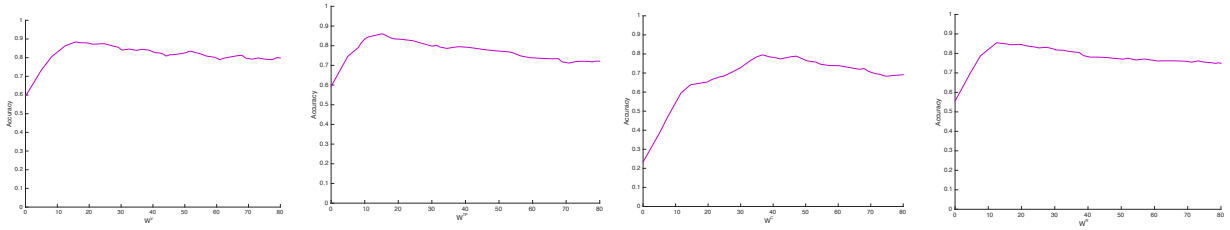
Predicting the Future: A Jointly Learnt Model for Action Anticipation: Supplementary Materials

Harshala Gammulle Simon Denman Sridha Sridharan Clinton Fookes
Image and Video Research Lab, SAIVT, Queensland University of Technology (QUT), Australia
{pranali.gammule, s.denman, s.sridharan, c.fookes}@qut.edu.au

1. Hyperparameter Evaluation

Hyper parameters of Eq. 14, W^V, W^{TP}, W^C, and W^R are evaluated experimentally using the validation set of UCF-101 [2] split 1, where we change the respective weight value while holding the rest of the parameters constant. The accuracy plots against the respective hyper parameters are presented in Fig. 1. Based on these observations we set hyper-parameters w^V, w^{TP}, w^c, w^R to 25, 20, 43 and 15, respectively.



(a) W^V (b) W^{TP} (c) W^C (d) W^R

Figure 1. Hyperparameter evaluations

2. Network Architectures

2.1. Generator, Discriminator and Classifier

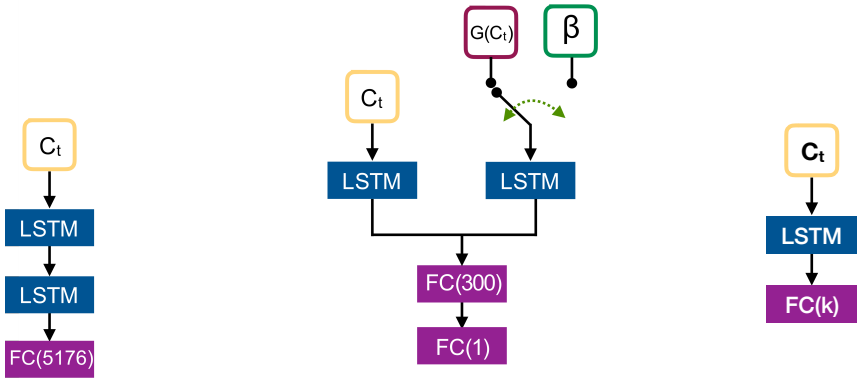
The architectures of the generator, discriminator and the classifier are presented in Fig. 2. For all LSTMs 300 hidden units are used.

2.2. Ablation models

The architectures of the non-GAN based ablation models from (a) to (c) are visually illustrated in Fig. 3 while in Fig. 4 illustrates the architectures of the non-GAN based ablation models with future representation generators (models from (d) to (g)).

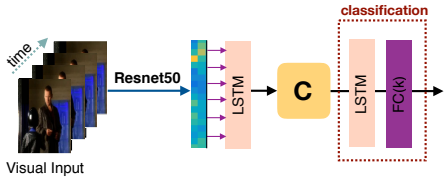
3. Qualitative Results

Synthesised visual and temporal representations from the proposed AA-GAN method along with ground truth information for 3 sample videos from the TV human interaction dataset [1] are given in Figs. 5, 6 and 7. In the heat maps (rows 2-6) yellow denotes high values and blue denotes low values. Considering the ground truth visual and temporal representations shown, it is clear that salient aspects of the input frames, such as humans, objects and their interactions, have been identified; and considering the synthesised representations the proposed model has been able to accurately anticipate these semantics allowing the proposed AA-GAN model to anticipate the future.

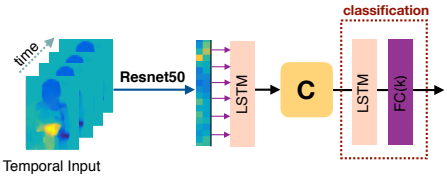


(a) G^V/G^{TP} (b) D^V/D^{TP} (c) Classifier

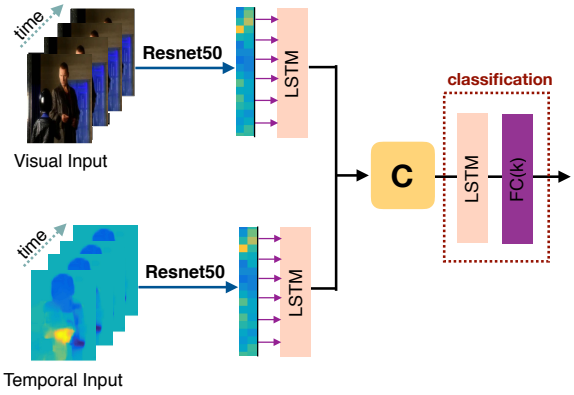
Figure 2. The architectures of generator (G^V/G^{TP}), discriminator (D^V/D^{TP}) and the classifier.



(a) $\eta^{C,V}$: A model trained to classify using the context feature extracted only from the visual input stream (V)



(b) $\eta^{C,TP}$: A model trained to classify using the context feature extracted only from the temporal input stream (TP)



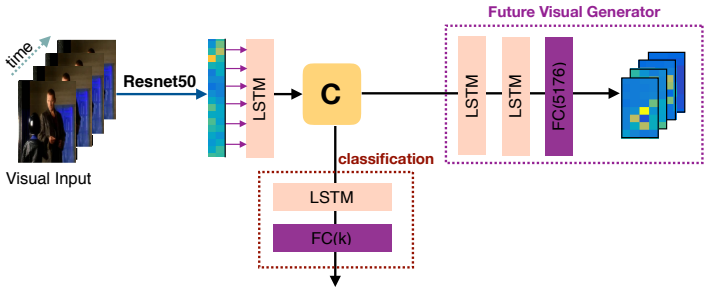
(c) $\eta^{C,(V+TP)}$: A model trained to classify using the context feature extracted from both visual and temporal streams

Figure 3. The architectures of the Non-GAN based ablation models: from (a) to (c).

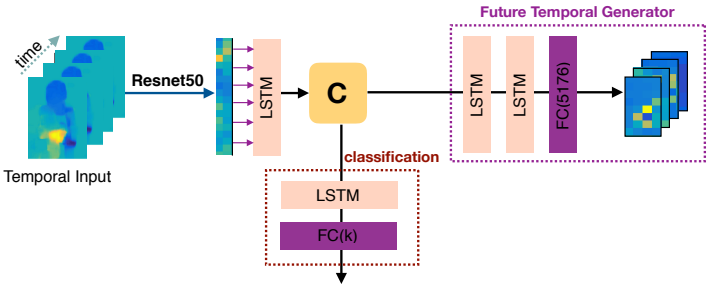
References

[1] Alonso Patron-Perez, Marcin Marszalek, Ian Reid, and Andrew Zisserman. Structured learning of human interactions in tv shows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2441–2453, 2012. 1, 4, 5, 6

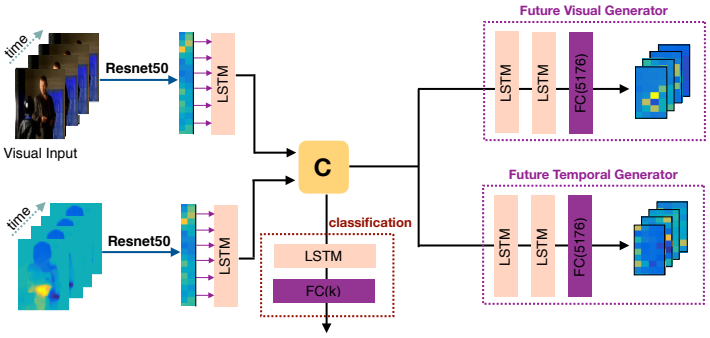
[2] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1



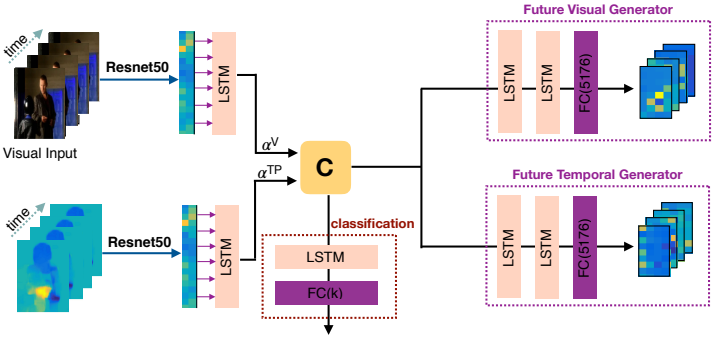
(d) $\eta^{C,V} + G^V$: Model with the future visual representation generator (G^V) and fed only with the visual input stream to train the classifier.



(e) $\eta^{C,TP} + G^{TP}$: Model with the future temporal representation generator (G^{TP}) and fed only with the temporal input stream to train the classifier.

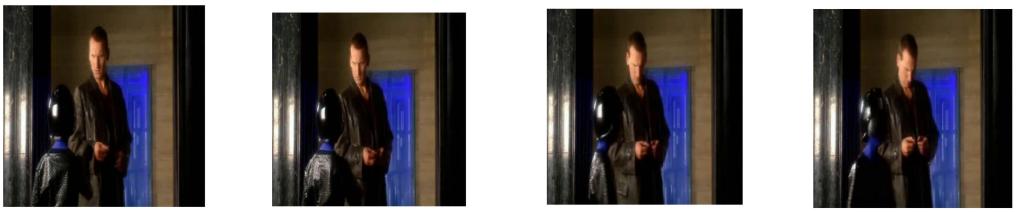


(f) $\eta^{C,(V+TP)} + G^V + G^{TP}$: Model with both future visual and temporal representation generators (G^V and G^{TP}) and fed with both visual and temporal input streams to train the classifier.



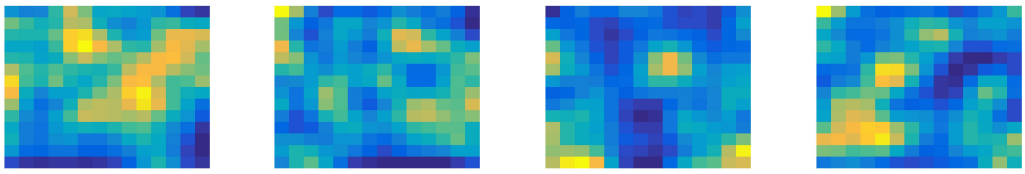
(g) $\eta^{C,(V+TP)} + G^V + G^{TP} + \mathrm{Att}$: Model with both future visual and temporal representation generators (G^V and G^{TP}) and fed with both visual and temporal input streams to train the classifier through attentions.

Figure 4. The architectures of the Non-GAN based ablation models with future representation generators: models from (d) to (g) .



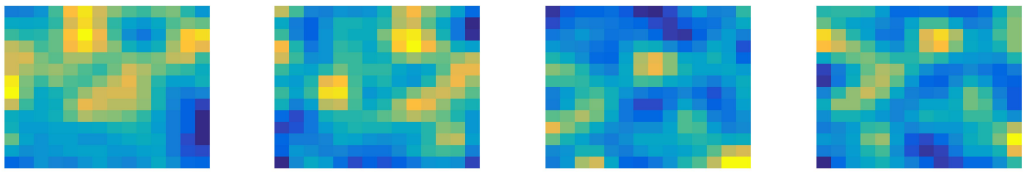
Frame 21 Frame 26 Frame 31 Frame 36

Ground truth future frames.



Frame 21 Frame 26 Frame 31 Frame 36

Ground truth future visual representation.

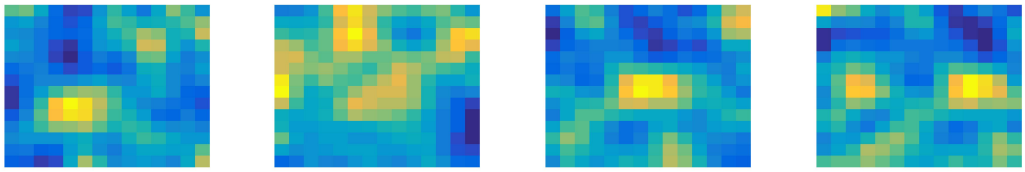


Frame 21 Frame 26 Frame 31 Frame 36

Predicted future visual representation.

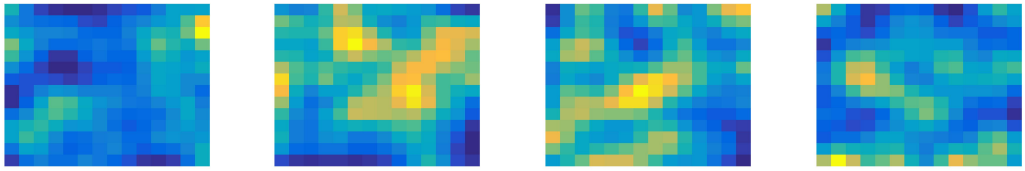
Frame 21 Frame 26 Frame 31 Frame 36

Ground truth optical flow for future frames.



Frame 21 Frame 26 Frame 31 Frame 36

Ground truth future temporal representation.



Frame 21 Frame 26 Frame 31 Frame 36

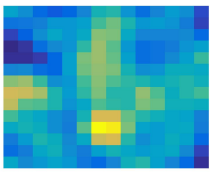
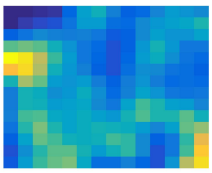
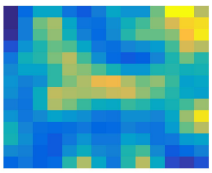
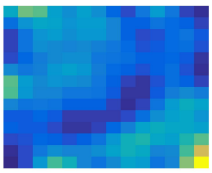
Predicted future temporal representation

Figure 5. Qualitative results for a sample video from TV human interaction dataset [1].



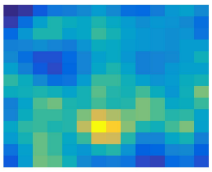
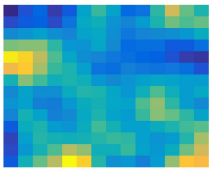
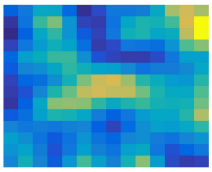
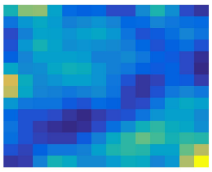

Frame 21 Frame 26 Frame 31 Frame 36

Ground truth future frames.

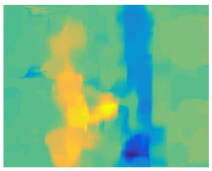
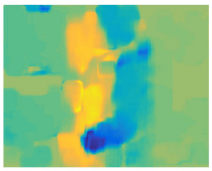
Frame 21 Frame 26 Frame 31 Frame 36

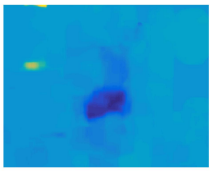
Ground truth future visual representation.

Frame 21 Frame 26 Frame 31 Frame 36

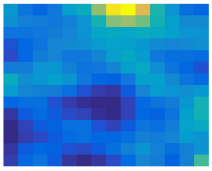
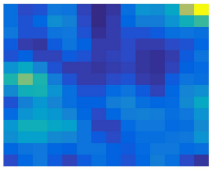
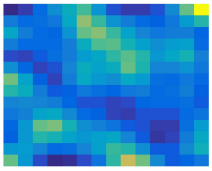
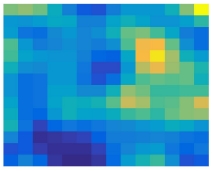
Predicted future visual representation.



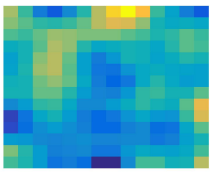
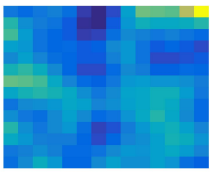
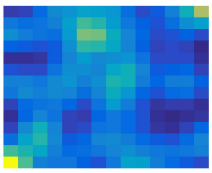
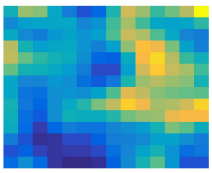
Frame 21 Frame 26 Frame 31 Frame 36

Ground truth optical flow for future frames.

Frame 21 Frame 26 Frame 31 Frame 36

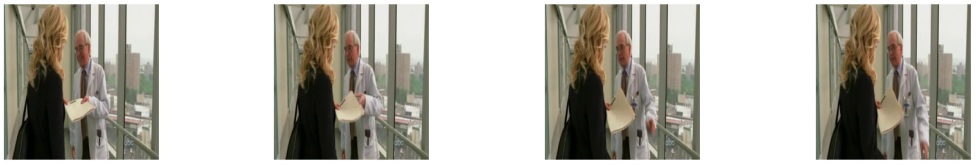
Ground truth future temporal representation.

Frame 21 Frame 26 Frame 31 Frame 36

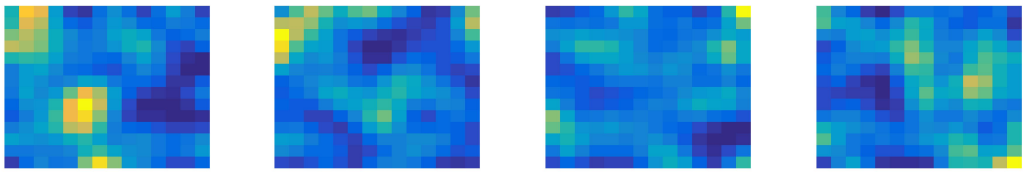
Predicted future temporal representation

Figure 6. Qualitative results for a sample video from TV human interaction dataset [1].



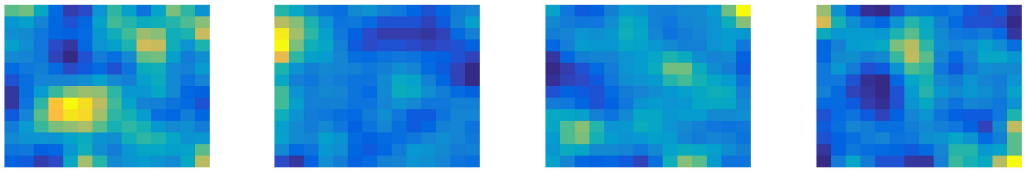
Frame 21 Frame 26 Frame 31 Frame 36

Ground truth future frames.



Frame 21 Frame 26 Frame 31 Frame 36

Ground truth future visual representation.

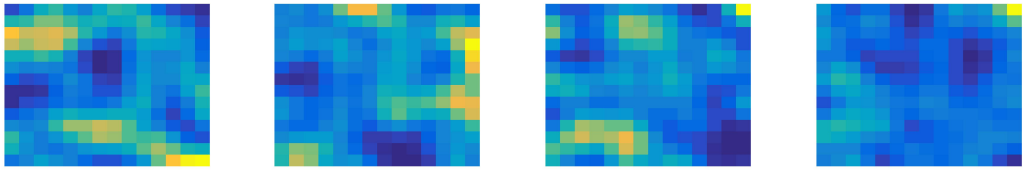


Frame 21 Frame 26 Frame 31 Frame 36

Predicted future visual representation.

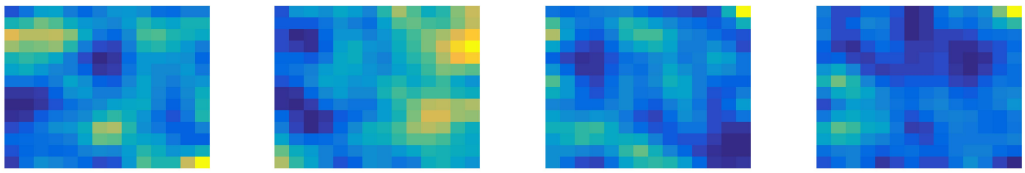
Frame 21 Frame 26 Frame 31 Frame 36

Ground truth optical flow for future frames.



Frame 21 Frame 26 Frame 31 Frame 36

Ground truth future temporal representation.



Frame 21 Frame 26 Frame 31 Frame 36

Predicted future temporal representation

Figure 7. Qualitative results for a sample video from TV human interaction dataset [1].