# A Unified Video Segmentation Benchmark: Annotation, Metrics and Analysis Supplementary Material

Fabio Galasso[1], Naveen Shankar Nagaraja[2], Tatiana Jiménez Cárdenas[2], Thomas Brox[2], Bernt Schiele[1]
[1] Max Planck Institute for Informatics, Germany
[2] University of Freiburg, Germany

This is supplementary material to [4]:

Fabio Galasso, Naveen Shankar Nagaraja, Tatiana Jiménez Cárdenas, Thomas Brox and Bernt Schiele.
A Unified Video Segmentation Benchmark: Annotation, Metrics and Analysis. In *ICCV*, 2013.

Please cite the publication above when referring to this supplementary material.

## 1. Experimental Study of Metrics

We report here an initial study which we made about the existing segmentation benchmark metrics of [1], recently extended to video by [3]. We started processing the video sequences of the BMDS dataset [2] with a number of video segmentation algorithms and noticed inconsistencies in the ranking which the *boundary precision-recall* (BPR) score was providing, as opposed to that given by the region metrics of *segmentation covering* (SC), *probabilistic Rand index* (PRI) and *variation of information* (VI). The consequent analysis, which we report here, confirmed experimentally the observation of [1], that the boundary BPR metric evaluates segmentation outputs better than the region metrics of SC, PRI and VI. We have found this analysis of potential interest to the research community and extended it to our proposed *volume precision-recall* (VPR) metric, with a particular emphasis on the *consistency* and *complementarity* between BPR and VPR. Please note that this analysis uses the dataset of [2] and not the one which we propose in the paper.

The analysis is based on ∼1000 coarse-to-fine video segmentation outputs, obtained by processing the video sequences of [2] with the algorithm of [3] under different operating setups. The considered segmentation outputs have varying accuracy levels, from degenerate segmentations, e.g. one label for the whole video, to qualitatively good ones.

First we analyze the covariance between the boundary BPR and the existing region metrics SC, PRI and VI. Figure 1 shows the scatter plots for the SC, PRI, VI and BPR measures and the covariance matrix among them. We observe a strong correlation between SC, PRI and VI (values of VI are inversely related to SC and PRI, as for the range $[0, \inf)$). By contrast BPR is poorly correlated with the region measures of SC, PRI and VI (c,d). While intuitively boundary and region scores should contain complementary information, it is rather critical that the scores are not correlated. This could explain the observation of [1] that BPR is better suited than SC, PRI and VI to evaluate segmentation quality.

Next we analyze the covariance of the novel volume metric VPR against all other metrics SC, PRI, VI and BPR. Both the covariance matrix in figure 1(d) and the scatter plot in figure 2(d) confirm a high correlation between VPR and BPR. By contrast, the comparison of VPR against SC, PRI and VI is not straightforward: a clear trend cannot be observed, but the figures show limited range of variation for SC and PRI, as opposed to the full ranges of VPR and BPR.

We investigate the relations between VPR and the other metrics BPR, SC, PRI and VI further by selecting 5 study cases from the scatter plots in figure 2, for which we illustrate ground truth annotations and segmentation outputs in figure 3, and corresponding numerical values in table 1. We make the following observations.

- SC allows for a single segmentation volume to be matched against a GT volume (constraint to one-to-one matches), and therefore penalizes the background labeling behind the actor in study case 1, figure 3(a-c). The VPR metric penalizes the case with low recall, but scores the case slightly higher for the accurate precision.

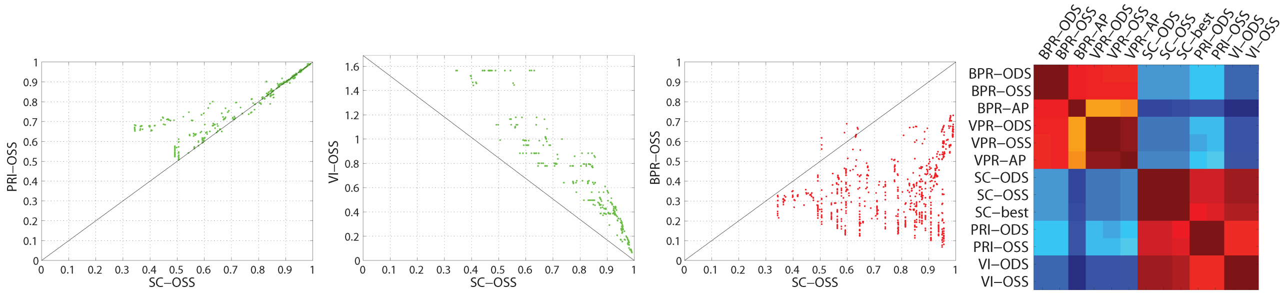
- SC, PRI and VI all suffer from their limited range of values due to their unnormalized scores. Study case 2 in figure 3(f,g) illustrates a failure segmentation where entire frames get just one label. There, a large background object grants nearly perfect scores for SC, PRI and VI. VPR correctly classifies these as failure cases. This limitation of SC, PRI and VI may also be observed by the degenerate score of “video=1label” in table 1: only VPR and BPR correctly score this case 0.

- the entropic formulation of VI over penalizes volumes spanning equal areas of multiple GT objects, as for maximum entropy, e.g. the green and yellow labels in figure 3(h,i) – study case 3.
- VPR and BPR are nicely complementary. VPR penalizes study case 4, figure 3(d-f), for the lack of temporal consistency and the merge of the two persons in figure 3(e) due to a missing boundary element. BPR penalizes study case 5, figure 3(k-l), for missing the person label, small in terms of pixels but large in importance as for the many boundary pixels describing it.

So in summary SC, PRI and VI do not have several of the properties, which we have highlighted in Section 4.3 of the manuscript for BPR and VPR, and which qualify good video segmentation benchmark metrics. Most importantly SC, PRI and VI do not satisfy **i.***non-degeneracy* and **v.***coarse-to-fine segmentations and working regimes*; SC and PRI additionally do not satisfy **iv.***adaptive accommodation of refinement*. From this analysis, SC, PRI and VI do not qualify as good segmentation metrics, as also noted in [1]. By contrast the boundary BPR and the volume VPR metrics satisfy these quality criteria and are complementary, as we have shown here, which makes it worthwhile to report both metrics in segmentation evaluations.

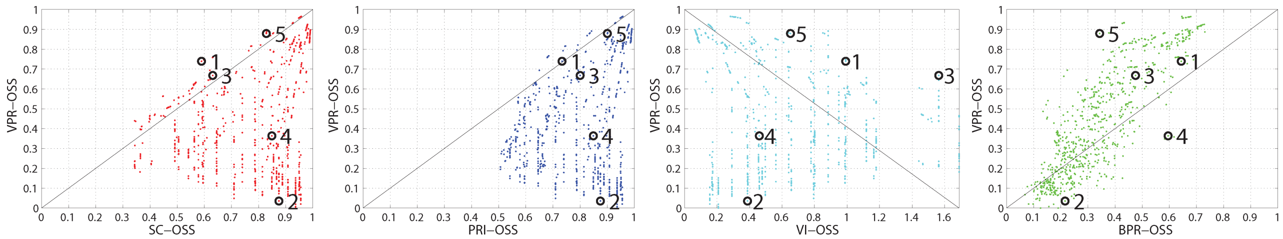
|  | BPR |  |  | VPR |  |  | SC |  |  | PRI |  | VI |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | ODS | OSS | AP | ODS | OSS | AP | ODS | OSS | Best | ODS | OSS | ODS | OSS |
| Study case 1 | 0.64 | 0.64 | 0.48 | 0.74 | 0.74 | 0.68 | 0.59 | 0.59 | 0.73 | 0.73 | 0.73 | 0.99 | 0.99 |
| Study case 2 | 0.22 | 0.22 | 0.10 | 0.22 | 0.22 | 0.00 | 0.88 | 0.88 | 0.88 | 0.88 | 0.88 | 0.39 | 0.39 |
| Study case 3 | 0.48 | 0.48 | 0.30 | 0.67 | 0.67 | 0.53 | 0.63 | 0.63 | 0.63 | 0.80 | 0.80 | 1.56 | 1.56 |
| Study case 4 | 0.60 | 0.60 | 0.52 | 0.36 | 0.36 | 0.31 | 0.85 | 0.85 | 0.88 | 0.85 | 0.85 | 0.46 | 0.46 |
| Study case 5 | 0.34 | 0.34 | 0.16 | 0.88 | 0.88 | 0.78 | 0.83 | 0.83 | 0.84 | 0.90 | 0.90 | 0.65 | 0.65 |
| Degenerate: video = 1 label | 0 | 0 | 0 | 0 | 0 | 0 | 0.74 | 0.74 | 0.74 | 0.72 | 0.72 | 0.71 | 0.71 |
| Degenerate: 1 pixel = 1 label | 0.02 | 0.02 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.28 | 0.28 | 19.3 | 19.3 |

Table 1. Quantitative evaluation of the study cases indicated in figure 2. This table additionally contains evaluation of degenerate cases “video = 1 label” and “1 pixel = 1 label”. Note: these results are on a dataset (BMDS [2]) different from the one proposed in the paper.



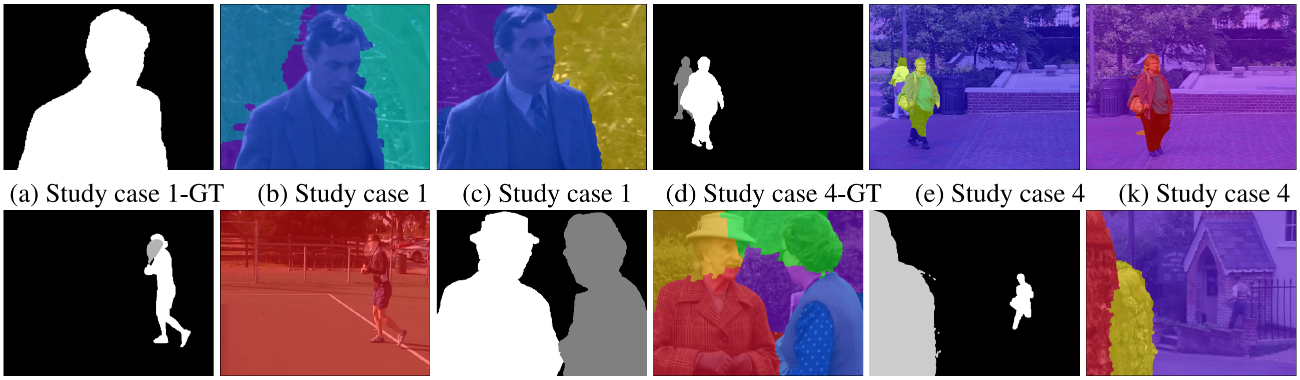
(a) SC – PRI scores (b) SC – VI scores (c) SC – BPR scores (d) Correlations

Figure 1. Scatter plots of measures obtained for ∼1000 video segmentation outputs and covariance matrix. (a-c) the region metric SC is compared to PRI, VI and BPR; (d) covariance matrix among all metrics SC, PRI, VI, BPR and VPR. Note: these results have been computed on a dataset (BMDS [2]) different from the one proposed in the paper.



(a) SC – VPR scores (b) PRI – VPR scores (c) VI – VPR scores (d) BPR – VPR scores

Figure 2. Scatter plots of measures obtained for ∼1000 video segmentation outputs. The novel metric VPR is compared to all others SC, PRI, VI and BPR. The small circles and numbers from 1 to 5 indicate the study cases which are described in section 1 of this supplementary material. Note: these results have been computed on a dataset (BMDS [2]) different from the one proposed in the paper.



(a) Study case 1-GT (b) Study case 1 (c) Study case 1 (d) Study case 4-GT (e) Study case 4 (k) Study case 4

(f) Study case 2-GT (g) Study case 2 (h) Study case 3-GT (i) Study case 3 (k) Study case 5-GT (l) Study case 5

Figure 3. Illustration of ground truth annotations and segmentation outputs for the study cases indicated in figure 2. Note: this figure shows examples from a dataset (BMDS [2]) different from the one proposed in the paper.

## 2. Further comparison cases of metrics

In this section, we consider additional study cases, which we extract from the scatter plot of figure 2. The purpose of such new cases is merely to provide additional comparison among BPR, SC, PRI, VI and VPR, and allow the interested reader to browse particular ones. All observations and conclusions drawn in this section were already discussed in section 1.

The new study cases are numbered 6 – 27, and illustrated in figure 4, which also reports the previous 5. The quantitative results for the new study cases are reported in table 2 and their qualitative evaluations are reported in figures 5, 6 and 7.

In the following, we consider groups of study cases.

**Study cases 12, 25, 11, 10.** These cases are granted very high scores by all previous region metrics SC, PRI and VI, while our proposed metric VPR scores them in a large range of values, as it can be noted in the plots in figures 4(a-c) and in table 2. We comment on the cases by considering the sample frames from the video sequences, the corresponding annotated ground truth (GT) and segmentation outputs in figures 5 and 7.

The coarse-to-fine segmentation output of study case 12 provides video volumes (clusters) which do not correspond to the objects in the video (background+walking person) at any granularity level (a sample frame in figure 5 provides the segmentation result for 10 clusters). By contrast, in study case 10 the foreground objects (car and van) are well segmented across the video and temporally consistent, as for a good video segmentation output.

Only our proposed VPR metric correctly scores the two study cases distinctly, assigning ∼0% to the segmentation output in study case 12, and ∼90% in study case 10. The two study cases are undistinguishable by the region metrics SC, PRI and VI: while the reason for the high performance in study case 10 is the quality of the segmentation output, in study case 12 the same high score is due to the normalization issues of SC, PRI and VI and the large size of the background label compared to the walking person size. SC, PRI and VI select and report performance of the coarsest level of the segmentation output – the whole video labelled as 1 label.

Study cases 25 and 11 present segmentation outputs which are inbetween in terms of performance: in study case 25 we have the same video sequence as in 12, but a different segmentation output, which addresses the walking person correctly at the initial frames, but not at later frames (two representative frames in figure 7); in study case 11 ( two sample frames in figure 5), the foreground object (car) is well segmented across all video, but the segmentation output shows cluster re-initialization, e.g. both the car and the background get different colors as for poor temporal consistency. Only our proposed metric VPR scores the two segmentation output correctly inbetween the performance of the two study case 12 and 10. The mentioned normalization issues makes the different performances undistinguishable for SC, PRI and VI.

**Study cases 18, 23, 25.** This group of cases presents the reverse situation as in group (12, 25, 11, 10): SC, PRI and VI provide a wide range of evaluation measures while VPR scores the segmentation outputs of this group cases with similar values (this can be seen from the plots in figures 4(a-c)).

Looking at figures 6 and 7, we see that the corresponding three tested machine segmentations correctly identify the foreground objects at some frames (first sampled frames for study cases 18 and 25) but miss the object in the second part of the video (second sampled frames for study cases 18 and 25), or only partially identify the foreground objects but with temporal consistence over the video (study case 23).

It is difficult to advocate the superiority of one result with respect to the others and it is intuitively wrong to evaluate study case 18 much lower than study case 23 and 25, which the metrics SC, PRI and VI do. The reason for the misjudgement is clear if we focus our attention on the size of the background object in the three cases: the background object size is similar to the foreground one in study case 18, but increasingly relatively larger in cases 23 and 25. The metrics SC, PRI and VI provide a plausible score for case 18, while the scores of cases 23 and 25 are hampered by normalization issues.

Our proposed VPR yields similar performance for the segmentation outputs of the cases, which is interestingly closest to the measures provided by SC, PRI and VI for study case 18.

**Study cases 6, 21, 16.** Similarly to study cases (12, 25, 11, 10), in these cases SC, PRI and VI yield similar performance measures while the proposed metric VPR scores the cases in a wide range.

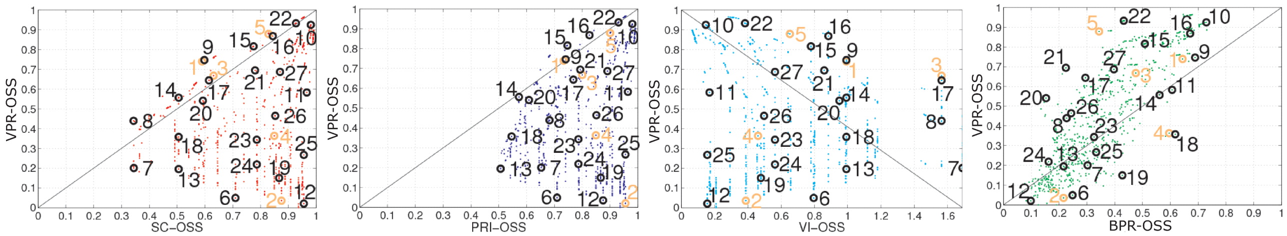
We observe from figure 5 that the segmentation output of study case 6 is a degenerate one, which only VPR correctly scores approximately 0 (as discussed above, SC, PRI and VI have normalization issues).

In the segmentation result of case 21 in figure 7, the two segments in the result can unambiguously and temporally-consistently be associated to the the wall and background objects, but the wall boundaries are off the GT edge and the small (in terms of pixels) person is missed. All the metrics SC, PRI, VI and VPR agree to score the case in the average, which seems intuitively correct (the boundary metric penalizes the case to a larger extent, as for its complementary – cases below offer more examples).

The segmentation result of study case 16 in figure 6 should then outperform the previous cases: the person is correctly assigned to a single temporally-consistent segment, few pixels are missing from the person segment and the background is consistently labelled, although some labels are re-initialized as for the shifting scene. This is however not the case for the metrics SC, PRI and VI. In particular, SC penalizes the background relabelling greatly (as we observe for study case 1 in section 1, the metric only allows one-to-one assignments) and scores therefore this case similarly to the previous 6 and 21; this also happens for the VI score, because VI penalizes largely the mixing labels at the bottom of the person and on the background (we explain this in study case 3 in section 1 in terms of its entropic formulation). Only the proposed metric VPR correctly score case 16 better than 6 and 21.

**Study cases 21,27,9.** As it may be observed from the plot in figure 4(d), VPR scores the cases to approximately the same value while BPR assigns them values in a large range. The sample frames in figures 5 and 7 illustrate that the segmentation output in study case 21 is not accurate but temporally consistent, while the ones in study cases 27 and 9 are more accurate on the car and person objects but fragment the background. Overall, it seems plausible to have similar measurements in terms of volume metrics, as VPR does. A clear distinction is provided however by the boundary metric BPR, which scores low the segmentation in study case 21 for the misaligned pixels and those missing from the person, but it scores high the one in study case 9 for matching most boundary pixels to the GT object boundaries. As discussed in study case 5 in section 1, BPR provides important complementary information to our proposed volume metric VPR.

**Study cases 18, 11, 16.** The scatter plot in figure 4(d) shows that the study cases have approximately the same BPR boundary scores but differ much in the volume VPR scores. Figures 5 and 6 illustrate that the machine segmentations in all three cases correctly preserve the GT object boundaries, explaining similar scores by BPR. The segmentation output in study case 18 has however very poor performance in terms of label consistency, i.e. the background label takes over the person label; the one in study case 11 shows better label consistency but issues with temporal consistency, i.e. the labels are re-initialized; the segmentation in study case 16 improves on both cases for label and temporal consistency. VPR correctly assigns different performance scores to the three cases and provides important complementary information to the boundary measures, as we comment for study case 4 in section 1.

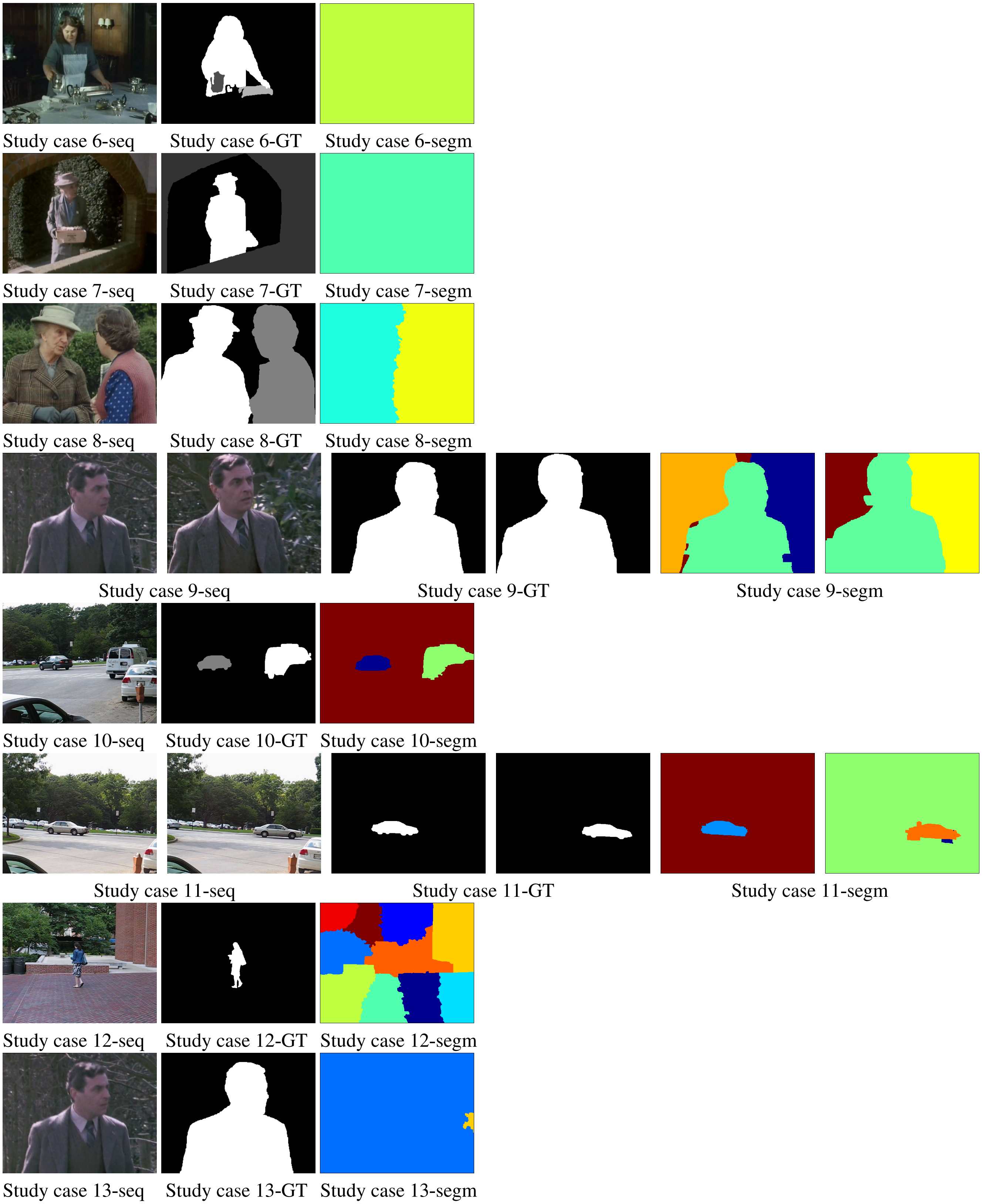


(a) SC – VPR scores (b) PRI – VPR scores (c) VI – VPR scores (d) BPR – VPR scores

Figure 4. Scatter plots representing the performance of ∼1000 coarse-to-fine video segmentation outputs. Each video segmentation result is plotted according to its score with respect to two metrics, as indicated. 27 study cases are identified across all plots: 1-5 (*orange*) are discussed in section 1, 6-27 (*black*) are additionally provided here. Note: these results have been computed on a dataset (BMDS [2]) different from the one proposed in the paper.

| | BPR | | | VPR | | | SC | | | PRI | | VI | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ODS | OSS | AP | ODS | OSS | AP | ODS | OSS | Best | ODS | OSS | ODS | OSS |
| Study case 6 | 0.25 | 0.25 | 0.08 | 0.05 | 0.05 | 0.00 | 0.71 | 0.71 | 0.71 | 0.71 | 0.71 | 0.80 | 0.80 |
| Study case 7 | 0.30 | 0.30 | 0.14 | 0.20 | 0.20 | 0.04 | 0.34 | 0.34 | 0.38 | 0.65 | 0.65 | 1.69 | 1.69 |
| Study case 8 | 0.23 | 0.23 | 0.07 | 0.44 | 0.44 | 0.30 | 0.34 | 0.34 | 0.38 | 0.68 | 0.68 | 1.56 | 1.56 |
| Study case 9 | 0.69 | 0.69 | 0.54 | 0.75 | 0.75 | 0.65 | 0.60 | 0.60 | 0.74 | 0.74 | 0.74 | 0.99 | 0.99 |
| Study case 10 | 0.73 | 0.73 | 0.48 | 0.92 | 0.92 | 0.84 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.15 | 0.15 |
| Study case 11 | 0.61 | 0.61 | 0.37 | 0.58 | 0.58 | 0.50 | 0.97 | 0.97 | 0.97 | 0.96 | 0.96 | 0.17 | 0.17 |
| Study case 12 | 0.10 | 0.10 | 0.04 | 0.02 | 0.02 | 0.00 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.16 | 0.16 |
| Study case 13 | 0.22 | 0.22 | 0.08 | 0.19 | 0.19 | 0.05 | 0.51 | 0.51 | 0.51 | 0.51 | 0.51 | 0.99 | 0.99 |
| Study case 14 | 0.56 | 0.56 | 0.42 | 0.56 | 0.56 | 0.48 | 0.51 | 0.51 | 0.53 | 0.57 | 0.57 | 0.99 | 0.99 |
| Study case 15 | 0.51 | 0.51 | 0.30 | 0.82 | 0.82 | 0.67 | 0.77 | 0.77 | 0.79 | 0.75 | 0.75 | 0.78 | 0.78 |
| Study case 16 | 0.67 | 0.67 | 0.59 | 0.87 | 0.87 | 0.79 | 0.84 | 0.84 | 0.85 | 0.83 | 0.83 | 0.88 | 0.88 |
| Study case 17 | 0.29 | 0.29 | 0.17 | 0.64 | 0.64 | 0.52 | 0.61 | 0.61 | 0.63 | 0.77 | 0.77 | 1.56 | 1.56 |
| Study case 18 | 0.62 | 0.62 | 0.41 | 0.36 | 0.36 | 0.28 | 0.51 | 0.51 | 0.53 | 0.55 | 0.55 | 0.99 | 0.99 |
| Study case 19 | 0.43 | 0.43 | 0.27 | 0.15 | 0.15 | 0.03 | 0.87 | 0.87 | 0.89 | 0.87 | 0.87 | 0.48 | 0.48 |
| Study case 20 | 0.15 | 0.15 | 0.05 | 0.54 | 0.54 | 0.40 | 0.59 | 0.59 | 0.69 | 0.61 | 0.61 | 0.95 | 0.95 |
| Study case 21 | 0.22 | 0.22 | 0.09 | 0.69 | 0.69 | 0.57 | 0.78 | 0.78 | 0.78 | 0.79 | 0.79 | 0.86 | 0.86 |
| Study case 22 | 0.43 | 0.43 | 0.20 | 0.93 | 0.93 | 0.85 | 0.93 | 0.93 | 0.93 | 0.93 | 0.93 | 0.38 | 0.38 |
| Study case 23 | 0.33 | 0.33 | 0.14 | 0.34 | 0.34 | 0.17 | 0.79 | 0.79 | 0.81 | 0.79 | 0.79 | 0.56 | 0.56 |
| Study case 24 | 0.16 | 0.16 | 0.08 | 0.22 | 0.22 | 0.09 | 0.79 | 0.79 | 0.81 | 0.79 | 0.79 | 0.56 | 0.56 |
| Study case 25 | 0.33 | 0.33 | 0.13 | 0.27 | 0.27 | 0.09 | 0.96 | 0.96 | 0.97 | 0.96 | 0.96 | 0.16 | 0.16 |
| Study case 26 | 0.24 | 0.24 | 0.14 | 0.46 | 0.46 | 0.35 | 0.85 | 0.85 | 0.87 | 0.85 | 0.85 | 0.50 | 0.50 |
| Study case 27 | 0.40 | 0.40 | 0.21 | 0.69 | 0.69 | 0.57 | 0.87 | 0.87 | 0.88 | 0.89 | 0.89 | 0.56 | 0.56 |
| Degenerate: video = 1 label | 0 | 0 | 0 | 0 | 0 | 0 | 0.74 | 0.74 | 0.74 | 0.72 | 0.72 | 0.71 | 0.71 |
| Degenerate: 1 pixel = 1 label | 0.02 | 0.02 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.28 | 0.28 | 19.3 | 19.3 |

Table 2. Detailed evaluation results of the additional study cases 6-27 selected. Note: these results have been computed on a dataset (BMDS [2]) different from the one proposed in the paper.



Study case 6-seq Study case 6-GT Study case 6-segm

Study case 7-seq Study case 7-GT Study case 7-segm

Study case 8-seq Study case 8-GT Study case 8-segm

Study case 9-seq Study case 9-GT Study case 9-segm

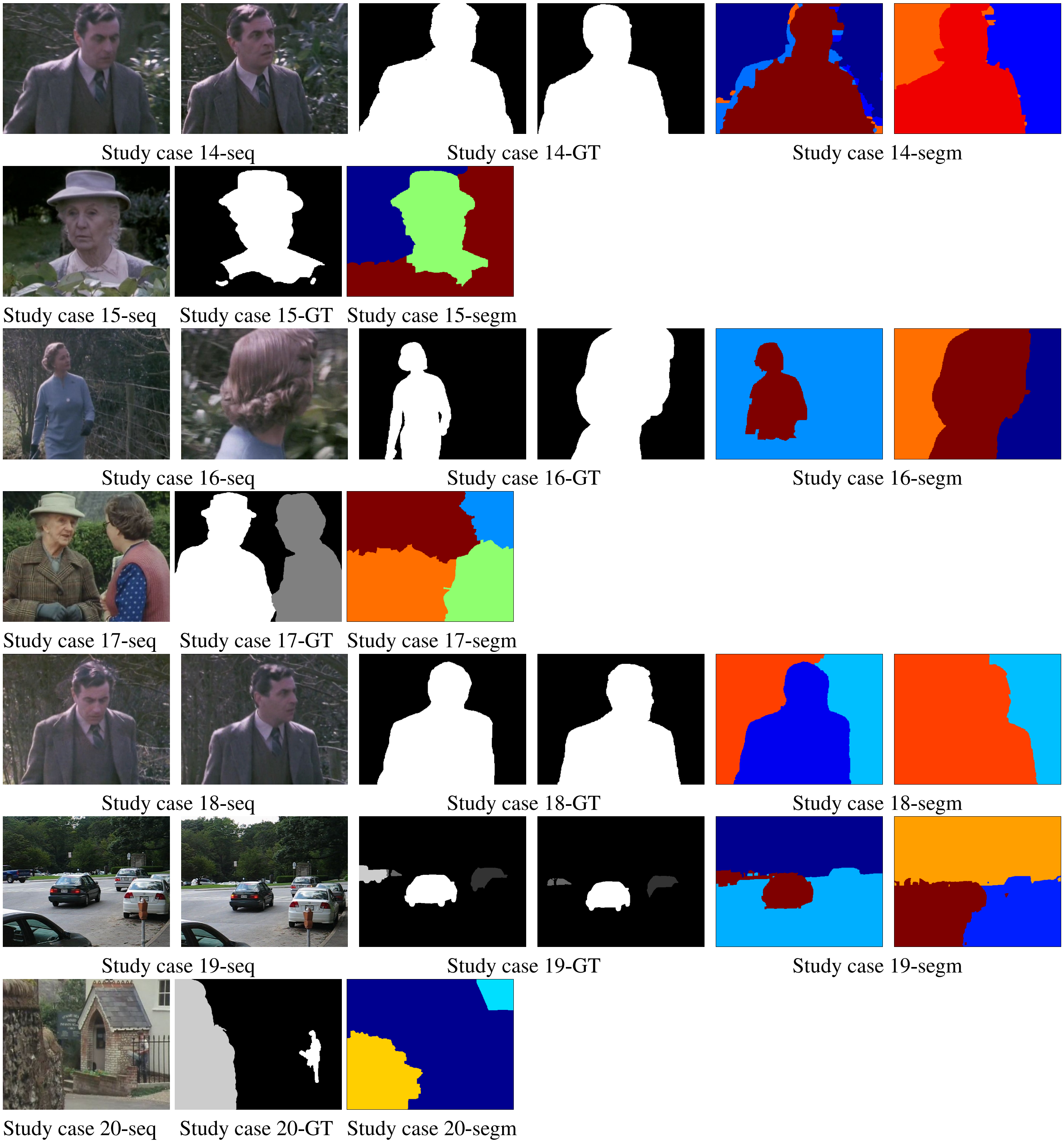
Study case 10-seq Study case 10-GT Study case 10-segm

Study case 11-seq Study case 11-GT Study case 11-segm

Study case 12-seq Study case 12-GT Study case 12-segm

Study case 13-seq Study case 13-GT Study case 13-segm

Figure 5. Sample frames from the video sequences, GT annotations and segmentation outputs for the additional study cases 1-13 selected in this supplementary material. Single frames are presented from each sequence when the illustrated result is representative over time. 2 frames are presented for study cases 9 and 11, to illustrate the segmentation output over time. Note: this figure shows examples from a dataset (BMDS [2]) different from the one proposed in the paper.



Study case 14-seq

Study case 14-GT

Study case 14-segm

Study case 15-seq

Study case 15-GT

Study case 15-segm

Study case 16-seq

Study case 16-GT

Study case 16-segm

Study case 17-seq

Study case 17-GT

Study case 17-segm

Study case 18-seq

Study case 18-GT

Study case 18-segm

Study case 19-seq

Study case 19-GT

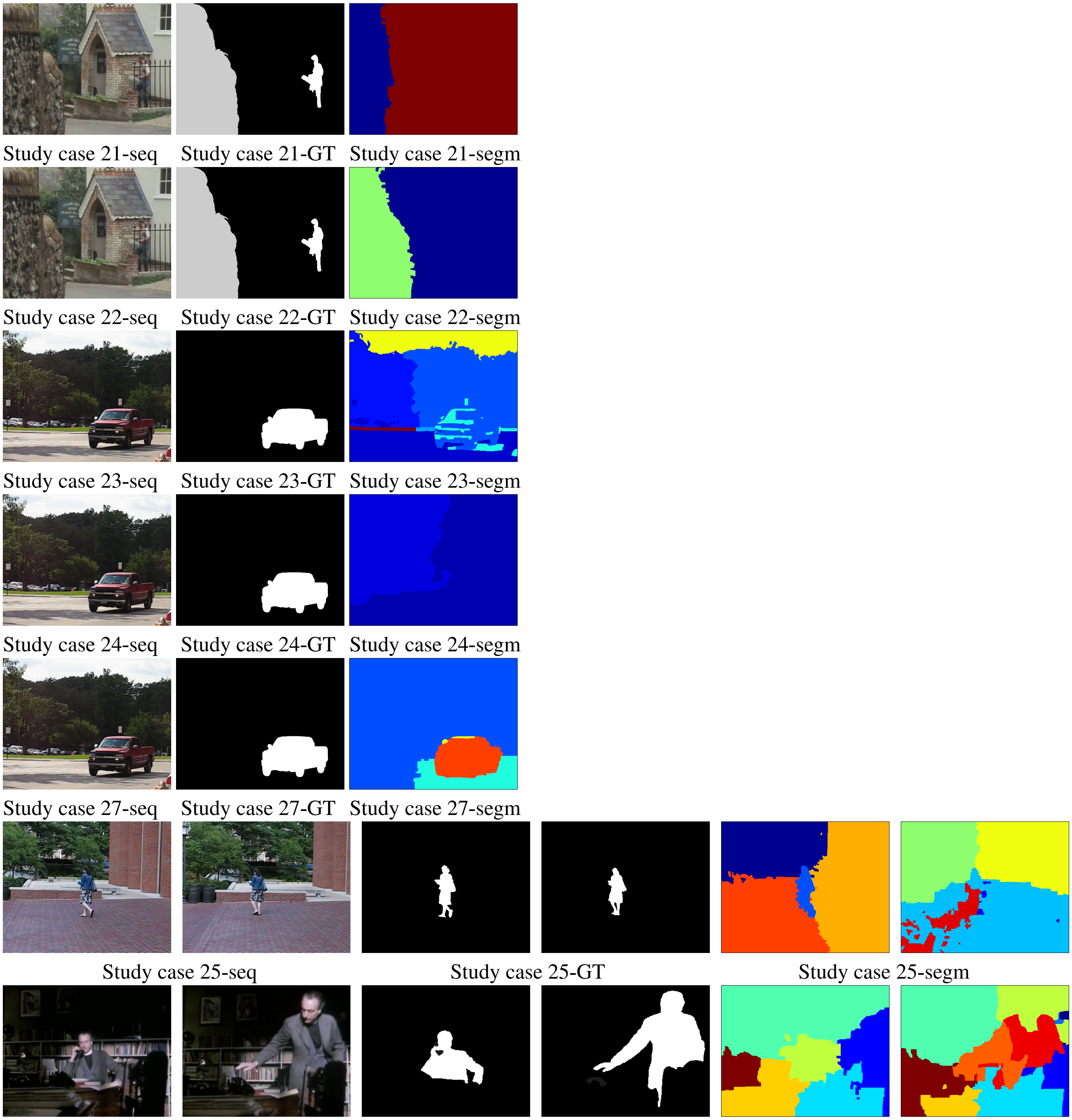
Study case 19-segm

Study case 20-seq

Study case 20-GT

Study case 20-segm

Figure 6. Sample frames from the video sequences, GT annotations and segmentation outputs for the additional study cases 14-20 selected in this supplementary material. Single frames are presented from each sequence when the illustrated result is representative over time. 2 frames are presented for study cases 14, 16, 18, 19, to illustrate the segmentation output over time. Note: this figure shows examples from a dataset (BMDS [2]) different from the one proposed in the paper.



Study case 21-seq Study case 21-GT Study case 21-segm

Study case 22-seq Study case 22-GT Study case 22-segm

Study case 23-seq Study case 23-GT Study case 23-segm

Study case 24-seq Study case 24-GT Study case 24-segm

Study case 27-seq Study case 27-GT Study case 27-segm

Study case 25-seq Study case 25-GT Study case 25-segm

Study case 26-seq Study case 26-GT Study case 26-segm

Figure 7. Sample frames from the video sequences, GT annotations and segmentation outputs for the additional study cases 21-27 selected in this supplementary material. Single frames are presented from each sequence when the illustrated result is representative over time. 2 frames are presented for study cases 25 and 26, to illustrate the segmentation output over time. Note: this figure shows examples from a dataset (BMDS [2]) different from the one proposed in the paper.

# References

[1] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 33(5):898–916, 2011.

[2] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010.

[3] F. Galasso, R. Cipolla, and B. Schiele. Video segmentation with superpixels. In *ACCV*, 2012.

[4] F. Galasso, N. S. Nagaraja, T. J. Cárdenas, T. Brox, and B. Schiele. A unified video segmentation benchmark: Annotation, metrics and analysis. In *ICCV*, 2013.