Appendix

Distribution of Segment Length

Figure 1 shows box plots of the segment lengths. The orange boxes refer to the segments selected for video summaries and blue boxes indicate the remaining ones. The segment boundaries are generated by two-peak segmentation in Section 4. As described in Section 4.3, Figure 1 shows that segment selection results in summaries composed of only short segments. The distribution of segment lengths is similar for DR-DSN and dppLSTM. This indicates that importance score prediction methods hardly affect segment selection.

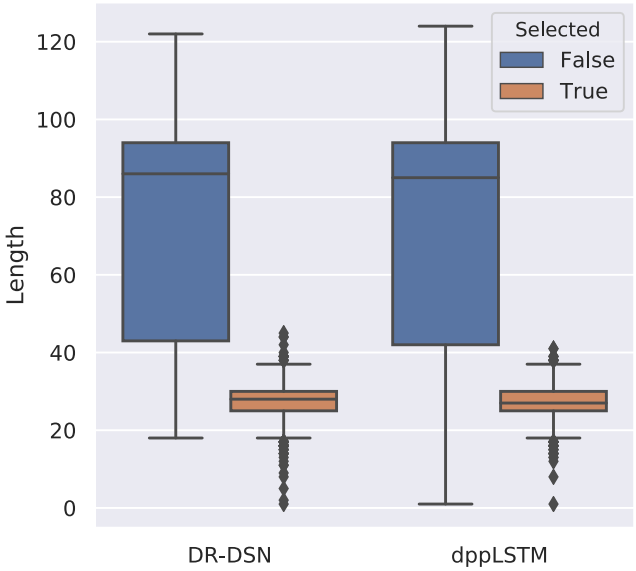

Figure 1. Box plot of segment length. Orange boxes represent segments selected for output summaries, and blue boxes indicate remaining ones.

Examples of Video Summaries

Figure 2 shows examples of video summaries by two recent methods [25, 28] and our randomized method in Section 4. All these methods uses KTS segmentation method. As described in 4.3, we can see that all methods result in similar output regardless of the importance scores.

Visualization of Importance Score Correlation

Figure 3 shows examples of correlation curves on TVSum dataset. The red lines show the a_t curve for each human annotator and the black dashed line is the expectation for a random importance scores. The blue and green curves show the corresponding results to dppLSTM and DR-DSN methods, respectively. Most examples show that human annotations, *i.e.*, red lines, have positive correlation.

Results with Different Summary Length

Figure 4 shows F1 scores of summaries generated with different summary length constraint; 15%, 25%, and 35% of original video length. The results are obtained by generating 10 summaries for each method. For all summary length constraint, the results show similar trends. The randomized summaries have similar performance to human annotations for two-peak, KTS, and randomized KTS segmentation methods. We also observed that F1 scores tend to get higher as summary length gets longer.

Human Evaluation

In the human evaluation, subjects watch two video summaries generated by two different methods and are asked, *Which video better summarizes the original video subjects?* We employed 30 subjects for each video summary pair on a crowdsourcing service. Subjects selected their answer from -2 (A is much more than B) to 2 (A is much less than B). Therefore, the averaged score larger than 0 indicates that method A is better otherwise, subjects prefer B better. The results comparing KTS and uniform segmentation with random importance scoring are show in Figure 5. Overall, the averaged score is 0.19, thus subjects prefer video summaries using uniform segmentation. In particular, subjects prefer uniform segmentation for videos recording long activity, *e.g.*, sightseeing of the statue of liberty and scuba diving. On the other hand, KTS works better for videos with notable events or activities. For such videos, the important parts have little ambiguity, therefore the F1 scores based on the agreement of selected frames can get higher.

Figure 6 shows the results of comparing video summaries generated with random and DR-DSN scoring. Both methods use KTS segmentation. Random approach obtained a slightly higher score than random DR-DSN, however, 46% of answers were that the summaries are equally good (bad). This result further supports our findings that the importance scoring hardly affects the performance with a certain approach.

Other results of comparing video summaries generated with random scoring and manually created summaries are shown in Figure 7. Overall, the averaged score is -0.17, thus subjects slightly prefer human summaries. However, it is important to note that the answers are diverse among subjects; most videos got both "A Much More Than B" and "A Much Less Than B." The results demonstrate the challenges in video summary evaluation. The quality of video summaries often depends on subjects. We also collected subject's free-form text feedback. We observed that some subjects focus on the coverage of events in the original video, and some value inclusion of many scenes with main objects or people. Evaluating a video summary from several aspects mentioned in the subjects' comments should be essential.

Figure 2. Comparison of summaries created by two recent methods and our randomized method. The blue line shows the segment level importance scores with respect to time (frames). The orange areas indicate the frames selected for the final summary. All of the three methods use the same segment boundaries by KTS.

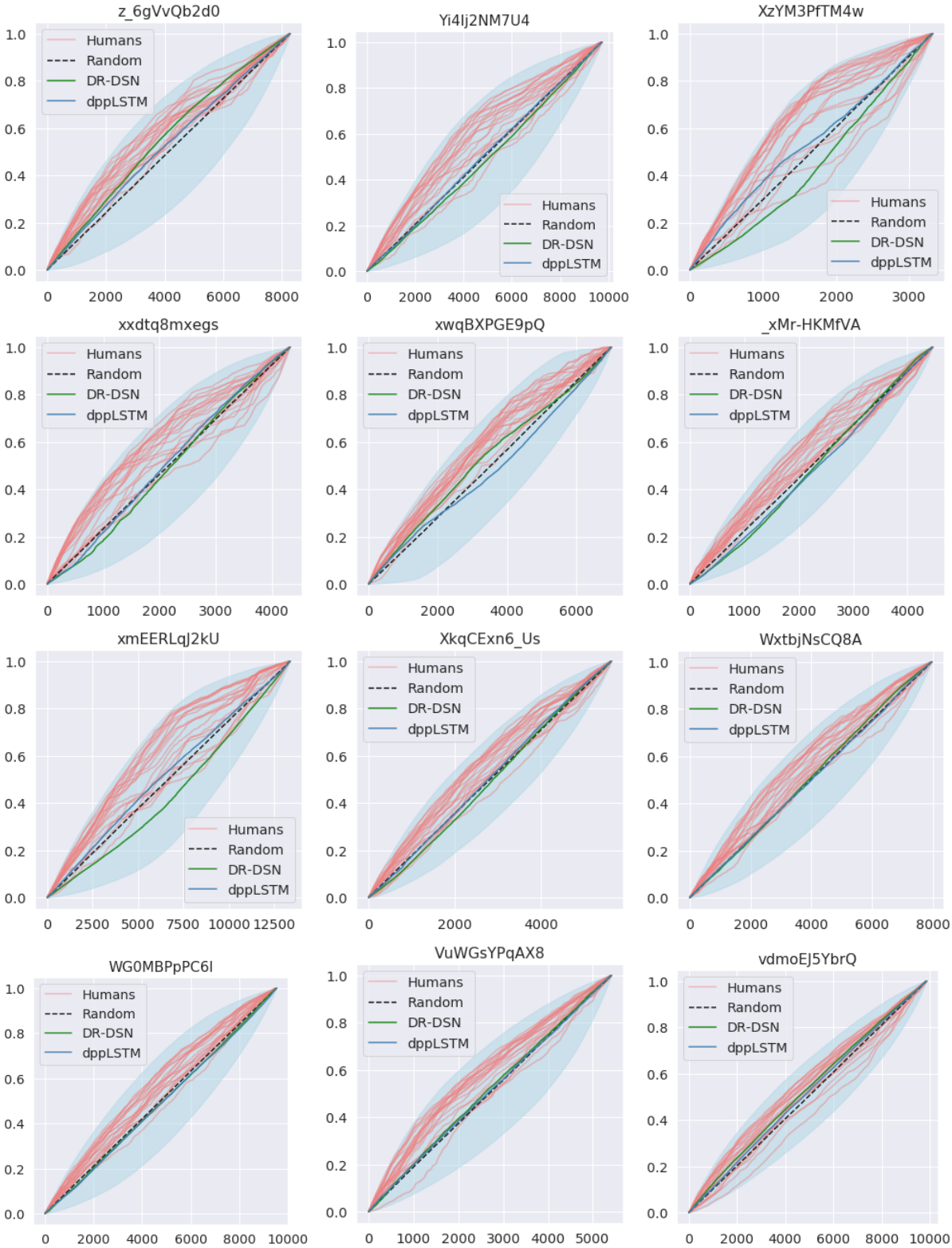

Figure 3. Example correlation curves. The red lines represent correlation curves for each human annotator and the black dashed line is the expectation for a random importance scores. The blue and green curves show the corresponding results to dppLSTM and DR-DSN methods, respectively.

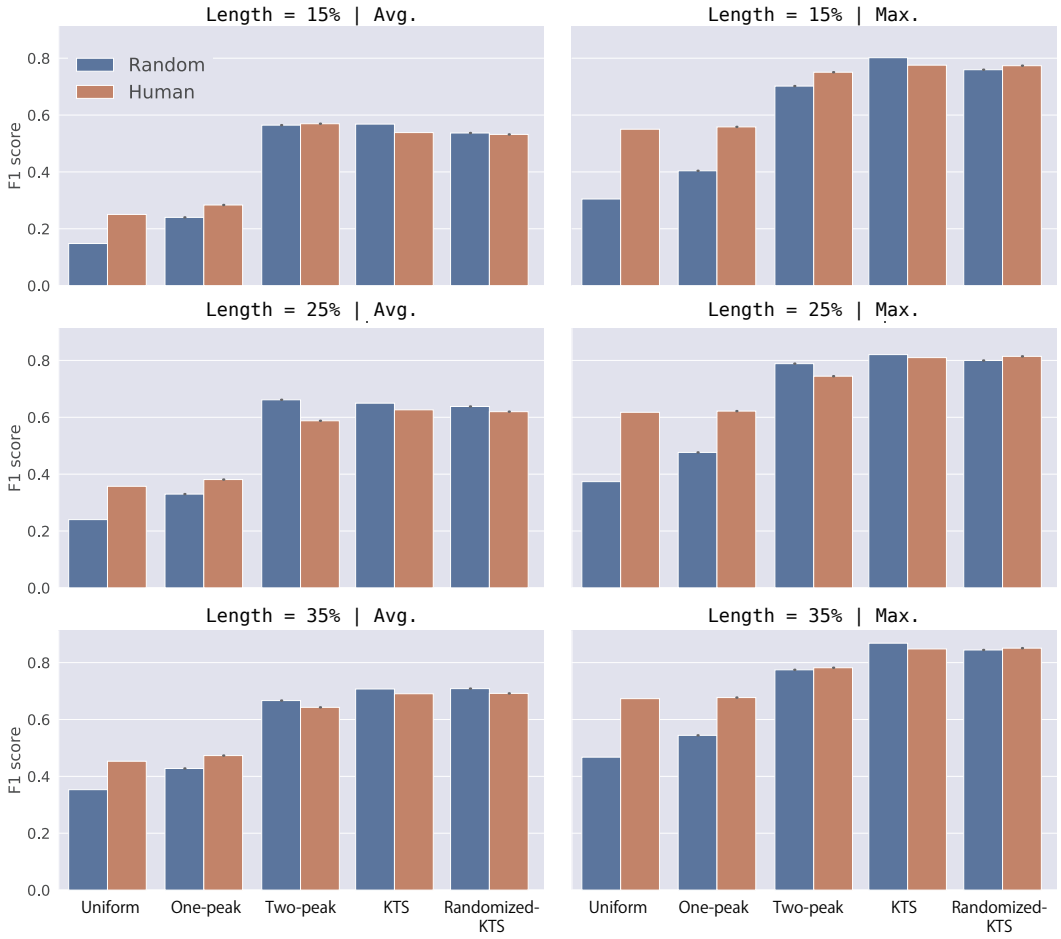

Figure 4. F1 scores for different segmentation methods combined to either random or human annotated importance scores (leave-one-out) for TVSum dataset. Blue bars refer to random scores and orange bars indicates human annotations. The results are computed with different summary length constraints, which are 15%, 25% and 35% of original video length.

Rank Order Statistics and Human Evaluation

One criticism for using rank order statistics for evaluating importance scores is that the metric might be irrelevant to the final quality of a video summary. We compute Spearman's ρ for DR-DSN on the SumMe dataset. For reference importance scores, we employ the ratio of human annotators who selected the frame for their manually-created summary. We split the video into two groups; one is for videos where DR-DSN shows a positive correlation to reference scores, and otherwise. For each group, we investigate the performance scores against randomized summaries. Videos with positive correlation are better rated than those with negative correlation. The average score of the group with positive correlation is 0.13, and the other is 0.22. In this result, a lower score means that video summaries generated by DR-DSN are better against ones with random scoring. This observation suggests that the higher correlation of importance scores can result in better quality of final output.

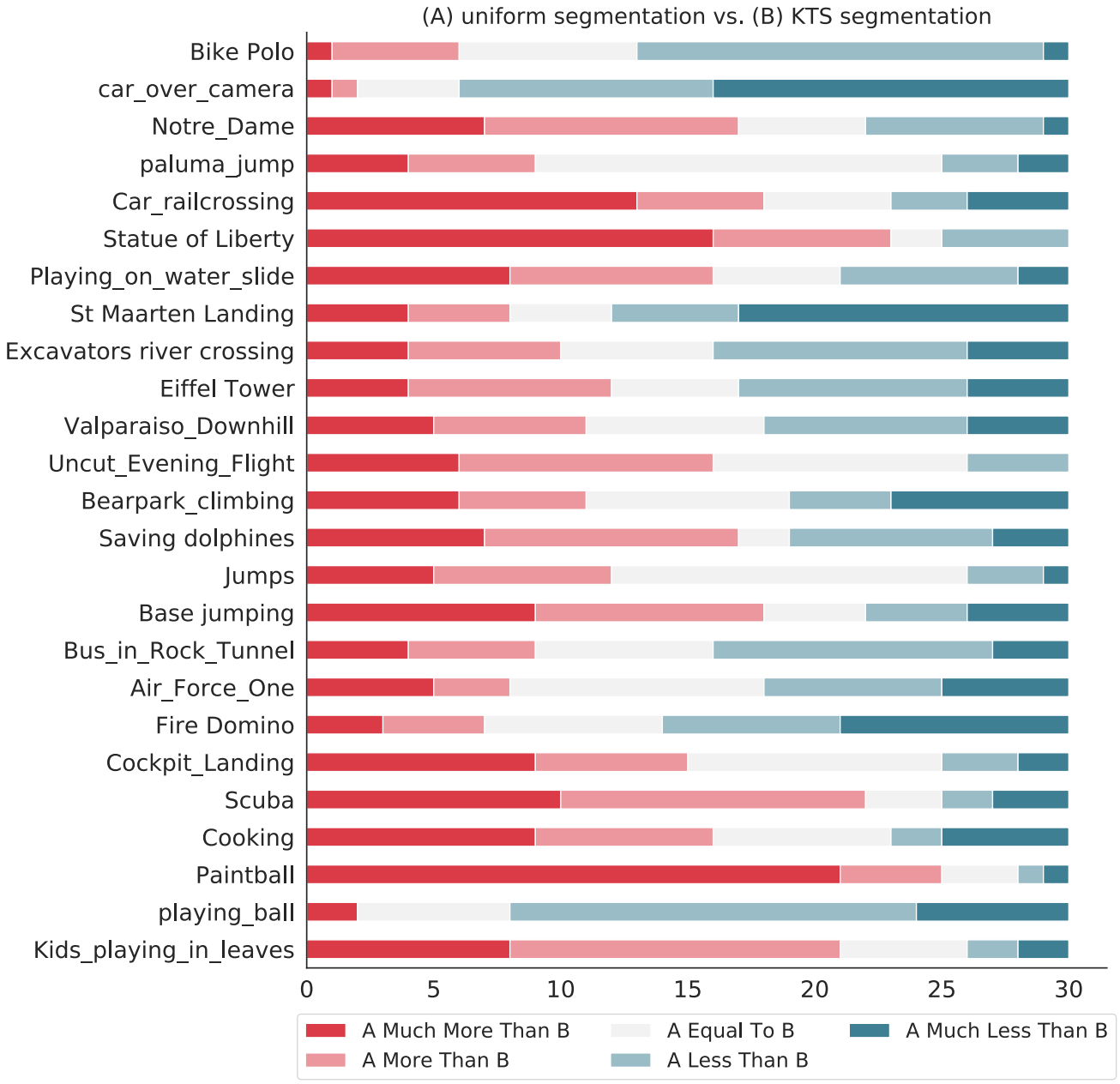

Figure 5. Comaprison of video summaries generated with A) uniform and B) KTS segmentation with random scores.

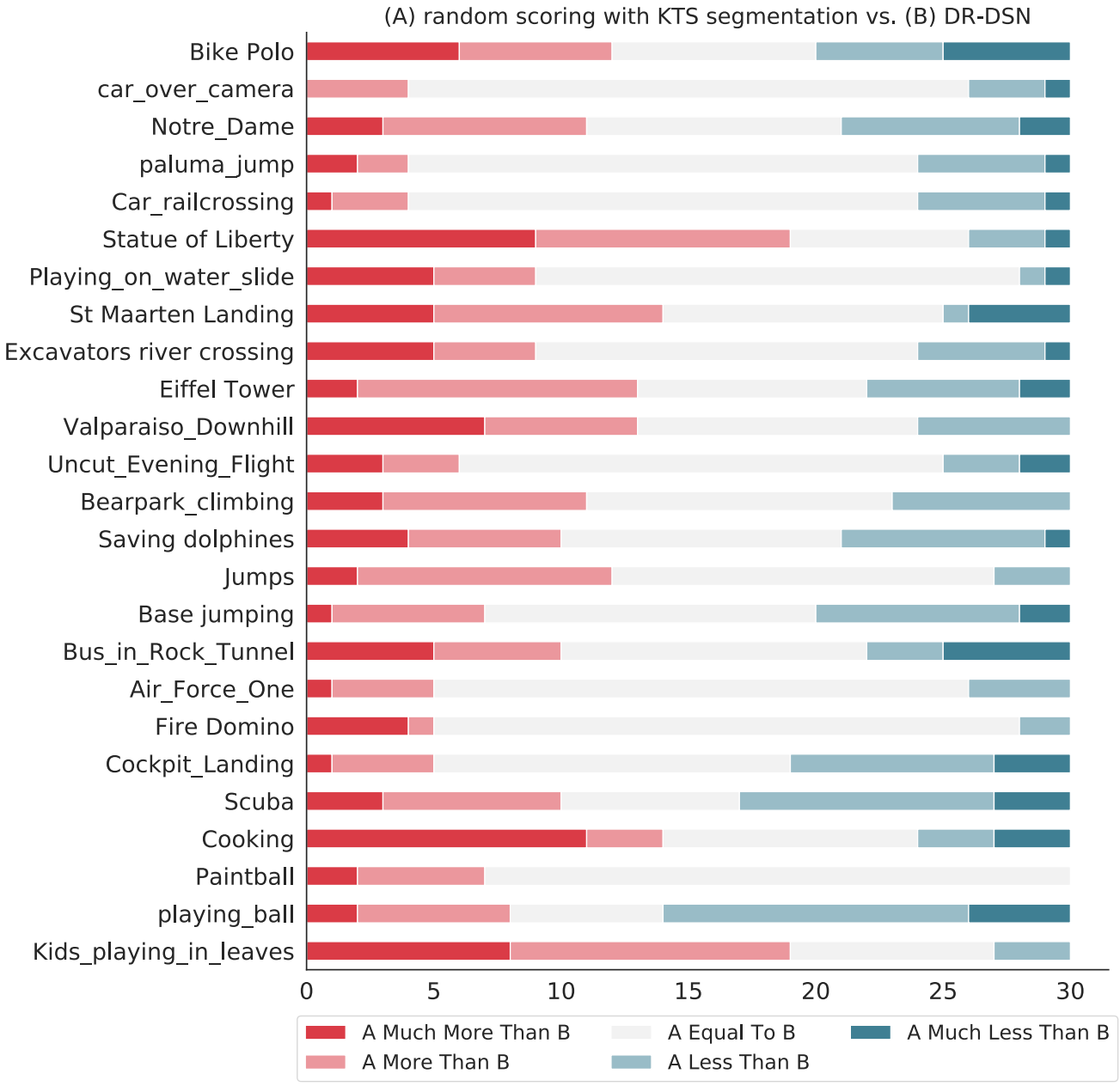

Figure 6. Comparison of video summaries generated with (A) random and (B) DR-DSN importance scores.

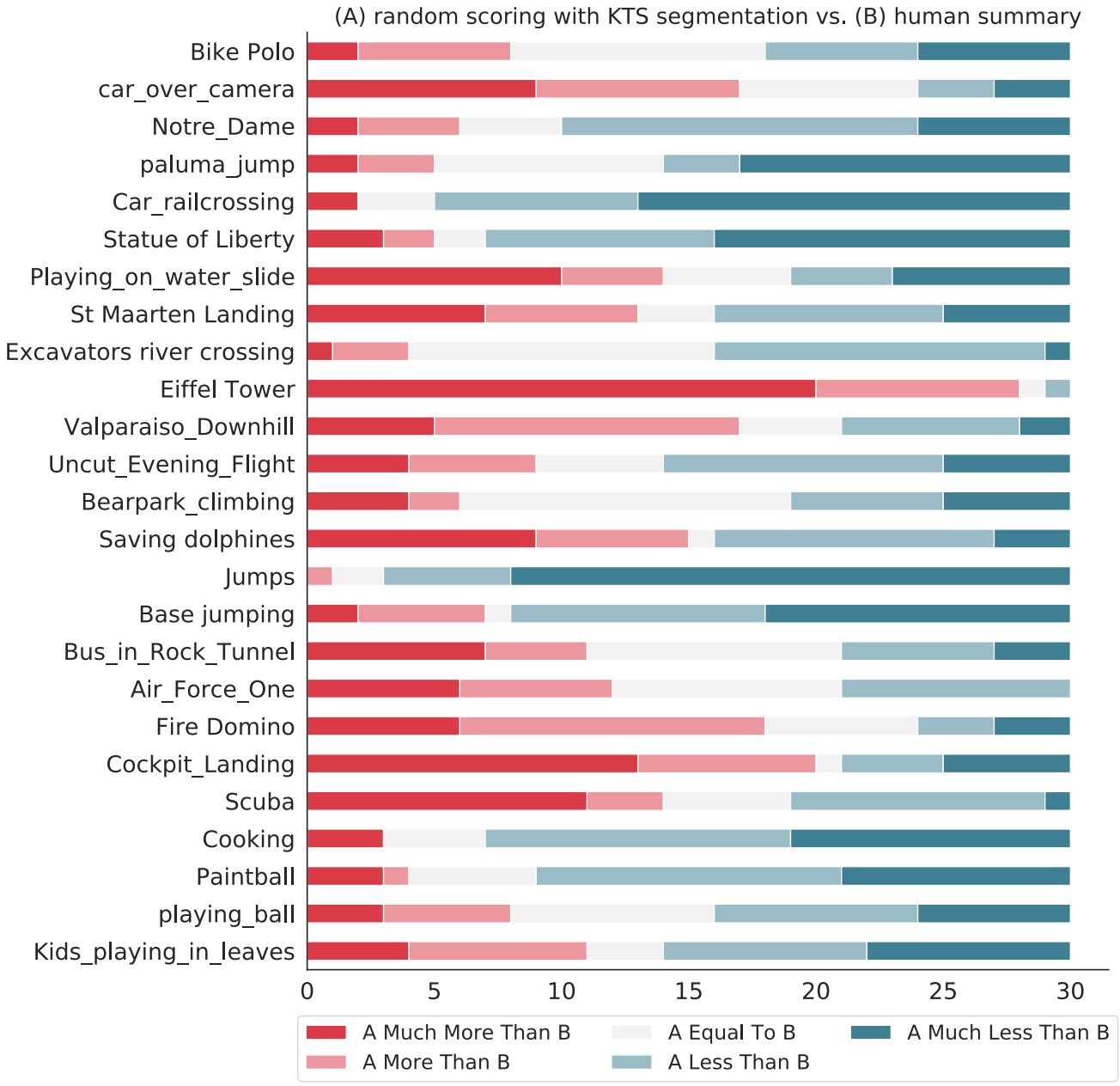

Figure 7. Comparison of video summaries generated with (A) random scoring with KTS segmentation and (B) human summaries.