Generative Adversarial Network for Future Hand Segmentation from Egocentric Video

Georgia Tech School of Interactive Computing, College of Computing

ECCV TEL AVIV 2022

Wenqi Jia*, Miao Liu*, James M. Rehg

Georgia Institute of Technology

* Indicates equal contribution

Source & Contact:

Motivation

- Problem of forecasting detailed egocentric hand movements remains unexplored
- Intentional motor behavior indicates how human prepare routine activities
- Applications in Augmented Reality (AR) and Human-Robot Collaboration

Where are the future hands?

Challenges:

- hands are nonrigid and can move fast
  --- Inherent Uncertainty
- head and hand motion are entangled
  --- Drastic Scene Context Change

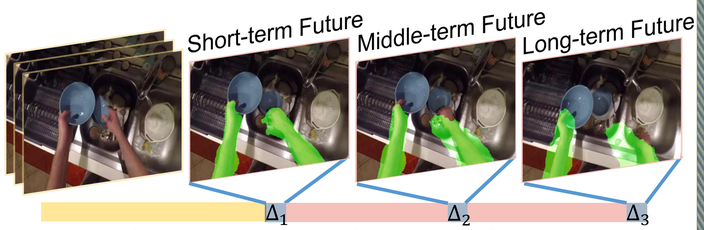
Key Idea:

*Hallucinating **future head motions** for **future hand mask segmentation**!*

Problem Formulation

Input: egocentric video sequence

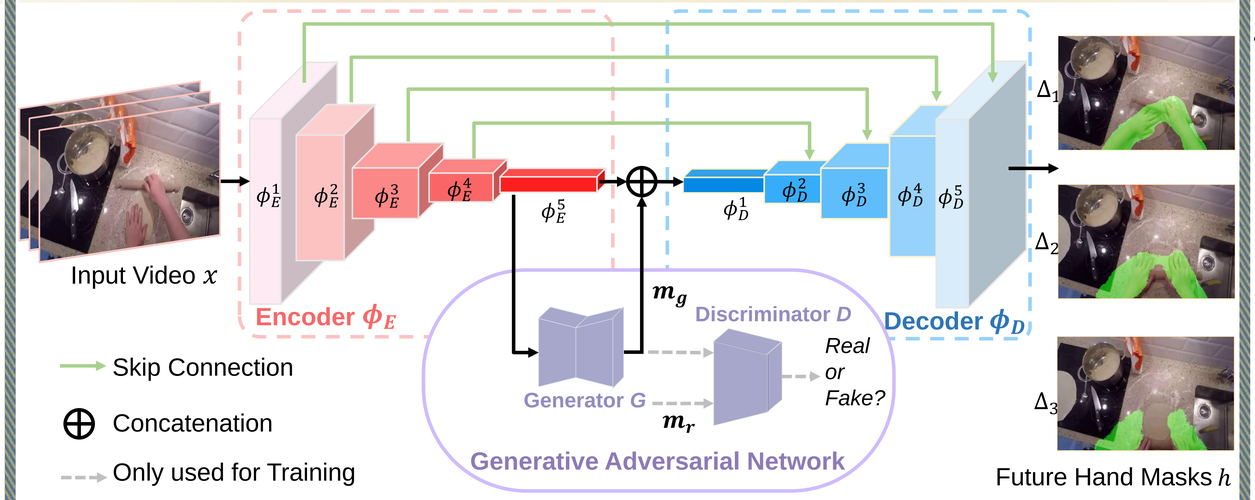
Output: hand masks of future video frames



Contribution

- A novel problem of anticipating future hand masks from egocentric videos
- First generative model that leverages egocentric motion cues for pixel-wise visual anticipation
- Consistent performance improvements on two egocentric video datasets

Method



3DFCN with **skip connection**: $\phi_D^{i+1}(x) = deconv(\phi_D^i(x) + \phi_E^{6-i}(x))$

GAN for **head motion generation**: $m_g = G(\phi_E(x))$

Joint model for **future hand mask** segmentation: $\phi_D^1(x) = deconv(\phi_E^5(x) \oplus m_g)$

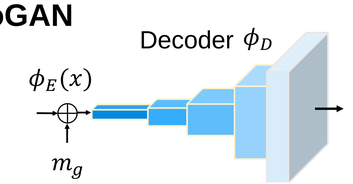
** real future head motion m_r is only used for training*

Learning Objective of EgoGAN

$$\mathcal{L}_d = \underbrace{\mathcal{L}_{ce}(D(m_r), 1) + \mathcal{L}_{ce}(D(m_g), 0)}$$

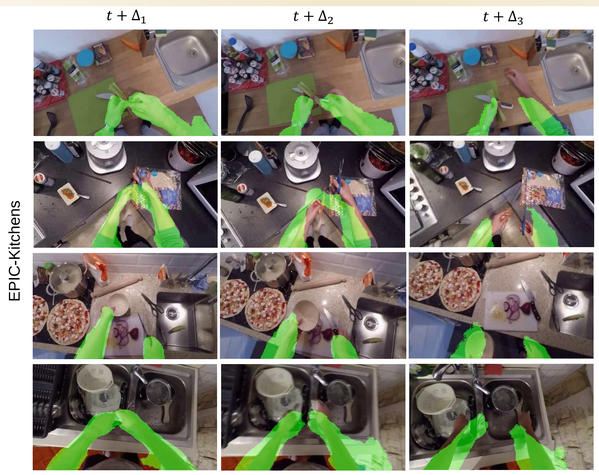
Tell whether the given sample is fake or real

$$\mathcal{L}_g = \underbrace{\mathcal{L}_{ce}(D(m_g), 1)}_{\text{Cross-Entropy loss to fool Discriminator}} + \underbrace{\lambda |m_g - m_r|}_{\text{Regularization for visual consistency}}$$

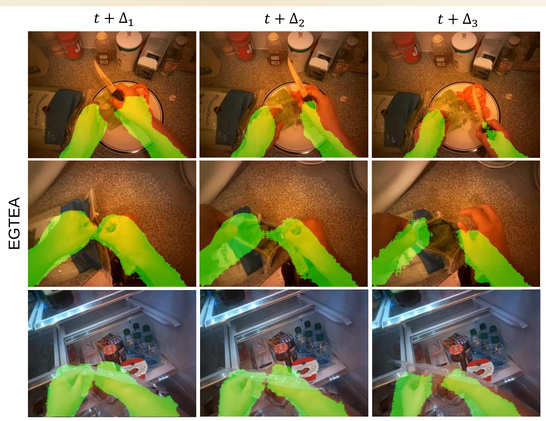


$$\mathcal{L}_{seg} = \underbrace{\mathcal{L}_{ce}(\phi_D(\phi_E(x), m_g), \hat{h})}_{\text{Segmentation Loss}}$$

Visualization



Visualization of dataset EPIC-Kitchens

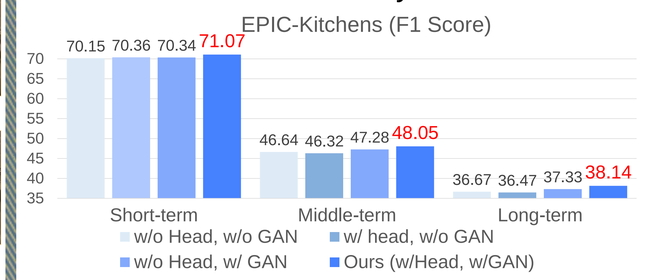


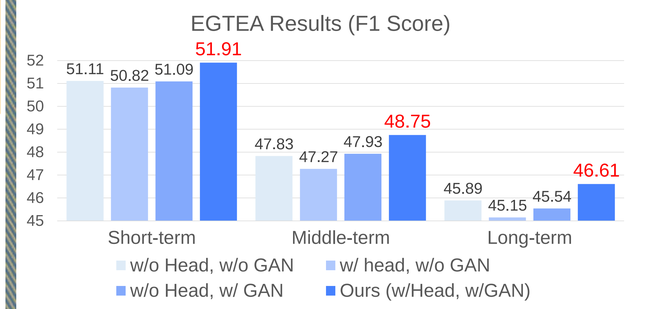
Visualization of dataset EGTEA Gaze+.

* Portions of this project were supported in part by a gift from Facebook.

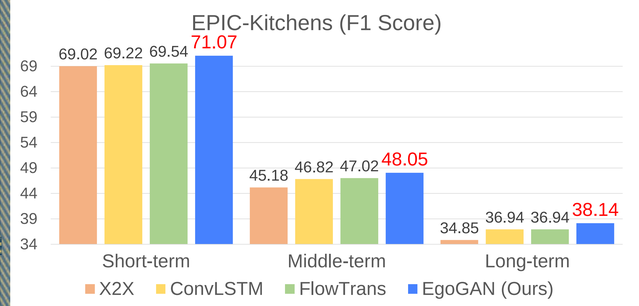
Experiments and Results

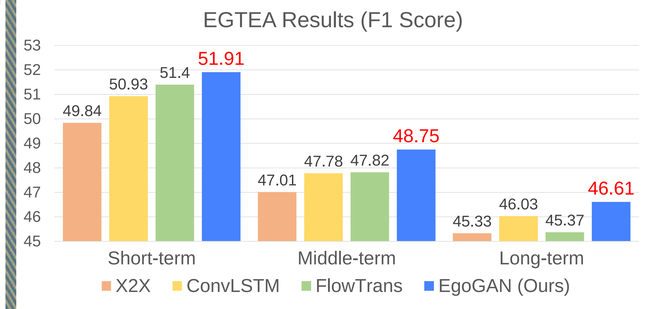
- Model Ablations and Analysis





- Comparison to SOTA Methods





Limitation

Our method does not differentiate left and right hands.