Motion and Context-Aware Audio-Visual Conditioned Video Prediction

Yating Xu
xu.yating@u.nus.edu
Conghui Hu
conghui@nus.edu.sg
Gim Hee Lee
gimhee.lee@nus.edu.sg

Department of Computer Science
National University of Singapore
Singapore

1 Qualitative Ablation Study

We conduct qualitative analysis to verify the effectiveness of MME and CAR.

Effectiveness of audio-motion correlation in MME. Fig. 1 presents the qualitative comparion of different motion estimation networks. 'GT Flow' is the ground truth optical flow. The baseline 'V' predicts optical flow without audio. 'V+Recall' *Recall* from audio motion memory, but does not condense it. 'MME' is our full model in stage 1 with both *condense* and *recall*. It is clear to see that only using visual modality ('V') to predict motion has the worst prediction. Although 'V+Recall' captures the correct motion in the short term, it fails at longer future due to its large-size motion memory. In contrast, MME has a better balance in both short-term and long-term predictions due to the condensed memory.

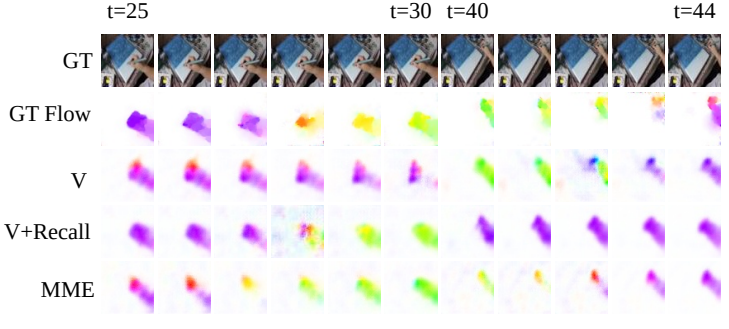

Figure 1: Qualitative comparison of different motion estimation networks.

Effectiveness of Context-Aware Refinement. Fig. 2 presents the qualitative comparison of different image refinement networks. 'Unet' only uses U-Net for refinement in stage 2. 'Unet+ContextEnc' adds a context encoder but does not perform affine transformation over

the context feature. It shows that CAR is able to provide and aggregate the feature of hand and pen to make the moving object more concrete.

Figure 2: Qualitative comparison of different refinement networks.

2 Qualitative Video Results

We provide a video ('video results.mp4') to show the prediction by our model and the ground truth in the supplementary zip file. The white denotes ground truth and the red denotes prediction. Video 1 and video 2 show 15 past frames and 30 future frames. Video 3 and 4 show 5 past frames and 20 future frames. Our prediction is temporally smooth and consistent with the ground truth audio and visual frames.

3 Architecture Details

We present the detailed architecture of our proposed framework.

Motion Encoder. The motion encoder consists of four convolutional layers with kernel size of 4×4 and stride of 2. The number of channels is $\{64, 64, 128, 128\}$. Each convolutional layer is followed by Batch Normalization and Leaky ReLU, except for the last layer which is followed by Batch Normalization and Tanh.

Motion Decoder. The motion decoder consists of four deconvolution layers with kernel size of 4×4 and stride of 2. The number of channels is $\{128, 64, 64, 2\}$. Each convolutional layer is followed by Batch Normalization and Leaky ReLU, except for the last layer.

Audio Encoder. The audio encoder consists of five convolutional layers with kernel size of 4×4 for the first four layers and with kernel size of 2×2 for the last layer. The number of channels is $\{64, 128, 256, 512, 128\}$. Each convolutional layer is followed by Batch Normalization and Leaky ReLU, except for the last layer which is followed by Batch Normalization and Tanh.

Condense and Recall. The operators Condense and Recall use attention mechanism as shown in Fig 3. Specifically, Operator Condense takes the motion memory MM as input and projects it into the query ($\mathbf{Q}$), key ($\mathbf{K}$) and value ($\mathbf{V}$) via three separate 1D convolutional layers with kernel size of 1, respectively. In Recall, the visual feature is sent into a 2D convolutional layer with kernel size of 1×1 and then is flattened along spatial dimensions as $\mathbf{Q}$. The condensed memory is sent into two separate 1D convolutional layers with kernel size of 1 and projected as $\mathbf{K}$ and $\mathbf{V}$, respectively.

Context Encoder. The context encoder has four blocks, where each block consists of two convolutional layers followed by Batch Normalization and ReLU. A max pooling layer is inserted between adjacent blocks.

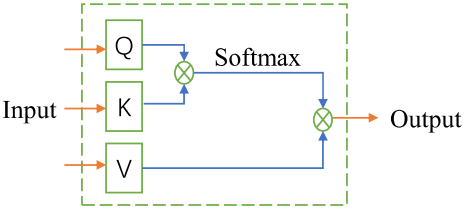

Figure 3: Illustration of the attention mechanism.