Autonomous Vehicles need a complete understanding of their surroundings to a pixel perfect level. So image segmentation can be used to detect lanes and other objects' information of the surrounding .



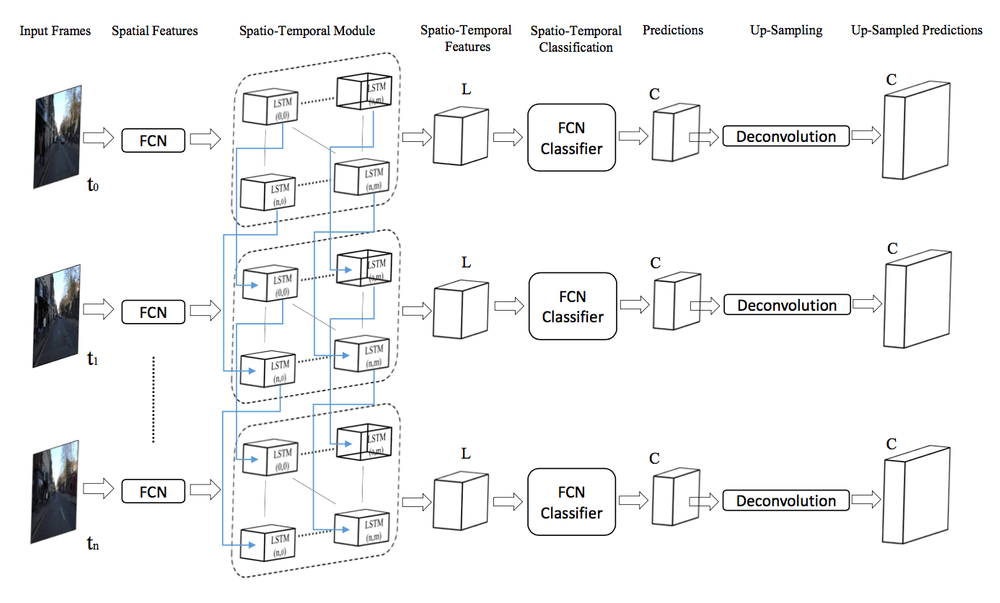
Source:- <https://medium.com/intro-to-artificial-intelligence/semantic-segmentation-udaitys-self-driving-car-engineer-nanodegree-c01eb6eaf9d>

For use cases like autonomous vehicles there is a need for real-time segmentation which comprises Faster prediction low latency response and with higher accuracy . There are a lot of overlapping scenes which can be used for faster processing on video . architectures which are specifically designed for videos:

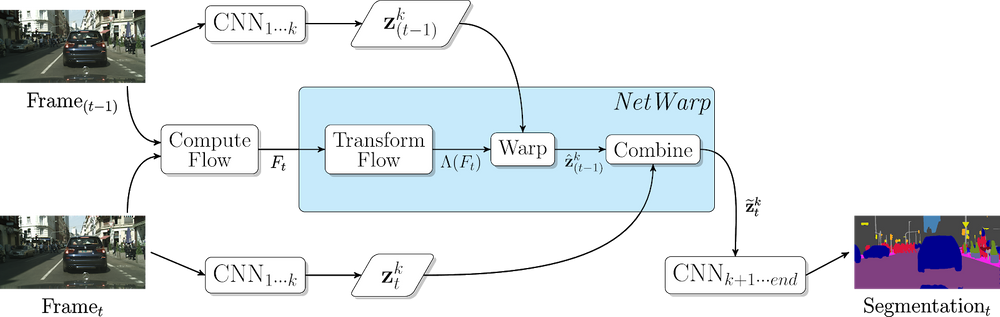
STFCN :

Spatio-Temporal FCN uses FCN and LSTM to do video segmentation . LSTM neural networks can capture sequential information over time. STFCN uses FCN with LSTM to capture both the spatial information and temporal information. Thus, the model is easily trained and memory efficient . Convolution architectures and spatial transformers support for video semantic segmentation.

Source:- <https://arxiv.org/abs/1608.05971>



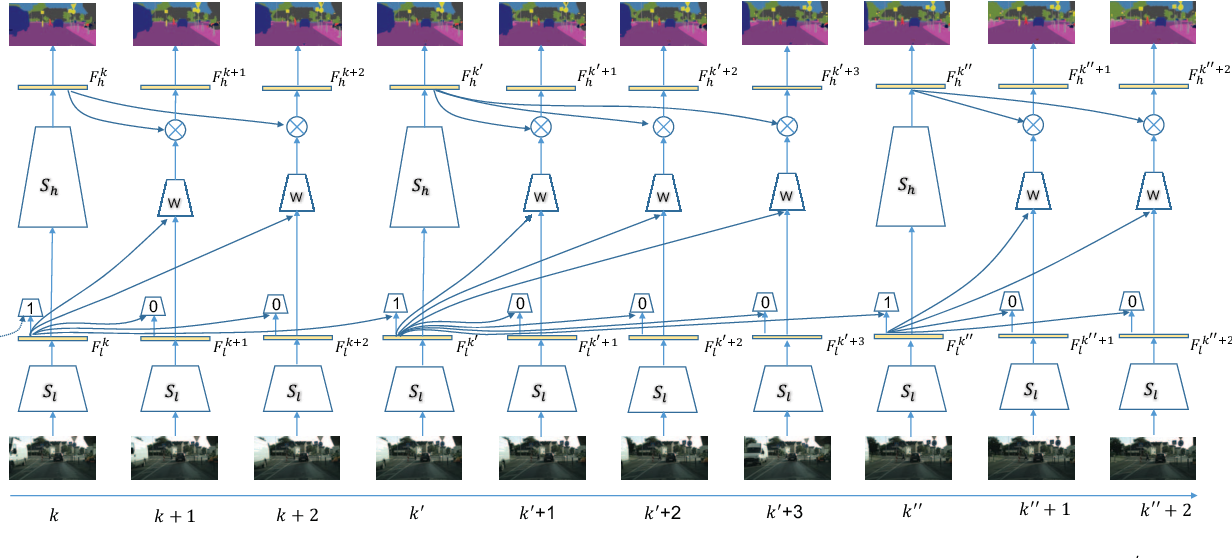
### **Semantic Video CNNs through Representation Warping**



Source:- <https://arxiv.org/abs/1708.03088>

This shows the proposed video CNN with NetWarp modules. This uses NetWarp modules in three different layers of an image CNN. The video CNN is applied in an online fashion which only looks one frame back. The CNN filter activations for the current frame are modified by the previous frame via NetWarap Modules.

### **Low-Latency Video Semantic Segmentation**



Source:- https://arxiv.org/abs/1804.00389

This paper introduced two neural network part low level features and high level features The cost of computing low level features in a network is much less compared to higher features. It uses the low level network features as an indicator for change in segmentation map. there is a strong correlation between low level features change and the segmentation map change. So to understand if there is a need to compute if the higher features are needed to be calculated, the lower features difference across 2 frames is found and is compared if it crosses a particular threshold. Since the network decision is based on the input frames the decision taken is dynamic compared to the above approach.