNet dataset [10]
- Stanford-2D-3D-Semantic dataset [1]
- SUN RGB-D dataset [36]
- Create own dataset by capturing the image and pose from mobile phone

Camera pose can be measured with the ARCore [23] API built inside the android phone. Some of the data contain intrinsic and extrinsic parameters of the camera.

T2.2 Data preprocessing

Preprocessing of data is carried out based on the input requirement of the model. The preprocessing step converts the raw sourced data into a format that enables successful model training—the image's input resolution is adjusted per the model requirement. The extrinsic parameters are in a roll, yaw, and pitch format, and data from a mobile phone in quaternions is transformed into a homogeneous transformation matrix.

WP3 Model implementation

This section explains the development and implementation of the model.

T3.1 Implementation of the existing model

The existing encoder-decoder-based U-Net is implemented with the cityscapes dataset. Once the model learns the weight, the temporal fusion approach is implemented with the dataset containing the intrinsic and extrinsic parameters.

T3.2 Implementation of temporal fusion for semantic segmentation

The idea of temporal fusion in the latent space for encoder-decoder-based depth estimation is studied and implemented for the segmentation task.

WP4 Evaluation

This package aims to evaluate the results based on the different metrics.

T4.1 Evaluation of the model

This task aims to compare the existing temporal fusion architecture to understand the impact of temporal fusion on the prediction. Later, the proposed temporal fusion segmentation task was evaluated with the collected data.

WP5 Android development

On an Android device, this package intends to implement the temporal fusion-based semantic segmentation technique.

T5.1 Implementation on android device
This task aims to implement semantic segmentation on an android device with the help of the chaquopy framework.

WP6 Project Report
This work package involves writing the project report. It is done in parallel with all previous work packages.

T6.1 Documentation of reviewed literature
In this task, a detailed state-of-the-art analysis is done, and all the findings are documented in the project report.

T6.2 Documentation of baseline results
In this task, the implementation result of the segmentation task without temporal fusion is achieved. It also documents the impact of different temporal fusion architectures with error metrics.

T6.3 Documentation of cross-domain application of temporal fusion approach
In this task, the result of the cross-domain application of the temporal fusion approach is performed.

5.2 Milestones

M1 Literature search - 01/07/2022 - 31/07/2022

M2 Data collection and preprocessing - 10/07/2022 - 31/07/2022

M3 Baseline, Comparison of results, and Segmentation - 01/08/2022 - 30/09/2022

M4 Evaluation - 01/10/2022 - 31/10/2022

M5 Android development - 01/11/2022 - 30/11/2022

M6 Project report - 01/12/2022 - 31/12/2022

5.3 Project Schedule

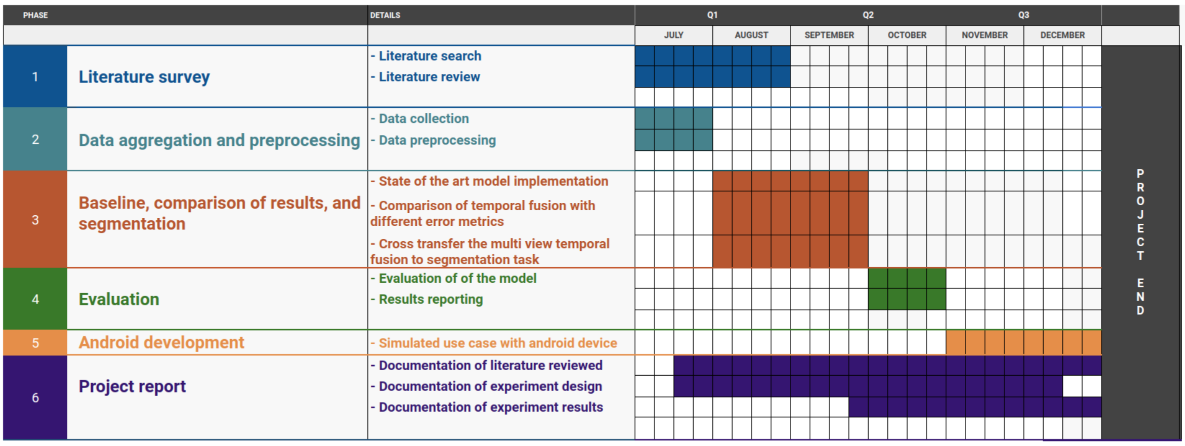

Figure 4: Timeline of the project

5.4 Deliverables

Minimum viable

- 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

Expected

- Compare performances of state-of-the-art temporal fusion techniques with different error metrics
- Cross transfer the temporal fusion architecture to the segmentation task
- Simple simulated use case of temporal fusion on the android device

Maximum

- Evaluation of the temporal segmentation method with different loss criteria
- Improved multi-view temporal fusion technique for segmentation task
- Performance of multi-view temporal fusion with different Gaussian kernels

References

[1] I. Armeni, A. Sax, A. R. Zamir, and S. Savarese. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*, February 2017.

[2] 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.

[3] V Bindhu. Biomedical image analysis using semantic segmentation. *Journal of Innovative Image Processing (JIIP)*, 1(02):91–101, 2019.

[4] Albert Bou. Deep learning models for semantic segmentation of mammography screenings, 2019.

[5] Lin Chen, Hua Yang, Ji Zhu, Qin Zhou, Shuang Wu, and Zhiyong Gao. Deep spatial-temporal fusion network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 63–70, 2017.

[6] Dan Ciresan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in neural information processing systems*, 25, 2012.

[7] Dan Ciresan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in neural information processing systems*, 25, 2012.

[8] Guy Barrett Coleman and Harry C Andrews. Image segmentation by clustering. *Proceedings of the IEEE*, 67(5):773–785, 1979.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPR Workshop on the Future of Datasets in Vision*, volume 2. sn, 2015.

[10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[11] Ruofei Du, Ming Chuang, Wayne Chang, Hugues Hoppe, and Amitabh Varshney. Montage4d: Interactive seamless fusion of multiview video textures. 2018.

[12] 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.

[13] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou. Multi-view video summarization. *IEEE Transactions on Multimedia*, 12(7):717–729, 2010.

[14] 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.

[15] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European conference on computer vision*, pages 297–312. Springer, 2014.

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

[17] Yuxin Hou, Muhammad Kamran Janjua, Juho Kannala, and Arno Solin. Movement-induced priors for deep stereo. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3628–3635. IEEE, 2021.

[18] 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.

[19] 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.

[20] Xinyang Jiang, Yifei Gong, Xiaowei Guo, Qize Yang, Feiyue Huang, Wei-Shi Zheng, Feng Zheng, and Xing Sun. Rethinking temporal fusion for video-based person re-identification on semantic and time aspect. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11133–11140, 2020.

[21] Rami N Khushaba, Erik Scheme, Ali H Al-Timemy, Angkoon Phinyomark, Ahmed Al-Taee, and Adel Al-Jumaily. A long short-term recurrent spatial-temporal fusion for myoelectric pattern recognition. *Expert Systems with Applications*, 178:114977, 2021.

[22] Tae-young Ko and Seung-ho Lee. Novel method of semantic segmentation applicable to augmented reality. *Sensors*, 20(6):1737, 2020.

[23] Micheal Lanham. *Learn ARCore-Fundamentals of Google ARCore: Learn to build augmented reality apps for Android, Unity, and the web with Google ARCore 1.0*. Packt Publishing Ltd, 2018.

[24] 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.

[25] 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.

[26] Bryan Lim, Sercan O Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *arXiv preprint arXiv:1912.09363*, 2019.

[27] Xiaolong Liu, Zhidong Deng, and Yuhan Yang. Recent progress in semantic image segmentation. *Artificial Intelligence Review*, 52(2):1089–1106, 2019.

[28] Yujian Mo, Yan Wu, Xinneng Yang, Feilin Liu, and Yujun Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493:626–646, 2022.

[29] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[30] Feng Ning, Damien Delhomme, Yann LeCun, Fabio Piano, Léon Bottou, and Paolo Emilio Barbano. Toward automatic phenotyping of developing embryos from videos. *IEEE Transactions on Image Processing*, 14(9):1360–1371, 2005.

[31] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015.

[32] Nils Plath, Marc Toussaint, and Shinichi Nakajima. Multi-class image segmentation using conditional random fields and global classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 817–824, 2009.

[33] 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.

[34] Muhammed Shafay, Taimur Hassan, Ernesto Damiani, and Naoufel Werghi. Temporal fusion based mutli-scale semantic segmentation for detecting concealed baggage threats. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 232–237. IEEE, 2021.

[35] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[36] 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*, pages 567–576, 2015.

[37] Qing Su and Shihao Ji. Chitransformer: Towards reliable stereo from cues. *arXiv preprint arXiv:2203.04554*, 2022.

[38] Markus Suchi, Timothy Patten, David Fischinger, and Markus Vincze. Easylabel: A semi-automatic pixel-wise object annotation tool for creating robotic rgb-d datasets. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6678–6684. IEEE, 2019.

[39] Michael Treml, José Arjona-Medina, Thomas Unterthiner, Rupesh Durgesh, Felix Friedmann, Peter Schuberth, Andreas Mayr, Martin Heusel, Markus Hofmarcher, Michael Widrich, et al. Speeding up semantic segmentation for autonomous driving. 2016.

[40] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 157–166, 2014.

[41] Shuo Wang, Dan Guo, Wen-gang Zhou, Zheng-Jun Zha, and Meng Wang. Connectionist temporal fusion for sign language translation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1483–1491, 2018.

[42] 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.

[43] Jianxiong Xiao and Long Quan. Multiple view semantic segmentation for street view images. In *2009 IEEE 12th international conference on computer vision*, pages 686–693. IEEE, 2009.

[44] Wei Yang, Ping Luo, and Liang Lin. Clothing co-parsing by joint image segmentation and labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3182–3189, 2014.

[45] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Light-field image super-resolution using convolutional neural network. *IEEE Signal Processing Letters*, 24(6):848–852, 2017.

[46] Fan Zhu, Ling Shao, and Mingxiu Lin. Multi-view action recognition using local similarity random forests and sensor fusion. *Pattern recognition letters*, 34(1):20–24, 2013.

[47] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8865, 2019.