Movie Genre Classification by Language Augmentation and Shot Sampling Supplementary

Zhongping Zhang[1] Yiwen Gu[1] Bryan A. Plummer[1]
Xin Miao[2] Jiayi Liu[2] Huayan Wang[2]
[1]Boston University [2]Kuaishou Technology
[1]{zpzhang, yiweng, bplum}@bu.edu [2]wanghy514@gmail.com

A. Expanded Experimental Results

A.1. Per-genre performance

In Figure 1, we present the performance of Movie-CLIP across each genre on the MovieNet dataset. We observe that Movie-CLIP achieves robust performance across a majority of genres. However, the accurate prediction of genres such as *Mystery*, *Biography*, or *History* is still challenging. This difficulty may arise due to the imbalanced label distribution among various genres. Therefore, exploring methods to solve the long-tail distribution issue can be promising to further improve our model's performance.

A.2. Low-level Visual Attribute Analysis.

In Figure 2, we analyze the distribution of brightness and warm-cold color ratio across genres to explore correlations between genres and low-level visual attributes. From the figure, we observe that *horror* genre exhibits the lowest brightness value, which is consistent with the intuition that *horror* films aim to evoke the fear emotion through dark and ominous atmospheres. In contrast, genres such as *family* and *animation* tend to favor higher brightness values. This trend corresponds with the inherent intent of these genres, which is to convey feelings of love and warmth to their audience.

Regarding the cold-warm color ratio, *western* films demonstrate the lowest values, while *Sci-Fi* achieves the highest values. This contrast can be attributed to the characteristics of these genres, *i.e.*, *Western* films often has a sepia tone due to scenes like desert, blazing sun, and dirt. On the other hand, *Sci-Fi* movies lean towards colder colors for scenes such as the universe, spacecraft, or robot armies, contributing to an impression of high-tech and sharpness

A.3. Ablation Study on Visual Features

To evaluate the effectiveness of our proposed mechanism when using various pretrained visual features, we present the experiment results where we replaced CLIP features with ResNet-50 in Table 1. We observe that our method continues to improve the performance compared to the base models (44.16 vs. 47.40 vs. 49.25 in macro-mAP), validating that the effectiveness of our method does not simply rely on CLIP features.

Method	macro-mAP	micro-mAP
Shot (ResNet-50)	44.16	54.29
Shot+Audio	47.40	57.16
Shot+Audio+Language	49.25	62.53

Table 1. **Applying ResNet-50 features as visual representations on MovieNet:** We still observe performance improvements by our proposed approach, even the visual representations are based on ResNet-50 features.

A.4. Additional Sound Event Analysis

Figure 7 provides sound events of various movie genres, supplementing the results from the main paper. The distinctive attributes of these sound events substantiate our motivation for incorporating them to augment our model.

A.5. Additional Keyword Analysis

We plot wordclouds in Figure 4 to supplement the main paper. These visualizations provide additional support for using keyword-aware documents to improve our model.

A.6. Additional Genre-based Shot Retrieval

Additional examples of genre-based shot retrieval, Titanic (1997) and Jurassic Park (1993), are presented in Figure 5 and Figure 6, demonstrating that Movie-CLIP effectively identified shots aligned with different genres across video content of hours duration.

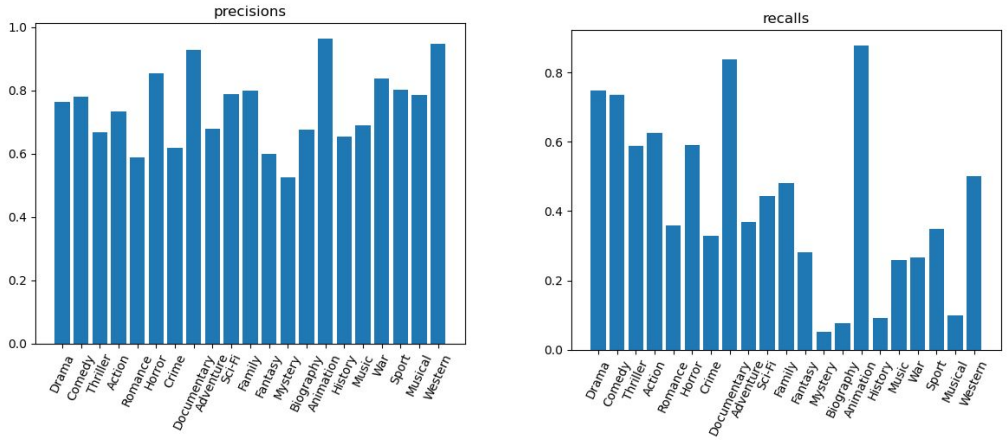

Figure 1. Per-genre performance of Movie-CLIP on MovieNet (Left: precision; Right: recall). See Appendix A.1 for discussion.

Datasets	MovieNet	Condensed Movies
Type of Video	Trailer	Movie Clip
Total	28,466	22,174
Training	19,926	15,521
Validation	2,846	2,217
Test	5,694	4,436

Table 2. Statistics of source videos on MovieNet and Condensed Movies.

B. Statistics of Videos

We present the statistics of source videos that we used in MovieNet and Condensed Movies in Table 2 and Figure 3. From Figure 3, we observe that the label distribution is remarkably imbalanced, validating the significance of using "micro" and "macro" metrics in our experiments.

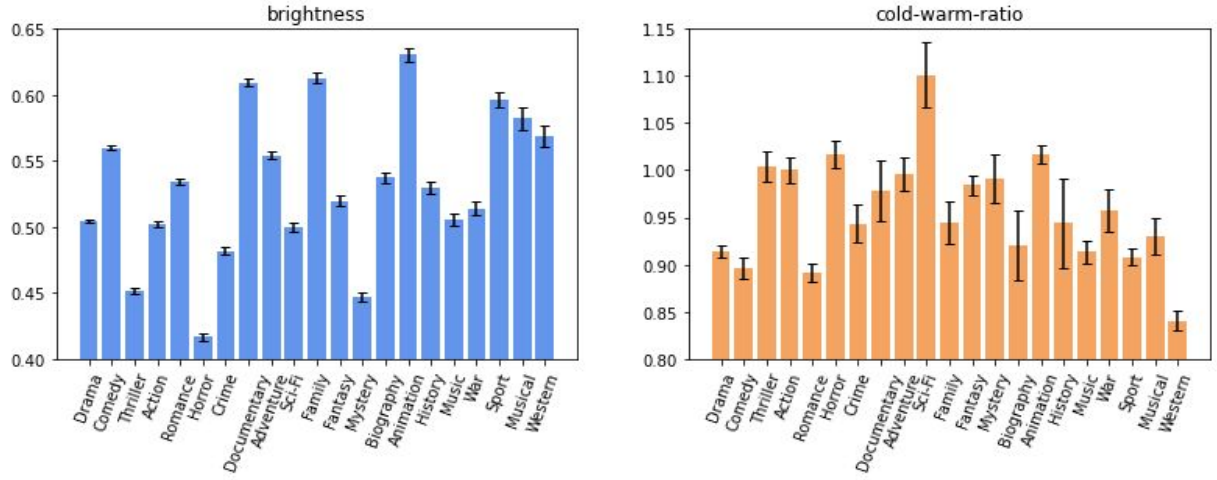

Figure 2. Low-level visual feature analysis across movie genres. Left: brightness with confidence interval; Right: cold-warm color ratio with confidence interval. See Appendix A.2 for discussion.

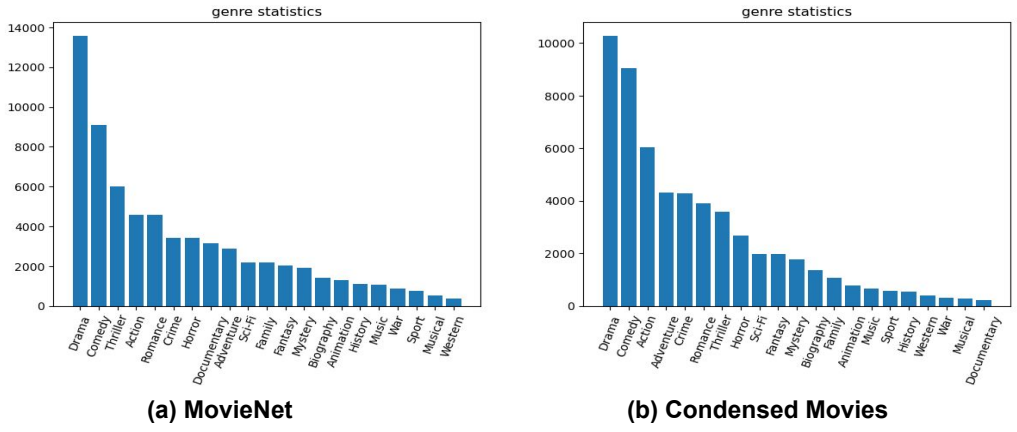

Figure 3. Distribution of genres on MovieNet (left) and Condensed Movies (right). See Appendix B for discussion.

Figure 4. Wordclouds of different movie genres on MovieNet. See Appendix A.5 for discussion.

romance

drama

music

adventure

thriller

western

Figure 5. Genre-based shot retrieval on “Titanic.” See Appendix A.6 for discussion.

adventure

sci-fi

comedy

horror

documentary

family

Figure 6. Genre-based shot retrieval on “Jurassic Park.” See Appendix A.6 for discussion.

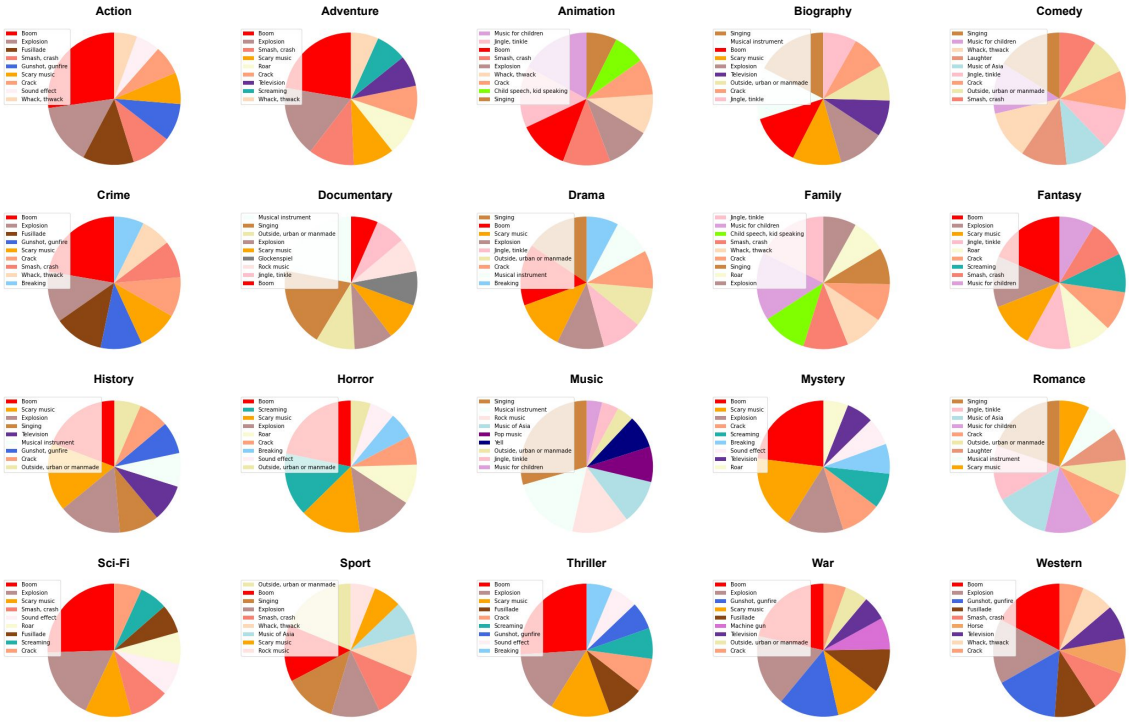

Figure 7. Sound events of different movie genres on MovieNet. See Appendix A.4 for discussion.