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

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

[1]TU Delft, Delft, The Netherlands
[2]INRIA/IRISA, Rennes, France
[3]INSA Rennes, Rennes, France
[4]CNRS, France

{vedran.vukotic, christian.raymond, guillaume.gravier}@irisa.fr
{S.L.Pintea, j.c.vangemert}@tudelft.nl

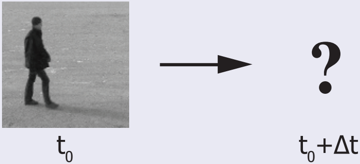
NCCV 2016

December $12^{th}$ 2016, Lunteren, NL

# Introduction

## Task

- given an image, predict its future appearance



# Previous Works / Approaches

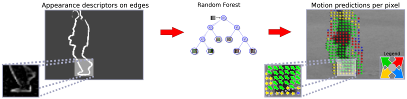
- predicting future motion

[1] S L Pintea, J C van Gemert, and A W M Smeulders. "Déja vu". In: *ECCV*. Springer. 2014, pp. 172–187.

[2] J Walker et al. "An uncertain future: Forecasting from static images using variational autoencoders". In: *ECCV*. Springer. 2016, pp. 835–851.

# Previous Works / Approaches

- predicting future motion
  - predicting optical flow[1]

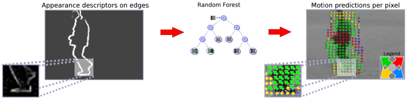



[1] S L Pintea, J C van Gemert, and A W M Smeulders. "Déja vu". In: *ECCV*. Springer. 2014, pp. 172–187.

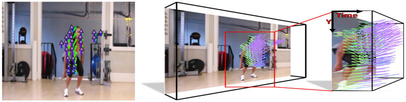
[2] J Walker et al. "An uncertain future: Forecasting from static images using variational autoencoders". In: *ECCV*. Springer. 2016, pp. 835–851.

# Previous Works / Approaches

- predicting future motion
  - predicting optical flow[1]




  - predicting trajectories[2]

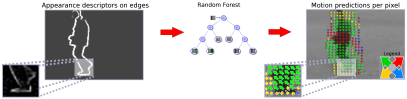



[1] S L Pintea, J C van Gemert, and A W M Smeulders. "Déja vu". In: *ECCV*. Springer. 2014, pp. 172–187.

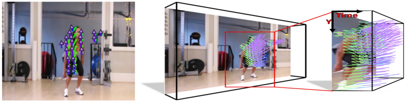
[2] J Walker et al. "An uncertain future: Forecasting from static images using variational autoencoders". In: *ECCV*. Springer. 2016, pp. 835–851.

# Previous Works / Approaches

- predicting future motion
    - predicting optical flow[1]




    - predicting trajectories[2]




- predicting future appearance

[1] S L Pintea, J C van Gemert, and A W M Smeulders. “Déja vu”. In: *ECCV*. Springer. 2014, pp. 172–187.

[2] J Walker et al. “An uncertain future: Forecasting from static images using variational autoencoders”. In: *ECCV*. Springer. 2016, pp. 835–851.

# Previous Works / Approaches II

## Predicting Future Appearance

- predicting an image

# Previous Works / Approaches II

## Predicting Future Appearance

- predicting an image
- multiple approaches:
  - generative methods[a]

# Previous Works / Approaches II

## Predicting Future Appearance

- predicting an image
- multiple approaches:
  - generative methods[a]
  - autoencoder methods
    - image in - image out
    - our approach

[a] A van den Oord, N Kalchbrenner, and K Kavukcuoglu. "Pixel Recurrent Neural Networks". In: *CoRR* (2016), A van den Oord et al. "Conditional image generation with pixelcnn decoders". In: *CoRR* (2016).

# Previous Works / Approaches II

## Predicting Future Appearance

- predicting an image
- multiple approaches:
  - generative methods[a]
  - autoencoder methods
    - image in - image out
    - our approach

[a] A van den Oord, N Kalchbrenner, and K Kavukcuoglu. "Pixel Recurrent Neural Networks". In: *CoRR* (2016), A van den Oord et al. "Conditional image generation with pixelcnn decoders". In: *CoRR* (2016).

## Autoencoder Methods

- predictions are typically obtained for a predefined temporal displacement

# Previous Works / Approaches II

## Predicting Future Appearance

- predicting an image
- multiple approaches:
  - generative methods[a]
  - autoencoder methods
    - image in - image out
    - our approach

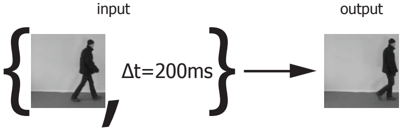
[a] A van den Oord, N Kalchbrenner, and K Kavukcuoglu. "Pixel Recurrent Neural Networks". In: *CoRR* (2016), A van den Oord et al. "Conditional image generation with pixelcnn decoders". In: *CoRR* (2016).

## Autoencoder Methods

- predictions are typically obtained for a predefined temporal displacement
- predictions at other (quantized!) intervals are obtained iteratively

# Goal

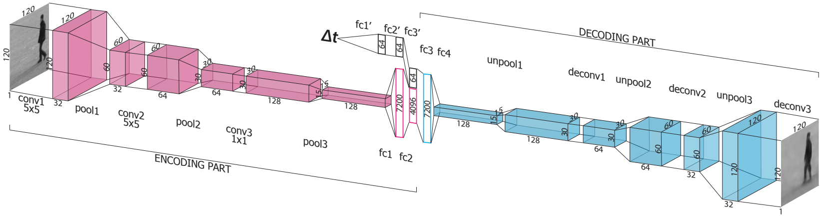
- given an image and a temporal displacement $\Delta t$, predict the future image




## Setup

- inputs:
  - image $I_0$ at current time $t_0$
  - temporal displacement $\Delta t$
- output:
  - anticipated image $I_{t_0+\Delta t}$ at time $t_0 + \Delta t$
- minimizing $MSE(I_{t_0+\Delta t}, I'_{t_0+\Delta t})$
- **one-step** predictions at **arbitrary** temporal displacements

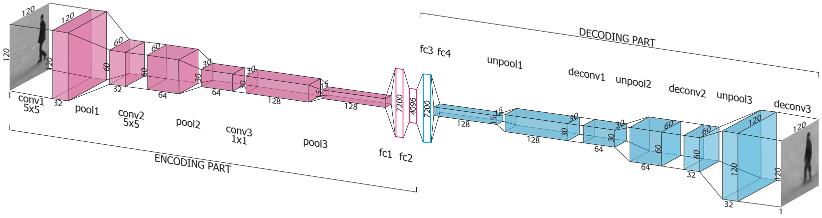
# Architecture




- encoder network
  - image encoding branch
  - time encoding branch (continuous input!)
- decoder network
- similar architecture used to generate object rotations[3]

[3] M Tatarchenko, A Dosovitskiy, and T Brox. “Multi-view 3D Models from Single Images with a Convolutional Network”. In: *ECCV*. Springer. 2016, pp. 322–337.

# Baseline




- analogous encoder-decoder architecture
  - no time modelling branch
  - one-step prediction for a fixed temporal displacement $\Delta t$
  - further predictions computed iteratively for $k\Delta t$

# Dataset

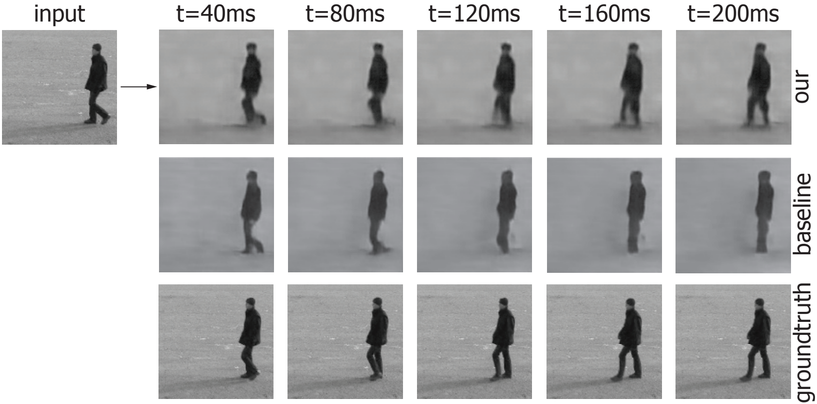
- KTH human action recognition dataset
  - 6 actions (*walking, jogging, running, hand-waving, hand-clapping, boxing*)
  - 25 actors; 4 recordings for each actor and action

# Dataset

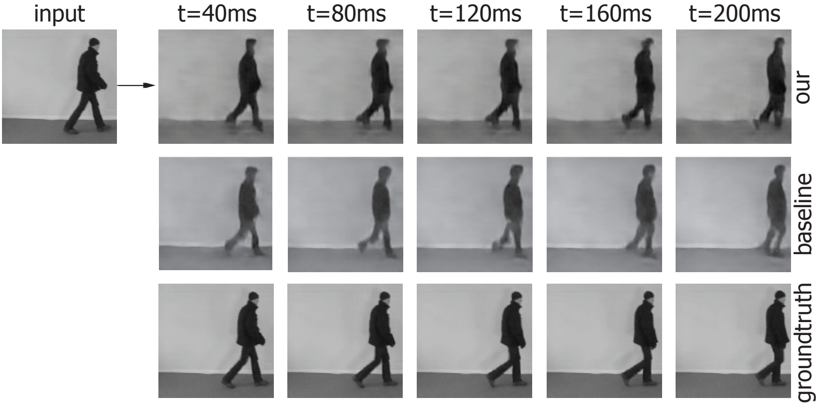
- KTH human action recognition dataset
  - 6 actions (*walking, jogging, running, hand-waving, hand-clapping, boxing*)
  - 25 actors; 4 recordings for each actor and action

- randomly split by actors
  - 80% - training set
  - 20% - testing set

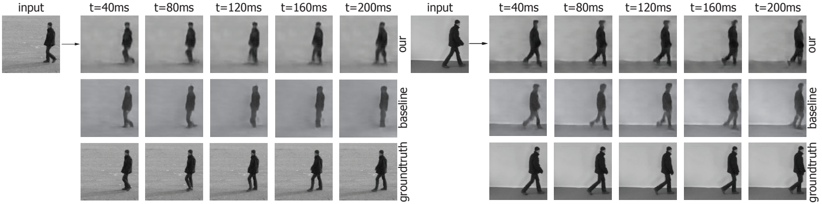
# Example Anticipations



# Example Anticipations



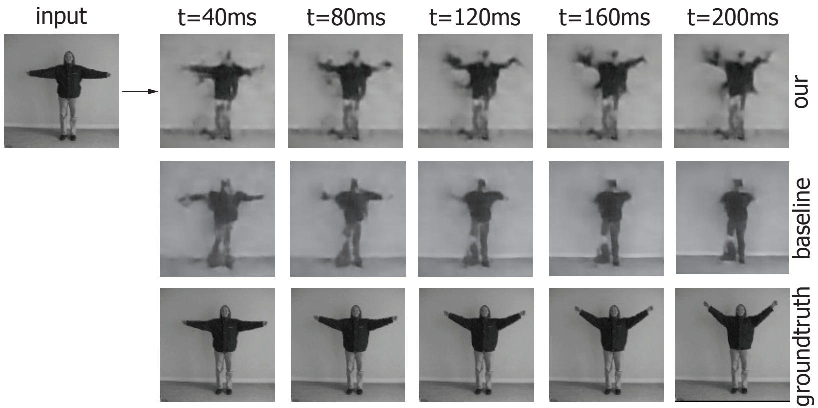
# Example Anticipations



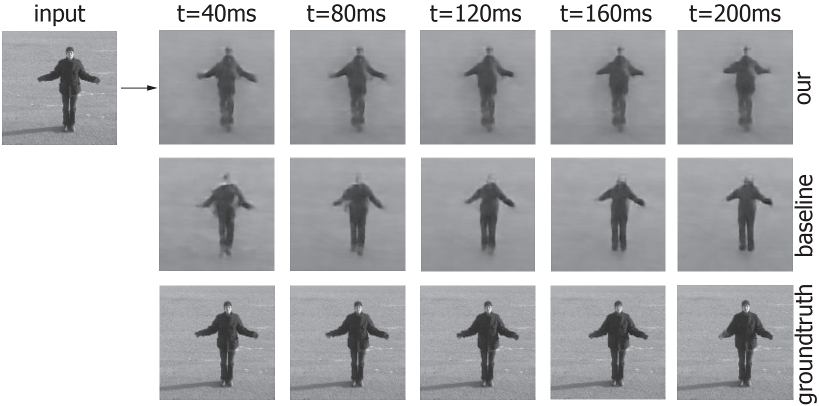

## The Architecture is:

- able to recognize location and pose
- able to anticipate spatial displacement and appearance
- able to understand orientation (*e.g.* walking left to right vs right to left)

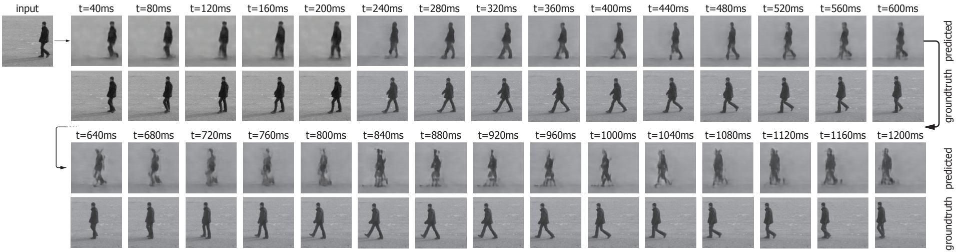
# Example Anticipations II



# Example Anticipations II

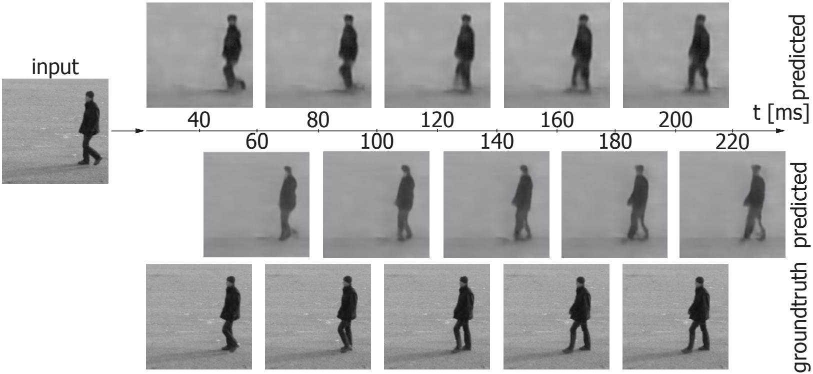


# Long-Distance Anticipations

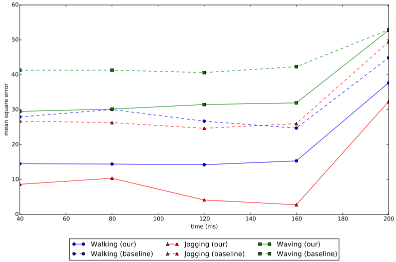



# Anticipating Unseen Temporal Displacements

- intervals during training dependent on the video framerate
- predicting unseen temporal displacements:



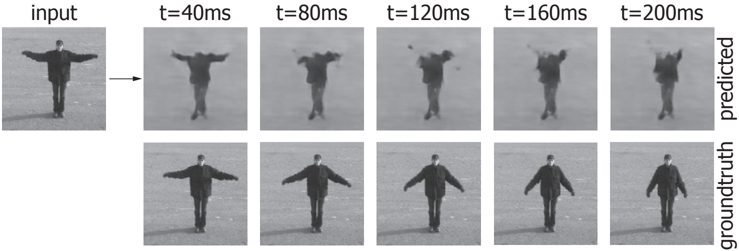
# Quality Estimations - MSE




| | 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** |

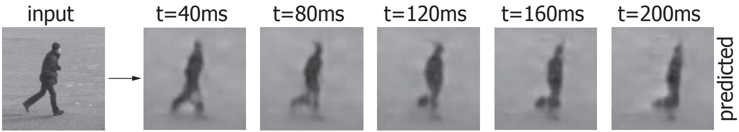
# Downsides

- artifacting and loss of details due to pose ambiguity:

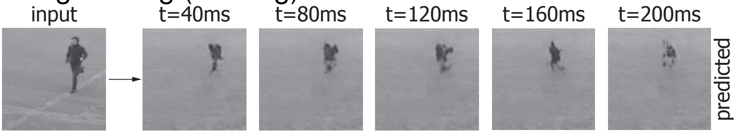


# Downsides II

- loss of details due to large frame differences during training (jogging):

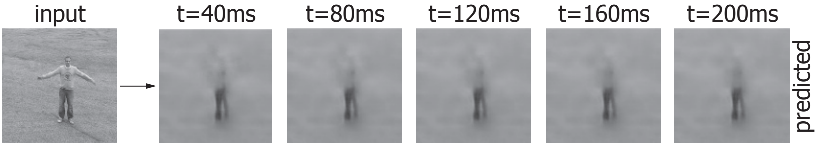



- extreme loss of details due to even larger frame differences during training (running):

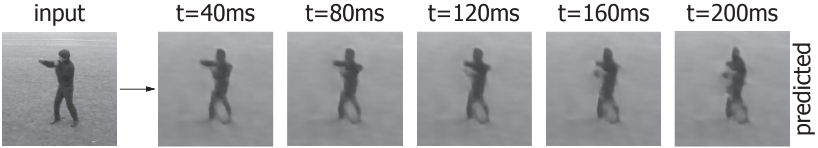


# Downsides III

- loss of details due to low fg/bg contrast:




- loss of details and artifacting due to small and sporadic movement:



# Conclusion

- anticipates future at arbitrary time displacements ✓
- does so in one step, with no iterations ✓
- outperforms iterative predicting in terms of MSE and visual analysis ✓

- ambiguities represent cannot be tackled by this architecture alone ✗
- bigger displacements and decreased contrast lead to artifacting and loss of details ✗

Thank you! Questions?