Learnable Optimal Sequential Grouping for Video Scene Detection

Supplementary Material

Daniel Rotman, Yevgeny Yaroker, Elad Amrani, Udi Barzelay, Rami Ben-Ari
[danieln@il.,yevgenyy@il.,elad.amrani@,udib@il.,ramib@il.]ibm.com
IBM Research
Haifa, Israel

A TRIPLET LOSS FOR VIDEO SCENE DETECTION

As mentioned in the paper, the triplet loss [5] learns a feature space embedding where samples from the same class are close in the feature space while samples from different classes are further apart. This is useful for a range of tasks, but for scene division this is doubly intuitive because the triplet loss causes samples (shots, in this case) to cluster together. In Figure 1 is a reduced 2-dimensional representation of shot feature vectors (using TSNE) from the video Meridian from the OVSD dataset. This video contains the smallest amount of shots, and offers the ability to visually and qualitatively inspect the distribution of the shot representations.

Despite the success of separating the shot representations into clusters, we can see that classic clustering algorithms might have trouble dividing correctly. Specifically we are referring to the single-shot scenes surrounding the large light blue square scene. In this instance, the OSG algorithm will likely be beneficial given the order and locations of the scenes, and the ability to make a decision based on the temporal order of the shots.

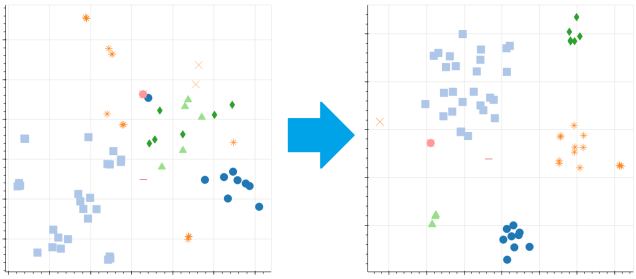


Figure 1: A reduced 2-dimensional representation of shot feature vectors (using TSNE) from the video Meridian from OVSD. Different marker types and colors represent scenes. After applying the triplet loss (right) the scenes are better separated into clusters.

B ESTIMATING THE NUMBER OF SCENES K

The number of divisions K is estimated using the log-elbow approach [3, 6]. To this end, the singular values of the distance matrix are computed, and the plot of the log values is analyzed. The point of plateau ('elbow') in the plot was shown to correspond to the number of blocks with intuition from performing a low-rank matrix approximation. The mathematical intuition is that given a distance matrix with a block-diagonal structure, we can see the rows of the matrix which belong to a specific block as being roughly linearly dependant. If the matrix were ideal (zeros on the block diagonal and ones outside), the rank of the matrix would be exactly the number of blocks in the block diagonal. Given a real noisy matrix, we expect the noise to act as the high frequency and low energy additions to the underlying inherent structure of the matrix. By identifying the plateau point of the singular values we can estimate the rank of the fundamental structure of the matrix.

Practically, this plateau point is located with an elbow estimation, as the point farthest from the diagonal running over the graph. Formally, if s is the log singular values of length N and we consider the index of each value as the first dimension, then the vector $I_i = [i, s_i]^T$ represents the values of the graph. The diagonal would be: $H = [N-1, s_N - s_1]^T$, with $\hat{H} = H/\|H\|$ the unit vector in the same direction, and using the euclidean distance to each point and projecting the vector I, we can identify the index of the plateau

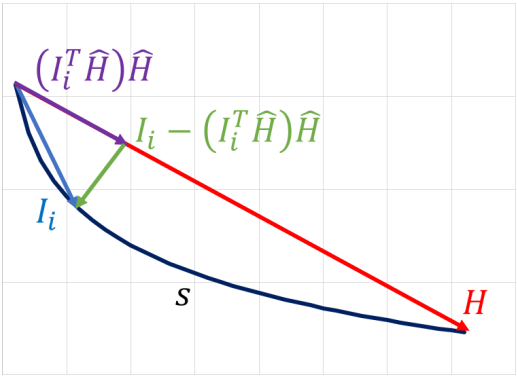


Figure 2: A depiction of the estimation of the log-elbow plateau point in the log graph of the singular values of the distance matrix.

point:

$$\text{log-elbow} = \arg\max_i \left\{ \|I_i - (I_i^T \hat{H})\hat{H}\| \right\}. \quad (1)$$

See Figure 2 for an illustration.

C OVSD DATASET

For video scene detection we used the OVSD dataset [4]. OVSD, is one of the only freely-available video scene detection datasets allowing both academic and industrial research use (creative commons licenses). To the extent of our knowledge, this dataset is the

only video scene detection dataset that has entire movies and is freely available with only minimal legal restrictions.

The dataset contains 21 full-length motion-picture films from a variety of genres with ground truth scene labeling. Table 1 presents the details of the OVSD dataset. 'Short Name' is the name presented in the results table in the paper to conserve space, and '# shots' is the number of shots as estimated using a shot boundary detection method [2]. Some videos are defined by a number of genres (as is acceptable with films). For the analysis per genre in the paper, the first genre was used to aggregate results, where Meridian was added to Crime (being the genre closest to Mystery).

D EVALUATION METRIC

We measure the performance of our OSG configurations on the OVSD dataset. For a metric, we use the widely accepted Coverage C and Overflow O [7], with a single value F-score for assessing the quality of division as the harmonic mean between C and $1 - O$.

Formally, as in [1], we denote $s_1, s_2, \ldots, s_m$ as the series of detected scenes, and $\tilde{s_1}, \tilde{s_2}, \ldots, \tilde{s_n}$ as the series of ground truth scenes, where each element s is a set of shots. The coverage C_t of ground truth scene $\tilde{s}_t$ is computed as:

$$C_t = \frac{\max_{i=1,\ldots,m} \#(s_i \cap \tilde{s}_t)}{\#\tilde{s}_t}, \tag{2}$$

where $\#(s)$ is the number of shots in scene s. Essentially, this is the relative amount of the ground truth scene that was allocated to a single scene in the proposed division. The overflow O_t for ground truth scene $\tilde{s}_t$ is computed as:

$$O_t = \frac{\sum_{i=1}^{m} \left[\#(s_i \setminus \tilde{s}_t) \cdot \min(1, \#(s_i \cap \tilde{s}_t))\right]}{\#(\tilde{s}_{t-1}) + \#(\tilde{s}_{t+1})}. \tag{3}$$

Essentially, $\min(1, \#(s_i \cap \tilde{s}_t))$ is a binary indicator whether scene s_i shares at least one shot with $\tilde{s}_t$, and $\#(s_i \setminus \tilde{s}_t)$ are the shots of these scenes which are not part of $\tilde{s}_t$. Therefore this measures how much the overlapping proposed scenes extend beyond the ground truth scene normalized by the number of shots in the neighboring scenes.

These measures for each ground truth scene are aggregated into video-wide metrics as the weighted average:

$$C = \sum_{t=1}^{n} C_t \cdot \frac{\#(\tilde{s}_t)}{\sum_i \#(\tilde{s}_i)}, \qquad O = \sum_{t=1}^{n} O_t \cdot \frac{\#(\tilde{s}_t)}{\sum_i \#(\tilde{s}_i)}. \tag{4}$$

Finally, as a single score for the quality of the scene detection, we compute the harmonic mean:

$$F = 2 \cdot \frac{C \cdot (1 - O)}{C + (1 - O)}. \tag{5}$$

E ADDITIONAL D EXAMPLES

In Table 2 we present various stages of D from visual features of the video La Chute D'une Plume from OVSD with the accompanied ground truth D^*, and in Table 3 the same for the video Big Buck Bunny. In Tables 4 and 5 we show how the D matrices and gradients evolve over a number of epochs for the videos La Chute D'une Plume and Big Buck Bunny respectively.

In general our observations are that OSG-Triplet manages to emphasize the small scenes better than large scenes, while OSG-Block is the reverse. OSG-Block-Adjacent gives a good trade-off of emphasizing the immediate off-diagonal, but results in some low distances in the far off-diagonal. In practice, these shouldn't affect the OSG algorithm if the intervening distances are large enough. OSG-Prob converges more slowly, learns from the boundary edges, and gives a good trade-off as well.

Regarding this last point, part of our motivation for OSG-Prob is to have a configuration which is specifically reliant on division locations as opposed to the block-diagonal. Such a structure would allow OSG to be integrated into a larger learning pipeline. For example, there are other temporal analysis tasks where division is only a part of the process. In the weakly-supervised regime there might not be ground truth divisions with which to perform OSG-Triplet or OSG-Block. OSG-Prob on the other hand, could be configured to perform backpropagation on a loss which reflects on the locations of division, and is inferred back to the distance values. In this respect, our continued research involves having this component as a plug-and-play module for other tasks which can act as temporal region proposal networks (see Figure 3).

F ADDITIONAL T EXAMPLES

In Figure 4 and 5 we show the progression of the values of $T(i)$ over a number of iterations for videos La Chute D'une Plume and Big Buck Bunny respectively. We can see that as the iterations progress, T raises the probability at the ground truth points of division. The probability is lowered for locations with no true division even though this is not specifically enforced by the loss but rather an outcome of the construction of T.

Specifically we note that in these instances the small beginning scenes proved difficult for the network to emphasize T on. We speculate that this is due to the formulation, where smaller values of n inspect longer and longer sequences (see the formulation in the paper). Possible future work could be to formulate an additional mirrored OSG which inspects the D matrix backwards.

G ADDITIONAL VISUAL RESULTS

Figures 6, 7, and 8, show results on sections of videos from the OVSD dataset. In general, we can see divisions which result in reasonable and often precise scene divisions. Using these divisions for applying video understanding and classification technologies will undoubtedly be superior over applying them on the entire video or on naive uniform divisions. Specifically, Figure 8 is a single scene from the video Tears of Steel which includes intricate character and setting changes. This portrays the complexity of the task and the challenges that the method needs to overcome. Despite the fact that all of the proposed scene divisions in this case are technically false, we note that they present a plausible division to story-units, and can be useful for a variety of downstream tasks.

REFERENCES

[1] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. 2015. Measuring scene detection performance. In *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 395–403.

[2] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. 2015. Shot and scene detection via hierarchical clustering for re-using broadcast video. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 801–811.

[3] Daniel Rotman, Dror Porat, and Gal Ashour. 2016. Robust and efficient video scene detection using optimal sequential grouping. In *2016 IEEE International Symposium on Multimedia (ISM)*. IEEE, 275–280.

[4] Daniel Rotman, Dror Porat, and Gal Ashour. 2017. Robust video scene detection using multimodal fusion of optimally grouped features. In *2017 IEEE 19th*

Figure 3: OSG-Prob as a plug-and-play temporal region proposal network.

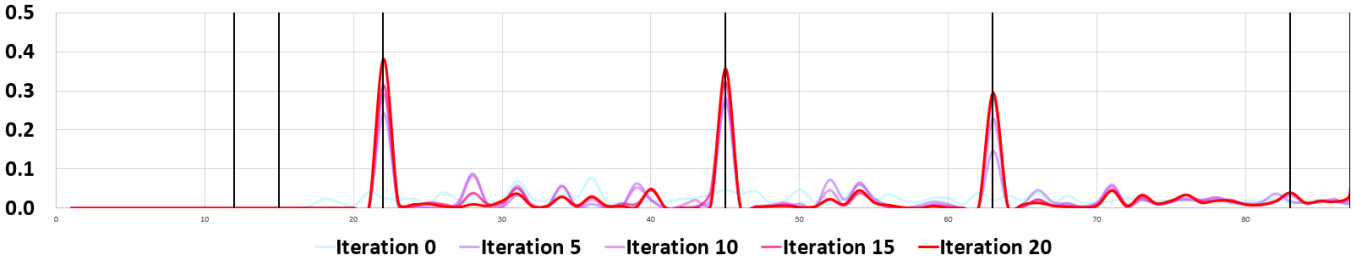


Figure 4: Progression of T as a function of i (shot number) over a number of iterations for video La Chute D'une Plume. Graphs go from translucent blue to opaque red as iterations progress (best viewed in color). Vertical black lines indicate ground truth divisions.

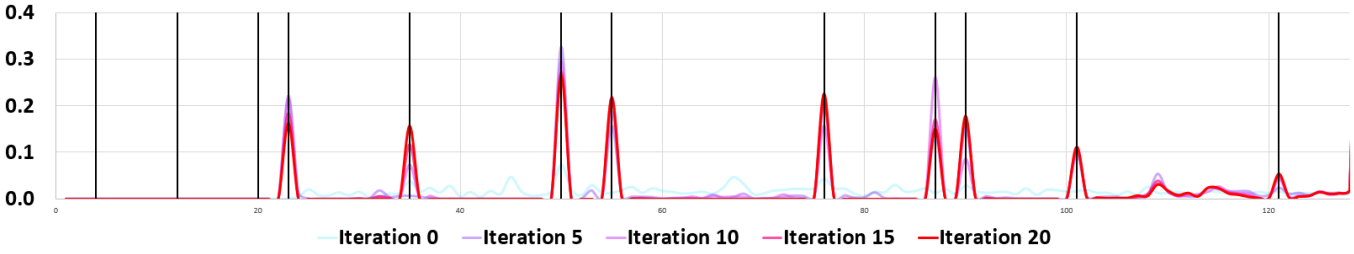


Figure 5: Progression of T as a function of i (shot number) over a number of iterations for video Big Buck Bunny. Graphs go from translucent blue to opaque red as iterations progress (best viewed in color). Vertical black lines indicate ground truth divisions.

International Workshop on Multimedia Signal Processing (MMSP). IEEE, 1–6.

[5] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[6] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. 2014. Storygraphs: visualizing character interactions as a timeline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 827–834.

[7] Jeroen Vendrig and Marcel Worring. 2002. Systematic evaluation of logical story unit segmentation. *IEEE Transactions on Multimedia* 4, 4 (2002), 492–499.

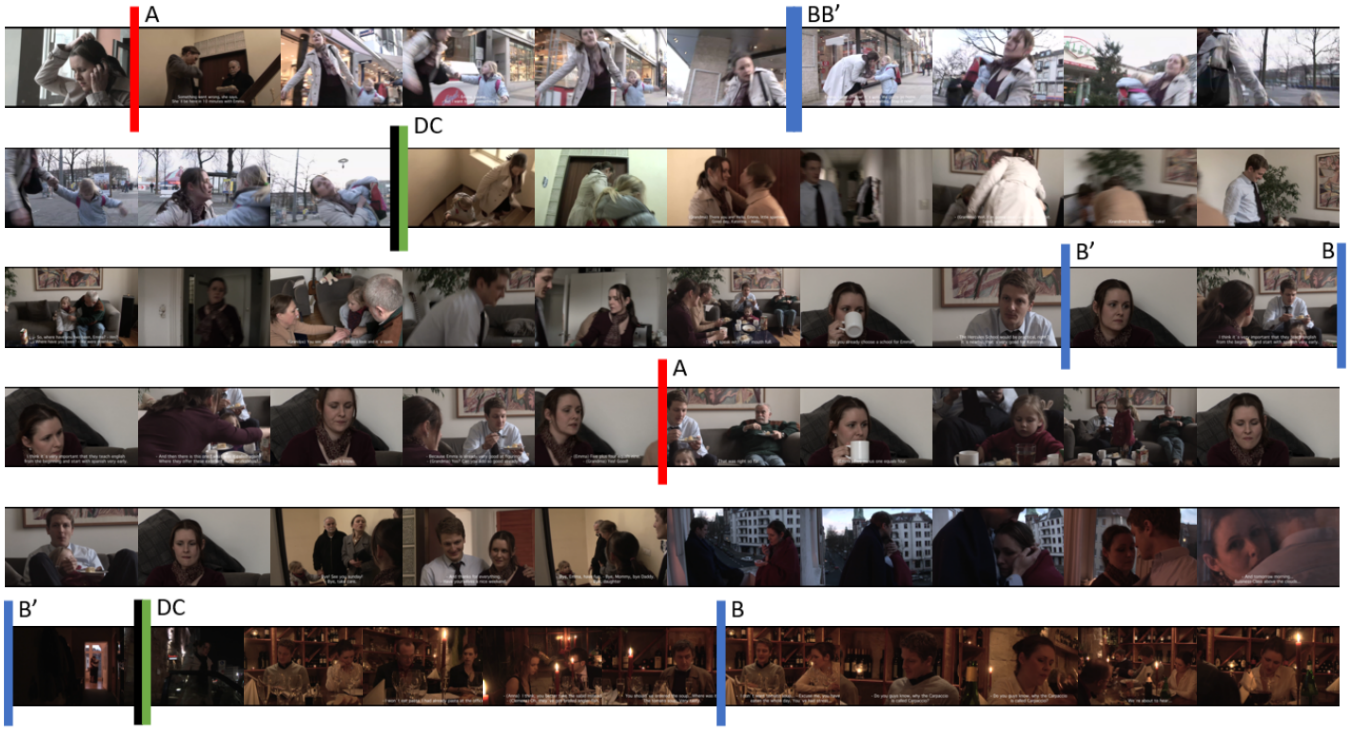


Figure 6: Qualitative results of configurations on shots 68 through 128 of the video 1000 Days from the OVSD dataset. Points of division marked by A. OSG-Triplet (red) B. OSG-Block (blue) B‘. OSG-Block-Adjacent (blue) C. OSG-Prob (green) and D. Ground truth (black).

Figure 7: Qualitative results of configurations on a section of the video Meridian from the OVSD dataset. Points of division marked by A. OSG-Triplet (red) B. OSG-Block (blue) B‘. OSG-Block-Adjacent (blue) C. OSG-Prob (green) and D. Ground truth (black).

Table 1: OVSD dataset details

Video Name	Short Name	Duration (minutes)	# Scenes	# Shots	Genre
1000 Days	1000	43	23	404	Drama
Big Buck Bunny	BBB	8	13	129	Animation
Boy Who Never Slept	BWNS	69	23	336	Comedy, Romance
CH7	CH7	86	45	1293	Crime
Cosmos Laundromat	CL	10	6	94	Animation
Elephants Dream	ED	9	8	128	Animation
Fires Beneath Water	FBW	76	63	411	Documentary
Honey	Honey	86	21	326	Drama
Jathia's Wager	JW	21	16	177	Drama, Sci-Fi
La Chute D'une Plume	LCDP	10	11	88	Animation
Lord Meia	LM	37	28	333	Crime, Comedy
Meridian	Meridian	12	10	64	Mystery, Sci-Fi
Oceania	Oceania	54	32	253	Drama, Mystery
Pentagon	Pentagon	50	32	305	Comedy, Drama
Route 66	Route 66	103	56	1357	Documentary
Seven Dead Men	SDM	57	35	167	Crime
Sintel	Sintel	12	7	198	Animation
Sita Sings the Blues	SStB	81	53	1384	Animation, Comedy
Star Wreck	SW	103	56	1439	Comedy, Sci-Fi
Tears of Steal	ToS	10	6	136	Drama, Sci-Fi
Valkaama	Valkaama	93	49	714	Drama

Table 2: An example D from the video La Chute D'une Plume from OVSD. On the left: Ground Truth (D^*), Orig (without an applied embedding), Epoch 0 (embedding before learning). On the right, trained examples after 20 epochs for: OSG-Triplet, OSG-Block, OSG-Block-Adjacent, and OSG-Prob, with corresponding gradients (bottom row)

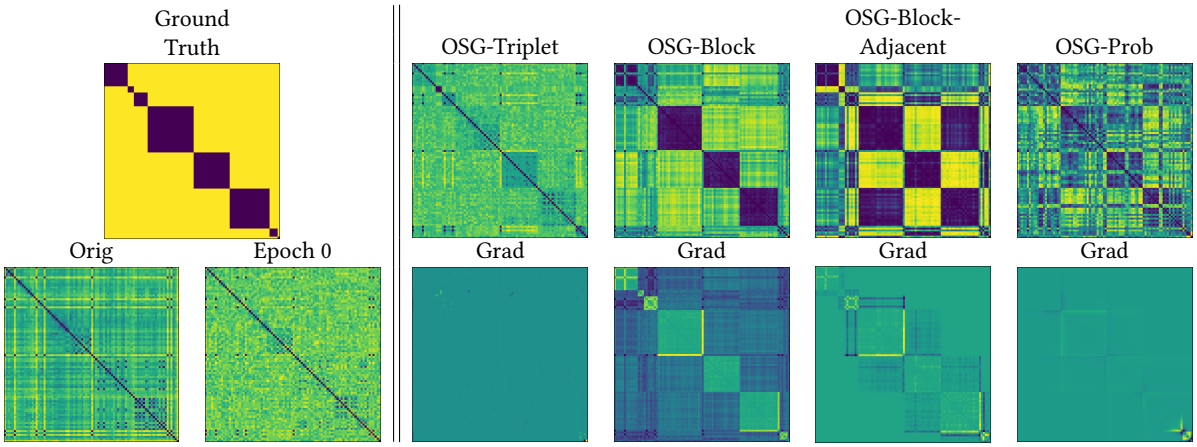


Table 3: An example D from the video Big Buck Bunny from OVSD. On the left: Ground Truth (D^*), Orig (without an applied embedding), Epoch 0 (embedding before learning). On the right, trained examples after 20 epochs for: OSG-Triplet, OSG-Block, OSG-Block-Adjacent, and OSG-Prob, with corresponding gradients (bottom row)

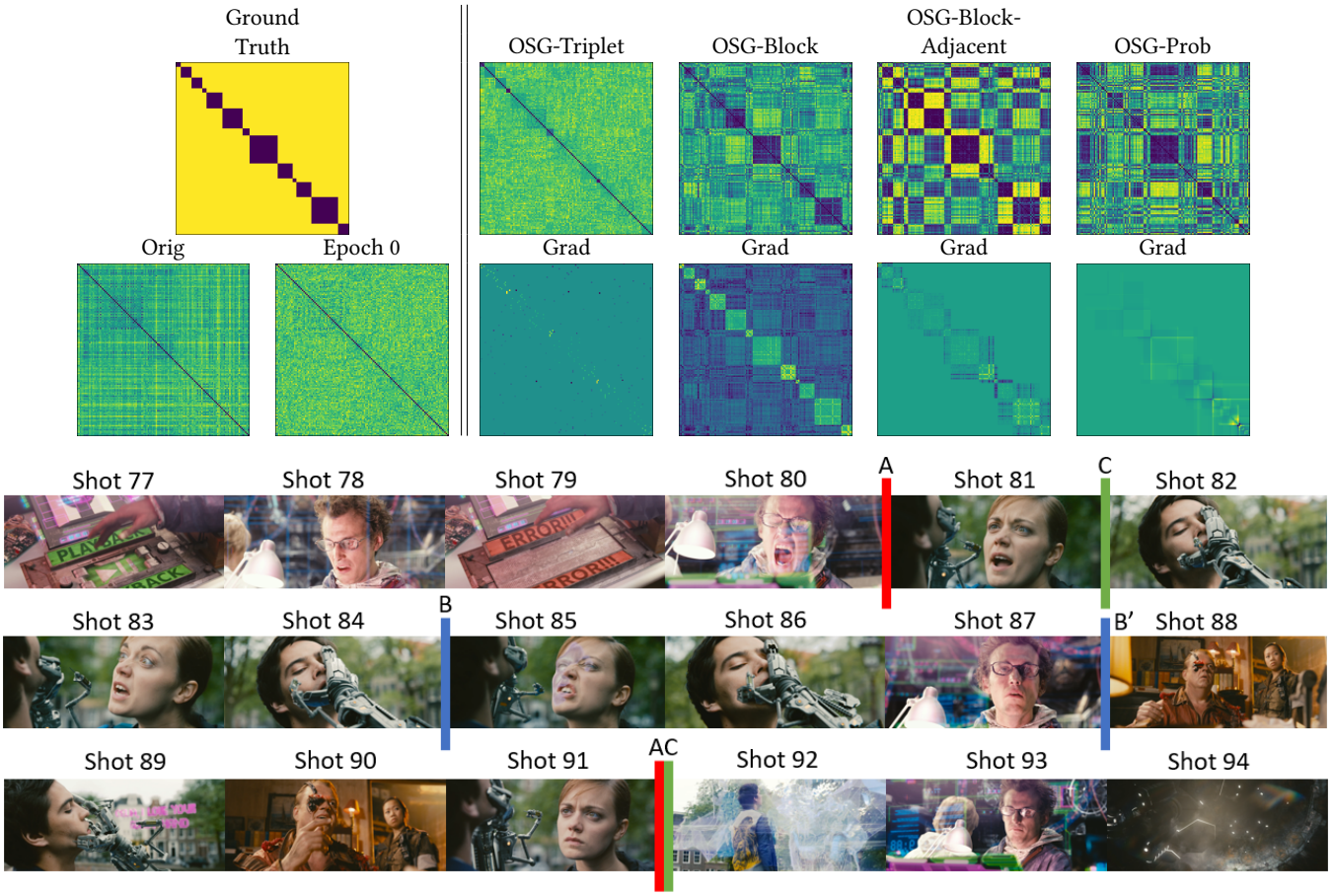
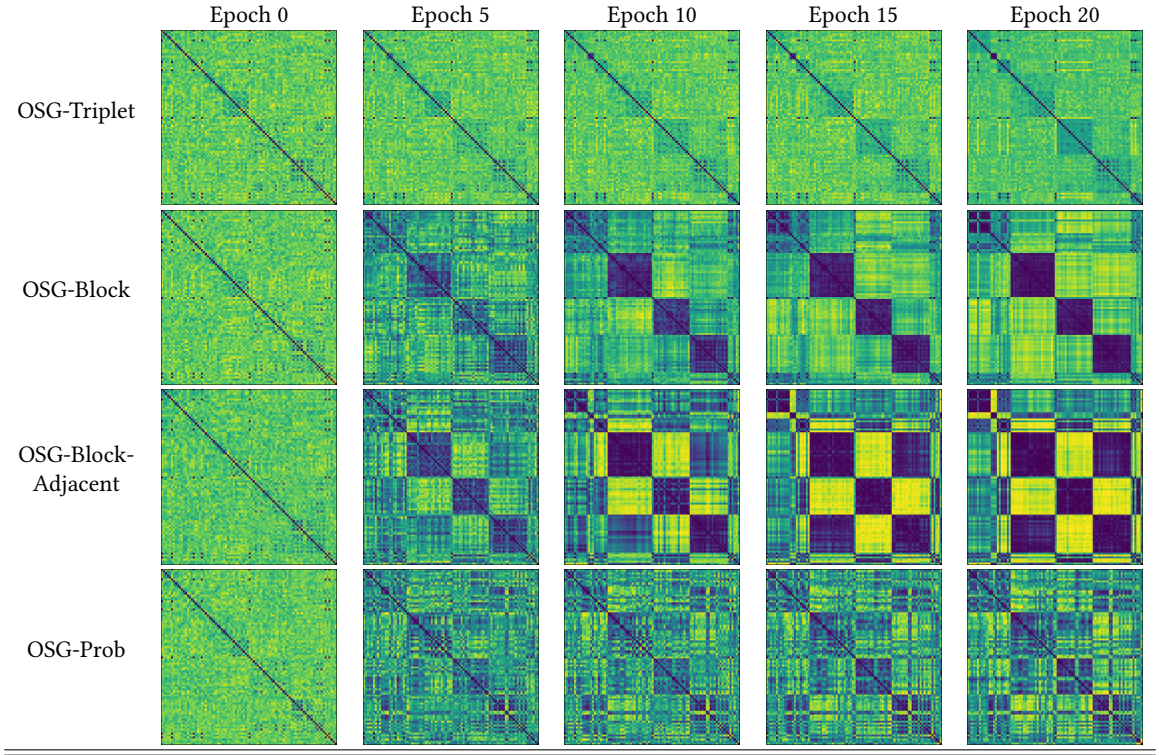


Figure 8: Qualitative results of configurations on a section of the video Tears of Steel from the OVSD dataset. Points of division marked by A. OSG-Triplet (red) B. OSG-Block (blue) B'. OSG-Block-Adjacent (blue) C. OSG-Prob (green). The shots are part of a single complex ground truth scene.

Table 4: D and gradients from the video La Chute D'une Plume from OVSD evolving over a number of Epochs



Gradients

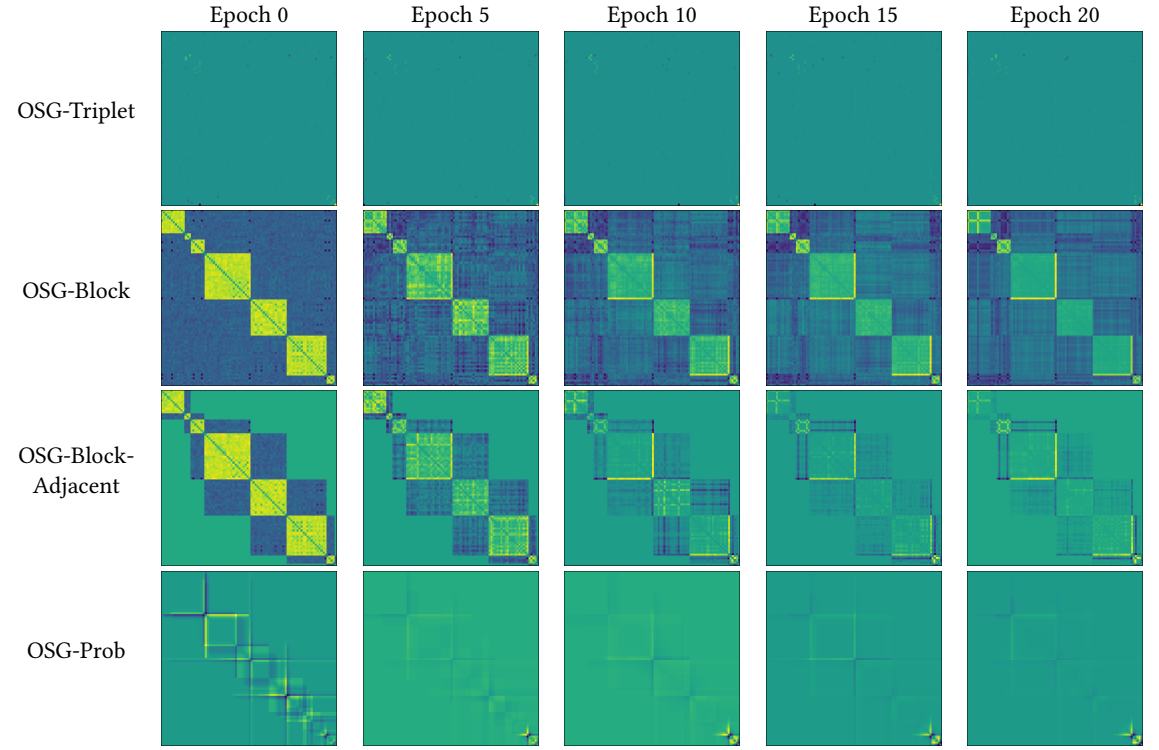


Table 5: D and gradients from the video Big Buck Bunny from OVSD evolving over a number of Epochs