

Unsupervised Image-to-Video Adaptation via Category-aware Flow Memory Bank and Realistic Video Generation

Anonymous Authors

1 OVERVIEW

In this document, we conduct ‘Integrating domain adaptation technique’ experiments with Full Model on $E \rightarrow H$ and $S \rightarrow U$. Moreover, we present the computational information of our method. Finally, we visualize the generated video frames. The source codes are also provided.

2 ADDITIONAL EXPERIMENTS

Further evaluations of integrating domain adaptation (DA) techniques. We further investigate the performance of combining our Full Model with DA methods on $E \rightarrow H$ and $S \rightarrow U$ benchmarks and the results are presented in Tab. 1. The results demonstrate that the integration with typical DA methods, significantly enhances the model’s recognition capability. The performance not only achieves new state-of-the-art results on both benchmarks but also attains comparable performance to supervised learning on $S \rightarrow U$ benchmark. The improvement is attributed to the effective mitigation of distribution discrepancies by employing classical DA methods. Our experimental results further emphasize the high compatibility between our approach and DA methods.

Table 1: Accuracies on $E \rightarrow H$ and $S \rightarrow U$, averaged over 3 random trials.

method	$E \rightarrow H$	$S \rightarrow U$
Full Model	77.4	97.3
Full Model + DAN	77.4	98.0
Full Model + MCC	78.4	99.2
Full Model + BNM	79.2	99.3

Analysis of computational costs. As shown in Table. 2, we present the computational information of video generation and our Full Model. Although the video generation requires more computational resources, it only needs to run once, and the generated frames can be used throughout the training process.

Table 2: Computational information

component	Params(M)	GFLOPs	Train/Inference time(s)
Video generation	266.13	371.85	- / 0.17
Full Model	24.59	65.62	0.16 / 0.13

3 VISUALIZATIONS OF GENERATED FRAMES

In this section, we visualize generated video frames for certain categories including ‘Hource Race’, ‘Running’, ‘Climb’ and ‘Ski-jet’. The visualization results are shown in Fig. 1. We exhibit two groups of generated frames for each category. The first column

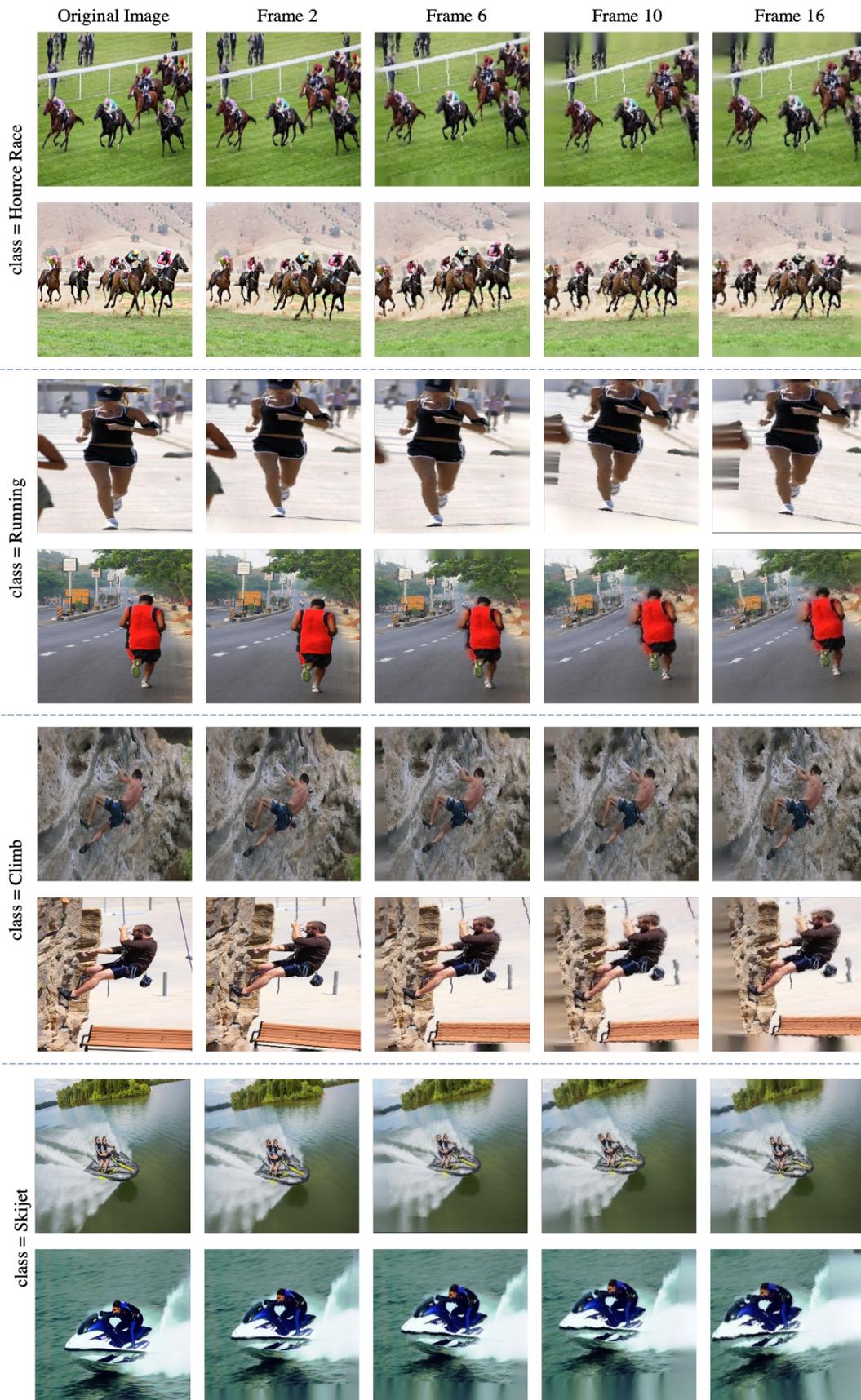
represents the input still images from source domain, and the subsequent columns show the sampled second, sixth, tenth and sixteenth frames, respectively.

Through visualization, we can observe that by simulating the arbitrary movements of the camera in 3D space, we can obtain more realistic video frames. These generated video frames are beneficial for learning a high-performance spatial-temporal model.

4 SOURCE CODES

We provide the source codes for training the spatial-temporal model. Details can be referred to readme.md in codes/.

117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174



175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232

Figure 1: Visualizations of generated frames.