

Supplementary Materials: Joint-Motion Mutual Learning for Pose Estimation in Video

Anonymous Authors

In this supplementary file, we first present additional results and analyses on the effects of different hyperparameters (Section 1). Then, we show more visual comparison results between state-of-the-art method TDMI [1] and our method JM-Pose on challenging scenes (Section 2). Finally, we demonstrate more visual results of our JM-Pose on benchmark datasets (Section 3).

1 SUPPLEMENTARY EXPERIMENTS

In this section, we conduct more experiments on the PoseTrack2017 validation set to study the effects of different hyperparameters, including the ratio α of the trade-off loss (in the main paper Eq. 16) and temporal span δ of supporting frames (in the main paper xxx).

Ablation study on loss ratio α . In the Subsection 3.3 of the main paper, we utilize α to balance the two loss terms in the total loss (in the main paper Eq. 16). We try utilizing different ratios and present the results in Table 1. We observe that performance drops from 86.4 mAP ($\alpha = 0.25$) to 85.5 mAP ($\alpha = 0.01$). This is consistent with our intuition that a smaller ratio hinder the effect of the proposed information orthogonality objective. Additionally, the model performance tends to plateau as α increases and the best results (86.4 mAP) are obtained at $\alpha = 0.25$.

Table 1: Ablation study on the loss ratio α .

Ratio α	$\alpha = 0.01$	$\alpha = \mathbf{0.25}$	$\alpha = 1$
Mean	85.5	86.4	85.9
Declines	↓ 0.9	-	↓ 0.5

Ablation study on temporal span δ . Furthermore, we investigate the impact of using different temporal spans δ , which constrains the number of supporting (neighboring) frames. We experiment with three different temporal span settings: (a) $\delta = 1$, (b) $\delta = 2$, and (c) $\delta = 3$. From the results in Table 2, we observe that mAP improves with increasing δ , namely 85.3 for $\delta = 1$, 86.4 for $\delta = 2$, and 86.4 for $\delta = 3$. This is in line with our intuition that more temporal information can be obtained by employing more supporting (neighboring) frames, which provides more cues for pose estimation. When $\delta = 2$ or $\delta = 3$, the performance of pose estimation tends to be similar, we adopt a temporal span of $\delta = 2$.

Table 2: Ablation study on the temporal span δ .

Span δ	(a) $\delta = 1$	(b) $\delta = 2$	(c) $\delta = 3$
Mean	85.3	86.4	86.4
Declines	↓ 1.1	-	-

2 MORE VISUAL COMPARISON RESULTS

As shown in Fig. 1, we visualize the comparison results for challenging scenes, such as *self-occlusions*, *rapid-motion*, and *pose-occlusions*. We try to investigate the robustness of our method. In figure 1, we observe that TDMI [1] fails to estimate some hard-to-detect joints since TDMI ignores the implicit joint information encoded in initial heatmaps. In contrast, our method JM-Pose displays more accurate pose estimation results in the heavily crowded and challenging scenarios.

3 MORE VISUAL RESULTS

As shown in Fig. 2, we present more visual results of our method on challenging datasets. From this figure, we observe that our method achieves accurate and robust pose estimation in various scenes.

REFERENCES

- [1] Runyang Feng, Yixing Gao, Xueqing Ma, Tze Ho Elden Tse, and Hyung Jin Chang. 2023. Mutual information-based temporal difference learning for human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17131–17141.

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

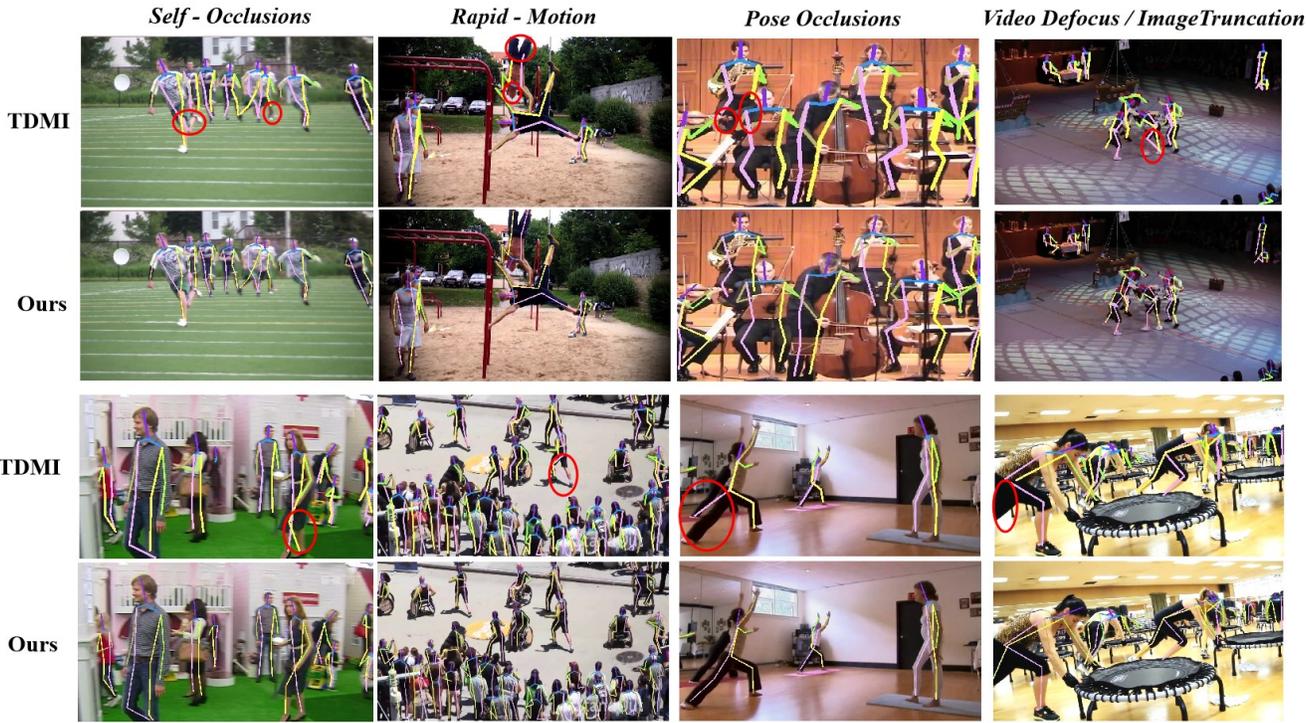


Figure 1: More visual comparisons of detection results obtained from state-of-the-art TDMI and our method JM-Pose on challenging scenes. Inaccurate detections are highlighted with red circles.



Figure 2: More visual results of our JMI-Pose on benchmark datasets. Challenging scenes such as fast motion or occlusions are involved.