Convolutional gated recurrent architectures for video segmentation

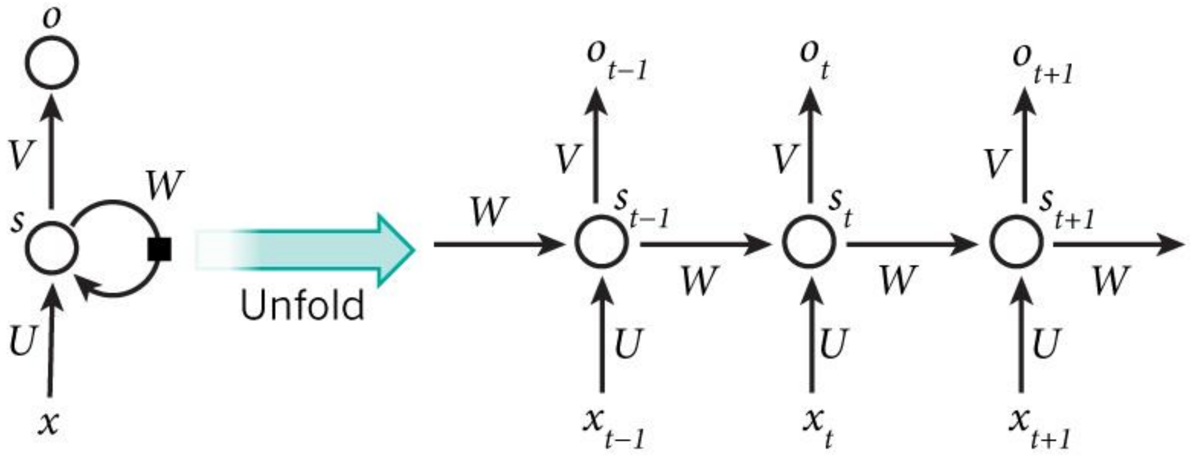
Mennatullah Siam, Sepehr Valipour, Martin Jagersand, Nilanjan Ray

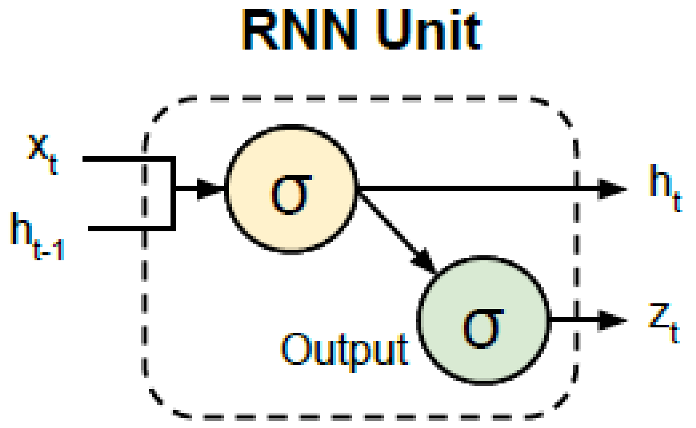
Abstract

- Previous work: semantic segmentation on still images
- Temporal data in video for semantic segmentation
- Fully-CNN and Gated Recurrent Unit (GRU)
- Sequence of consecutive frames and outputs segmentation of last frame
- GRU to preserve spatial connectivity
- Around 5% improvement for Dice respect to F-CNN

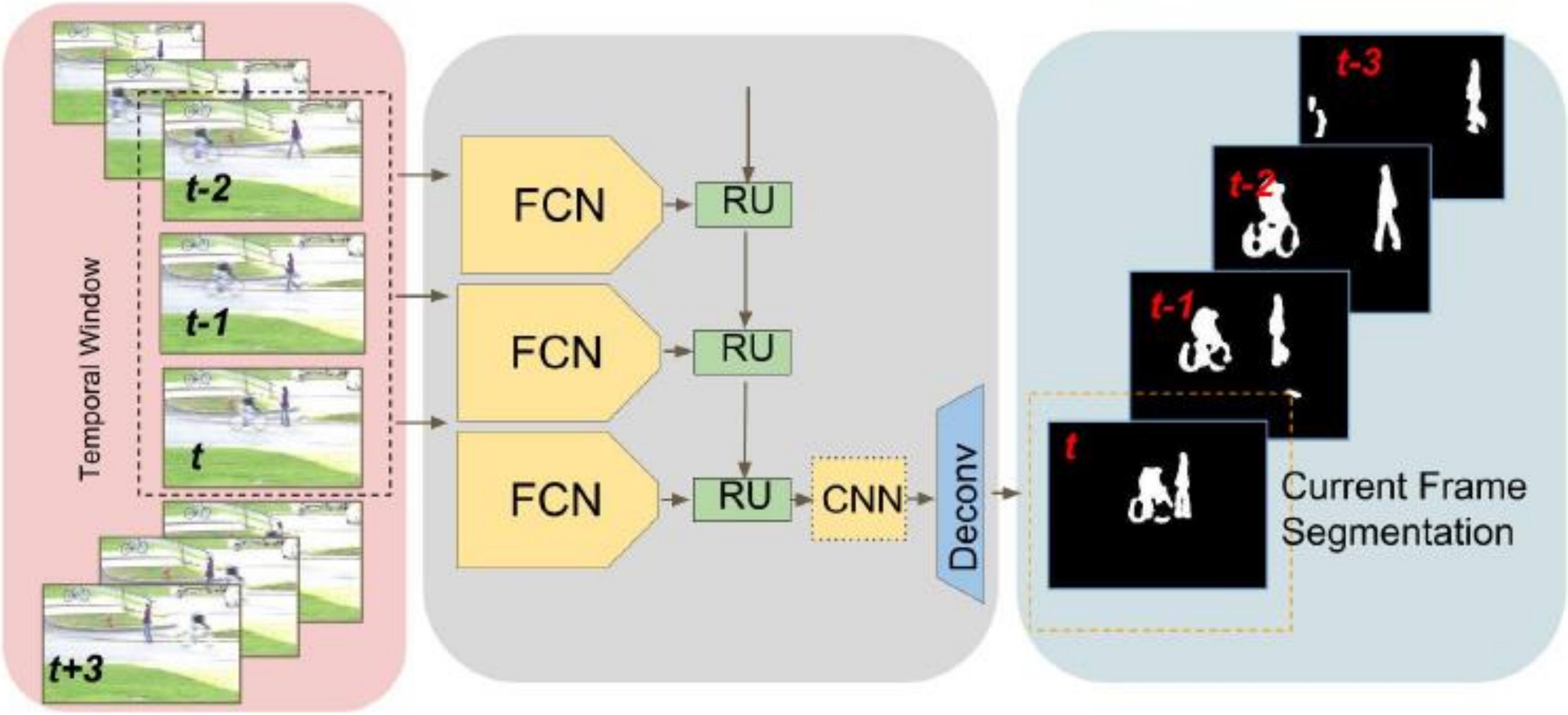
Method

- Recurrent neural networks:
  - Long short term memory (LSTM)
  - Gated recurrent unit (GRU): reduced number of gates and parameters



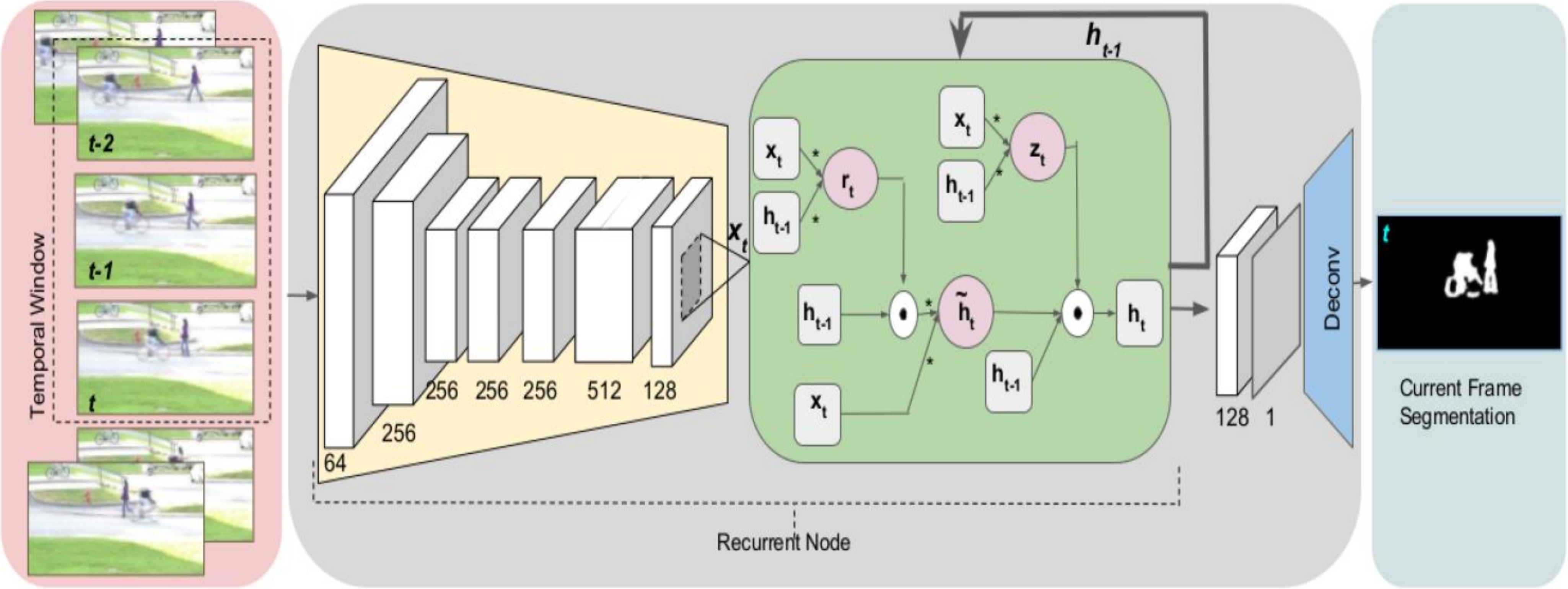


Overview

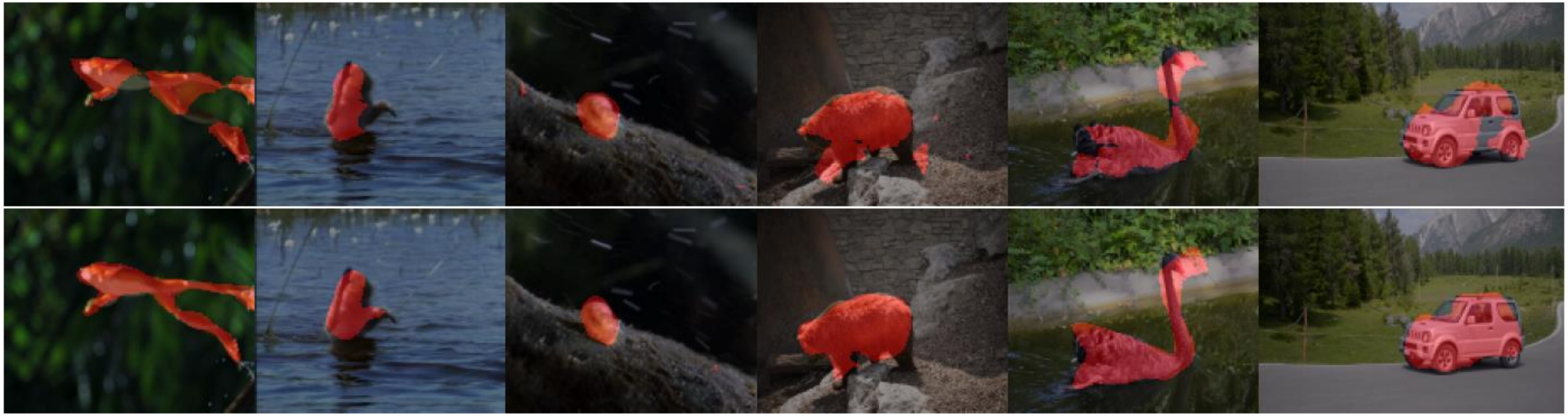


Temporal Window
t-2
t-1
t
t+3
FCN
FCN
FCN
RU
RU
RU
CNN
Deconv
t-3
t-2
t-1
t
Current Frame
Segmentation

Recurrent Fully Convolution



Qualitative Results



Quantitative Results

Table 2: Comparison of RFC-VGG with its baseline counterpart on DAVIS and SegTrack

		Precision	Recall	F-measure	IoU
SegTrack V2	FC-VGG	0.7759	0.6810	0.7254	0.7646
	RFC-VGG	**0.8325**	0.7280	**0.7767**	**0.8012**
	FC-VGG Extra Conv	0.7519	**0.7466**	0.7493	0.7813
DAVIS	FC-VGG	0.6834	0.5454	0.6066	0.6836
	RFC-VGG	**0.7233**	**0.5586**	**0.6304**	**0.6984**

Qualitative Results

Table 3: Semantic Segmentation Results on Synthia Highway Summer Sequence for RFC-VGG compared to FC-VGG

	Mean Class IoU	Per-Class IoU								
		Car	Pedestrian	Sky	Building	Road	Sidewalk	Fence	Vegetation	Pole
FC-VGG	0.755	0.504	0.275	0.946	0.958	0.840	0.957	0.762	0.883	0.718
RFC-VGG	**0.812**	**0.566**	**0.487**	**0.964**	**0.961**	**0.907**	**0.968**	**0.865**	**0.909**	**0.742**

Summary

- Time correlation is effective overall
- More effective for smaller objects (car, pedestrian) than large objects(sky, buildings)
- Higher complexity