

Supplementary Materials: Decoupling Heterogeneous Features for Robust 3D Interacting Hand Poses Estimation

Anonymous Authors

1 INTRODUCTION

This supplemental material provides:

- More qualitative results to showcase the effectiveness and practicality of our method. These results consist of in-the-wild images or videos captured using a smartphone or from the internet in Sec. 1.1,
- More detailed qualitative analysis of how the proposed decoupling strategy works in Sec. 1.2,
- Details of the Feature Interaction between Joints and the Image in Sec. 1.3,
- More ablation experiment. in Sec. 1.4.

Please note that all the notation and abbreviations in this supplementary material are consistent with those in the main manuscript.

1.1 More qualitative results

In this section, we provide more qualitative results on images captured by a smartphone Fig. 2 and sourced from the Internet Fig. 3. Our model is only trained on the InterHand2.6M dataset. We compare our method with the previous state-of-the-art model-free work [2]. We obtained the result by utilizing a hand detector to extract the hand region from an in-the-wild image while preserving the aspect ratio. Our method achieves robust results even in cases of severe self-similarity compared to [2] (For more qualitative results, please refer to the video demo).

1.2 More qualitative Analysis

More heatmaps are provided to illustrate the effectiveness of the decoupling strategy further in Fig. 1. Due to the entanglement of the two types of features, the baseline method struggles to precisely focus on the positions of the joints in the presence of severe self-similarity. After incorporating our decoupling strategy, the two types of features mutually enhance by leveraging visual and spatial cues. As a result, both the position and appearance features can localize the positions of the joints in the image.

1.3 Details of the Feature Interaction between Joints and the Image

In this section, we provide more mathematical details about the implementation of feature interaction between joints and the image. In the main paper, we fuse the relationships between position and appearance and use these relationships to guide the enhancement of two types of features:

$$\begin{aligned} \mathbf{Q}_t &= \text{Concat}(\mathbf{J}_t^a, \mathbf{J}_t^p), \mathbf{K}_t = \mathbf{V}_t = \text{Concat}(\mathbf{F}_t^a, \mathbf{F}_t^p), \\ \mathbf{J}_{t+1}^a, \mathbf{J}_{t+1}^p &= \text{Split}(\text{Attn}(\mathbf{Q}_t, \mathbf{K}_t, \mathbf{V}_t)). \end{aligned} \quad (1)$$

where $\text{Attn}(q, k, v)$ denotes multi-head attention [4]. Split and Concat represent the operations of separating and concatenating along the last dimension, respectively. Assuming the number of heads in the multi-head attention is 1, the attention mechanism

Table 1: Ablation study on InterHand2.6M [3].

ID	MPJPE (mm)↓		
	Single	Two	All
Best model	7.35	9.82	8.67
w/o Multi-scale	7.54	9.84	8.77
w/o Identity Info	7.56	9.86	8.78
w/o All	7.62	9.88	8.82

first applies linear transformations to the position and appearance features using weight matrices. For brevity of exposition, we assume that the weight matrices are identity matrices similar to the proof in [5]. We keep the notation for the position and appearance features unchanged after the identity mapping. Then, the attention mechanism performs similarity calculations as:

$$\mathbf{A} = \text{Softmax}(DP(\left[\mathbf{J}_t^a, \mathbf{J}_t^p\right], \left[\mathbf{F}_t^a, \mathbf{F}_t^p\right]^T)). \quad (2)$$

where $DP(\mathbf{M}_1, \mathbf{M}_2)$ denotes Dot Product computation representing the pairwise dot product operation between the row vectors of matrix \mathbf{M}_1 and the column vectors of matrix \mathbf{M}_2 . $\mathbf{A} \in \mathbb{R}^{2J \times (H_t W_t)}$ is the normalised attention map. For simplification, we disregard constants. We can rewrite it as:

$$\begin{aligned} \mathbf{A} &= \text{Softmax}(DP(\mathbf{J}_t^a, \mathbf{F}_t^a) + DP(\mathbf{J}_t^p, \mathbf{F}_t^p)) \\ &= \text{Softmax}(\mathbf{A}_a + \mathbf{A}_p) \end{aligned} \quad (3)$$

where the differences in constants are disregarded. This equation represents the extraction of respective relationships from each feature and the fusion of them. Afterwards, the relationships are used to aggregate each feature.

1.4 Ablation Study

In our best model, we incorporate multi-scale features and part segmentation to improve performance. Multi-scale features allow the model to capture information at different scales. Part segmentation, on the other hand, helps provide identity information for individual pixels, improving the model’s ability to classify different hand parts as mentioned in [1]. As indicated in Table 1, both designs resulted in a performance improvement of approximately 0.1mm. If both designs are removed, the MPJPE drops 0.15mm.

REFERENCES

- [1] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J Black, and Otmar Hilliges. 2021. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 1–10.
- [2] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. 2023. A2J-Transformer: Anchor-to-Joint Transformer Network for 3D Interacting Hand Pose Estimation from a Single RGB Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8846–8855.
- [3] Gyeongsik Moon, Shouo-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020. InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image. In *European Conference on Computer Vision*. 548–564.

117	[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N	175
118	Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.	176
119	<i>Advances in neural information processing systems</i> 30 (2017).	177
120		178
121		179
122		180
123		181
124		182
125		183
126		184
127		185
128		186
129		187
130		188
131		189
132		190
133		191
134		192
135		193
136		194
137		195
138		196
139		197
140		198
141		199
142		200
143		201
144		202
145		203
146		204
147		205
148		206
149		207
150		208
151		209
152		210
153		211
154		212
155		213
156		214
157		215
158		216
159		217
160		218
161		219
162		220
163		221
164		222
165		223
166		224
167		225
168		226
169		227
170		228
171		229
172		230
173		231
174		232

233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290

291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348

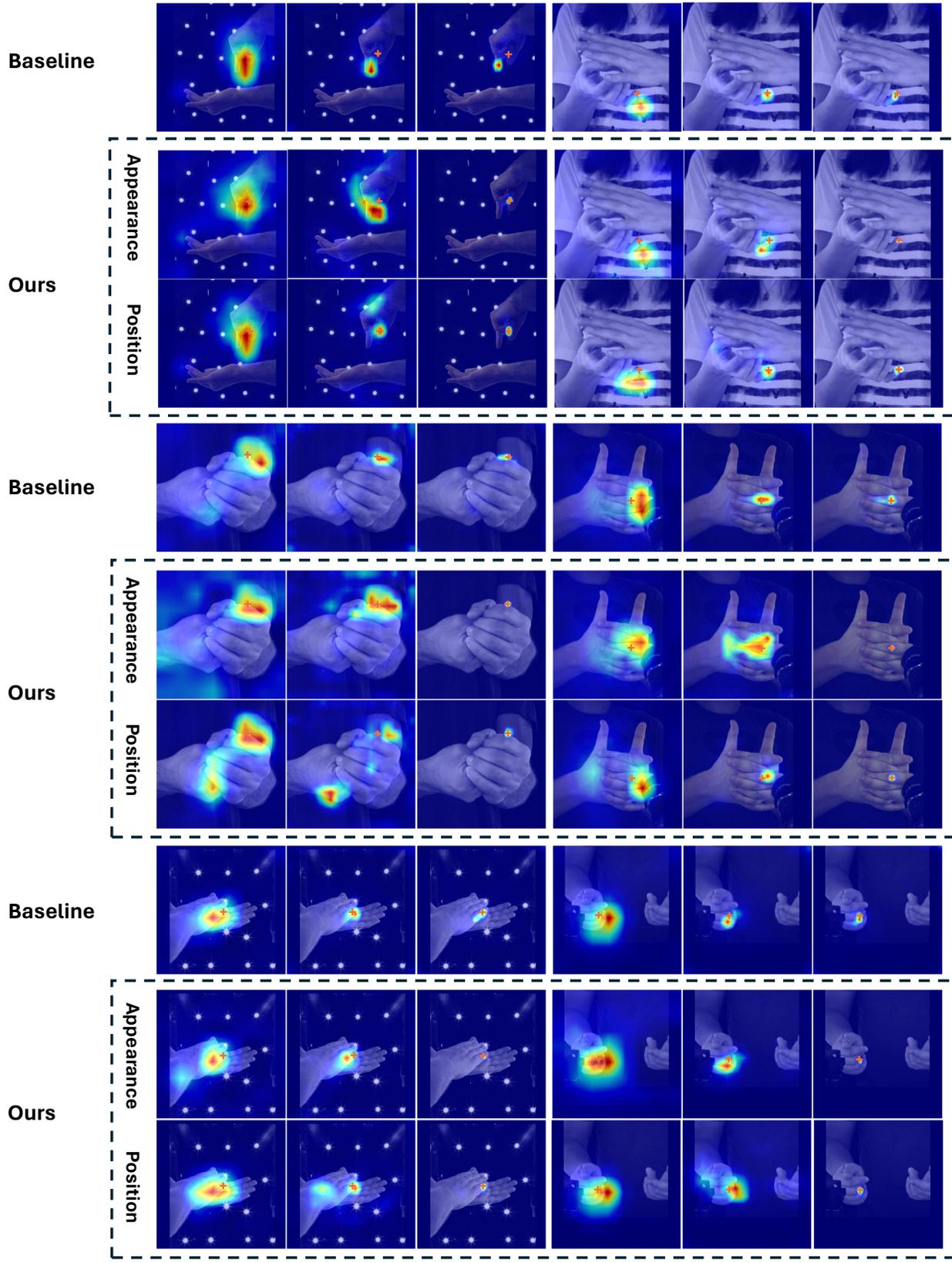


Figure 1: Illustration of the 3-iteration attention maps from the feature interaction between joints and image. The heatmaps from the baseline method without the decoupling strategy are presented first, followed by the heatmaps of the appearance and position from our method. The ground truth joint positions are marked with an orange cross.

349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406

407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464

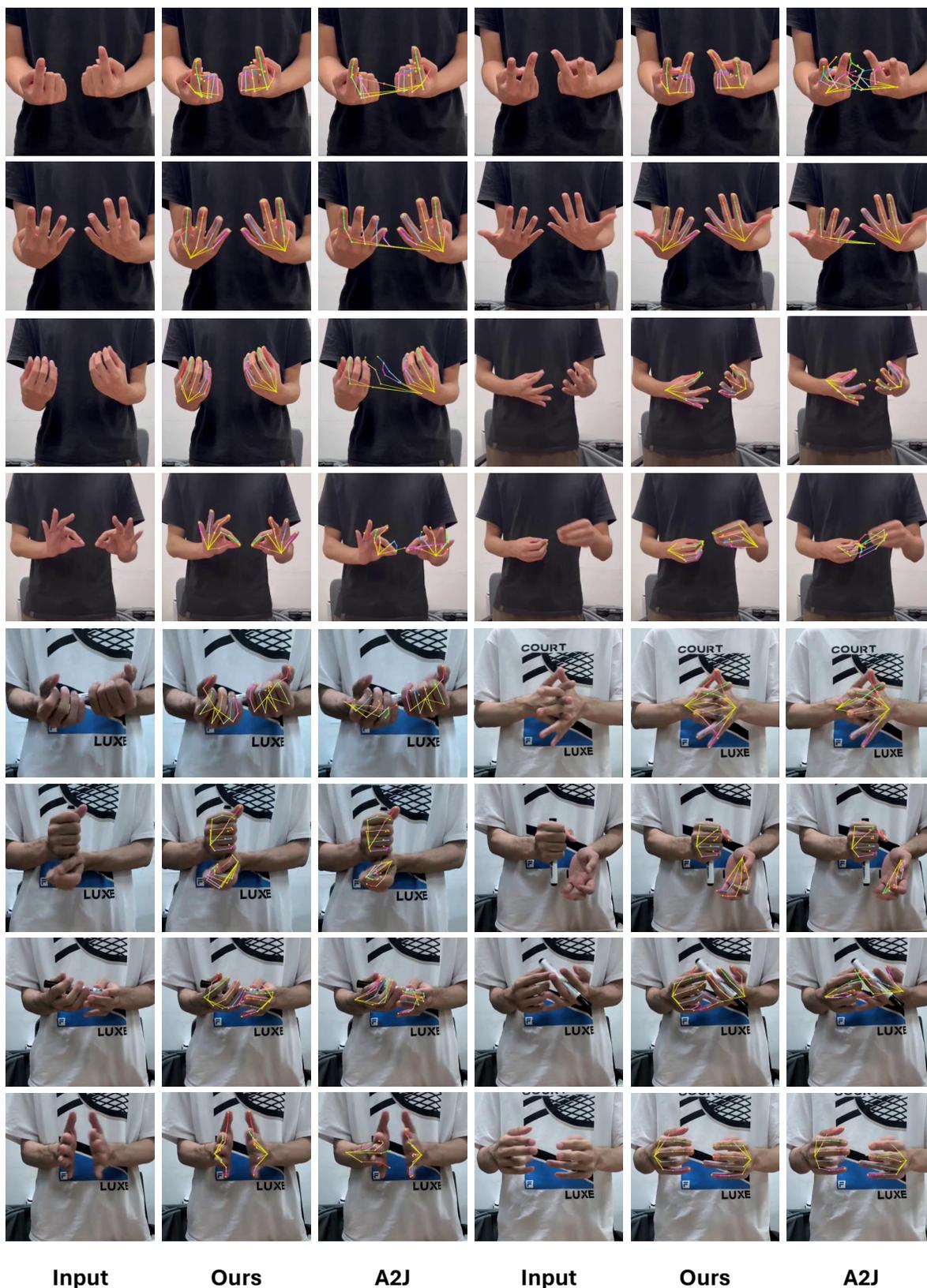
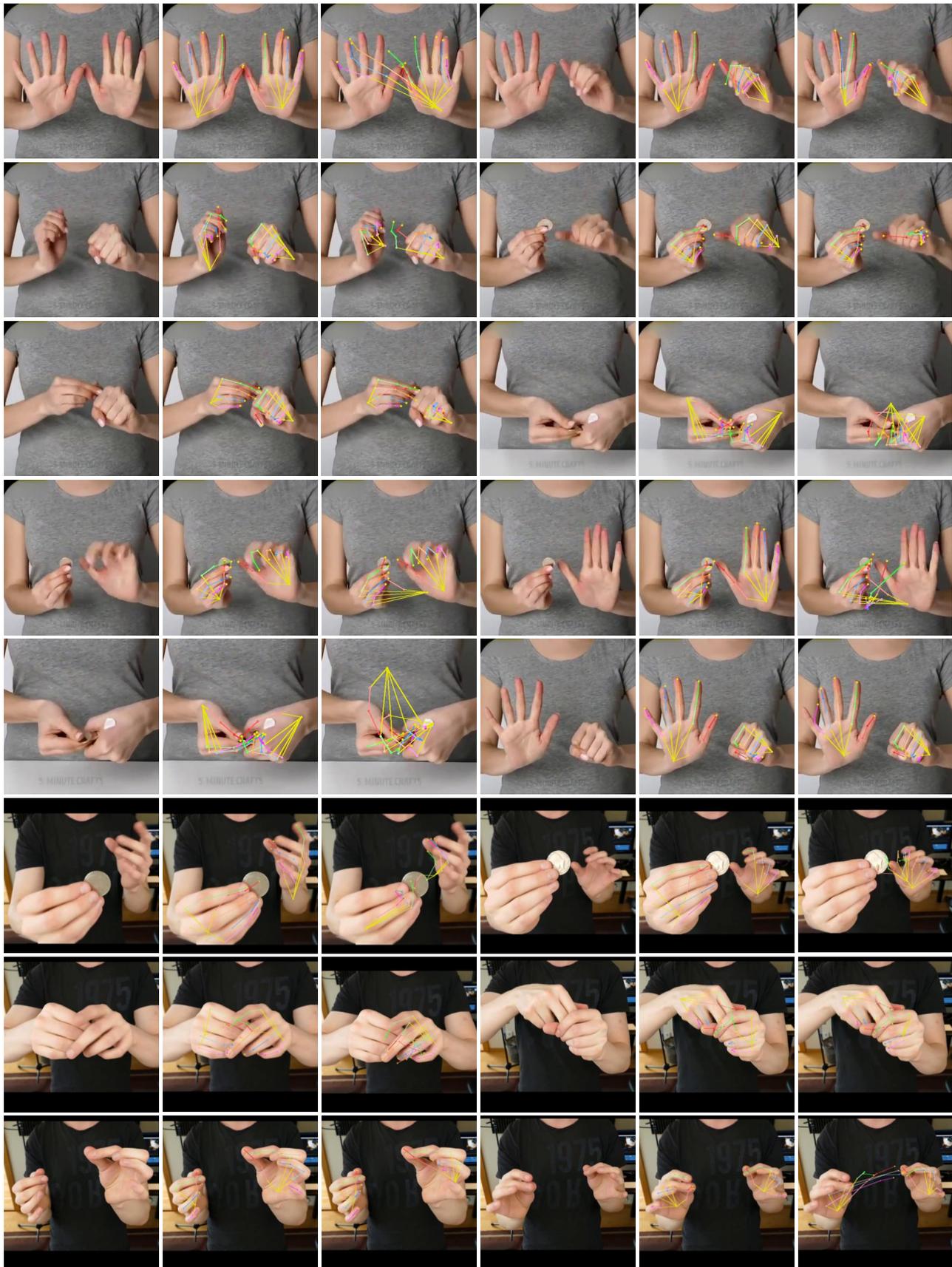


Figure 2: Qualitative results of A2J [2] and ours on images captured using a smartphone.

465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522



523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580

Input Ours A2J Input Ours A2J

Figure 3: Qualitative results of A2J [2] and ours on images sourced from the Internet.