MotoSum: A Video Summarization Experiment

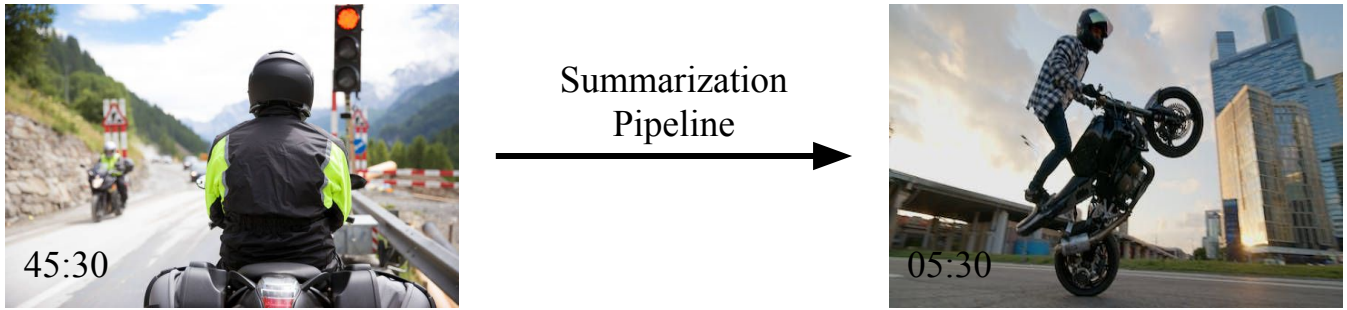
Student
Luis Vitor Zerkowski

Supervisors
Prof. Dr. Luc Van Gool
Prof. Dr. Flavio Soares

Co-supervisor
Dr. Simon Hecker

Proposal

- The modern world and its massive data production
- Summarization allows more objective information consumption
- Motorcycle rides domain
- *MotoSum*: Three ways of summarizing videos



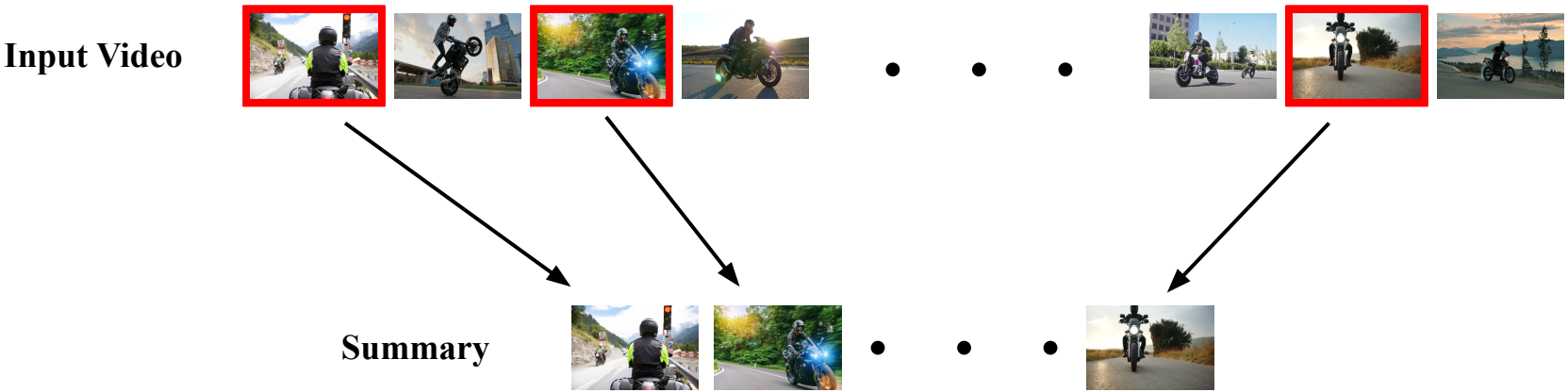
Input Video **Summary**

Evaluation Methods

- Analyzing an intrinsically subjective task and comparing models
- Ground-truth comparison
- Agent-based evaluation
- Independent evaluation
  - Diversity
  - Representativeness
  - Image Quality Assessment

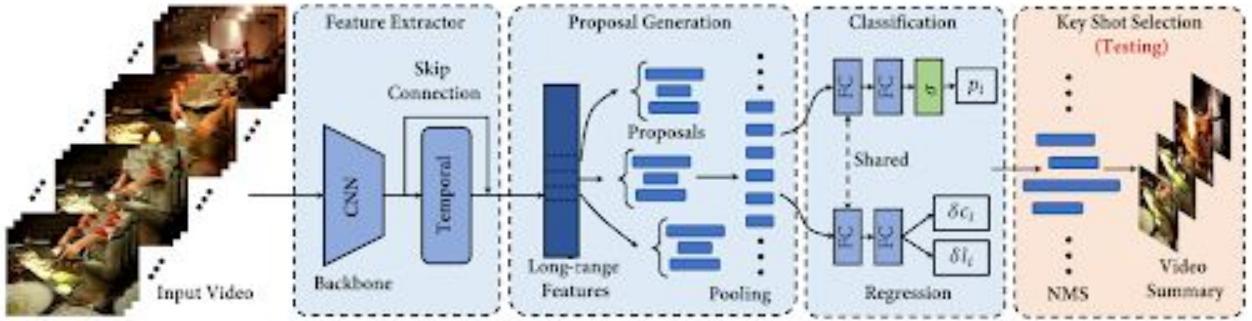
A Reference Model

- The most basic summarization: uniform sampling
- Select keyshots, not keyframes
- Split the video in 100 segments and choose 15 of them



Pretrained DSNet

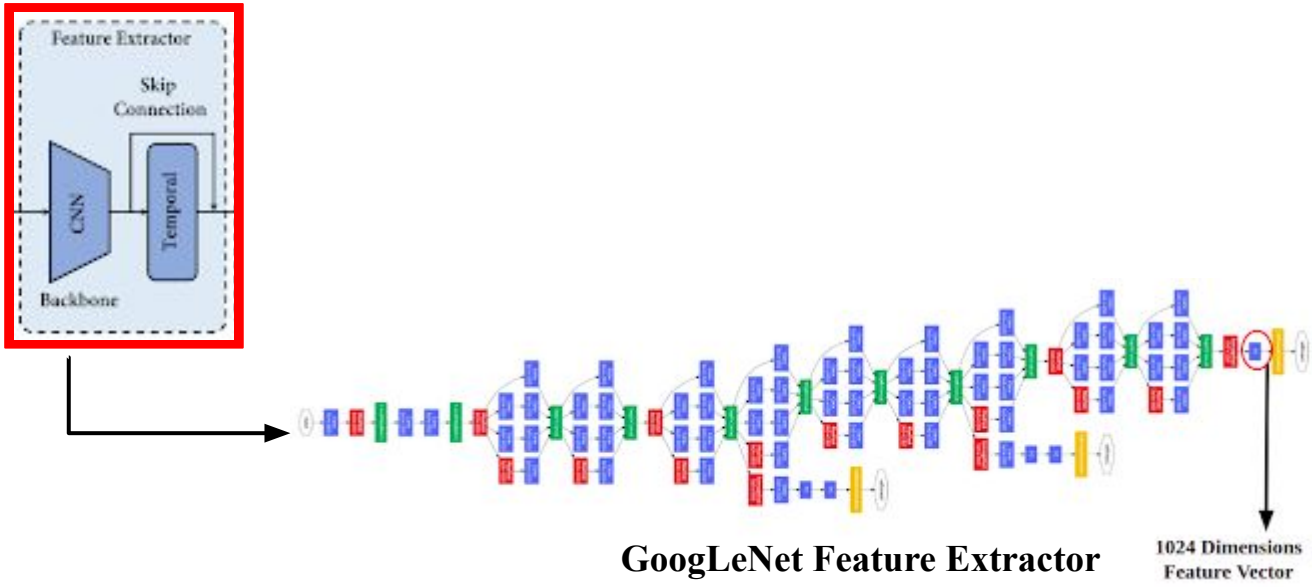
- DSNet: supervised learning neural network
- TVSum dataset: 50 videos, 10 categories, 20 annotations each



Network Architecture

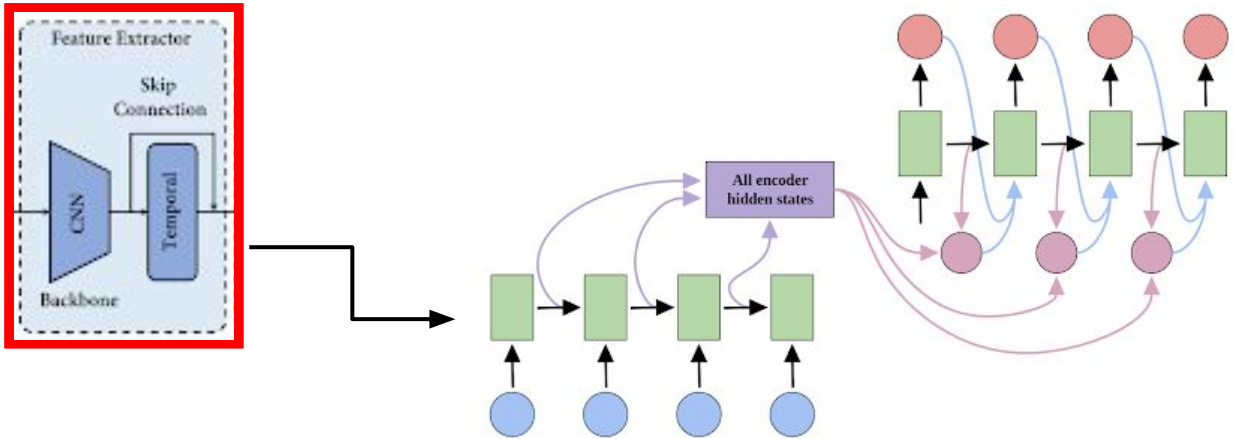
Pretrained DSNet

- DSNet: supervised learning neural network
- TVSum dataset: 50 videos, 10 categories, 20 annotations each



Pretrained DSNet

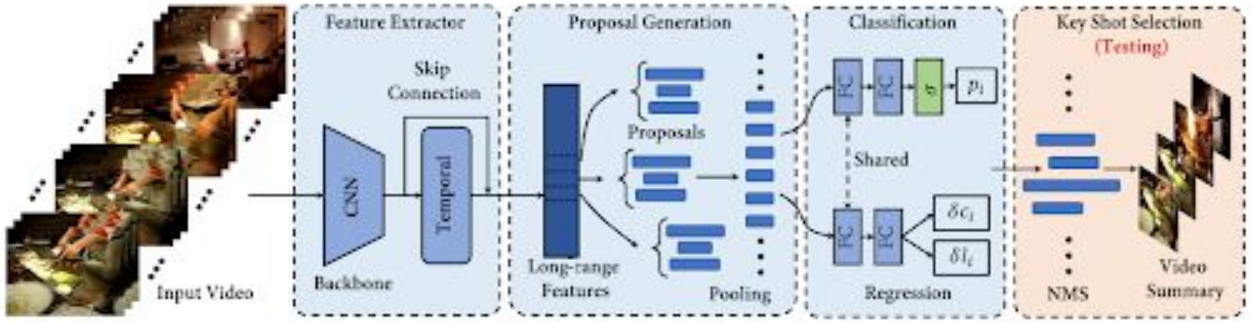
- DSNet: supervised learning neural network
- TVSum dataset: 50 videos, 10 categories, 20 annotations each



Attention Mechanism

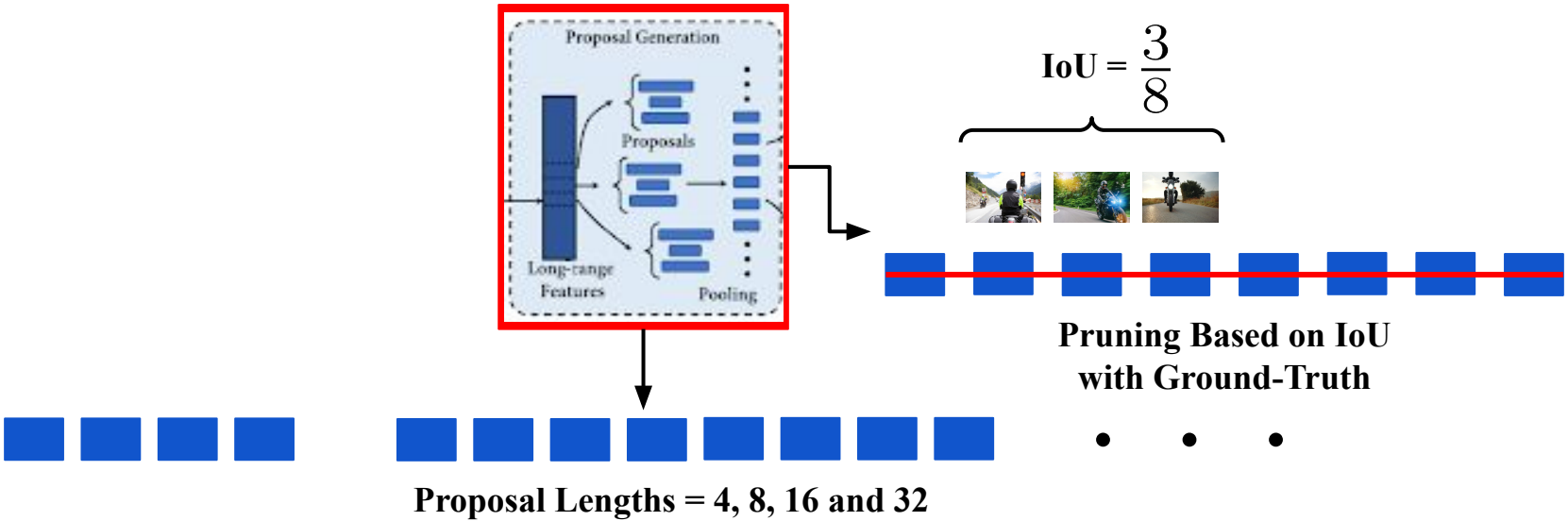
Pretrained DSNet

- DSNet: supervised learning neural network
- TVSum dataset: 50 videos, 10 categories, 20 annotations each



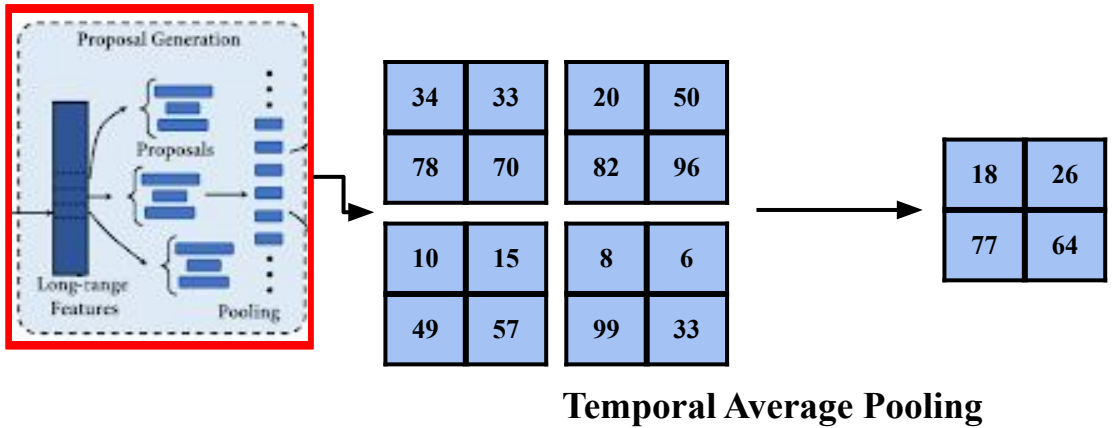
Pretrained DSNet

- DSNet: supervised learning neural network
- TVSum dataset: 50 videos, 10 categories, 20 annotations each



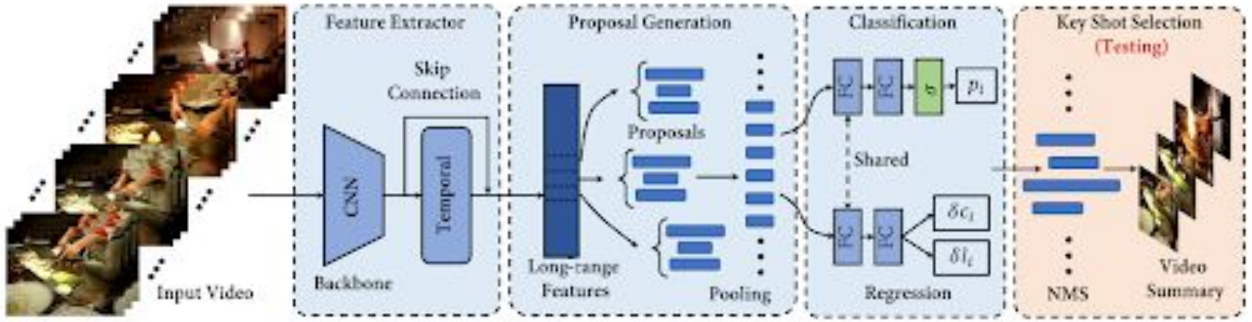
Pretrained DSNet

- DSNet: supervised learning neural network
- TVSum dataset: 50 videos, 10 categories, 20 annotations each



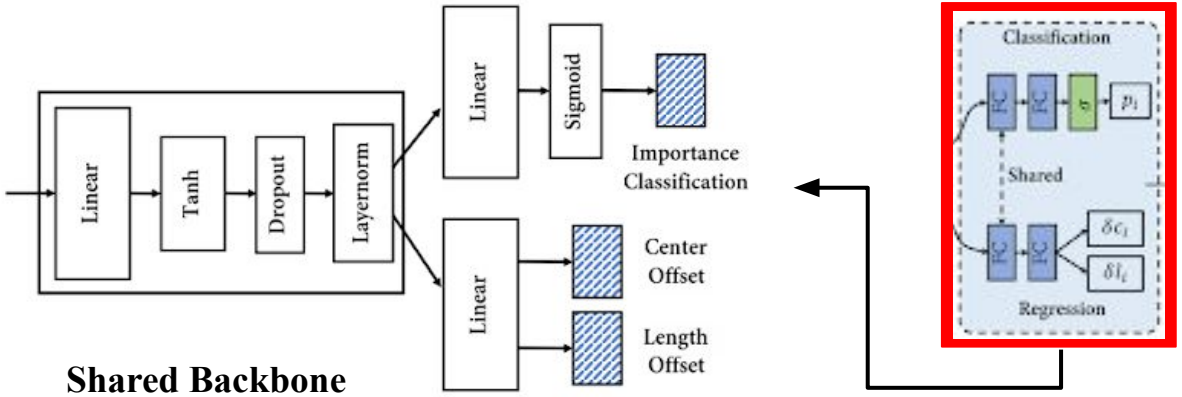
Pretrained DSNet

- DSNet: supervised learning neural network
- TVSum dataset: 50 videos, 10 categories, 20 annotations each



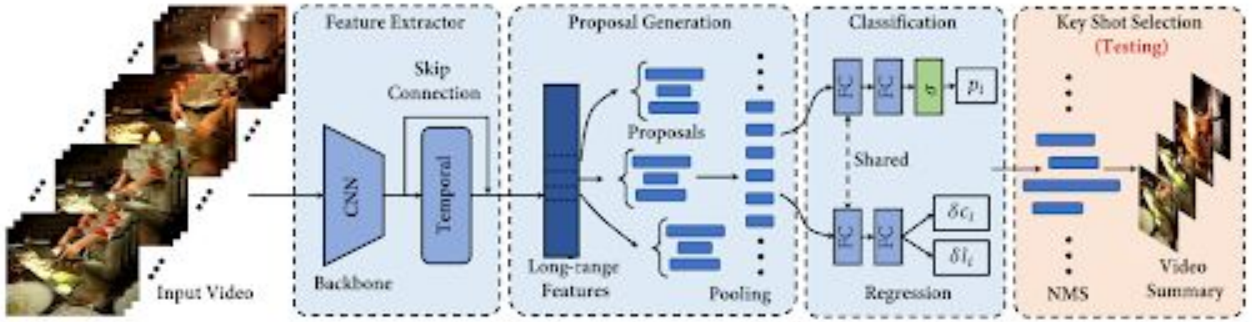
Pretrained DSNet

- DSNet: supervised learning neural network
- TVSum dataset: 50 videos, 10 categories, 20 annotations each



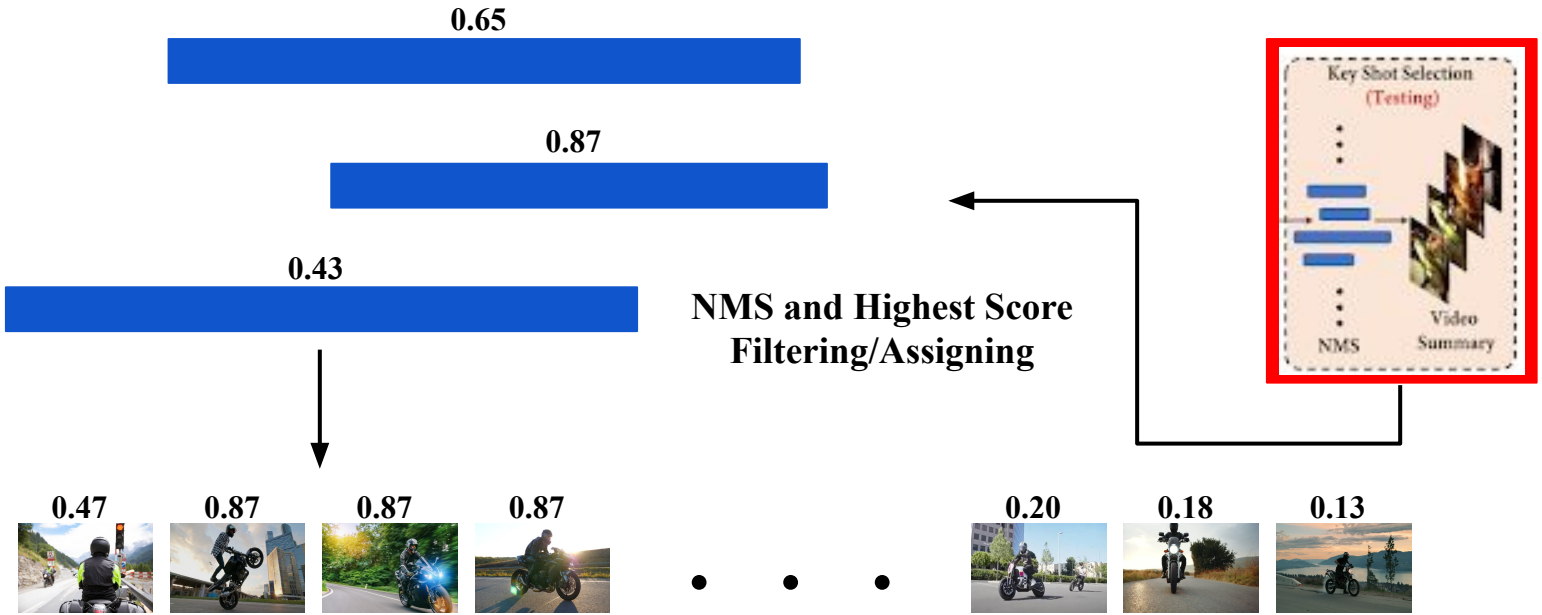
Pretrained DSNet

- DSNet: supervised learning neural network
- TVSum dataset: 50 videos, 10 categories, 20 annotations each



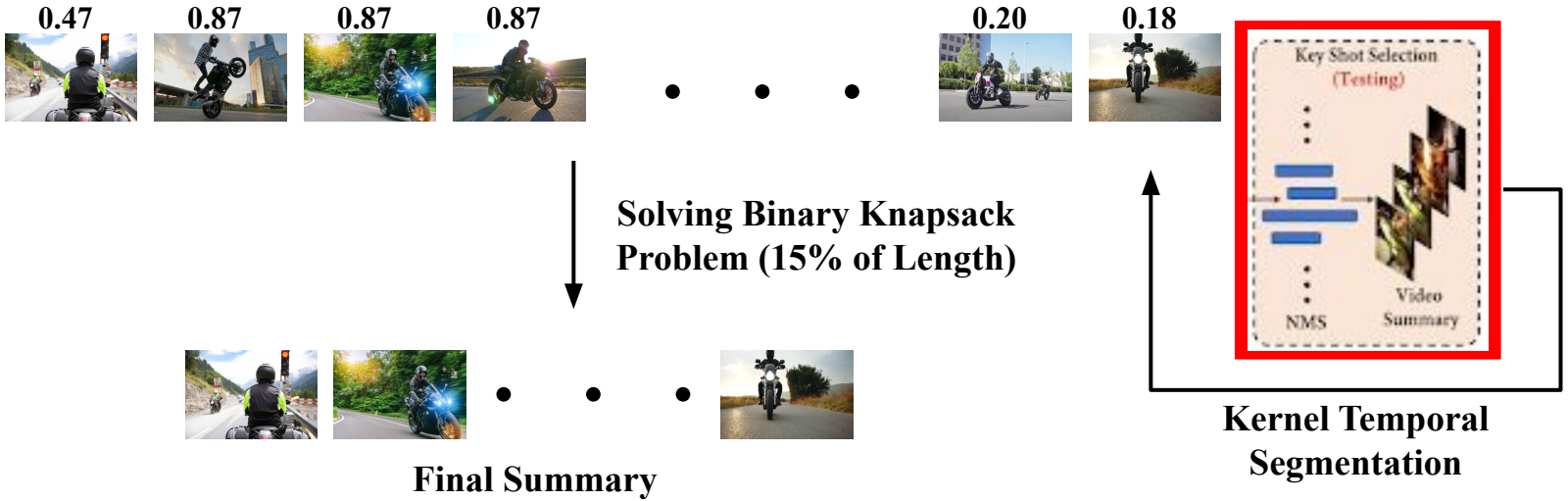
Pretrained DSNet

- DSNet: supervised learning neural network
- TVSum dataset: 50 videos, 10 categories, 20 annotations each



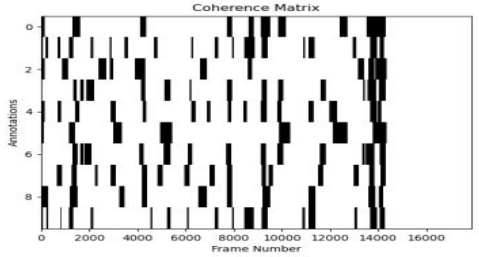
Pretrained DSNet

- DSNet: supervised learning neural network
- TVSum dataset: 50 videos, 10 categories, 20 annotations each

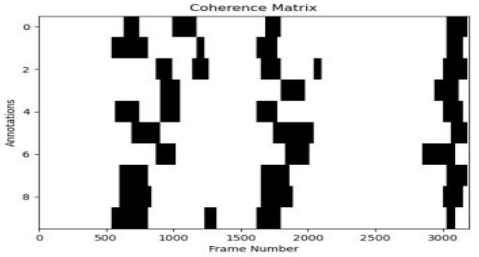


Retrained DSNet

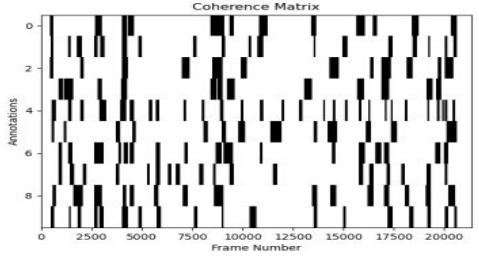
- The *MotoSum* dataset: 34 videos, 1 category, 10 annotations each



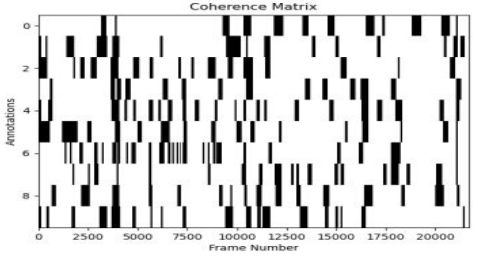
(a) *Coherence matrix plot for the annotations from video 26. The f-score and Cronbach alpha values are $\bar{F} = 0.37$ and $\alpha = 0.78$ respectively, above the MotoSum average.*



(b) *Coherence matrix plot for the annotations from video 28. The f-score and Cronbach alpha values are $\bar{F} = 0.47$ and $\alpha = 0.84$ respectively, the maximum among the videos in MotoSum.*



(c) *Coherence matrix plot for the annotations from video 12. The f-score and Cronbach alpha values are $\bar{F} = 0.30$ and $\alpha = 0.64$ respectively, below the MotoSum average.*

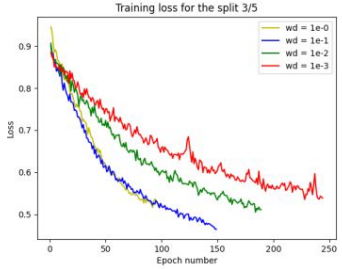


(d) *Coherence matrix plot for the annotations from video 1. The f-score and Cronbach alpha values are $\bar{F} = 0.26$ and $\alpha = 0.49$ respectively, the minimum among the videos in MotoSum.*

Retrained DSNet

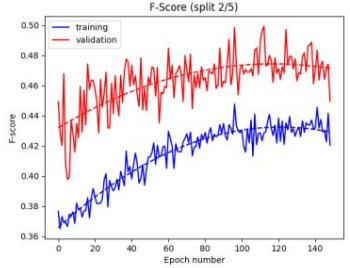
- Cross-validation and hyperparameters optimization

(a) *Graph showing training loss for four different training processes with different learning rates.*

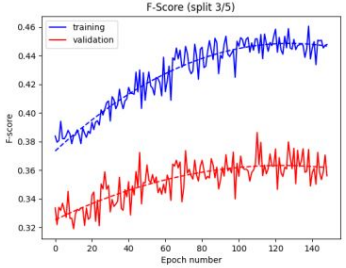


(b) *Graph showing training loss for four different training processes with different weight decay values.*

- Training and validation F-Scores



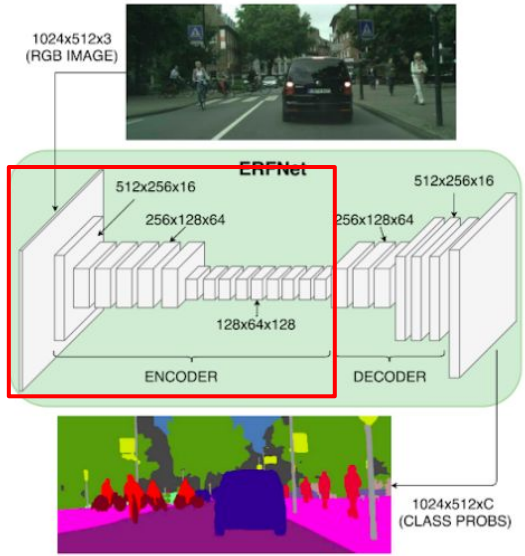
(a) *Graph showing f-score for validation and training sets during training process of the second split.*



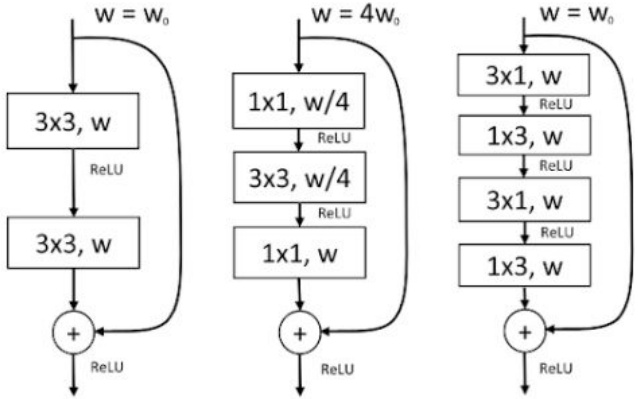
(b) *Graph showing f-score for validation and training sets during training process of the third split.*

DSNet + ERFNet

- Changing feature extractor
- ERFNet: Semantic Segmentation for Urban Landscape



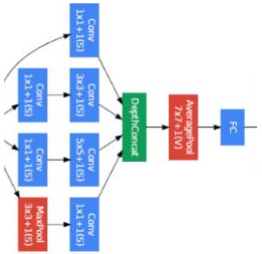
ERFNet Encoder



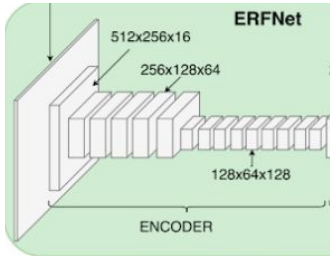
Residual Blocks Comparison

DSNet + ERFNet

- Changing feature extractor
- ERFNet: Semantic Segmentation for Urban Landscape
- Much higher dimension feature space



→ **1024 Dimensions**



→ **128 × 64 × 128 = 1048576 Dimensions**

Numerical Analysis

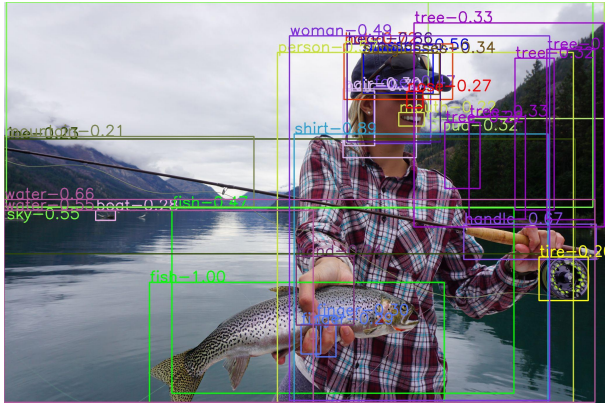
- Comparing models scores

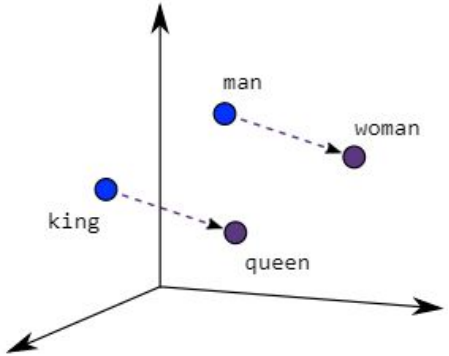
Model	Diversity	Representativeness	BRISQUE	NIQE
Reference Model	0.67951	0.33452	52.71482	11.94220
Pre-trained Agent	0.71394	0.53834	52.98820	12.03526
Retrained Agent	0.71358	0.51296	51.87517	12.07534

- Improvement over the reference model
- Feature extractor perspective

Query Filtering

- Summarize videos based on video title - text query
- Word2vec and Embedded Semantic Space
- Canonical Correlation Analysis (CCA)
- VinVL ResNext-152 C4: object detection network





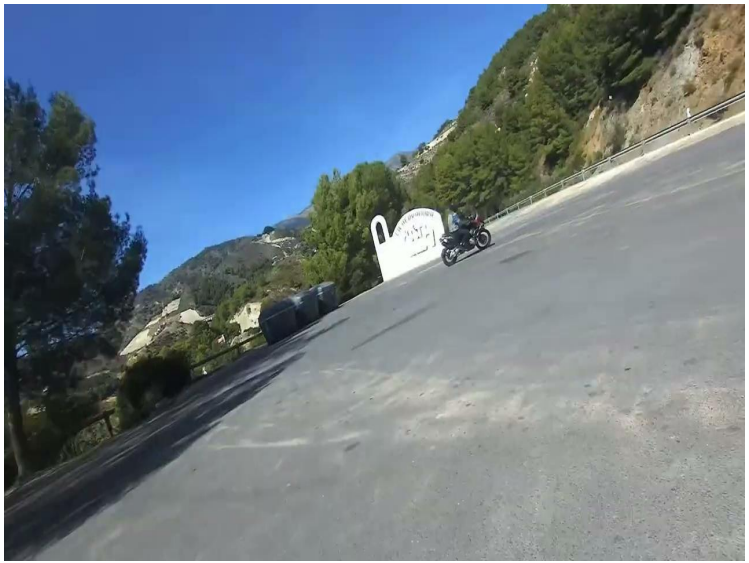
Stat-Based Filtering

- Feature filtering approach
- Speed and acceleration/deceleration
- Altitude and ascent/Descent
- Summary flexibility and interactiveness
- Data preprocessing for smoothness

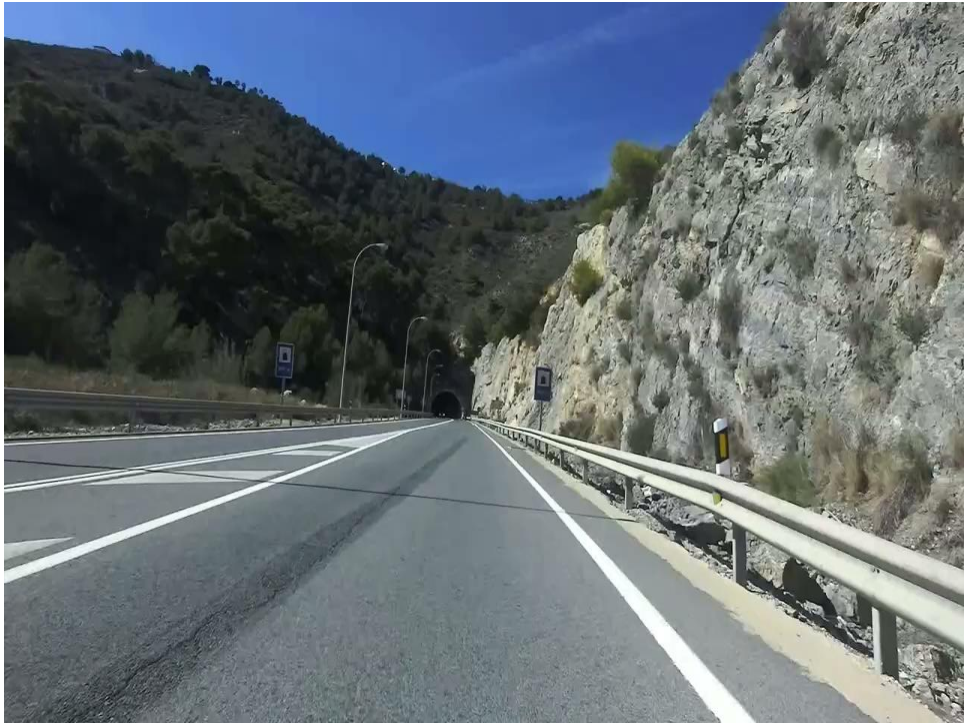
Wrapping Up

- Broad but deep overview on video summarization
- Pedagogical journey through multiple computer science fields
- Building knowledge from fundamentals to state-of-the-art techniques
- Self-contained project that can be widely expanded
- There is still much to be done to turn it into the final product

Summaries



Summaries



Thank You For Your Attention :)

Contact: luis.zerkowski@usp.br

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA UNIVERSIDADE DE SÃO PAULO

ETH zürich