Towards Learning and Generating Audience Motion from Video

Kenneth Chen, Norman I. Badler

Problem

There has been an explosion of interest in creating large-scale shared virtual spaces for multiplayer content as Metaverse experiences Rendering player-controllable avatars in real-time causes networking issues when scaling to thousands of players.

We introduce a human audience video dataset to support applications in deep learning-based 2D video audience simulation, bypassing the need for background 3D virtual humans. The dataset contains hand-annotated labels across 10 different audience classes spanning sports, concerts, and political rallies, etc..

We apply deep learning tasks on this data based on video prediction techniques, and propose a novel method for 2D audience simulations usable as animated texture maps.

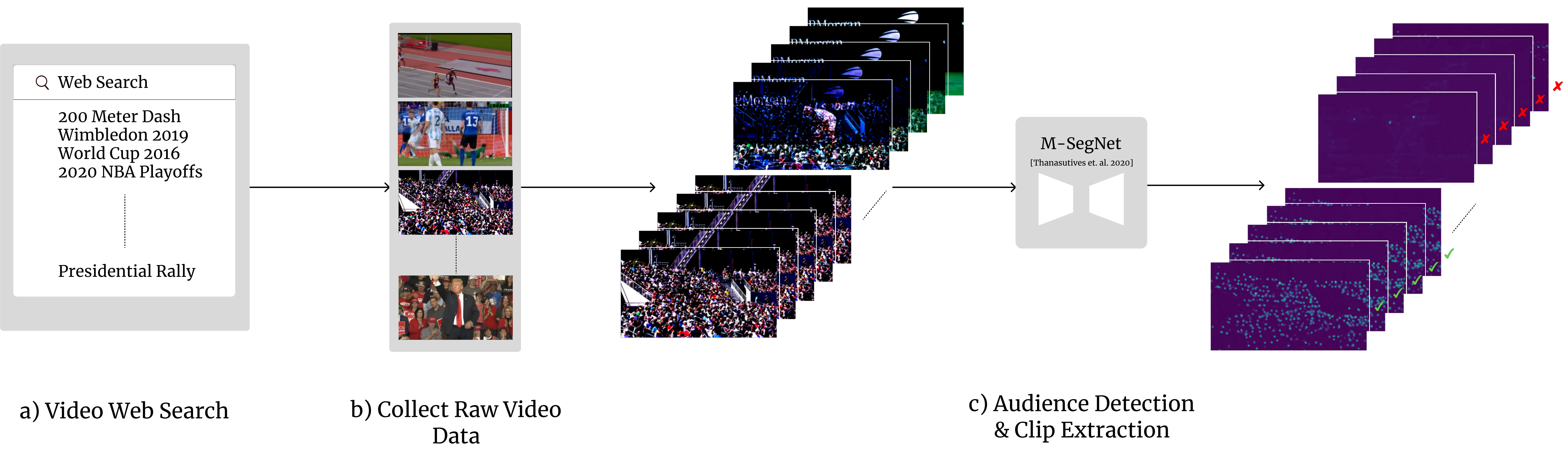
Method

1) Find web videos using an implicit search term (i.e., "Wimbledon 2019 Replay").

2) Feed each frame of a video to a crowd counting deep neural network (M-SegNet pre-trained on the Shanghai Tech Dataset[1], [2]).

3) Accept frames if density surpasses a predefined threshold and crowd count exceeds 60 people.

4) Extract clips which are composed of ≥15 consecutive accepted frames.



Web Search

200 Meter Dash
Wimbledon 2019
World Cup 2016
2020 NBA Playoffs

Presidential Rally

M-SegNet
[Thanasutives et. al. 2020]

a) Video Web Search

b) Collect Raw Video Data

c) Audience Detection & Clip Extraction

Our Approach

We collect a diverse dataset of audience videos spanning 10 classes by using implicit queries and a crowd counting method.

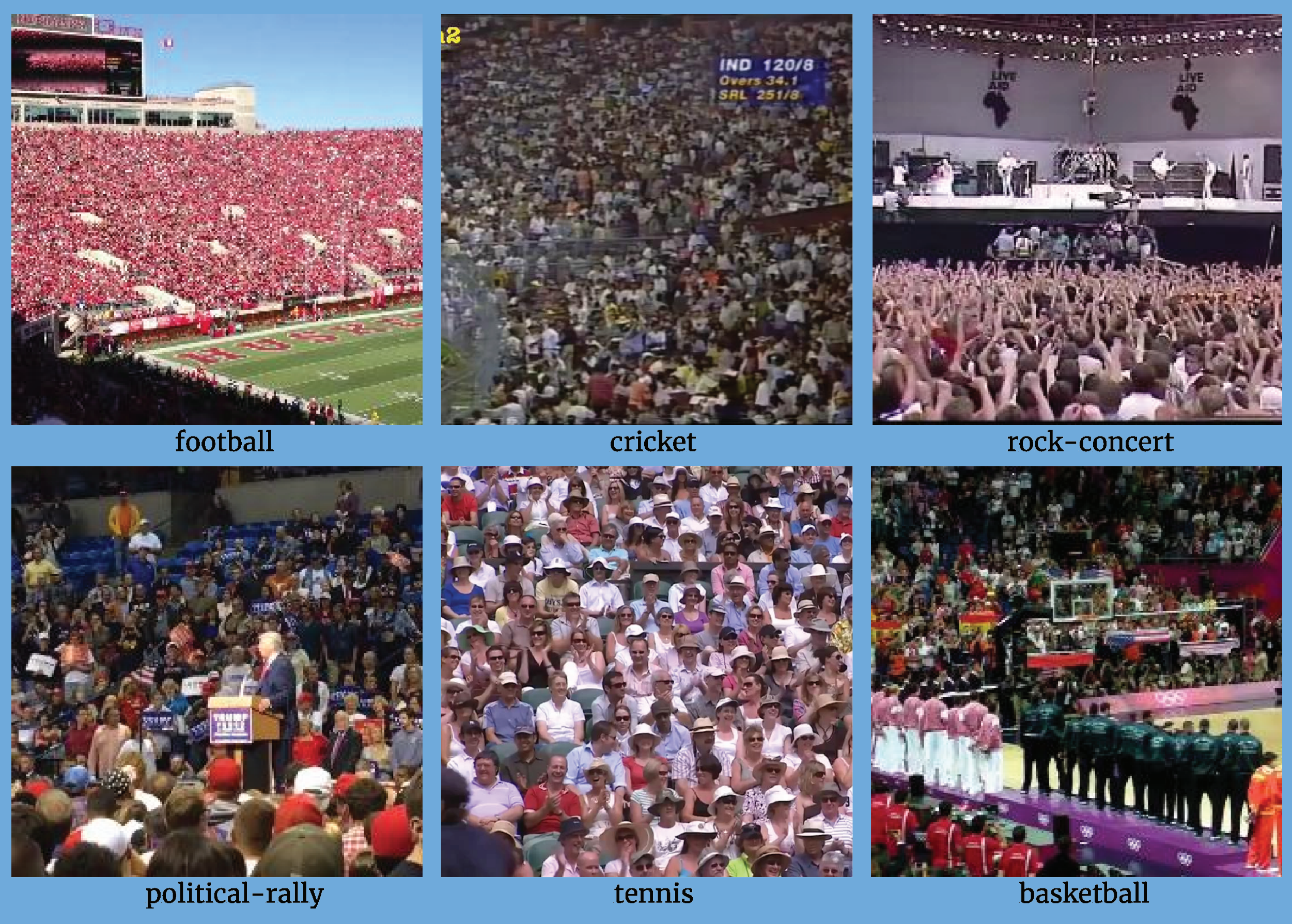


Fig 1) Example frames from out dataset.

Audiences exhibit distinct behavioral patterns in different contexts. We hypothesized that we might be able to automatically differentiate audience types to identify their context, without overtly observing that context itself. We study this idea by training a classifier to predict audience video class (Fig 2).

We finally train a model [3] to generate audience motion given 4 input frames.

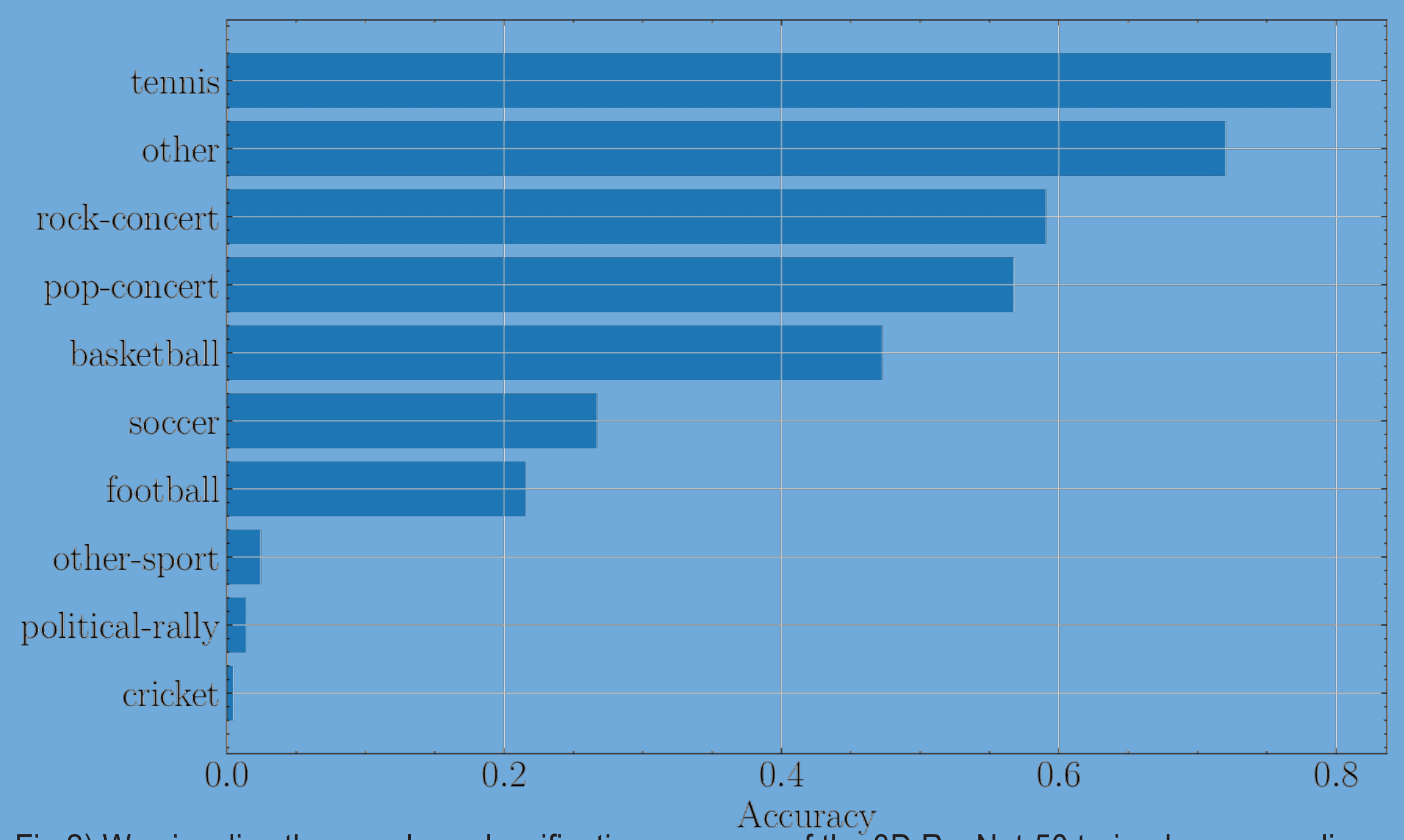


Fig 2) We visualize the per-class classification accuracy of the 3D ResNet-50 trained on our audience video dataset.

Results

We adapt our dataset to a video prediction task for audience behavior synthesis. Our method renders realistic background audience video, conditioned on several input frames, at a similar speed regardless of audience size.

We display qualitative prediction results of training a VRNN conditioned on 4 input frames and predicting 10 future frames in Table 1. Example video prediction results can also be found in our supplementary video.

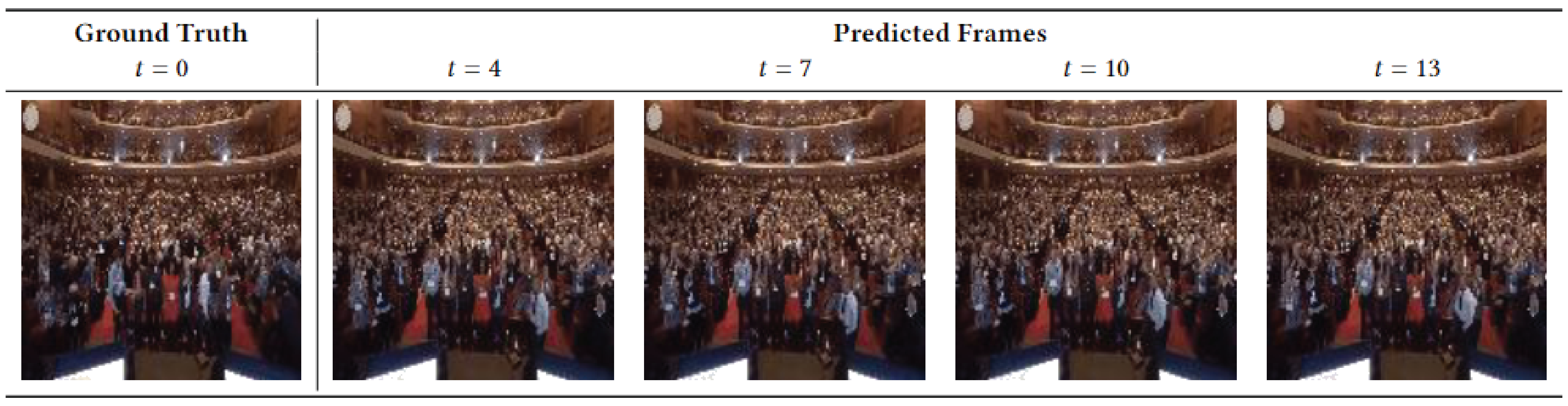


Table 2) An example prediction result for a single scene trained using a VRNN conditioned on 4 input frames.

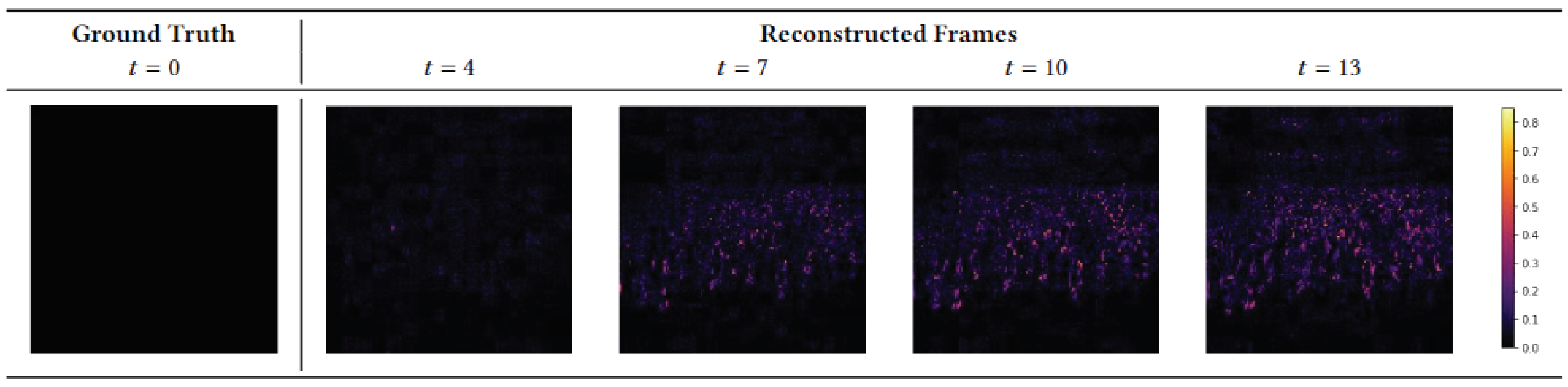


Table 2) We illustrate the absolute difference between ground truth and reconstructed video. Times t=0-3 are input frames to the model.

References

[1] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-Image Crowd Counting via Multi-Column Convolutional Neural Network, In CVPR 2016.
[2] P. Thanasutives, K. Fukui, M. Numao, and B. Kijsirikul.
Encoder-Decoder Based Convolutional Neural Networks with Multi-Scale-Aware Modules for Crowd Counting. In 2020 ICPR
[3] L. Castrejon, N. Ballas, and A. Courville. 2019. Improved Conditional VRNNs for Video Prediction. In The IEEE International Conference on Computer Vision (ICCV)