Learning to Cut by Watching Movies
Supplementary Material

1. Learning to Cut by Users- Toy Example

Please visit: ***https://alejandropardo.net/publication/learning-to-cut/*** for code and complete supplementary material.

To illustrate to the reader a toy example of Learning to Cut we included a folder called Spot-the-Real-Cut. We encourage the reader to open the *html* file contained in this folder and try to choose the more suitable cuts. You will have to wait around 15 seconds for the link to load all the videos. There are going to be 30 examples, each of them showing a pair of cuts. One of them breaks continuity, while the other is an actual cut made by a professional editor. The task is simple: **choose the cut that is real**. To play the video, click on top of it. To decide what you consider is the real cut, click on the button "This cut is real" below the clip. At the end of the study, you will see what percentage of cuts from the ones chosen were actually real. The purpose of this toy example is to illustrate that there is a signal that a model could learn to Learn how to cut. Such a signal is the one that Learning to Cut is aiming to leverage.

2. DatasetStatistics

Additional statistics are shown below. Figure 1 shows the distribution of number of shots per genre along the dataset. Figure 2 shows the shots-duration's distribution. Most of the shots in the movies are shorter than 2 seconds. This challenging property comes from the fast-pace edits of action scenes, where the shot duration is tipically short.

3. Generalization to unedited set.

Qualitative results. In table 1, we report the qualitative number of our method on the unedited set. We see that the real task is really challenging as the number drop significantly from the proxy task. However, we observe the same trend in the results, our method outperforms the baselines. This results show how challenging is the tasks in a real-world scenario. Thus, future methods have to put effort to solve first the proxy as this results will be reflected in the real task.

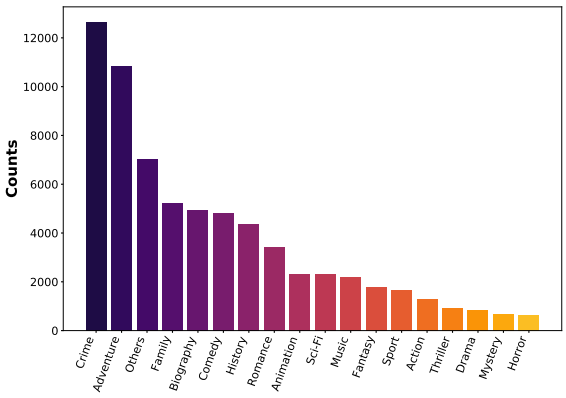


Figure 1: **Distribution of shots per genre.**

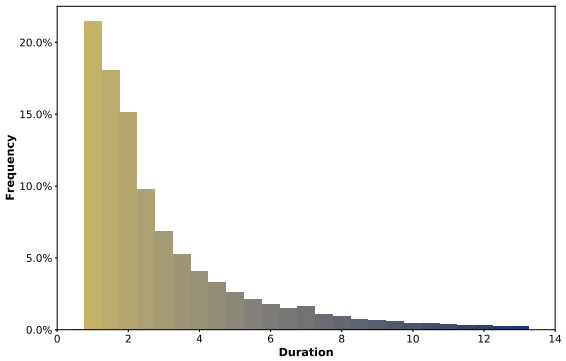


Figure 2: **Shots' duration distribution.**

4. Additional Ablation Study

Ablation Study. In Table 2, we report the all metrics for the ablation study shown as Table 2 in the main manuscript.

Impact of M. Figure 3 shows the impact of the top $M\%$ parameter in the two-stage prediction. We can see different peaks according to the metric that we are looking at; however, there is a clear pattern top 20% favors the best $R@1K$, top 30% $R@5K$, and top 40% $R@10K$, no matter the distance. For the Unedited set we chose top 30%, since we were using the top-5 predictions.

Model	$d=1$			$d=2$			$d=3$		
	R@1K	R@5K	R@10K	R@1K	R@5K	R@10K	R@1K	R@5K	R@10K
Random	0.83	3.33	3.33	1.67	8.33	15.00	3.33	17.50	30.00
Audio-visual	1.67	4.17	5.00	2.50	5.83	10.83	2.50	11.67	20.00
OURS	0.83	2.50	8.33	1.67	10.83	30.83	5.83	17.50	34.67

Table 1: **Generalization to unedited videos.** We showcase our model's results in a real-case scenario where it processes raw unedited footage. We report the same metrics as before by comparing our results with the cuts of professional editors.

Model	$d=1$			$d=2$			$d=3$		
	R@1K	R@5K	R@10K	R@1K	R@5K	R@10K	R@1K	R@5K	R@10K
OURS	8.18	24.44	30.59	15.30	48.26	59.83	19.18	64.30	79.87
w/o visual	7.82	24.65	32.82	14.99	48.21	63.56	18.96	64.07	84.13
w/o audio	6.30	22.65	31.88	12.61	44.56	61.85	16.54	59.37	82.13
w/o auxiliary	4.91	20.64	23.23	10.08	43.95	48.85	13.78	61.29	67.95

Table 2: **Ablation study.** We evaluate our method against its variants: without the visual stream, without the audio stream, and without the auxiliary task.

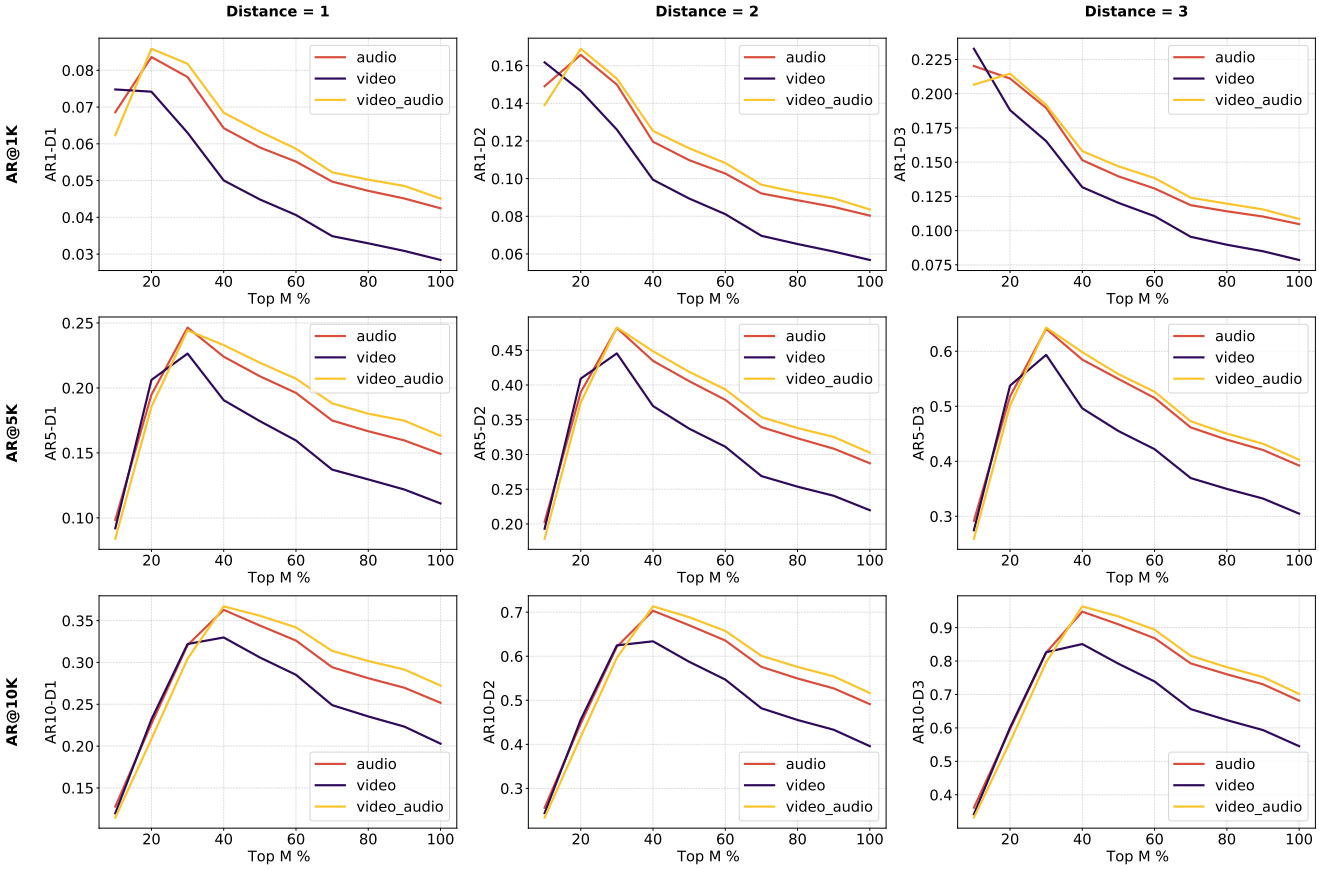
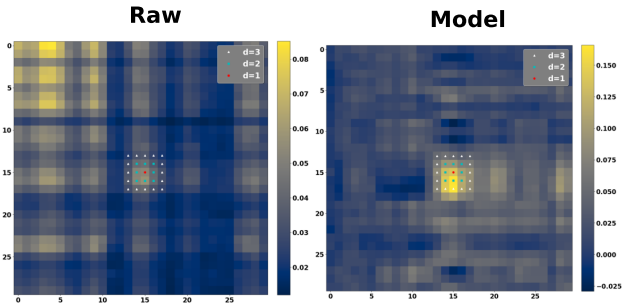


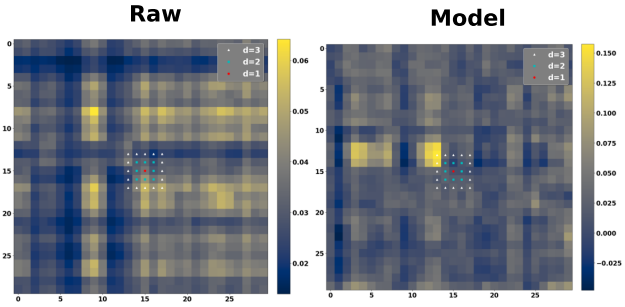
Figure 3: **Influence of top M% per metric.**

5. Qualitative Results

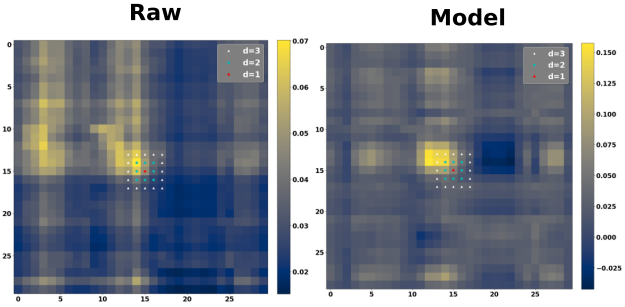
In Figure 4 and Figure 5 we show the feature temporal similarity of two candidate videos to be stitched together. The columns of the matrix represent snippets of video one, and the rows represent snippets of video 2. We show the similarity between these two set of snippets before (Raw) and after (Model) Learning to Cut. In this case all the cuts are in $(15, 15)$, we show in red the region of the ground-truth with distance $d = 1$, in cyan the edge of the region for ground-truth with $d = 2$, and in white the edge of the region for ground-truth with $d = 3$. On the one hand, we observe in figure 4 that our model tends to localize the similarities around the actual cutting place and the ground-truth regions. In Figure 4a, the most salient region is overlapping the ground-truth region after our model was applied, before it, the similarity spikes where located in a complete different place. Thus, our model was able to transform the video features such that the similarities spike around the cutting points. We observe similar behavior for Figure 4b and Figure 4c; however, the center of the spike region is a couple of spaces off the ground-truth region. On the other hand, we can observe in Figure 5 some examples in which the spike of the similarities do not match the ground-truth region. Interestingly Figure 5a and Figure 5b show that the cutting point for one of the videos was predicted correctly (the spike happens along the 15th row); yet, the model was not able to find a cutting point for the second video that would match the ground-truth. This does not necessarily mean that the cutting point found the model is not correct. It means that did not match the cut made by the professional. The Figure 5c shows a spike on a region that does not correspond with the ground truth. In this case, the model was not able to move the features away from the initial state, since the features were already spiking in a similar region before the model (raw column). Regardless of the ground-truth region, Figure 4 and Figure 5 show that our model helps to sharpen feature similarities in specific regions across a pair of videos. The similarity spike is not as blur anymore as it was in the original features (Raw). **Additional qualitative results with the actual clips ranking can be found on the attached files and slides. In the examples' files (Qualitative.zip) the videos are name after their ranking, the real cut is also included.**



(a) Correct prediction - Example 1

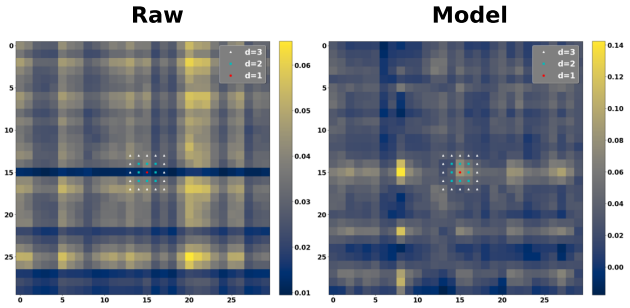


(b) Correct prediction - Example 2

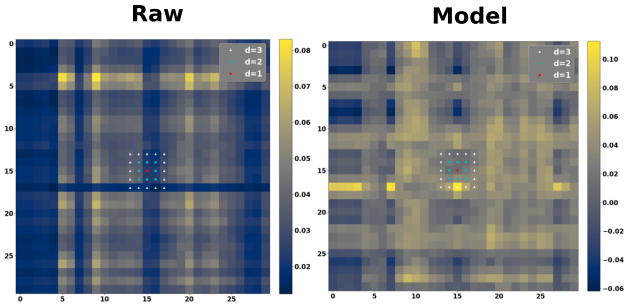


(c) Correct prediction - Example 3

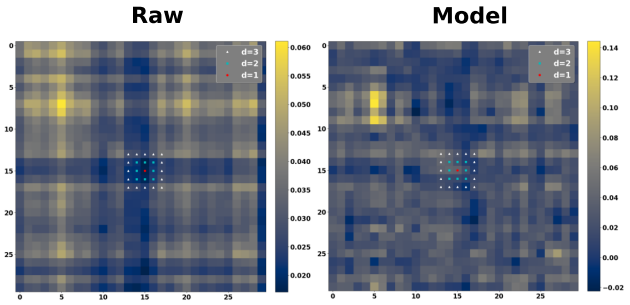
Figure 4: **Feature similarities before and after Learning to Cut.** Examples where the feature similarities were moved correctly to the ground-truth region. Please zoom in for a better view.



(a) Incorrect prediction - Example 1



(b) Incorrect prediction - Example 2



(c) Incorrect prediction - Example 3

Figure 5: **Feature similarities before and after Learning to Cut.** Examples where the feature similarities were moved out of the the ground-truth region. Please zoom in for a better view.