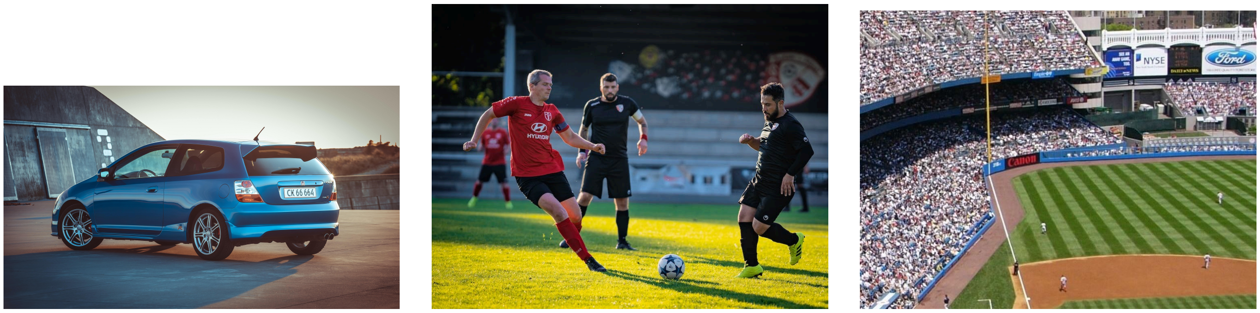
Appendix



(a) (b) (c)

Figure 6. (a) An intrusive insertion example where a commercial is played during game break, (b) an non-intrusive insertion example where the “Hyundai” logo displayed on the shirt of the player, and (c) another non-intrusive insertion example where the ”Ford” and ”Canon” brand names displayed on the billboard. All the images shown here are publicly available.

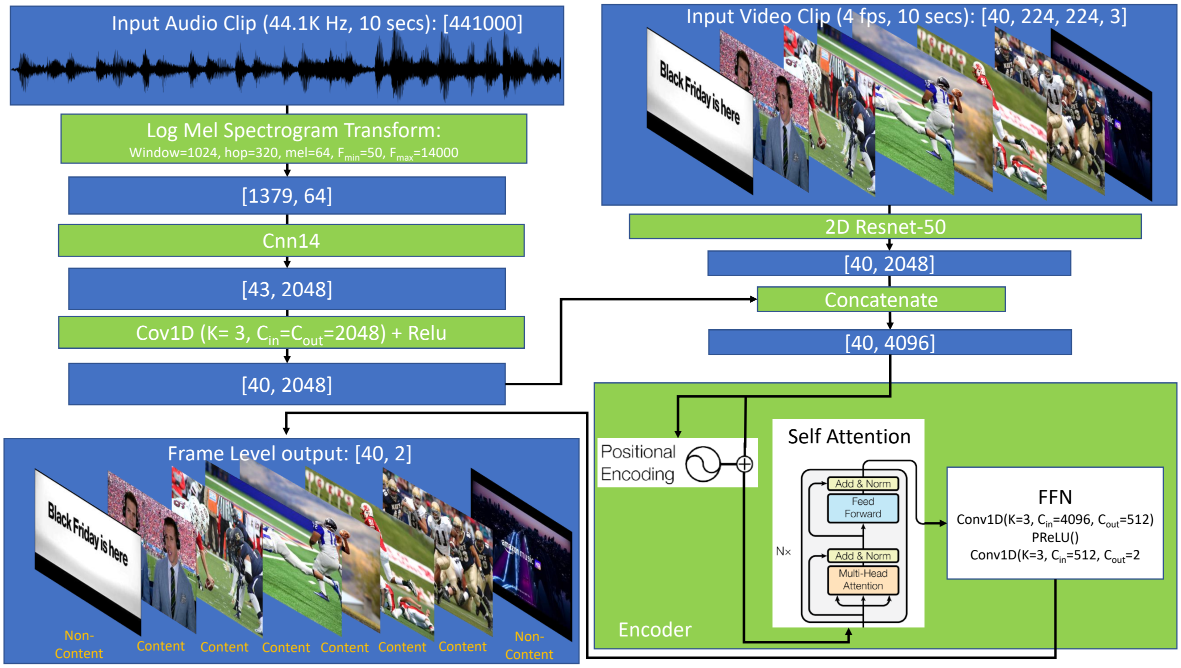

Input Audio Clip (44.1K Hz, 10 secs): [441000]
Log Mel Spectrogram Transform:
Window=1024, hop=320, mel=64, F_{min}=50, F_{max}=14000
[1379, 64]
Cnn14
[43, 2048]
Cov1D (K= 3, C_{in}=C_{out}=2048) + Relu
[40, 2048]
Input Video Clip (4 fps, 10 secs): [40, 224, 224, 3]
Black Friday is here
2D Resnet-50
[40, 2048]
Concatenate
[40, 4096]
Frame Level output: [40, 2]
Black Friday is here
Non-Content
Content
Content
Content
Content
Content
Content
Non-Content
Positional Encoding
Self Attention
Add & Norm
Feed Forward
Nx
Add & Norm
Multi-Head Attention
FFN
Conv1D(K=3, C_{in}=4096, C_{out}=512)
PReLU()
Conv1D(K=3, C_{in}=512, C_{out}=2
Encoder

Figure 7. The workflow of the short-term frame level classifier where the input is an 10 seconds long video+audio clip and the output is the classification result of every frame. The images listed here are publicly available online.

A. Feature Histogram Study on League Promotion Classifier

In Sec. 3.2 we proposed a hypothesis that *league promotion segments typically have more shot changes than content segments*. To empirically validate this assumption, we sampled 94 sport games and 86 league promotion segments, compute the feature histogram, and plot the average value in Fig. 8. Note that here we dropped the last bin which is the sum of the first nine bins, in order to better visualize. We can see that the non-content (green) class has higher values than the content (red) class in all bins, which supports our hypothesis.

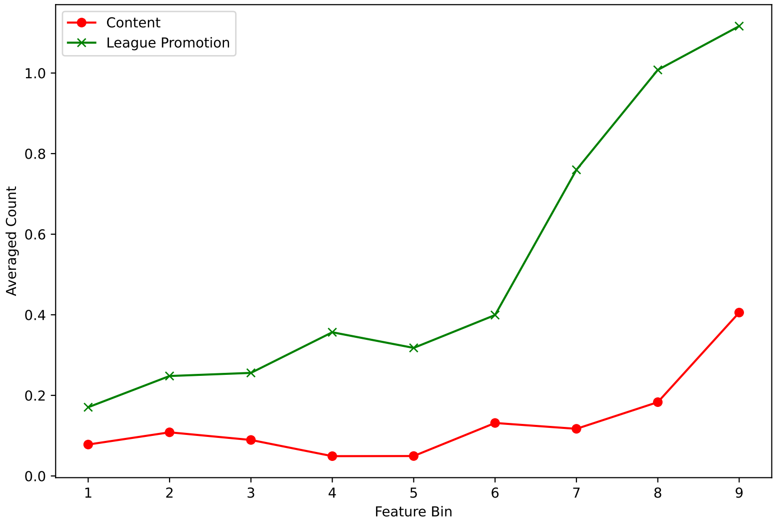


Figure 8. A sample feature histogram used as the input of the League Promotion Classifier as defined in Sec. 3.2 (Fig. 3).

B. Short-term Classifier Video/Audio Encoders Training Process

In Section 4.5, we studied the performance of different types of encoders in the short-term classifier. The results are presented in Table 6. To ensure a fair comparison, we retrained all the encoders using the same dataset as defined in Section 4.1. We employed transfer learning and initialized the weights with pretrained models obtained from [15,27,28]. Specifically, we used the following models for initialization: vit_base_patch16_224_in21k for the Vit model, swin_base_patch4_window7_224 for the Swin2D model, facebook/wav2vec2-base-960h for the Wav2Vec2 model, and facebook/data2vec-audio-base-960h for the Data2Vec Audio model.

The model complexity metrics in Table 6, namely *Memory Space* and *Running Time*, were measured by running the models in inference mode on an AWS P3 instance with a single Tesla V100 GPU.

C. Accuracy Metric Definition

The accuracy of a classifier is commonly measured by three metrics: precision, recall and $f1$ score. The *precision* metric, defined in Eq. 2, measures the accuracy of positive predictions made by a classifier. It calculates the proportion of true positive predictions among all the positive predictions made by the model.

$$Precision = \frac{\text{\# of correct positive prediction}}{\text{\# of total positive predictions}} \tag{2}$$

The *recall* metric, defined in Eq. 3, measures the ability of a classifier to identify all positive instances correctly. It calculates the proportion of true positive predictions among all the actual positive instances.

$$Recall = \frac{\text{\# of correct positive predictions}}{\text{\# of ground truth positive instances}} \tag{3}$$

The $f1$ score as defined in Eq. 4 combines precision and recall into a single metric that balances their contributions. It is the harmonic mean of precision and recall, providing a comprehensive evaluation of a classifier's performance. It is particularly useful when there is an uneven distribution between positive and negative instances in the dataset.

$$f1 = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision + Recall}} \tag{4}$$