Surveillance Video Parsing with Single Frame Supervision

Si Liu¹, Changhu Wang², Ruihe Qian¹, Han Yu¹, Renda Bao¹, Yao Sun¹
¹ Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
² Toutiao AI Lab

Introduction:

- We develop a Single frame Video Parsing (SVP) method which requires only one labeled frame per video in training stage to parse one Surveillance video.

Function:

- Segment the video frames into several labels, e.g., face, pants, left-leg.

Train & Test:

- During training, only a single frame per video is labeled (red box).
- During testing, a parsing window is slided along the video. The parsing result (orange box) is determined by itself, the long-range frame (green box) and the short-range frame (blue box).

Network:

- Feature learning, correspondence mining and the temporal fusion are updated in a unified optimization process and collaboratively contribute to the parsing results.

Approach:

- Contribution:
  - Single Frame Supervision: the first attempt to segment the human parts in the surveillance video by labeling single frame per training video.
  - Good performance: the feature learning, pixelwise classification, correspondence mining and the temporal fusion are updated in a unified optimization process and collaboratively contribute to the parsing results.
- Applicable: the proposed SVP framework is end-to-end and thus very applicable for real usage.

Frame Parsing Sub-network:

- Video V = {I₁, ..., Iₜ}. The single labeled frame is Iₛ. The frame parsing sub-network produces the rough label maps for the triplet, donated as \( \{ P_{l-}, P_{c-}, P_{r-} \} \).

Optical Flow Estimation Sub-network:

- Estimate the dense correspondence between adjacent frames on the fly.
  \( F_{l-} = o(I₁, I₋₁) \)
- Temporal Fusion Sub-network:
  - Apply the obtained optical flow \( F_{l-} \) and \( P_{l-} \) to \( P_{c-} \) and \( P_{r-} \), producing \( P_{l-} \) and \( P_{r-} \) to \( P_{c-} \) and \( P_{r-} \). To alleviate the influence of imperfect optical flow, the pixel-wise flow confidence \( C_{l-} \) and \( C_{r-} \) are estimated.

Experiment:

- The test image, the groundtruth label, results of the EM-Adapt and SVP are shown sequentially.