

Hierarchical Long-term Video Prediction without Supervision^[2018]

SMARCLE 신도현

Contents

- 1. Introduction(Abstract)
- 2. Related Works
- 3. Background
- 4. Architecture
- 5. Experiments
- 6. Conclusion

0. Abstract

Abstract

Much of recent research has been devoted to video prediction and generation, yet most of the previous works have demonstrated only limited success in generating videos on short-term horizons. The hierarchical video prediction method by Villegas et al. (2017b) is an example of a state-of-the-art method for long-term video prediction, but their method is limited because it requires ground truth annotation of high-level structures (e.g., human joint landmarks) at training time. Our network encodes the input frame, predicts a high-level encoding into the future, and then a decoder with access to the first frame produces the predicted image from the predicted encoding. The decoder also produces a mask that outlines the predicted foreground object (e.g., person) as a by-product. Unlike Villegas et al. (2017b), we develop a novel training method that jointly trains the encoder, the predictor, and the decoder together without high-level supervision; we further improve upon this by using an adversarial loss in the feature space to train the predictor. Our method can predict about 20 seconds into the future and provides better results compared to Denton and Fergus (2018) and Finn et al. (2016) on the Human 3.6M dataset.

- Previous video prediction
: ground truth annotation of high-level structures



Hierarchical Long-term Video
Prediction without Supervision

- predict 20 seconds into the future
- better results (compared to Denton/Fergus Human datasets)

1. Introduction

1. Introduction

pixel-level noise. However, it is common for the previously mentioned methods to generate quality predictions for the first few steps, but then the prediction dramatically degrades until all of the video context is lost or the predicted motion becomes static.

- Previous methods : prediction dramatically **degrades**

Video Prediction

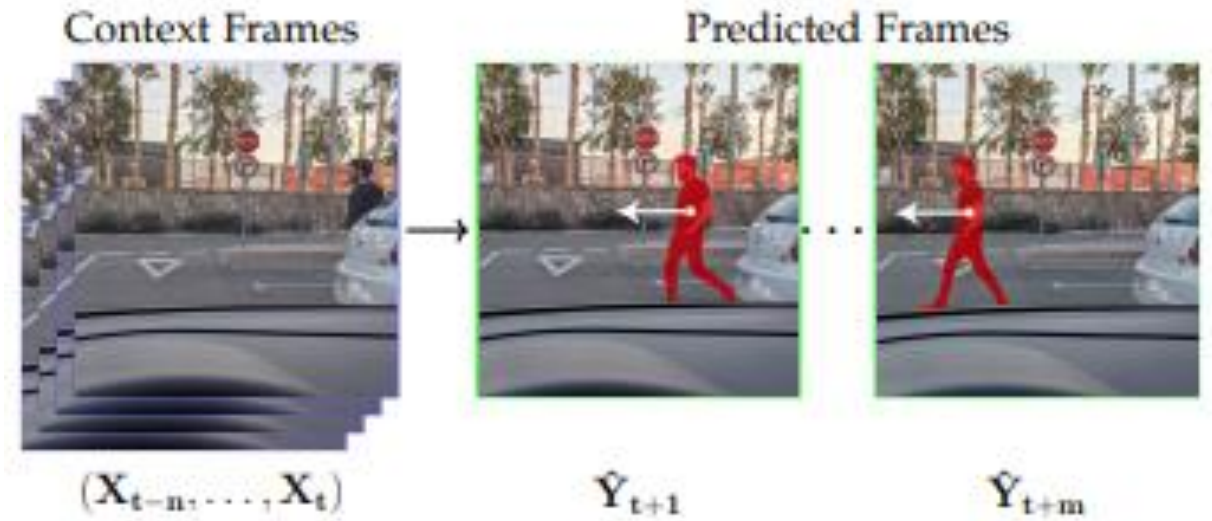
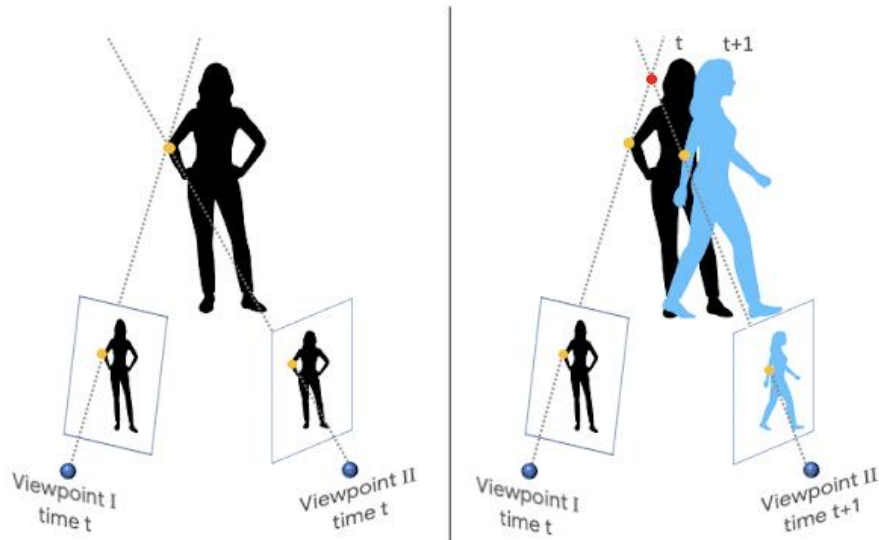
- Input: x_1, \dots, x_t frames
- Output: x_{t+1}, \dots, x_{t+T} frames



Learning to Generate Long-term Future via Hierarchical Prediction(2017)

Application

- Self-driving(Autonomous driving)
- Crime Prediction
- Robot



2. Related Works

- Early : Patch-level prediction

Video prediction을 위한 최초 방법

문제점: 예측 프레임에 가로/세로 줄이 생기는 현상 발생

- Frame-level prediction realistic videos

픽셀의 움직임을 직접 예측하여 해당 움직임이 반영된 frame 생성

Encoder-decoder 구조 사용

- 문제점

예측 거리가 멀어질수록 예측한 프레임을 입력으로 또 다시 프레임을 예측하기 때문에 error 증폭

-> **blurriness / degradation** 현상 발생



Background

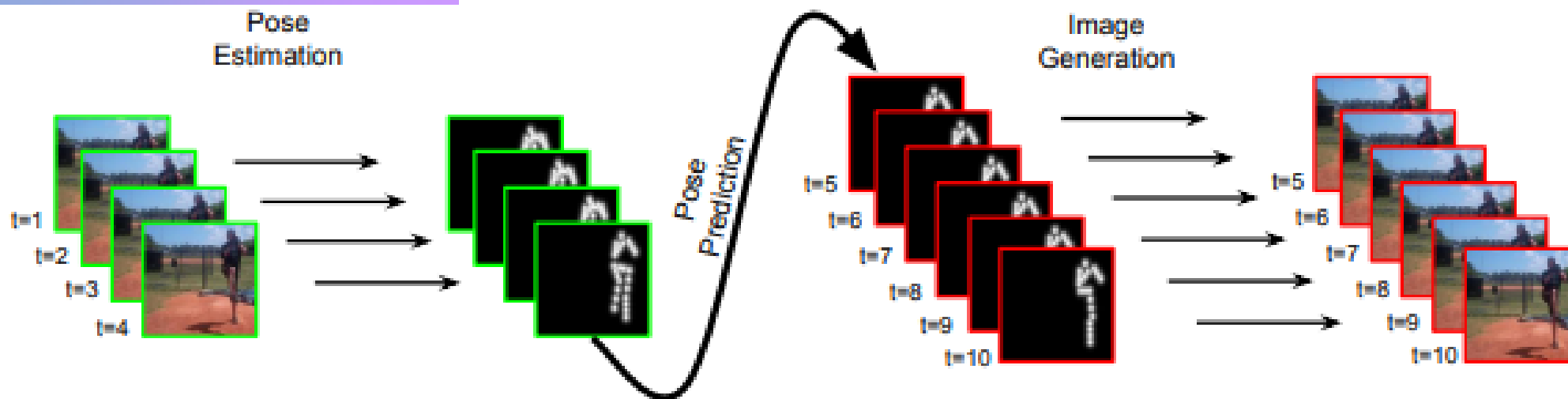
Learning to Generate Long-term Future via Hierarchical Prediction

- Inference Pipeline
: 3 steps



Hierarchical Long-term Video Prediction without Supervision

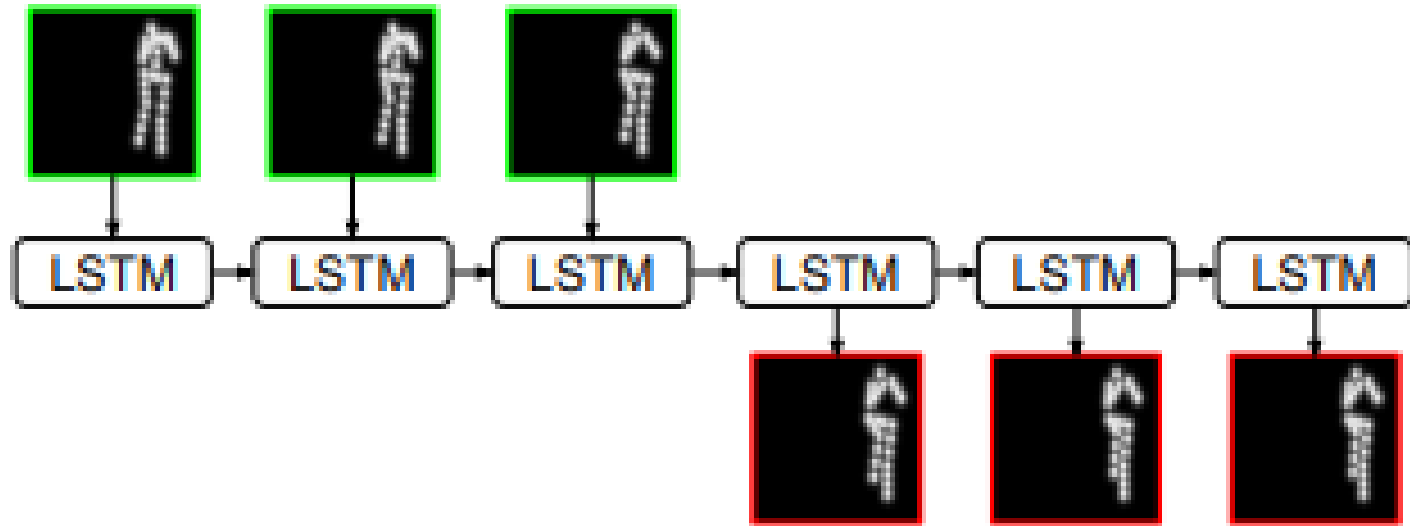
Background



- 1. 이전 프레임 $X_{1:t-1}$ 에 대하여 pose estimation을 수행
-> high-level feature (pose)를 추출

* Pose estimation: 이미지 / 비디오에서 사람의 관절이 어떻게 구성되어 있는지 위치를 측정하고 추정하는 문제

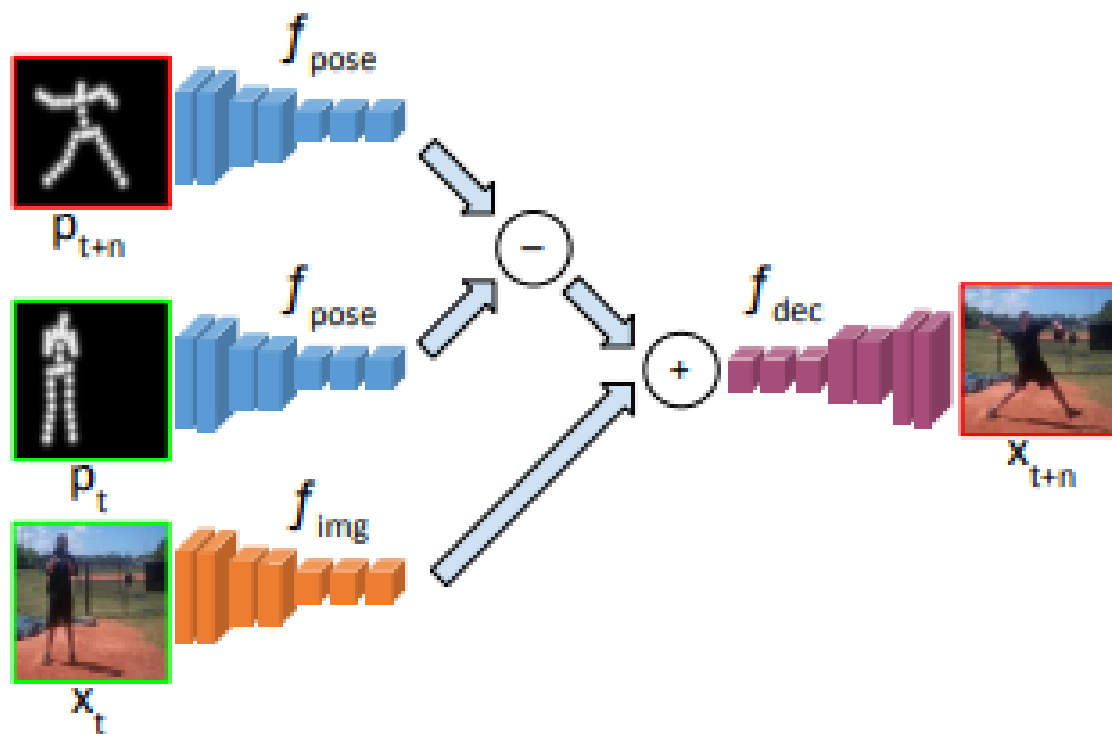
Background



2. 추출된 pose sequence 정보를 LSTM에 입력

→ future pose sequence 를 예측

Background



3. 현재 프레임 X_t
 X_t 의 pose 정보 p_t
예측된 pose 정보 p_{t+n}
을 VAN*에 입력

→ 예측 프레임 X_{t+n} 생성

* VAN : Visual Analogy Network

Background

- Contribution

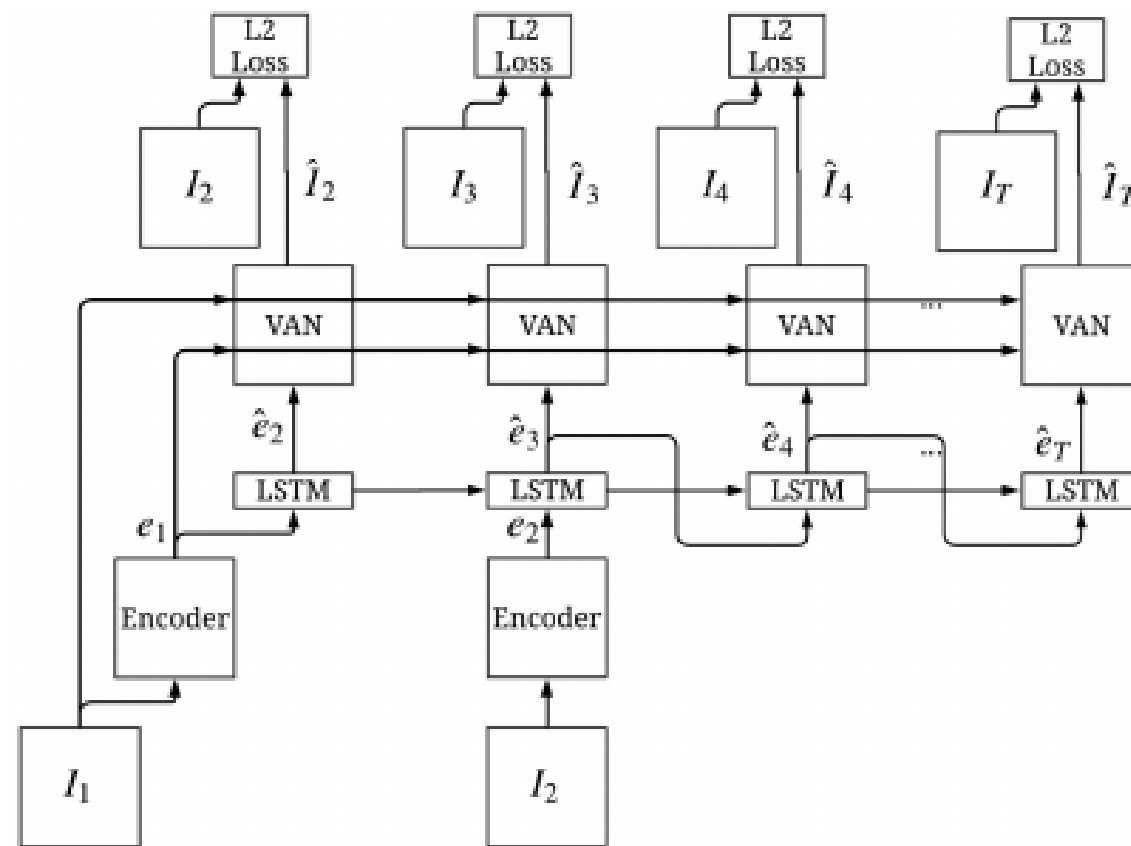
1. 예측 거리가 먼 경우에도 예측된 프레임이 아닌 실제 프레임만을 가지고 프레임을 생성 -> degradation 문제 완화
2. High level feature를 통해 영상의 동적인 모션을 효과적으로 포착 가능

- Limitation

1. Pose estimation을 지도 학습 방식으로 학습 -> labeling 작업 필요
2. Pose estimator, LSTM, VAN을 각각 별도로 학습

4. Architecture

1. Encoder
2. LSTM Predictor
3. VAN(Visual Analogy Network)



Architecture

1. Encoder

현재 프레임 I_t 를 입력으로 하여 general feature vector e_{t-1} 출력

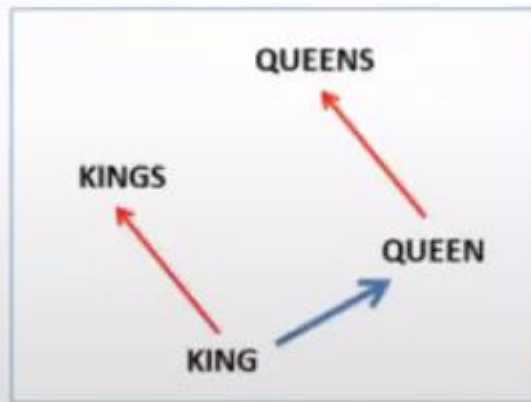
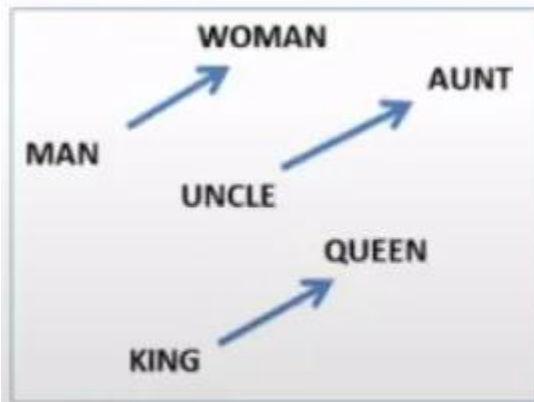
$$2. \text{LSTM Predictor} \begin{cases} [\hat{e}_t, H_t] = LSTM(e_{t-1}, H_{t-1}) & \text{if } t \leq C \\ [\hat{e}_t, H_t] = LSTM(\hat{e}_{t-1}, H_{t-1}) & \text{if } t > C, \end{cases}$$

관찰된 프레임($t \leq C$) : encoder 의 출력 \rightarrow LSTM $\rightarrow \hat{e}_t$ 예측

관찰되지 않은 프레임($t > C$) : LSTM의 출력 \rightarrow LSTM $\rightarrow \hat{e}_t$ 예측

VAN (Visual Analogy Network)

- 이미지에 대한 Analogical reasoning을 구현하기 위한 기법
- Word2Vec 와 같은 word embedding 기법에서 영감을 얻은 것으로, 더하기/빼기와 같은 단순 벡터 연산을 통해 관계를 유추



Queen + Kings - King = Queens

(Mikolov et al., NAACL HLT, 2013)

VAN

$$\bar{I}_t, M_t = VAN(e_1, \hat{e}_t, I_1) = \text{Queen} + \text{Kings} - \text{King} = \text{Queens}$$

$$f_{dec}(f_{enc}(\hat{e}_t) + T(f_{img}(I_1), f_{enc}(e_1), f_{enc}(\hat{e}_t))),$$

$$\hat{I}_t = \bar{I}_t \odot M_t + (1 - M_t) \odot I_1,$$

$$T(x, y, z) = f_{analogy}([f_{diff}(x - y), z]),$$

Training Strategies

- 1. End-to-End Prediction

- 단순 전략으로 LSTM / VAN을 함께 학습 – 실제 프레임과 예측 프레임 간의 L2 loss를 minimize

$$\min(\sum_{t=1}^T L_2(\hat{I}_t, I_t))$$

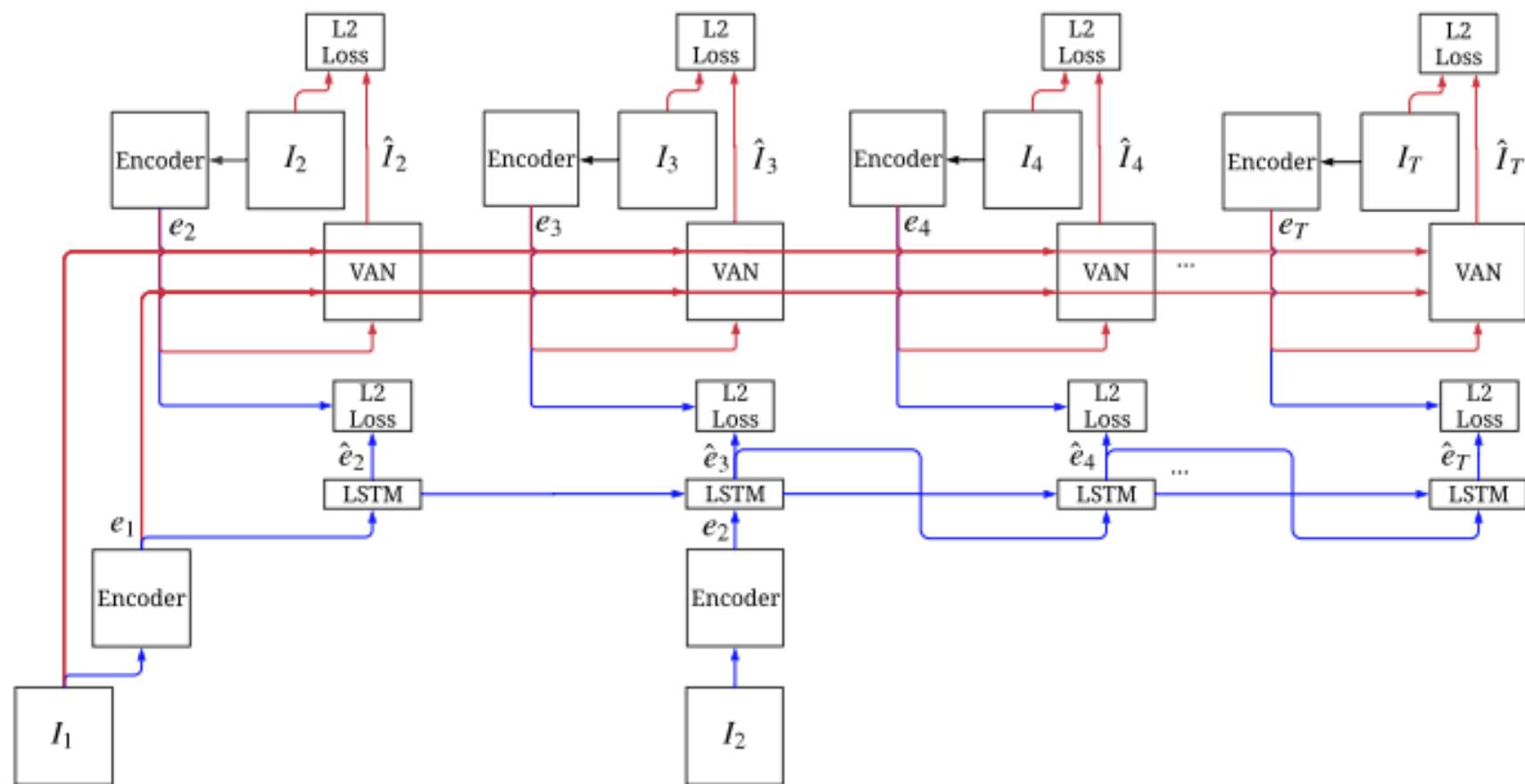
$$* \text{L2 loss} = \sum_{i=1}^n (y_i - f(x_i))^2$$

- 2. EPVA (Encoder Predictor with Visual Analogy)

- Encoder에서 출력한 feature vector와 LSTM이 예측한 feature vector가 같도록 하는 제약을 적용

$$\min(\sum_{t=1}^T L_2(\hat{I}_t, I_t) + \alpha L_2(\hat{e}_t, e_t)),$$




























EPVA



Experiments 1. Toy Dataset

Table 1. Crowd-sourced human preference evaluation on the moving shapes dataset.

Method	Shape has correct color	Shape has wrong color	Shape disappeared
EPVA	96.9%	3.1%	0%
CDNA Baseline	24.6%	5.7%	69.7%

	t=1	t=2	t=3	t=256	t=257	t=258	t=1020	t=1021	t=1022
EPVA									
Finn et al. (2016)									
G.T.									

Experiments 2. Human 3.6M

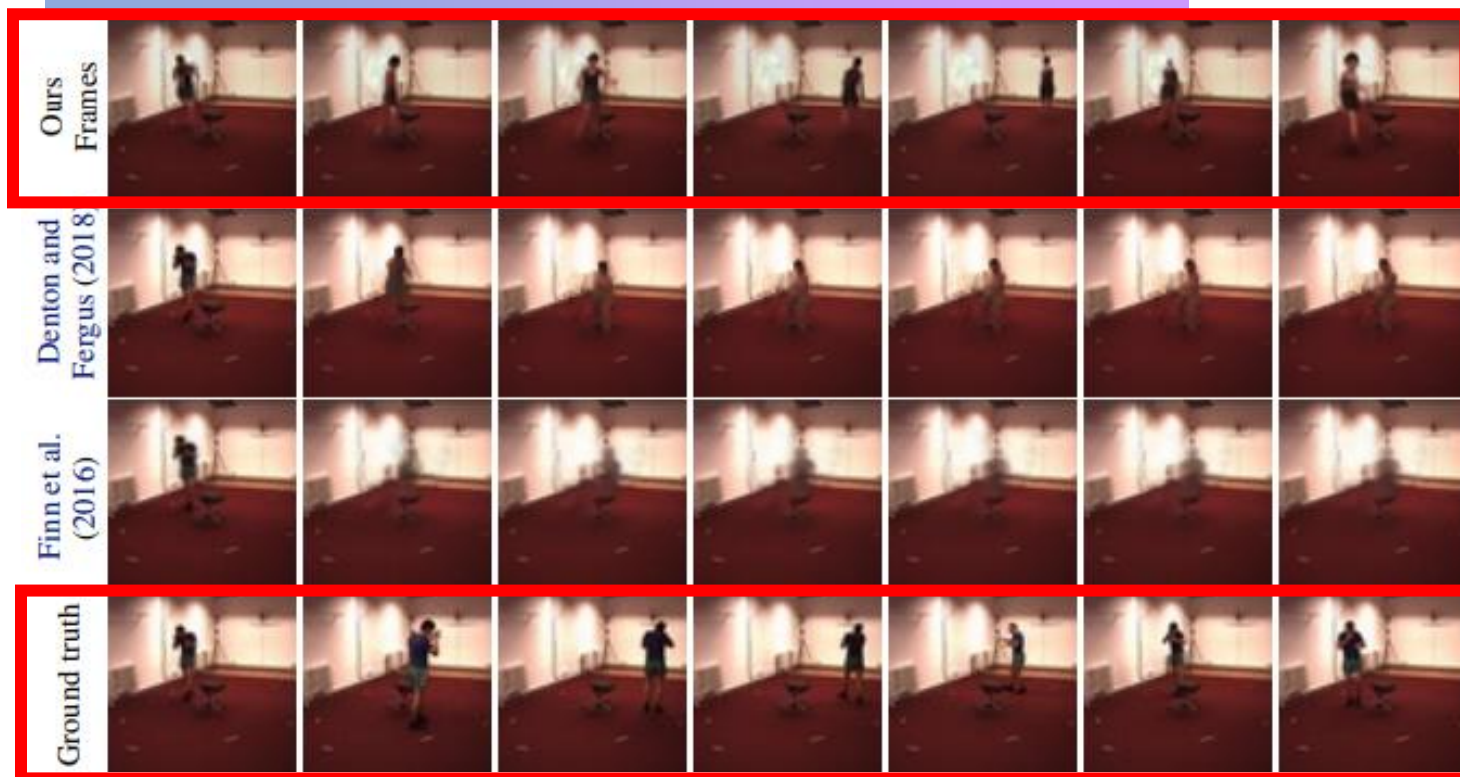
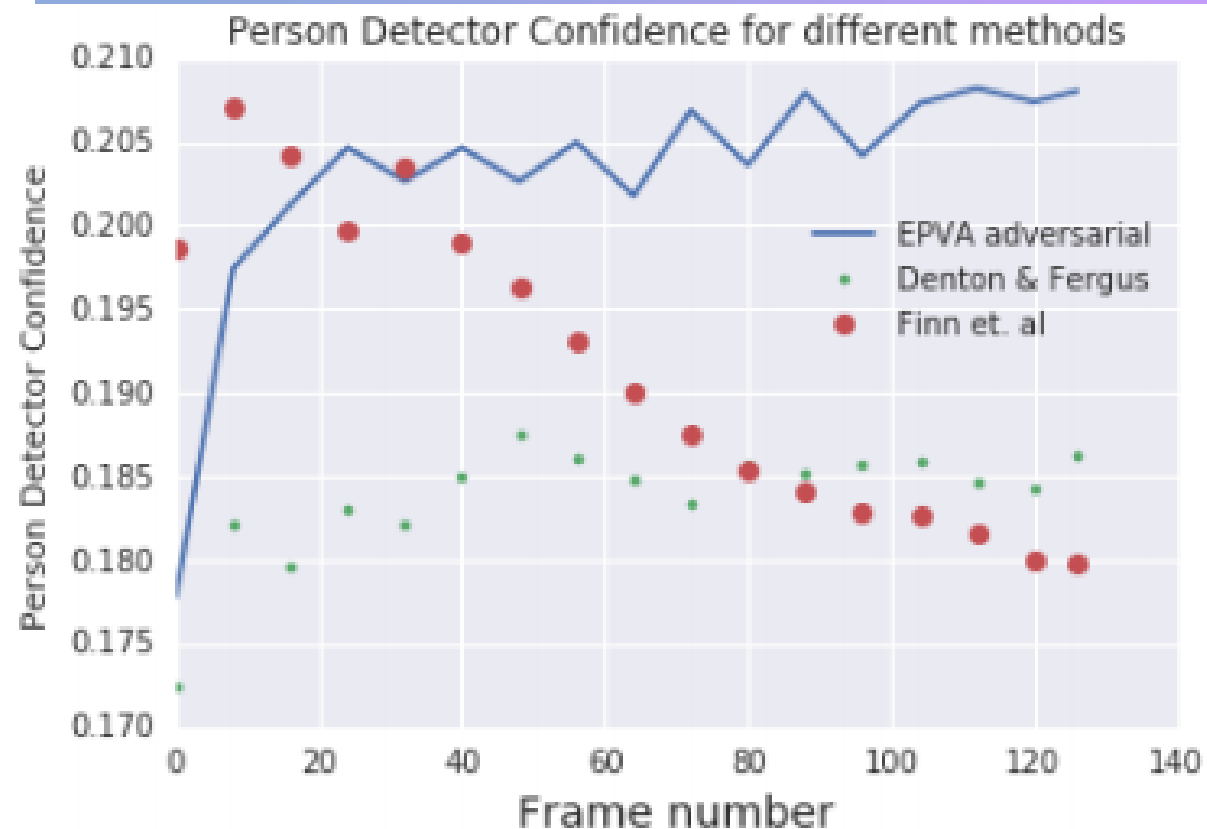


Table 2. Crowd-sourced human preference evaluation on the Human3.6M dataset.

Comparison		Ours is better	Same	Baseline is better
EPVA	1-127 vs Finn et al. (2016) 1-127	46.4%	40.7%	12.9%
EPVA ADV	1-127 vs Finn et al. (2016) 1-127	73.9%	13.2%	12.9%
EPVA ADV	63-127 vs Finn et al. (2016) 1-63	67.2%	17.5%	15.3%
EPVA ADV	5-127 vs Denton and Fergus (2018) 5-127	58.2%	24.0%	17.8%

Experiments 3. Person Detector Evaluation



6. Conclusion & discussion

The contributions of our work are summarized below:

- An unsupervised approach for discovering high-level features necessary for long-term future prediction.
- A **joint training strategy** for generating high-level features from low-level features and low-level features from high-level features simultaneously.
- **Use of adversarial training** in feature space for improved high-level feature discovery and generation.
- Long-term pixel-level video prediction for about 20 seconds into the future for the Human 3.6M dataset.

6. Conclusion

We presented hierarchical long-term video prediction approaches that do not require ground truth high-level structure annotations. The proposed EPVA method has the limitation of the predictions occasionally disappearing, but it generates sharper images for a longer period of time compared to Finn et al. (2016), and the E2E method. By applying adversarial loss in the higher-level feature space, our EPVA ADVERSARIAL method generates more realistic predictions compared to all of the presented baselines including Finn et al. (2016) and Denton and Fergus (2018). This result suggests that it is beneficial to apply an adversarial loss in the higher-level feature space. For future work, applying other techniques in feature space such as the variational method described in Babaeizadeh et al. (2018) could enable our network to generate multiple future trajectories.

Thank you



References

- Learning to Generate Long-term Future via Hierarchical Prediction[2017] (<https://arxiv.org/pdf/1704.05831.pdf>)
- Video Prediction ! Hierarchical Long-term Video Frame Prediction without Supervision paper review(deep-learning Image Team Eungi Hong)