INFORMATIK

augmented VISION DFKI

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

# 2D Image Processing & Augmented Reality
# Winter Semester 2019/2020
# Survey on Face Tracking with Deep Learning


Vinay Balasubramanian

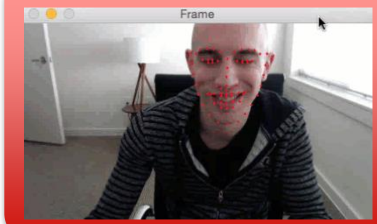
v_balasubr18@cs.uni-kl.de

Supervisor: Jilliam Diaz Barros

# Outline

# Quick recap

# Face Tracking




Tracking a face across all frames of a video.

Bounding-box or landmark based.

Applications - Face analysis, Person identification, Activity recognition, Expression analysis, Face modeling etc.


Challenges - Videos can be captured in unconstrained conditions.

May have illumination variations, large head poses, occlusions, etc.

Image source: Google

# General Idea

- Video – Sequence of frames with temporal connection
- Sequence data? (Use RNN)

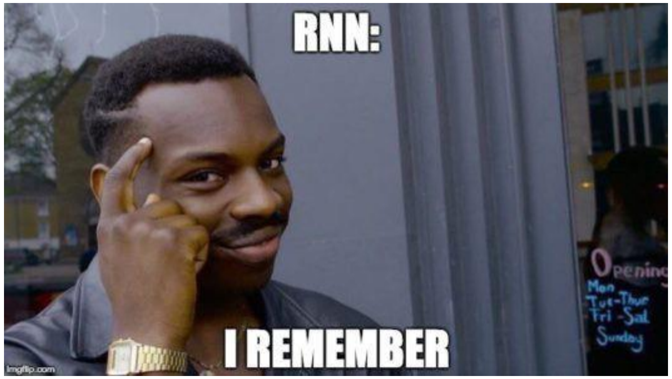
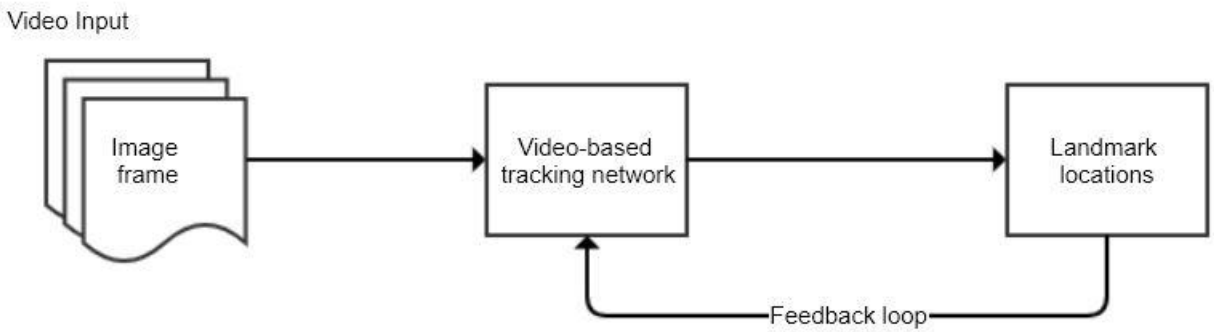



Image source: https://towardsdatascience.com/pytorch-basics-how-to-train-your-neural-net-intro-to-rnn-cb6ebc594677

# General Idea

- Video – Sequence of frames with temporal connection
- Sequence data? (Use RNN)
- Give frames in temporal order, detect landmarks, feedback along with next frame.



# Why Deep Learning?

- State-of-the-art in image processing tasks
- Operate directly on data
- Learn more generic features directly from data
- Domain knowledge is never obsolete

# Datasets

- AFLW – Around 25k annotated face images. 21 points
- COFW – 1007 occluded face images. 29 points
- Helen – 2000 training and 330 test images. 194 points
- IBUG – 135 images with difficult poses and expressions. 68 points
- LFPW – 1432 images. 29 points
- LFW – 13233 images of 5749 people. 7 points
- 300-W – 600 images in the wild. 68 points
- 3D Menpo – 12k images, 280k video frames, with 2D and 3D landmarks. 84 points
- FM – 2150 images of 6 videos. 68 points
- RWMB – 20 videos with motion blur. 68/98 points
- TF – 5000 frames of a person engaged in a conversation. 68 points
- 300-VW – 114 videos with 218,595 frames. 68 points

# Approaches

- **RNN** [Jinwei Gu et al., 2017]
- **Two-stream network** [Hao Liu et al., 2017]
- **LSTM** [Qiqi Hou et al., 2017]
- **Encoder-Decoder network** [Xi Peng et al., 2018]
- **Deep reinforcement learning** [Minghao Guo et al., 2018]

# Using RNNs [Jinwei Gu et al., 2017]

- **RNNs and Bayesian filters are operationally similar**

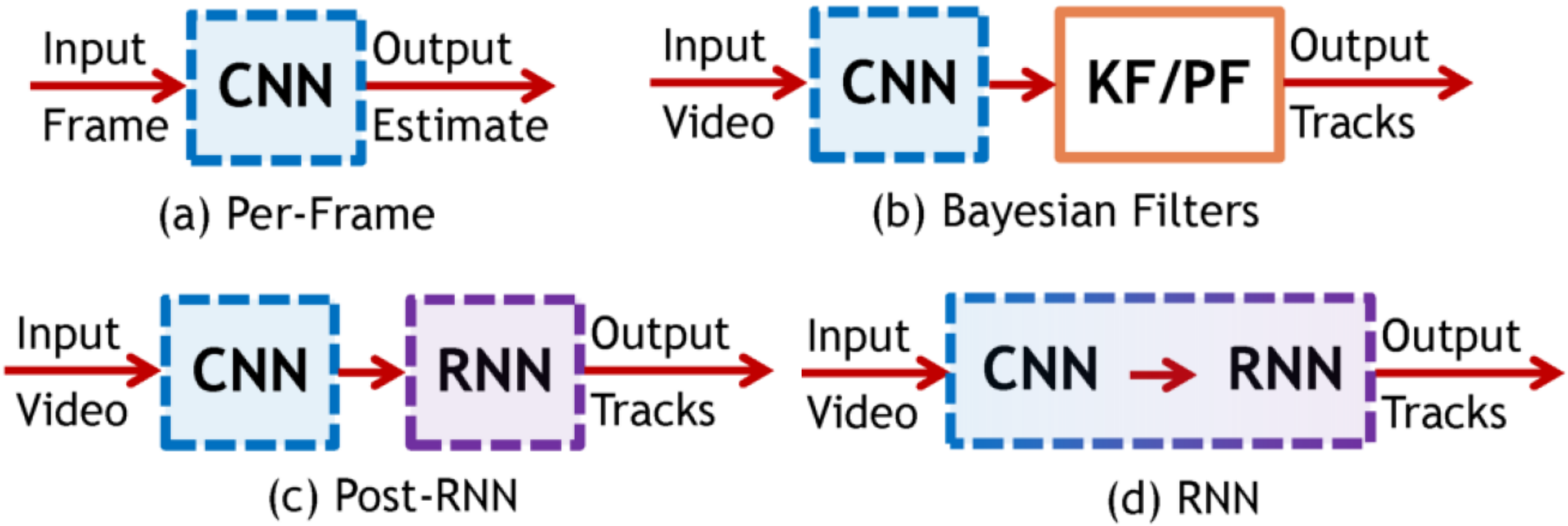



Image source: Jinwei Gu, Xiaodong Yang, Shalini De Mello, Jan Kautz: Dynamic Facial Analysis: From Bayesian Filtering to Recurrent Neural Network(2017).

# Using RNNs [Jinwei Gu et al., 2017]


(d) RNN

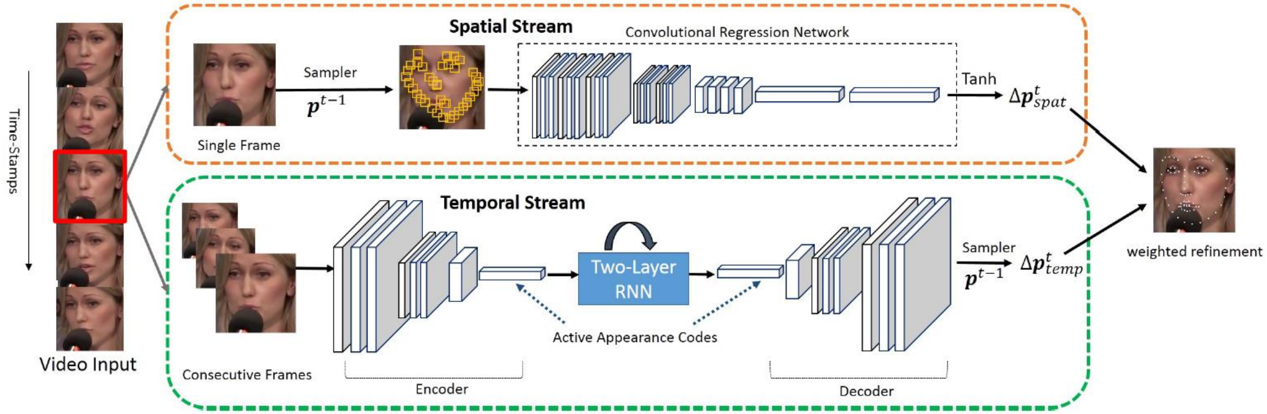
- **FC-RNN is used to retain generalization of pre-trained CNN**
- **Trained end-to-end**
- **300-VW dataset for facial landmark localization**
- **L2 loss function**
- **Evaluation - Area Under the Curve (AUC), Failure Rate (FR %)**

Image source: Jinwei Gu, Xiaodong Yang, Shalini De Mello, Jan Kautz: Dynamic Facial Analysis: From Bayesian Filtering to Recurrent Neural Network(2017).

# Two-stream network [Hao Liu et al., 2017]

- **Exploit both appearance information from still frames(spatial) and temporal information across frames(temporal)**
  - Spatial stream – Image pixels (still) -> landmark locations
  - Temporal stream – Compress video as active appearance codes

# Two-stream network [Hao Liu et al., 2017]




- **Spatial stream transforms local facial patches to shape residuals used to refine current face shape from previous.**
- **Temporal stream – Encoder-decoder network with 2-layer RNN. Capture facial dynamics in temporal dimension**
- **Final prediction is a weighted fusion of spatial and temporal streams shape updates**

Image source: Hao Liu, Jiwen Lu, Jianjiang Feng, Jie Zhou: Two-Stream Transformer Networks for Video-based Face Alignment(2017).

# Two-stream network [Hao Liu et al., 2017]

- Tested on 300-VW and TF datasets
- Evaluation – Normalized RMSE and Cumulative Error Distribution plots
- Weighted fusion – $\beta_1$ and $\beta_2$ = 0.5 yields the best performance

# LSTMs [Qiqi Hou et al., 2017]

- **LSTM is used to exploit spatial and temporal information**

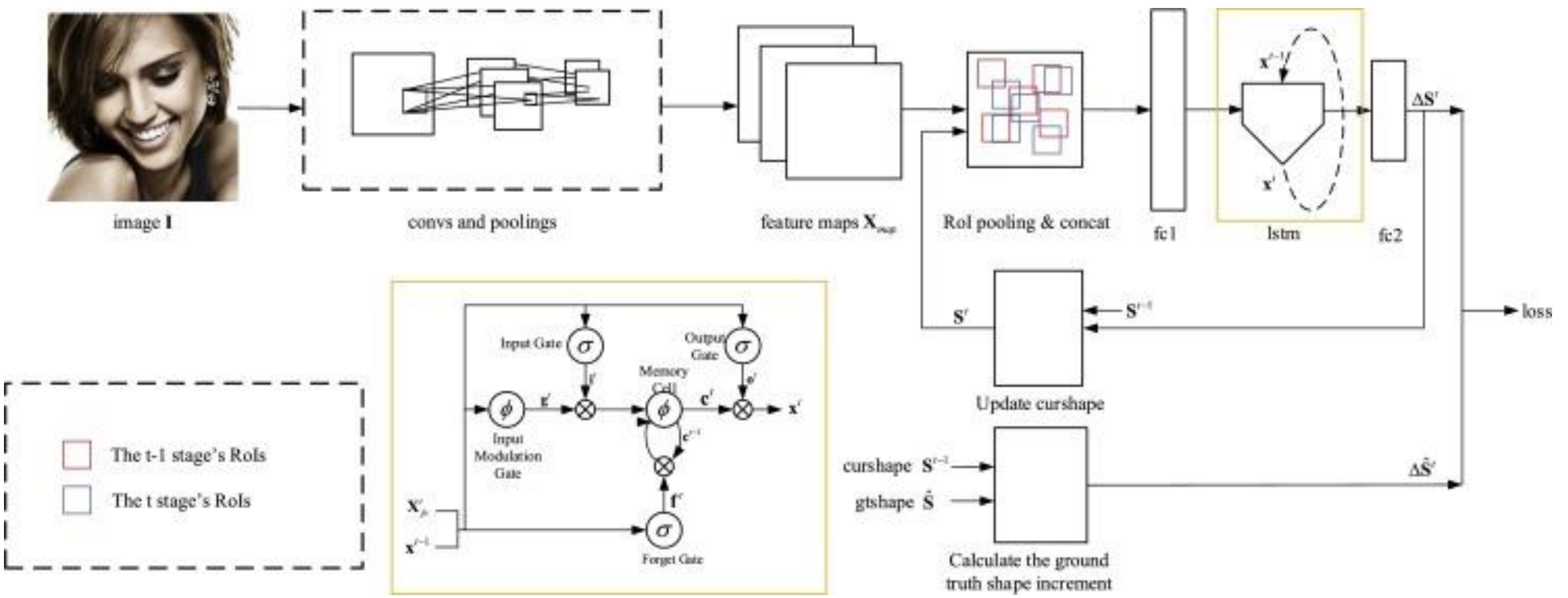
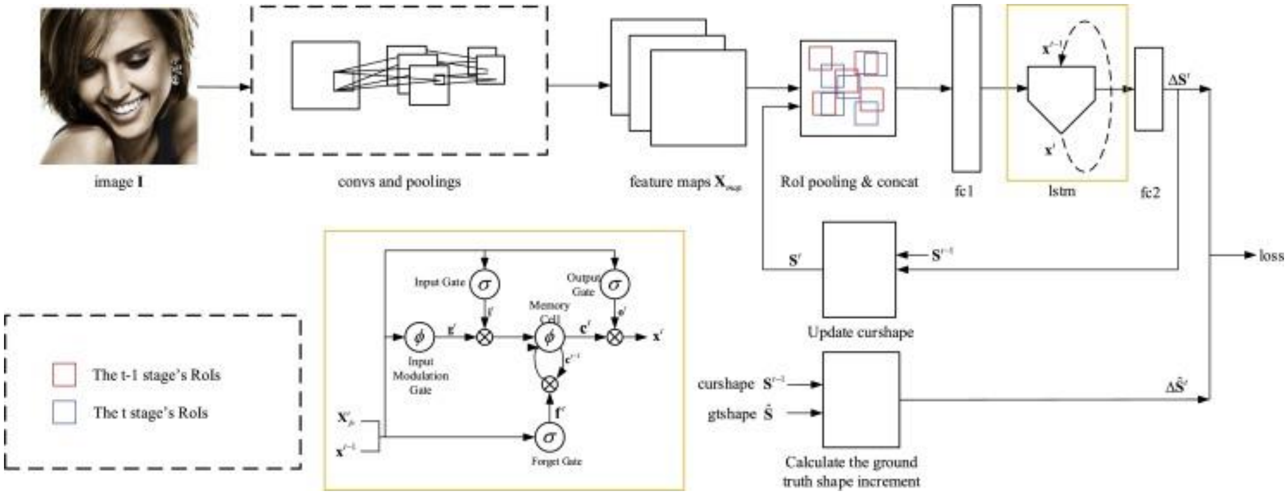





Image source: Qiqi Hou, Jinjun Wang, Ruibin Bai, Sanping Zhou, Yihong Gong: Face Alignment Recurrent Network(2017).

# LSTMs [Qiqi Hou et al., 2017]




- **Input – Image and Initial face shape**
- **Output – Predicted shape increment for the initial face shape**
- **Input -> several conv + max pooling -> ROI pooling for initial face shape -> concat -> FC layer -> LSTM -> predicted shape increment**
- **Update initial shape according to predicted shape increment**



Image source: Qiqi Hou, Jinjun Wang, Ruibin Bai, Sanping Zhou, Yihong Gong: Face Alignment Recurrent Network(2017).

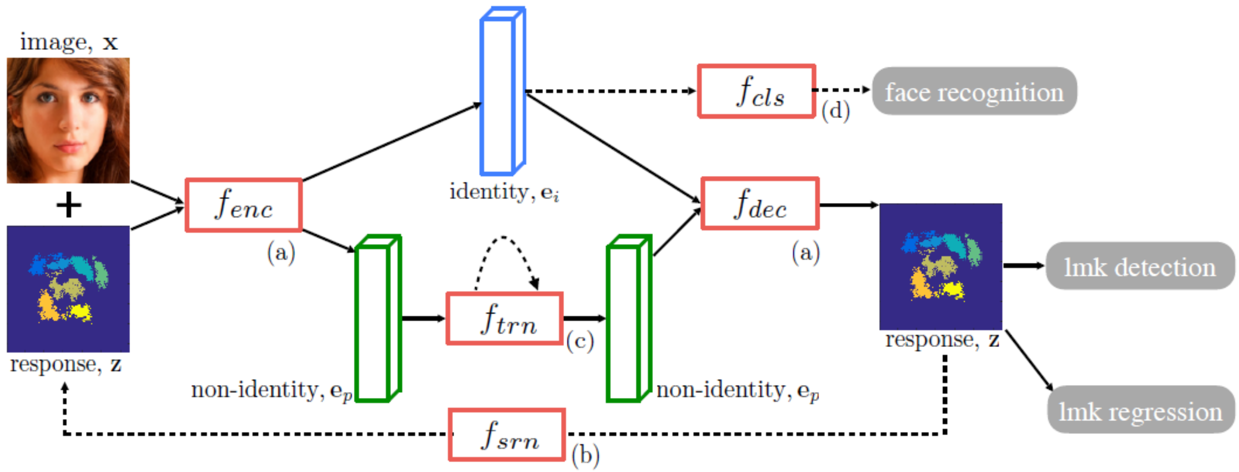
# LSTMs [Qiqi Hou et al., 2017]

- Landmark detection method
- Trained on COFW, LFPW, Helen, AFW
- Evaluated on – COFW, Helen, 300-W, 300-VW
- Evaluation – Point-to-point RMSE
- Runtime – 18ms

# Encoder-Decoder Network [Xi Peng et al., 2018]

- **Encoder – Image pixels -> Low dimensional feature space**
- **Decoder – Features in low dimensional space -> facial landmark heatmaps**
- **Feedback loop between the output(facial points) and the input**

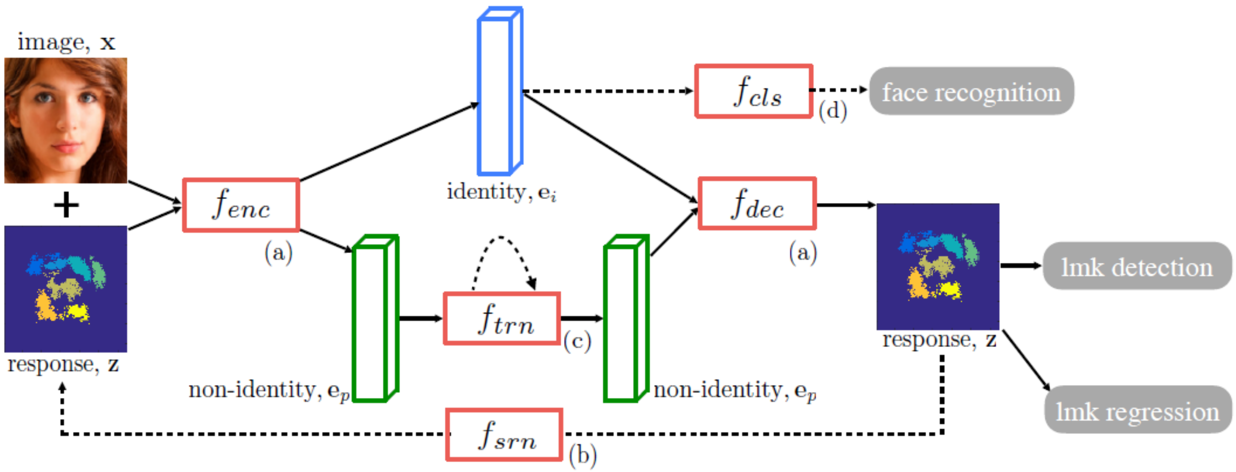
# Encoder-Decoder Network [Xi Peng et al., 2018]




- Encoder -> Feature disentangling on low dimensional representation -> Decoder
- Disentangle temporal-variant and temporal-invariant factors
- Temporal-invariant: Person identity
- Temporal-variant: Pose, expression, illumination

Image source: Xi Peng, Rogerio S. Feris, Xiaoyu Wang, Dimitris N. Metaxas: RED-Net: A Recurrent Encoder-Decoder Network for Video-based Face Alignment(2018).

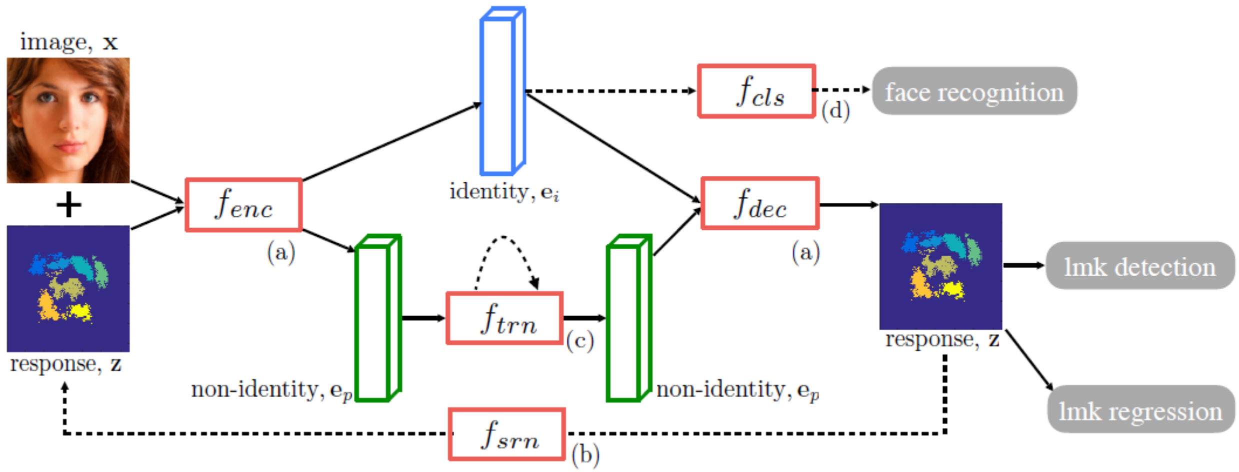
# Encoder-Decoder Network [Xi Peng et al., 2018]




- Spatial recurrent learning: Coarse-to-fine landmark search
- Feedback loop: Previous prediction + image
- Landmark detection: Detect major landmarks
- Landmark regression: Refine predicted locations from previous detection step

Image source: Xi Peng, Rogerio S. Feris, Xiaoyu Wang, Dimitris N. Metaxas: RED-Net: A Recurrent Encoder-Decoder Network for Video-based Face Alignment(2018).

# Encoder-Decoder Network [Xi Peng et al., 2018]




- **Temporal recurrent learning: Model non-identity factors (temporal-variant) using LSTM**

Image source: Xi Peng, Rogerio S. Feris, Xiaoyu Wang, Dimitris N. Metaxas: RED-Net: A Recurrent Encoder-Decoder Network for Video-based Face Alignment(2018).

# Encoder-Decoder Network [Xi Peng et al., 2018]

- **Evaluated on – AFLW, LFW, Helen, LFPW, TF, FM, 300-VW**
- **Evaluation – Inter-ocular distance normalized RMSE**

# Reinforcement Learning [Minghao Guo et al., 2018]

- **Bounding box generation and facial landmark detection are heavily dependent**
- **Two agents – bounding box Tracking and facial landmark Alignment**

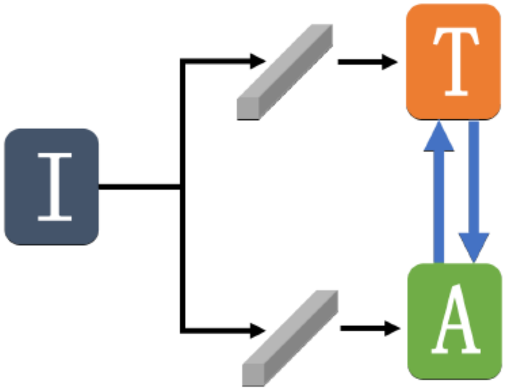
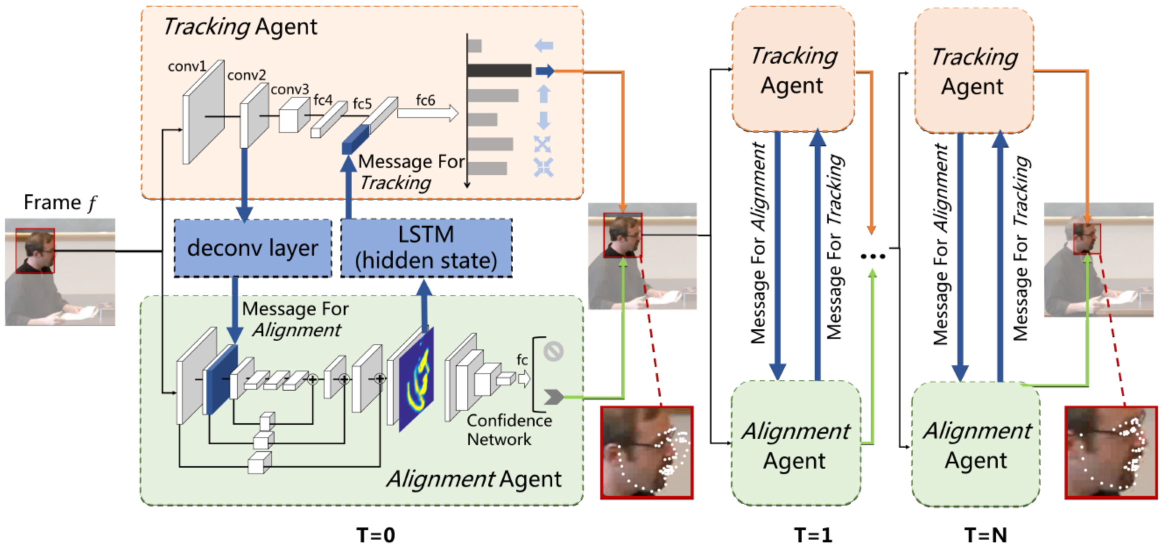



Image source: Minghao Guo, Jiwen Lu, and Jie Zhou: Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking(2018).

# Reinforcement Learning [Minghao Guo et al., 2018]




- **Communication channels between agents (Deep Q-learning)**
- **Current state initialized to terminal state of previous frame**
- **Agents decide sequence of actions based on observed state and received messages**

Image source: Minghao Guo, Jiwen Lu, and Jie Zhou: Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking(2018).

# Reinforcement Learning [Minghao Guo et al., 2018]

- **Go to next frame when landmarks are finalized**
- **State – Current image region extracted by bounding box**
- **Action – Tracking agent(Movement), Alignment agent(stop/continue iterations)**


- **Reward – Landmark detection accuracy improvements**

Image source: Minghao Guo, Jiwen Lu, and Jie Zhou: Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking(2018).

# Reinforcement Learning [Minghao Guo et al., 2018]

- Evaluated on Category 3 of 300-VW
- Supervised learning stage –
  - Alignment agent trained on 300-W
  - Tracking agent trained on 300-VW
- Reinforcement learning stage –
  - Whole network trained on 300-VW
- Evaluation – Normalized RMSE and cumulative error distribution plots
- DADRL-3D – Trained with 3D data from 3D Menpo. 3D landmarks.

# Comparison

| Approach | Evaluated on dataset | Evaluation metrics |
|---|---|---|
| RNN | 300-VW | RMSE, AUC, FR |
| Two-stream network | TF, 300-VW | RMSE, CED plot |
| LSTM | COFW, Helen, 300-W, 300-VW | RMSE |
| Encoder-decoder network | TF, 300-VW, FM | RMSE |
| Reinforcement Learning | 300-VW | RMSE and CED plot |

$$RMSE_i = \frac{1}{Pd_i} \sum_{p=1}^{P} \sqrt{(x_{i,p} - \hat{x}_{i,p})^2 + (y_{i,p} - \hat{y}_{k,p})^2}$$

# Comparison

| Approach | 300-VW | | TF | | |
|---|---|---|---|---|---|
| | RMSE (68 landmarks) | RMSE (7 landmarks) | RMSE (68 landmarks) | RMSE (7 landmarks) | Runtime (ms) |
| RNN | 6.16 | | | | |
| Two-stream network | 5.59 | | | **2.13** | 33 |
| LSTM | 5.9 | | | | **18** |
| Encoder-decoder network | 5.15 | **5.29** | **2.77** | 2.89 | 40 |
| Reinforcement Learning | **3.09** | | | | |

# Conclusion

- **RNN based approach**
  - FC-RNN to model temporal information.
- **Two-stream network**
  - Spatial and temporal streams
  - Least error on TF dataset for 7 landmarks
- **LSTM approach**
  - LSTM to model temporal information
  - Performs in real-time – 18ms
- **Encoder-Decoder approach**
  - Separate temporal-variant and temporal-invariant factors.
- **Reinforcement Learning approach**
  - Bounding box tracking ⇔ Facial alignment.
  - Least error on the category-3 of 300-VW

# References

1. Jinwei Gu, Xiaodong Yang, Shalini De Mello, Jan Kautz: Dynamic Facial Analysis: From Bayesian Filtering to Recurrent Neural Network(2017).

2. Hao Liu, Jiwen Lu, Jianjiang Feng, Jie Zhou: Two-Stream Transformer Networks for Video-based Face Alignment(2017).

3. Qiqi Hou, Jinjun Wang, Ruibin Bai, Sanping Zhou, Yihong Gong: Face Alignment Recurrent Network(2017).

4. Xi Peng, Rogerio S. Feris, Xiaoyu Wang, Dimitris N. Metaxas: RED-Net: A Recurrent Encoder-Decoder Network for Video-based Face Alignment(2018).

5. Minghao Guo, Jiwen Lu, and Jie Zhou: Dual-Agent Deep Reinforcement Learning for Deformable Face Tracking(2018).

# THANK YOU

## QUESTIONS?