One-Step Time-Dependent Future Video Frame Prediction with a Convolutional Encoder-Decoder Neural Network

[1] TU Delft [2] IRISA [3] INSA Institut National des Sciences Appliquées Rennes Inria [4] cnrs

Vedran Vukotić[1,2,3], Silvia-Laura Pintea[1], Christian Raymond[2,3], Guillaume Gravier[3,4], Jan van Gemert[1]

vedran.vukotic@irisa.fr s.l.pintea@tudelft.nl christian.raymond@irisa.fr guillaume.gravier@irisa.fr j.c.vangemert@tudelft.nl

Problem

- given the current video frame at time t_0 and an arbitrary temporal displacement t predict the frame at $t_0 + t$

Goal:

- anticipate future motion-induced appearance change

Means:

- creating representation that encodes appearance changes over time
- embedding the input image and a continuous time variable
- translating back to the image space to visualize the anticipated video frame

Existing Work

Predicting Future Motion:

- given an image, predict optical flow at the next timestep
- given an image predict motion trajectories

Predicting Future Appearance:

- hallucinating possible images (conditioned GANs)
- predicting future pixels from previus pixels (Pixel Networks)
- autoencoding methods - predicting the future image at the next timestep

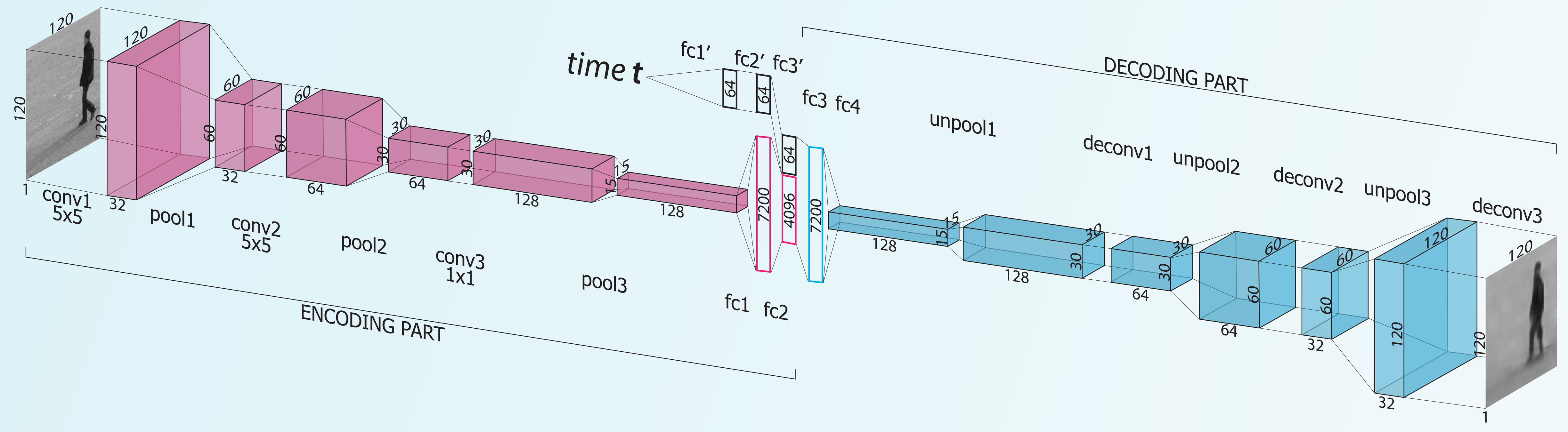
Evaluation

- KTH human action recognition dataset (randomly split by actors - 80% in training set, 20% in testing set)
- anticipating six actions: walking, jogging, running, hand-clapping, hand-waving and boxing
- compared to sequential analogous encoder-decoder baseline not conditioned on time
- visual evaluation and **Mean Square Error (MSE)** along the edges:

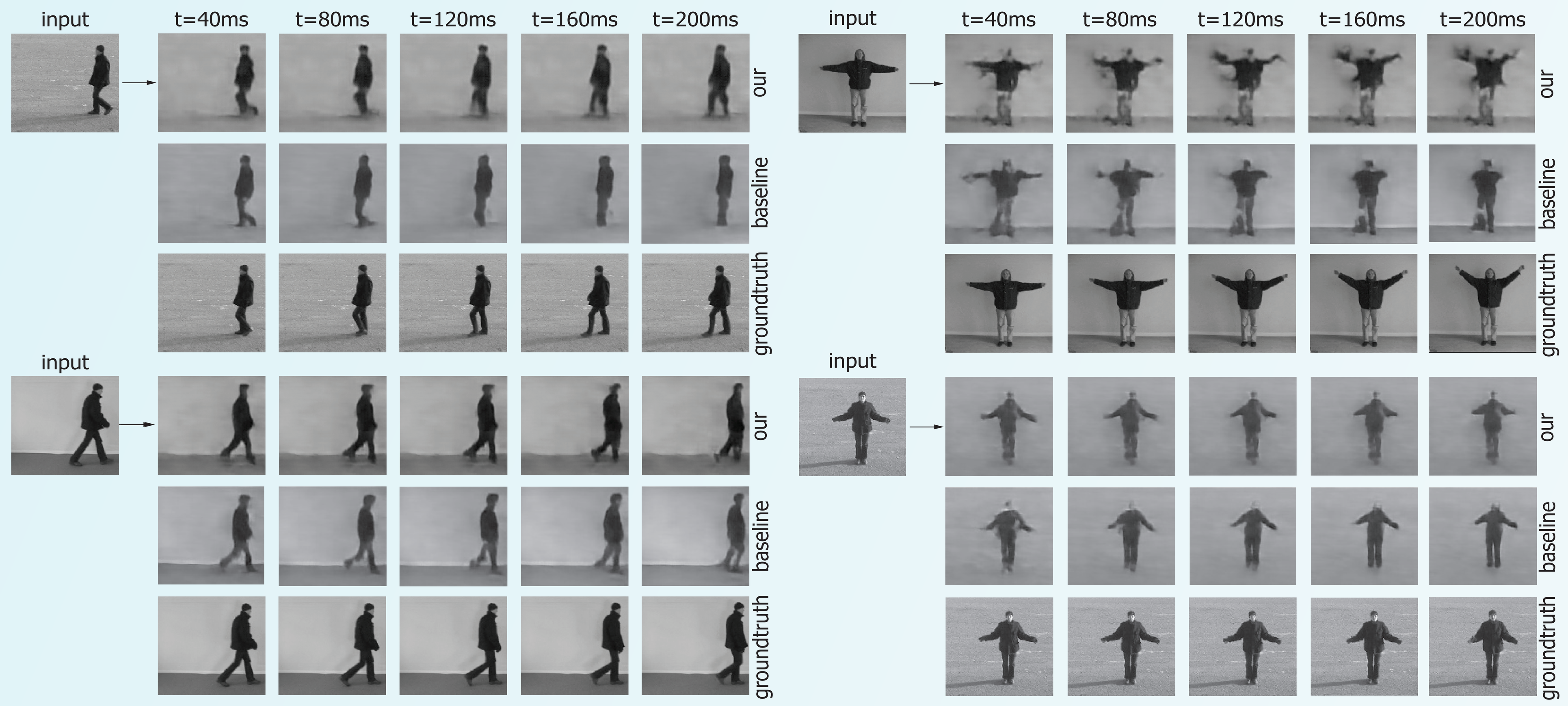
	Mean Squared Error	
Action	Baseline	Our Method
Jogging	30.64	**11.66**
Running	40.88	**17.35**
Walking	30.87	**19.26**
Hand-clapping	43.23	**33.93**
Hand-waving	43.71	**35.19**
Boxing	46.22	**37.71**
Mean MSE	39.26	**25.85**

NCCV 2016 Lunteren, NL

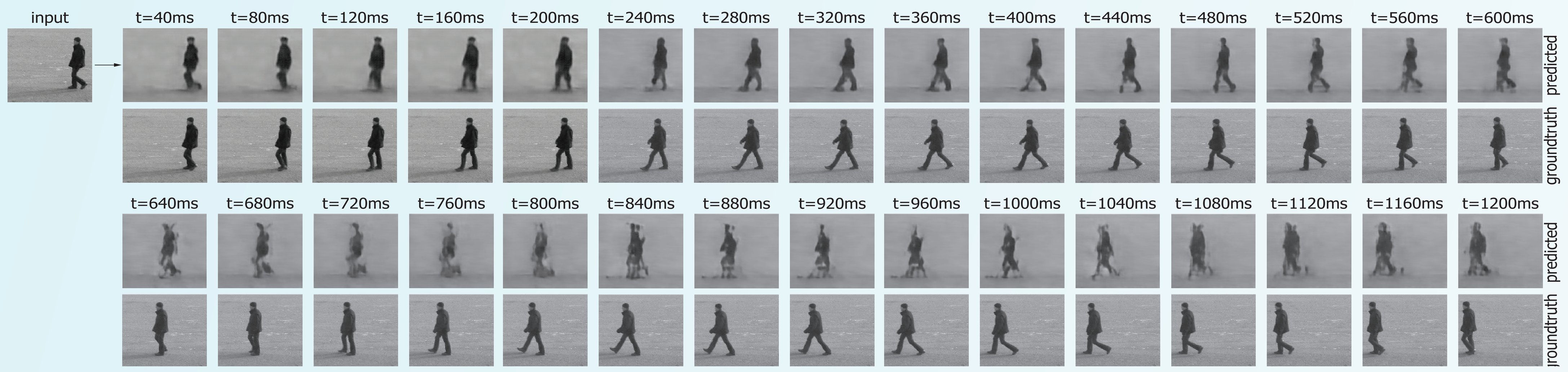
Proposed Architecture



Example Anticipations

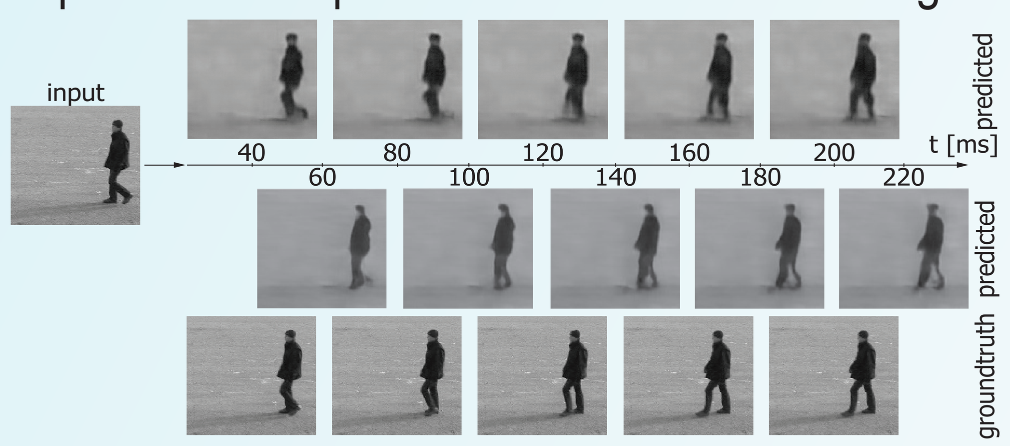


Long-Term Anticipations



Unseen Time Displacements

- anticipations at temporal distances not seen during training:



Conclusion

Successes:

- successfully predicts future frames at **arbitrary** temporal displacements, including temporal displacements **never seen during training**
- predictions are done directly, in one step

Downsides:

- unable to tackle ambiguities; artifacting and loss of details caused by addressable issues in videos

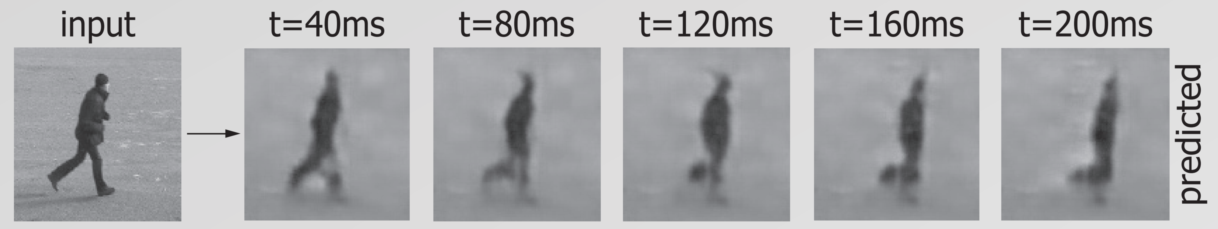
Idea

- **encoding network**
  - image encoding branch
  - time encoding branch (time modeled as a continuous variable t)
- **decoding network**
- inspired by the architecture in [1]

Downsides

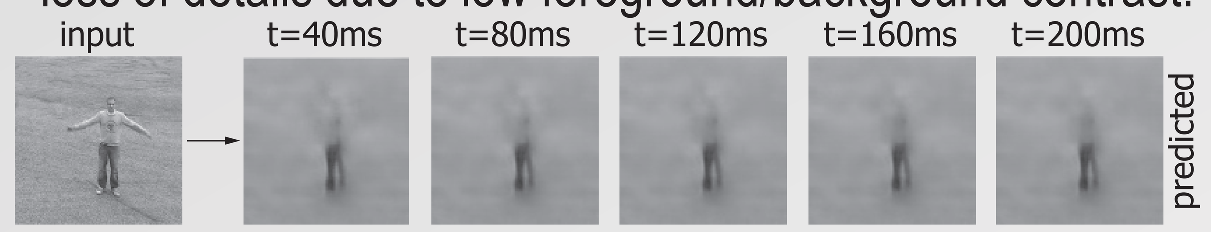
- loss of details and artifacting due to motion ambiguity:

- loss of details due to larger differences between two consecutive frames:



- extreme loss of details and artifacting in due to even larger differences between two consecutive frames:

- loss of details due to low foreground/background contrast:



- artifacting due to small and sporadic movements:

References

1.M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3D Models from Single Images with a Convolutional Network. *In ECCV 2016, Amsterdam, The Netherlands, 2016.*