I. Introduction

Task:
Semantic segmentation on video streams.

Approaches:
- Most methods suffer from either High Latency or Decreased Accuracy.
- Our methods achieve both Low Latency and High Accuracy.

Contribution & Novelty:
- Temporally distributed network architecture
- Attention Propagation Mechanism
- Grouped knowledge distillation
- SOTA accuracy and efficiency

II. Motivation


Distributed computation:
Sub-feature groups are extracted independently.

Application to video:
Run only one subnetwork and reuse previous feature.

III. Temporally Distributed Networks

- Challenge:
Pixelwise tasks are sensitive to the spatial misalignment caused by motion between frames.

- Attention Propagation:
\[ V'_i = \varphi \left( \text{Softmax} \left( \frac{Qk_{i-1}}{\sqrt{d_h}} \right) V_{i-1} \right) + V_i \]

- Attention Downsampling:
Saving computation by downsample the reference data in attention.

IV. Grouped Knowledge Distillation

Distilling knowledge at feature group level:
\[ \text{Loss} = CE(\pi_S(V'_i, g)) + \alpha \cdot KL(\pi_S(V'_i)||\pi_T(\sum f)) + \beta \cdot KL(\pi_S(V_i)||\pi_T(f_i)) \]

V. Experimental Results

- Comparison with state-of-the-art
Citiescapes

- Effect of knowledge distillation

- Different temporal aggregation

- Contribution of sub-paths

- Visualization of attention

- Qualitative results