

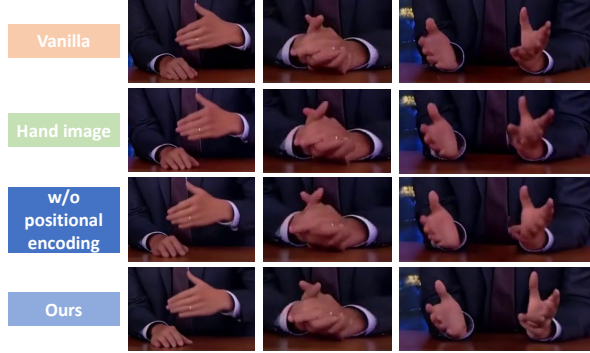


Figure 1: Ablation experiments of hand module.

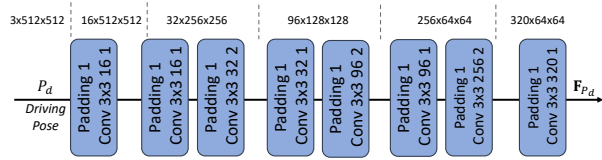


Figure 2: The pose encoder structure consists of 8 convolutional layers, where (Conv  $3 \times 3$  16 1) represents the kernel size is  $3 \times 3 \times 16$ , and the stride is 1. Except for the last convolutional layer, each convolution is followed by GroupNormalization and the activation function silu.



Figure 3: The generation results of adjacent frames.

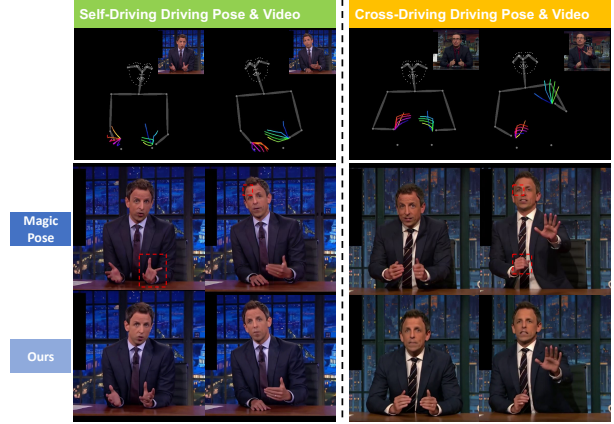


Figure 4: The results of MagicPose and our ShowMaker.

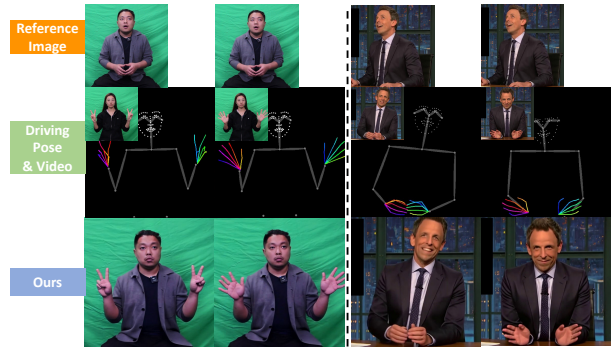


Figure 5: The generated results when the body shape and pose difference is obvious.

| IDs    | The number of clips in the training set | The number of clips in the test set |
|--------|-----------------------------------------|-------------------------------------|
| ID1    | 49                                      | 5                                   |
| ID2    | 57                                      | 6                                   |
| ID3    | 56                                      | 6                                   |
| ID4    | 65                                      | 7                                   |
| ID5    | 47                                      | 5                                   |
| ID6    | 64                                      | 7                                   |
| ID7    | 55                                      | 6                                   |
| Seth   | 3306                                    | 100                                 |
| Oliver | 6746                                    | 200                                 |

Table 1: Dataset details. ID1-7 are datasets recorded indoors. Seth and Oliver belong to the talkshow dataset. All training and test sets do not overlap.