

000	054
001	055
002	056
003	057
004	058
005	059
006	060
007	061
008	062
009	063
010	064
011	065
012	066
013	067
014	068
015	069
016	070
017	071
018	072
019	073
020	074
021	075
022	076
023	077
024	078
025	079
026	080
027	081
028	082
029	083
030	084
031	085
032	086
033	087
034	088
035	089
036	090
037	091
038	092
039	093
040	094
041	095
042	096
043	097
044	098
045	099
046	100
047	101
048	102
049	103
050	104
051	105
052	106
053	107
	Supplementary Material: Annotating Object Instances with a Polygon-RNN
	Anonymous CVPR submission
	Paper ID 2187
1. Additional examples of full image annotation	
We show additional qualitative results from our approach in Fig. 1, 2. In the first column, we show the GT provided by the Cityscapes dataset [1], and in the second column, we show results from our approach, obtained without any corrections (i.e., 0 clicks).	
2. Additional crop examples	
In Fig. 3, 4, we show visualizations of the instances inside the crop of the GT boxes. In the first column, we show the GT annotation, while in the second column, we show the output from SharpMask [2]. In the third column we report the our predictions without any human intervention. Finally, in the fourth column, we show a corrected prediction. Since the output of SharpMask is a dense pixel labeling, we draw the boundary based on the connectivity in the 8-neighborhood.	
3. Performance analysis	

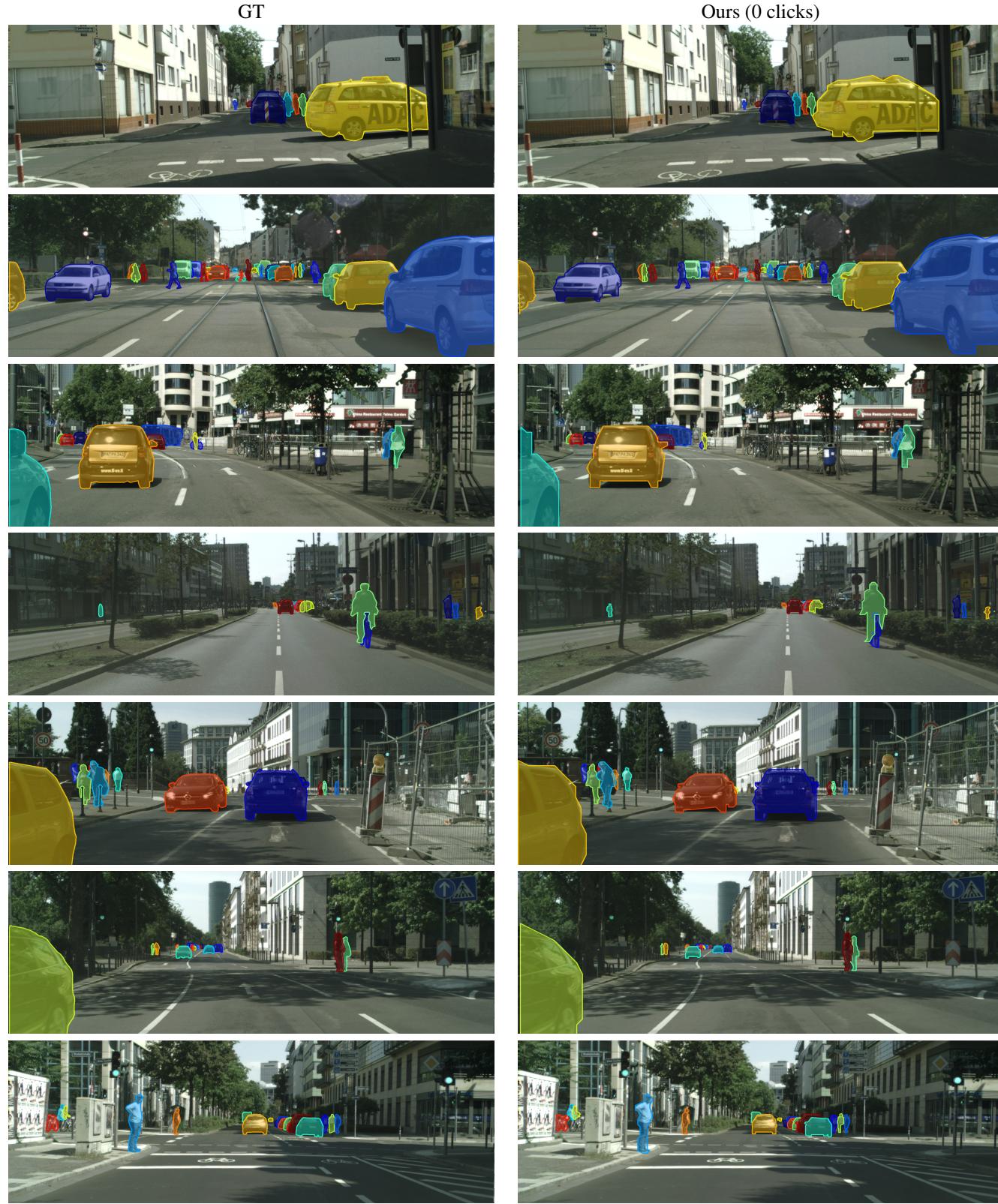
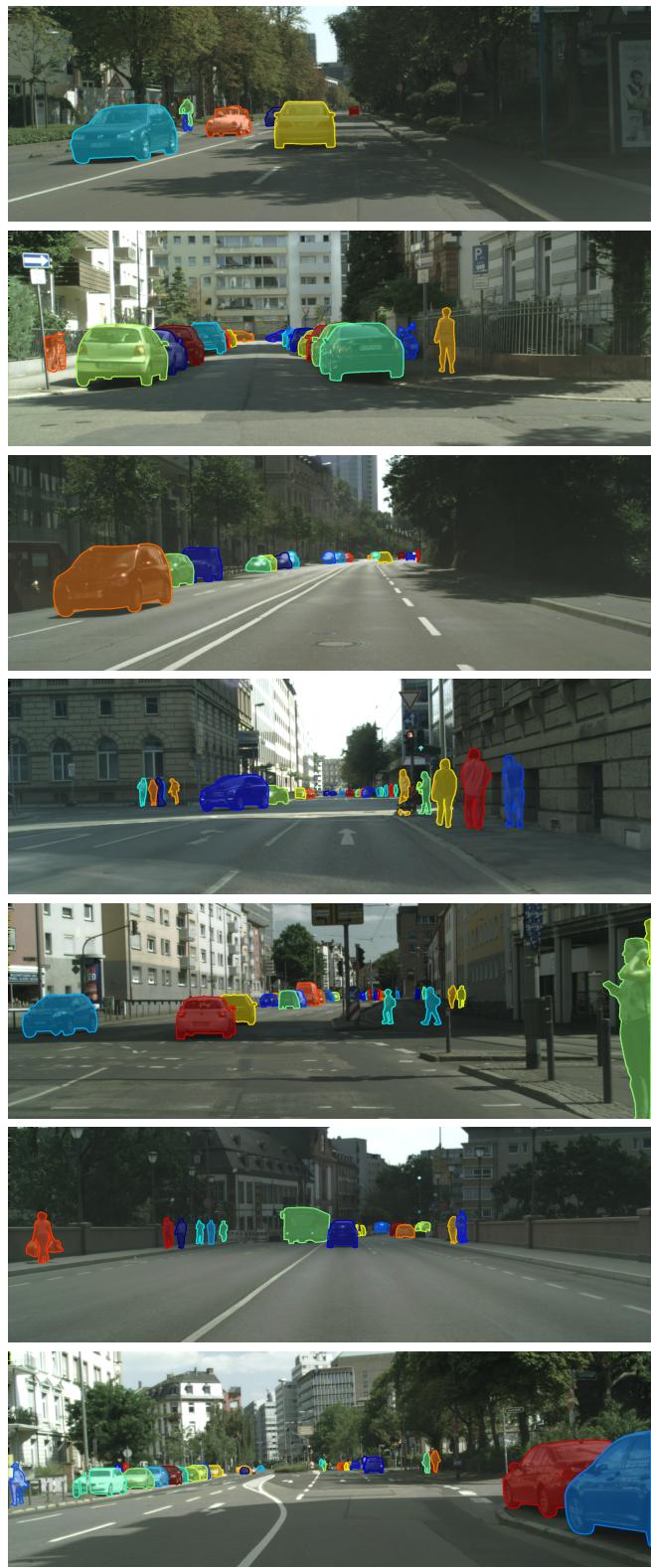


Figure 1. Here we show all our polygons (for all classes) in the original image. We remind the reader that that our approach exploited (ground-truth) boxes to be provided as input. On the **left** we show the ground-truth labeling of the image, while on the **right** we show our polygons in the 0-click regime (running in automatic prediction mode). The GT images contain 38, 12 and 28 instances and required 985, 308 and 580 clicks respectively from their Cityscapes annotators.

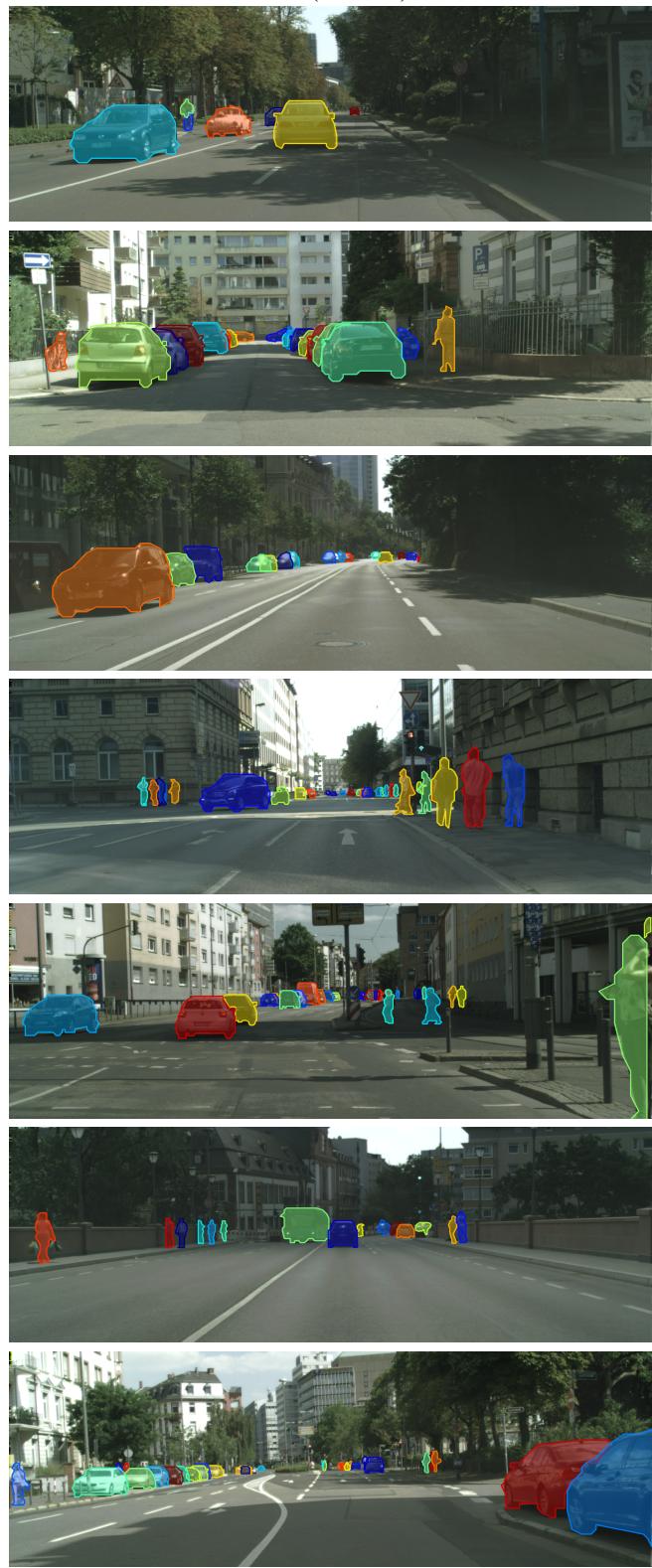
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
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  
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

GT



Ours (0 clicks)



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

Figure 2. Here we show all our polygons (for all classes) in the original image. We remind the reader that that our approach exploited (ground-truth) boxes to be provided as input. On the **left** we show the ground-truth labeling of the image, while on the **right** we show our polygons in the 0-click regime (running in automatic prediction mode). The GT images contain 38, 12 and 28 instances and required 985, 308 and 580 clicks respectively from their Cityscapes annotators.

324  
325  
326  
327  
328329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339340  
341  
342  
343  
344  
345  
346  
347  
348  
349350  
351  
352  
353  
354  
355  
356  
357  
358  
359360  
361  
362  
363  
364  
365  
366  
367  
368  
369370  
371  
372  
373  
374  
375  
376  
377378  
379  
380  
381  
382383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393394  
395  
396  
397  
398  
399  
400  
401  
402  
403404  
405  
406  
407  
408  
409  
410  
411  
412  
413414  
415  
416  
417  
418  
419  
420  
421  
422  
423424  
425  
426  
427  
428  
429  
430  
431

Figure 3. Here we look at a few instances in more detail. On the **first column** we show the GT annotation, while on the **second column**, we show the output from SharpMask. On the **third column** we report the PolygonRNN prediction without human intervention. Finally, on the **fourth column** we show a corrected prediction, in which we can observe how the segmentation is refined to better surround the car mirrors or their wheels.

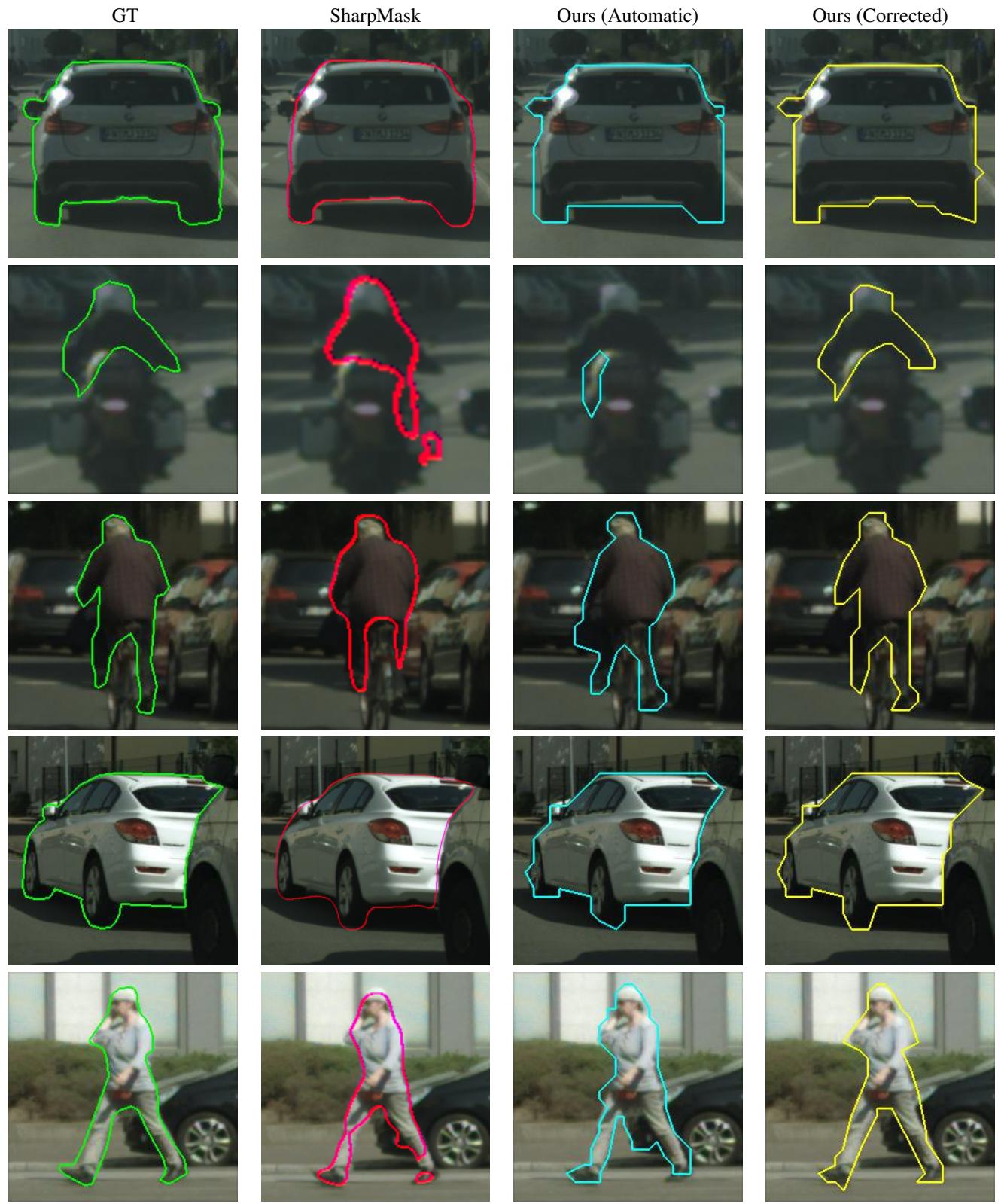


Figure 4. Here we look at a few instances in more detail. On the **first column** we show the GT annotation, while on the **second column**, we show the output from SharpMask. On the **third column** we report the PolygonRNN prediction without human intervention. Finally, on the **fourth column** we show a corrected prediction, in which we can observe how the segmentation is refined to better surround the car mirrors or their wheels.

540	<b>References</b>	594
541		595
542	[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler,	596
543	R. Benenson, U. Franke, S. Roth, and B. Schiele. The	597
544	cityscapes dataset for semantic urban scene understanding. In	598
545	<i>CVPR</i> , 2016. 1	599
546	[2] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár. Learning	600
547	to refine object segments. <i>ECCV</i> 2016, 2016. 1	601
548		602
549		603
550		604
551		605
552		606
553		607
554		608
555		609
556		610
557		611
558		612
559		613
560		614
561		615
562		616
563		617
564		618
565		619
566		620
567		621
568		622
569		623
570		624
571		625
572		626
573		627
574		628
575		629
576		630
577		631
578		632
579		633
580		634
581		635
582		636
583		637
584		638
585		639
586		640
587		641
588		642
589		643
590		644
591		645
592		646
593		647