Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

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
MSc. Deebul Sivarajan Nair

July 2022

1 Introduction

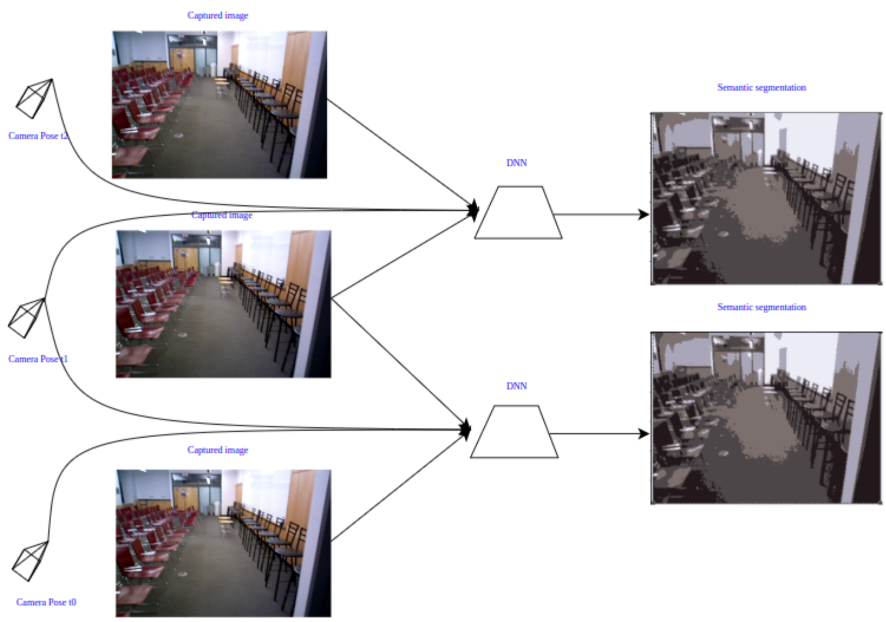

Figure 1: Semantic segmentation from pair of images and pose

Prediction of variables of interest at different future timestamps with integration of information at each step is a common temporal fusion problem. Fusing of information at every step helps to optimize the target prediction in the future. Fusion of information in temporal fashion is done in both classical as well as in deep learning architecture. A image-pose pair encoder decoder based architecture for disparity estimation fuse information in the latent space successively with the help of Gaussian process at the latent space [13]. A end-to-end video based person re-identification network fuse feature in spatial-temporal fashion. In this work both the spatial and temporal feature of input image sequence are calculated by collecting the feature vector $f^{(t)}$ from the convolutional neural network (CNN) connects the temporal pooling layer and aggregated feature are formed into a single feature vector that is averaged over the entire sequence [4]. Information can be fused at the single stage output of multi stage encoder decoder network. An online multi-view feature prediction network on the posed video frame, with fusing richer scene geometry feature calculated at the last step is fused into the current step in a efficient manner. Geometric information from the output of previous encoder decoder network is fused to compute a cost volume and feed it as input to the second stage encoder decoder architecture, also output of the second stage network is fed back to the convolutional long short-term memory (ConvLSTM) network in the latent space by warping [9].

Computer vision aims to understand the surrounding environment using various mathematical modeling techniques. And semantic segmentation is the key problem in the computer vision that targets to complete scene understanding. The semantic segmentation segments the input images based on the semantic information and categorizes each pixel to a class from the list of labels [24]. Semantic segmentation can be applied to 2D images, volumetric data or video. Number of field take advantage of understanding the surrounding environment such as human-machine interaction [27], computational photography [40], image search engines [35], augmented reality [19], autonomous driving [34], biomedical imaging [2], fashion [39]. Traditional semantic segmentation is performed with conditional random fields [28], clustering [7]. With the development of deep learning architecture, semantic segmentation problem are solved with CNNs [26], [5], [12]. A semantic segmentation network can be thought of as an encoder decoder network, usually with pre-trained classification network in the encoder side and semantically projects the learnt discriminative feature onto the feature space to get dense classification.

A general framework of semantic segmentation involves applying a deep image segmentation model onto every image in a sequence independently. However, these approaches result in information loss by not considering the correlation among the consecutive frames. To tackle this problem information from the previous step is fused at the current step in a temporal manner resulting in an efficient semantic segmentation. Estimation of semantic segmentation from the unconstrained monocular camera images is a challenging task. State of the art temporal semantic segmentation is based on deep learning. A temporally distributed network designed for semantic segmentation finds the correlation between the frames and the attention propagation module effectively combining the distributed feature groups [15]. Efficient semantic segmentation can be performed with fusion of current and previous camera pose information onto the latent space in an encoder decoder based architecture.

The goal of this thesis is to study the state of the art temporal fusion techniques and compare the results from these findings. And 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 to fuse 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, one such field is the segmentation task. Semantic segmentation aims to label each pixel in an image and it is an ill-posed problem. Over the past years, a large number of algorithms and architectures have been proposed to find the semantic segmentation of the scenes. Conventional semantic segmentation algorithms compute semantic maps for the individual frame and do not take into account the geometric correlation between the consecutive frames, thereby losing important overlapping information in the successive images. Images can be obtained from a stereo camera or a monocular camera. Estimation of semantic segmentation from the unconstrained monocular camera image is a challenging task, however there are potential benefits from using monocular camera images. Firstly, reduced cost, and flexibility in the movement of the monocular camera, thereby gathering rich information of the surrounding environment. Secondly, multiple varying point images are able to 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 a reasonable lightweight architecture with fair performance needs to be developed. This work concentrates on the literature survey of state of the art temporal fusion techniques and comparison of results from these findings, and cross transfer of temporal fusion techniques to the segmentation task. Finally end the thesis with deployment of the model on a low computational android device.

3 Related Work

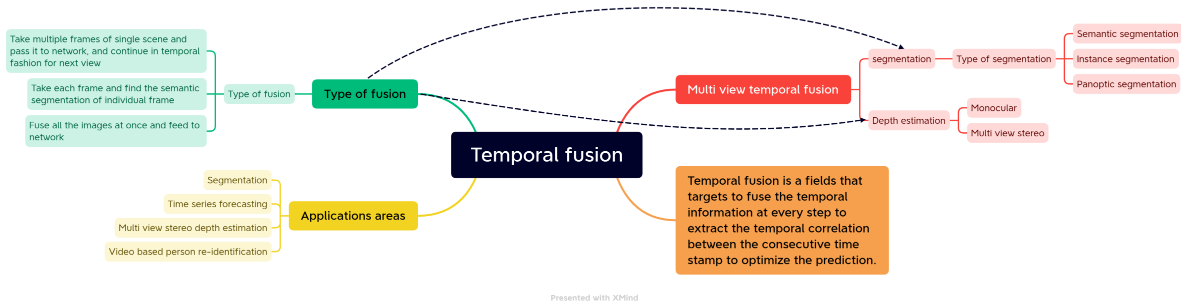

Figure 2: Multi view stereo mind map

Temporal fusion is a technique to fuse the gathered information in the time axis, thereby making the model robust and reliable. The rich correlation between the consecutive data sources is gathered and combined in temporal fusion fashion. Data collected over time can be fused in a number of different ways. For example, combine the collected data over time and feed it to the detection module [36], or at each time stamp we can take the previous step data and pass it onto the prediction module to make the current step prediction efficient [9], another variant would be fusing at different location of the prediction model [14], [13], [9]. In stereo vision, the stereo pair is passed onto the pair of ResNet-50 followed by a bunch of convolutional layers. The tokens are afterwards reassembled in different locations with variable resolution and fused progressively [32]. Number of field take advantage of the fusion technique, such as multi camera video surveillance [37], myoelectric interfaces to decode the Electromyogram (EMG) time series data [18], time series forecasting [22], multi view stereo depth estimation [9], video based person re-identification [17], [4], segmentation [20], [16], [30].

Multi view temporal fusion specifically deals with the fusion of image information taken at different time stamps. Combining the data from consecutive frames makes the next prediction step robust and reliable. Work by Fan Zhu et al solves the multi-view action recognition problem with local segment of silhouettes similarity voting scheme followed by multi-sensor fusion method. This approach gives impressive results by fusing the spatial multi sensor inputs [41]. A multi view summarization work by Yanwei Fu et al consumes large number of video files and find the important information quickly from these data by representing the multi-view video structure with spatio-temporal shot graph since it carries the information of the shots and at the same time find the correlation among the shots [10]. A real time texture montage for dynamic multi-view reconstruction takes advantage of the dilated depth discontinuities and majority voting from Holoportation to suppress the ghosting effect while blending textures. Sudden change in the viewpoint results in rapid change of texture weights fields, a temporal texture weights is deployed to accommodate the smooth transition of texture weights [8]. A online based multi-view depth prediction approach on the posed video streams which propagate the scene geometry information computed in the previous step onto the current step in a geometrically feasible way. A light weight 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 [9]. 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 [21].

Segmentation is the one of the sub-problems of the computer vision domain to classify the pixels or segments of an image to a particular class. Image segmentation is a super-set of image classification by not only detecting the objects but specifying the location of objects present in the image. Traditional segmentation methods made use of region growing and snake algorithms to compare the pixel values to get the segment map. With 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 three groups namely semantic segmentation, Instance segmentation, and Panoptic segmentation. Semantic segmentation is a task of classification of pixels in an image into a semantically meaningful class [23]. Instance segmentation is a method that classifies the pixel on the basis of instance rather than classes. Panoptic segmentation is a task that can be expressed as the combination of semantic and instance segmentation. In this each segment class is identified.

Different fields use semantic segmentation such as unmanned aerial vehicle

(UAV), Autonomous robots, medical imaging and diagnosis, facial recognition. Classical semantic segmentation is based on the decision trees [31] or markov random fields [38]. Ciresan et. al [6] 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 that was much faster than any previously developed method. The success of FCN paved the way for development of efficient segmentation architecture. In the context of medical imaging a state of the art encoder decoder based semantic architecture was developed named as U-Net [29]. The spatial dimension is reduced on the encoder side to find what information to capture and the same is up-sampled using decoder to find the where information. Following years a similar architecture to U-Net known as SegNet [1] was introduced. And it does not entirely transfer feature maps from encoder to the decoder, rather transfer only the max pooling indices [3]. Work by Yi Zhu et al 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 models ability to segment is exploited to predict the unlabelled frames by label propagation and joint label propagation [42]. 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 [11]. The filter representation of the present image is changed by the respective representation in the previous frame by NetWarp. There was a lot of work related to the semantic segmentation, however the problem of temporal fusion of camera pose with semantic segmentation is rarely addressed. NYU-Depth v2 [25] is a video sequence from a variety of indoor scenes recorded by the RGB and depth camera. The dataset contains a densely labeled pair of RGB and depth images. The roll, yaw, pitch and tilt angle of the accelerometer device is captured during each frame. Intrinsic camera parameter of camera also provided. Annotation tool [33] 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 focus on real time application. This thesis work aims to incorporate the pose of the camera in a temporal fashion to learn the overlapping geometric information from the consecutive image frames to efficiently improve the semantic segmentation prediction.

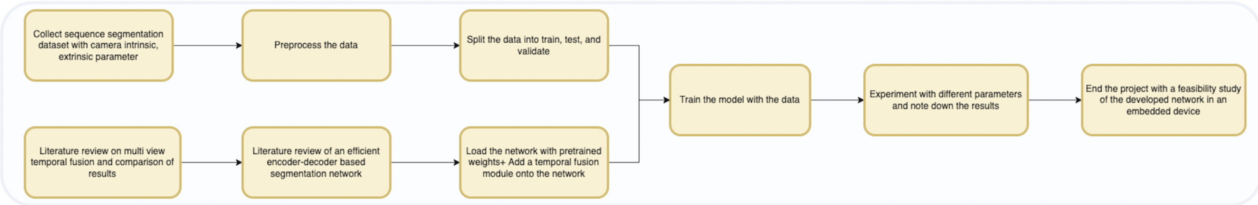

Figure 3: Pipeline of master thesis project

4 Research questions

Three research questions are 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 nonparametric fusion to the other tasks, such as object detection or semantic segmentation?

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

RQ3.2 What is the performance of semantic segmentation with respect to different Gaussian kernels?

5 Project Plan

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

5.1 Work Packages

The bare minimum will include the following packages:

WP1 Literature Search

This section aims to extensively search for references to papers that are related to multi-view stereo.

T1.1 Literature review

In this task collection of literature related to multi-view stereo is done and conceptual understanding of the 3D geometry from images.

WP2 Data aggregation and preprocessing

This section explains the data collection and data preprocessing.

T2.1 Data collection

In this section, data is collected from multiple sources, and the nature of the data is examined and analyzed using visualization tools or statistical methods. An analysis is carried out to ensure data is diverse, unbiased, and abundant in nature.

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 training of the model.

WP3 Model implementation

This section explains the development and implementation of the model.

T3.1 Evaluation of the model

This task aims to reproduce the Multi-view Stereo by Temporal Nonparametric Fusion architecture results.

T3.2 Cross application of the temporal fusion to the segmentation
In this section, the extension of temporal fusion architecture to the other application areas of computer vision is carried out.

WP4 Evaluation
This package aims to evaluate the results based on the different metrics.

T4.1 Results reporting
In this task, the output of the evaluation is reported.

WP5 Project Report
This work package involves writing the project report. It is done in parallel with all previous work packages.

T5.1 Documentation of reviewed literature
In this task, a detailed analysis of the state of the art is done and all the findings are documented in the project report.

T5.2 Documentation of baseline results
In this task, the implementation result of Temporal Fusion baseline is done.

T5.3 Documentation on the results for the different temporal fusion This task documents the result of different temporal fusion architecture with different error metric is found.

T5.4 Documentation of cross-domain application of temporal fusion approach
In this task, result of the cross-domain application of the temporal fusion approach is performed.

5.2 Milestones

M1 Literature search

M2 Data collection and preprocessing

M3 Building a baseline

M4 Experimental Analysis

M5 Development

M6 Report submission

5.3 Project Schedule

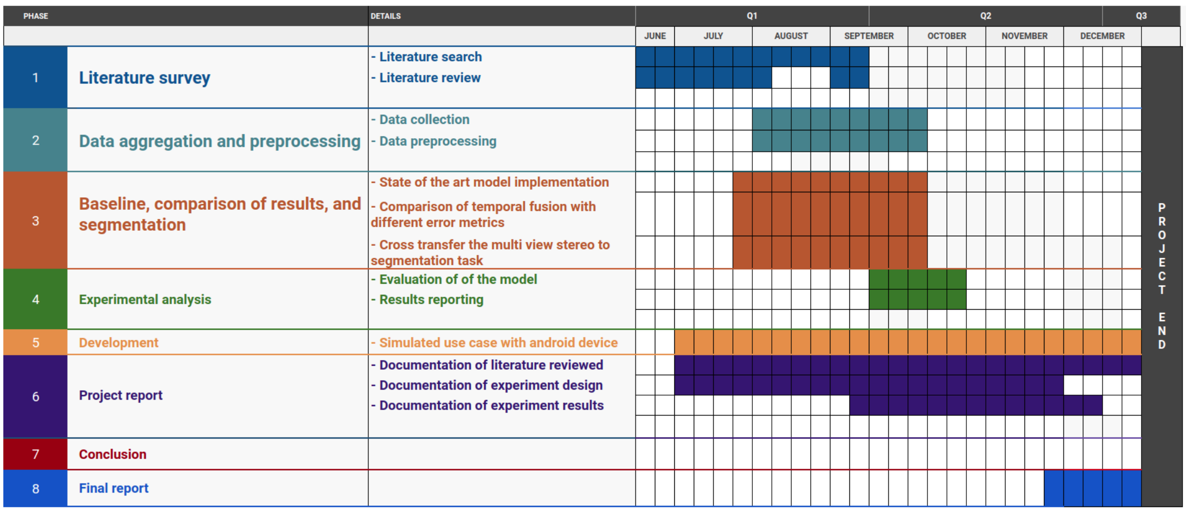

Figure 4: Timeline of the project

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 images from monocular camera

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 monocular image multi view temporal fusion technique
- Performance of multi view temporal fusion with respect to different Gaussian kernels

References

[1] 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.

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

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

[4] 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.

[5] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

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

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.