

# MovieCLIP: Visual Scene Recognition in Movies

Digbalay Bose<sup>1</sup>, Rajat Hebbar<sup>1</sup>, Krishna Somandepalli<sup>2</sup>, Haoyang Zhang<sup>1</sup>, Yin Cui<sup>2</sup>, Kree Cole-McLaughlin<sup>2</sup>, Huisheng Wang<sup>2</sup>, and Shrikanth Narayanan<sup>1</sup>

<sup>1</sup>University of Southern California, Los Angeles, CA <sup>2</sup>Google

<sup>1</sup>{dbose@, rajatheb@, zhangh21@, shri@ee.}usc.edu <sup>2</sup>{ksoman@, yincui@, kree@, huishengw@}google.com

## Abstract

This document contains the supplementary material associated with the submission: *MovieCLIP: Visual Scene Recognition in Movies*.

## 1. Introduction

In this document, we provide supplementary material associated with the submission: **MovieCLIP: Visual Scene Recognition in Movies**.

### 1.1. Scene classes distribution wrt sources

Here **Movie Slugline** refers to the set of labels obtained exclusively from sluglines in movie scripts. **Common Label** refers to the set of common labels between taxonomy considered in HVU [6] and **Movie Slugline**. Here HVU [6] refers to the set of labels obtained exclusively from the taxonomy used for curating HVU [6] dataset. **Human expert** refers to the set of labels added by the human expert during the taxonomy refinement procedure.

- **Movie Slugline:** *tent, computer room, truck, study, gas station, cafe, shuttle, courthouse, elevator, tower, dorm, station, club, lobby, mall, salon, prison, bus, stairs, theater, car, booth, locker room, hangar, closet, farmhouse, post office, townhouse, ship, loft, yard, zoo, funeral, art gallery, castle, subway, lounge, train, morgue, museum, wagon, manor, mansion, library, pool, cellar, cab, safe house, classroom, helicopter, police station, courtroom, city hall, fire station, corridor, control room, airport, cabin, war room, plane, press room, cottage, residence, penthouse, inn, church, suburban, interrogation room, conference room*
- **Common label:** *tunnel, bakery, shack, building, baseball field, hotel, desert, factory, bathroom, downtown, restaurant, village, playground, boxing ring, gym, bridge,*

*beach, workshop, cave, clinic, arena, garden, stage, office, attic, bowling alley, apartment, deck, cockpit, dining room, basketball court, grove, ballroom, forest, house, barn, alley, park, bay, golf course, chapel, home, parking, bar, kitchen, school, swamp, basement, walkway, bedroom, garage, lake, bank, living room, room, auditorium, street, valley, casino, hall, waterfall, warehouse, tennis court, farm, hospital, palace, estate, river*

- **HVU:** *archaeological site, shore, batting cage, animal shelter, plaza, hot spring, harbor, bullring, sandbank, town, mountain, retail, courtyard, sea, road, shooting range, pond, stadium, foundry, skyline, amusement park, market, laboratory, race track, kindergarten, ice rink*
- **Human expert:** *agriculture field, makeup studio, grassland, construction site, graveyard, automotive repair, overpass, studio, boat, fair, balcony, battlefield, banquet, phone booth, concert hall, meadow*

### 1.2. Hierarchy discovery in visual scene classes:

We extract dense 384D representations using the MiniLM-L6-v2 sentence transformers model [16] of the visual scene labels. Instead of manually binning the visual scene classes into subgroups, we use Affinity propagation clustering [9] with the the given parameter settings to obtain the clusters among the labels:

- *max<sub>iter</sub>* : 500
- *random<sub>state</sub>* : 12345
- *affinity* : euclidean
- *damping* : 0.5

The clusters are shown in Table 1. We can see from table 1, the clusters provide set of semantic grouping between visual scene classes that are closely related. Examples include cluster 5, where the visual scene classes associated with study i.e. *classroom, school, study, kindergarten* are listed together. In cluster 6, the visual scene classes associated

Cluster id	Visual scene labels
0	walkway, overpass, alley, courtyard, bridge, stairs, tunnel, corridor
1	funeral, graveyard, morgue
2	airport, station, train, fire station, subway, police station, post office, gas station
3	closet, conference room, apartment, press room, room, cabin, locker room, war room, computer room, lounge, bedroom, living room, dorm, kitchen, interrogation room, control room, dining room, ballroom, bathroom
4	pond, waterfall, river, hot spring, swamp, pool, lake
5	classroom, school, study, kindergarten
6	race track, batting cage, basketball court, tennis court, baseball field, golf course
7	attic, cellar, barn, cave, basement, forest
8	courtroom, courthouse, auditorium, stage, prison, theater
9	casino, ice rink, stadium, gym, arena, concert hall
10	safe house, cottage, yard, manor, castle, animal shelter, house, hall, shack, residence, farmhouse, estate, mansion, penthouse, home, garage
11	booth, tent, phone booth
12	parking, automotive repair, cab, car, truck, bus, wagon
13	laboratory, clinic, hospital
14	archaeological site, chapel, library, zoo, museum, hangar, art gallery
15	plaza, village, street, road, downtown, city hall, townhouse, town, suburban, market
16	playground, park, fair, amusement park
17	bay, beach, sandbank, shore, sea
18	bullring, boxing ring, battlefield, bowling alley
19	harbor, boat, ship, deck
20	tower, elevator, loft, lobby, construction site, skyline, palace, mall, church, building, balcony
21	cockpit, shuttle, plane, helicopter
22	shooting range, valley, desert, mountain
23	agriculture field, grove, grassland, garden, farm, meadow
24	café, bar, hotel, bakery, club, inn, bank, restaurant, banquet
25	foundry, makeup studio, warehouse, studio, retail, salon, office, factory, workshop

Table 1. Automatic clusters discovered using Affinity propagation clustering of the visual scene classes in the 384 D space.

with sports like *batting cage, basketball court, golf course* etc are grouped together. We also use tSNE [18] to project the 384 D embeddings to 2 dimensions and plot the layout of various visual scene classes in Fig 1. Certain groups of visual scene classes that are semantically close are encircled in Fig 1. Certain groupings of visual scene classes that are clustered in Table 1 are present in Fig 1. Examples include the visual scene classes associated with **study** (Cluster 5 in Table 1), **water bodies** (Cluster 4 in Table 1), **medical scene classes** (Cluster 13 in Table 1), **natural landforms** (Cluster 22 in Table 1), **automobile based scene classes** (Cluster 12 in Table 1), **dining scene classes** (Cluster 24 in Table 1).

## 2. MovieCLIP dataset

### 2.1. Shot statistics

We use PyScenedetect to extract shots from the movieclips provided with Condensed Movies. The average

number of shots per year and distribution of shots wrt different genres is shown in Fig 2.

## 3. CLIP based visual scene labeling

### 3.1. Prompt design choices

In terms of prompt designs, we consider the following templates customized for generic background information:

- *A photo of {}, a type of background location*
- *A photo of {}, a type of location*

Since multiple shots consider people in focus, we also consider the following people-centric prompt templates:

- *People at {}, a type of background location*
- *People at {}, a type of location*

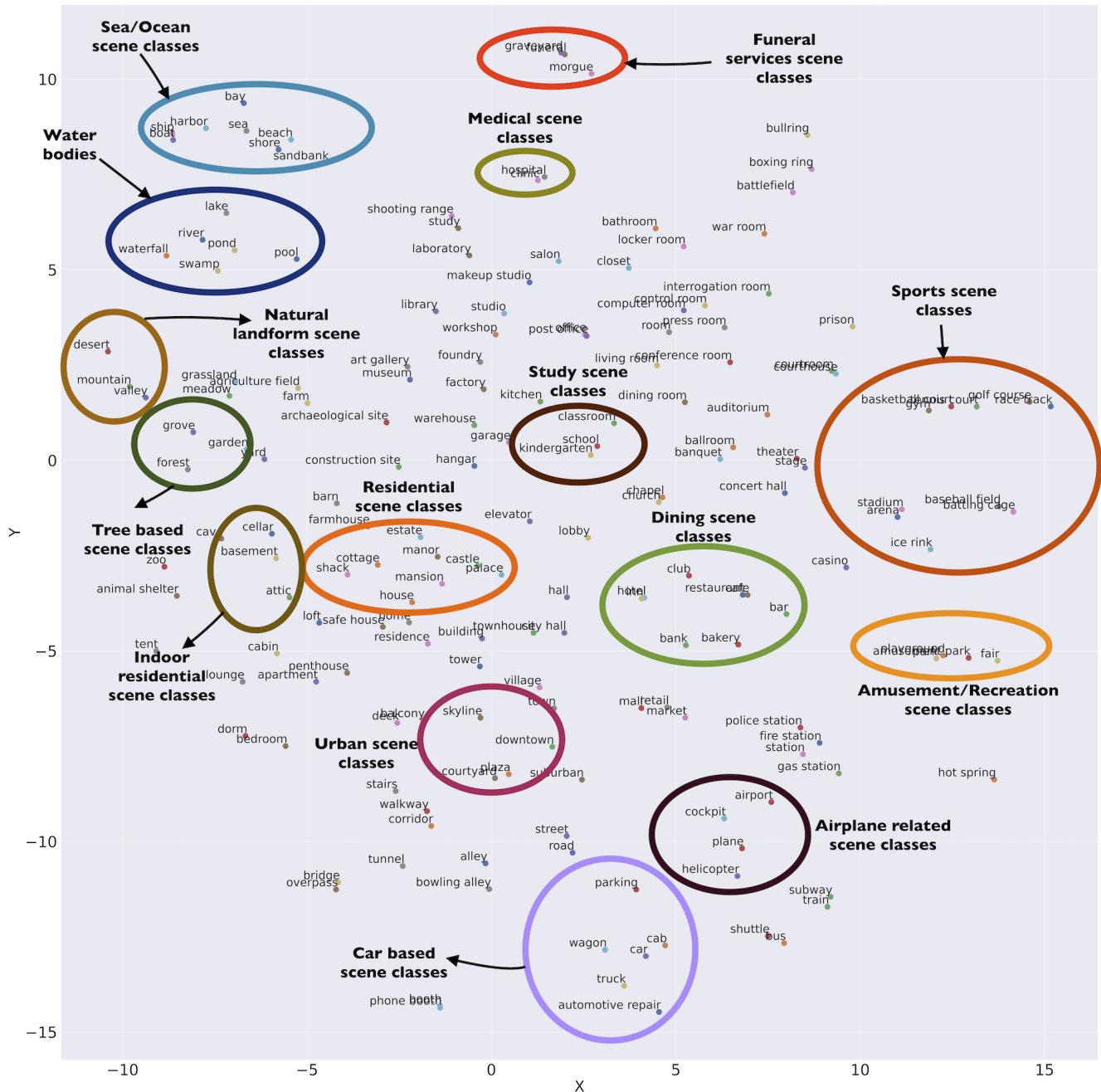


Figure 1. TSNE plot of the 179 scene classes in the taxonomy. Certain representative groups of visual scene labels that are semantically close to each other are enclosed by circular/oval shapes.

As shown in the Fig 3, the prompts associated with generic background information tend to perform better in terms of associating visual scene labels to movie shots with higher confidence, when compared with people-centric prompts. Further inclusion of the contextual phrase "a type of background location" tends to perform better than "a type of location" in associating top-1 visual scene labels with higher CLIPSceneScore values. When people-

centric prompts are used, the CLIP based labeling scheme can result in incorrect associations like interrogation room in Fig 3 (a) and cockpit in Fig 3 (c). Hence we consider *A photo of {}, a type of background location* as our final prompt choice.

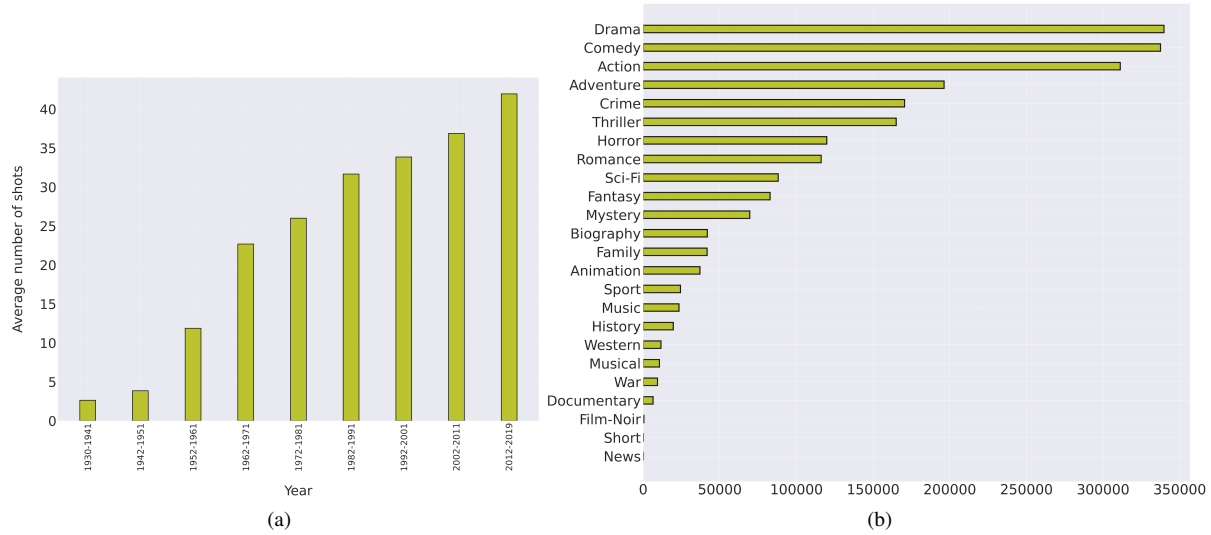


Figure 2. (a) Distribution of average number of shots per year (b) Distribution of number of shots wrt different genres

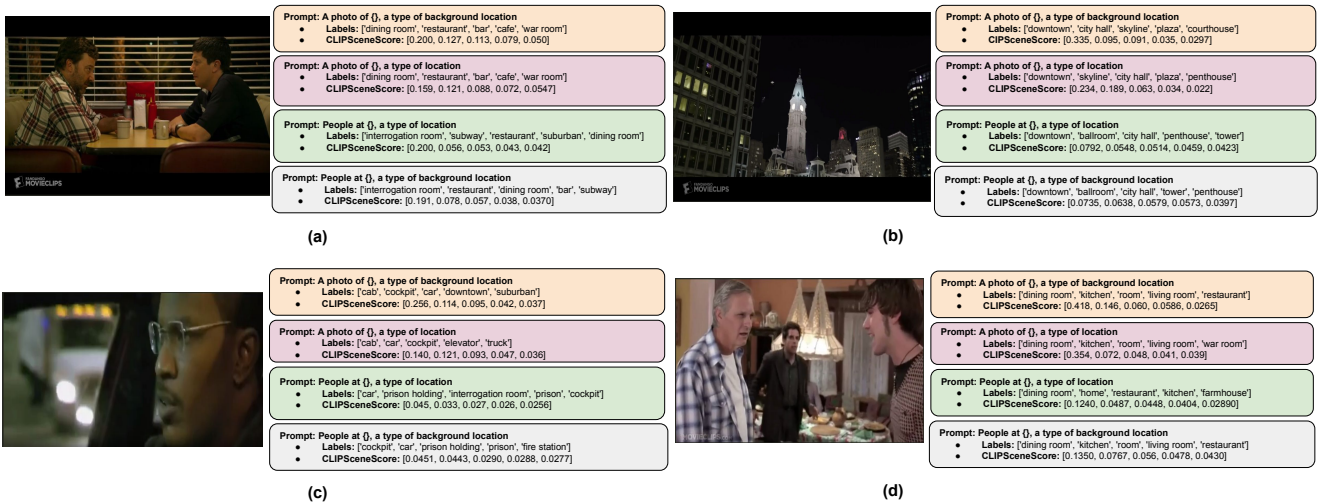


Figure 3. Examples of various prompt templates and associated CLIPSceneScores for top-k visual scene labels. Here k=5

### 3.2. CLIP score distribution

We analyze the distribution of top-k (k=1 to 5) scores provided by CLIP for tagging shots in MovieCLIP dataset. The distribution of scores is shown in Fig 4.

### 3.3. Analysis of CLIP labeling

#### 3.3.1 Qualitative analysis:

We show shot samples from MovieCLIP dataset with top-1 scene classes assigned by CLIP in Fig 5. From Fig 5, we can see diversity of shots being tagged by CLIP for individual classes like *mountain*, *airport*, *desert*, *restaurant*, *stadium* etc. In Fig 6, we show the multilabel nature of visual scenes tagged by CLIP. For multi-label tagging by CLIP, we consider visual scene classes having top-1 CLIPSceneScore

values  $\geq 0.4$  and visual scene classes in top-k(k=2 to 5) having CLIPSceneScore values  $\geq 0.1$ .

#### 3.3.2 Shot type distribution:

We analyze the distribution of shot types in terms of scale for samples extracted from Condensed Movies dataset [1]. We consider the shot scale type taxonomy in MovieShots [15] dataset as follows:

- **Long shot (LS):** Shot taken from long distance like quarter of a mile and having least person close up.
- **Full shot (FS):** Shot showing the entire human body.
- **Medium shot (MS):** Shot showing the humans from knees or waist up.

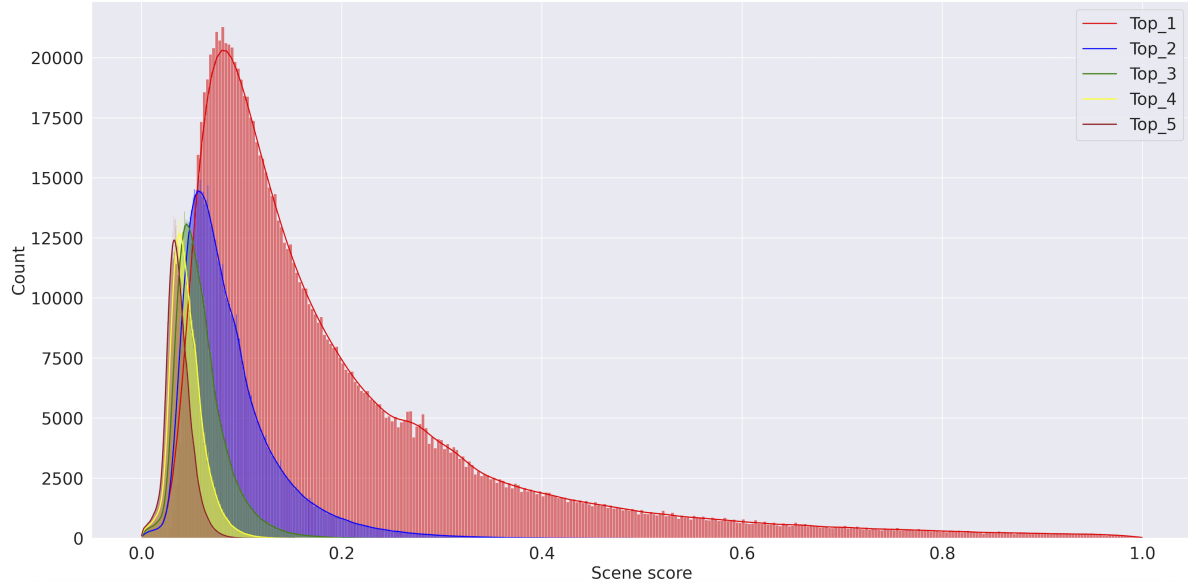


Figure 4. Distribution of Top-k scores of CLIP tagging for shots in MovieCLIP dataset(1.12 m shots across 32k movie clips). Here k=5 is considered.



Figure 5. Diversity within various visual scene classes in the curated taxonomy. 15 sample classes from our scene taxonomy

- **Close-up shot (CS):** Shot focussed on a small segment like face of a person.
- **Extreme close-up shot (ECS):** Shot showing smaller regions like image of eye or mouth.

We use the samples in MovieShots dataset to train a 2 layer LSTM [11] network (hidden dimension = 512) to classify between 5 scale classes. For individual shots, we extract frame-wise features from pretrained ViT-B/16 [7] network at 4 fps for inputs to the LSTM network. For training, validation and testing we use a split of 23557, 6764 and 3332 shots. We use the trained model to predict the shot

scale labels for the samples in MovieCLIP dataset.

From Figure 7 (a), we can see that for low-confidence shots in MovieCLIP dataset having top-1  $CLIPSceneScore \leq 0.2$ , significant proportion (86%) have person close-up ranging from moderate (MS) to very high (ECS). When combined, Close-up (CS) and extreme close-up (ECS) shots constitute 50% of total samples having top-1  $CLIPSceneScore \leq 0.2$ . However for high confidence shot samples in MovieCLIP, whose top-1  $CLIPSceneScore$  values are greater than 0.4, the combined share of CS and ECS scale labels decreases to 24%.

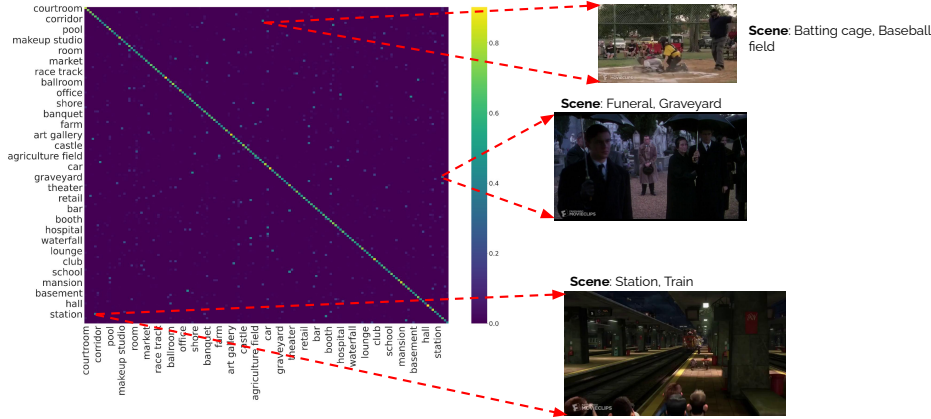


Figure 6. Multi label organization of 150 labels from CLIP’s automatic tagging with threshold limits of 0.4 and 0.1 in MovieCLIP dataset.

Further, the combined share of **FS** and **LS** rises from 14% in Fig 7 (a) to 35% in Fig 7 (b). This indicates that a major share of the shot samples having low CLIPSceneScore values have high(**CS**) to very high (**ECS**) person close-up. Whereas the increasing share of **FS** and **LS** scale labels in Fig 7 shows that CLIP needs more background information in the shots to tag visual scenes with high confidence.

### 3.3.3 Human verification experiment

We conduct human verification through Amazon Mechanical Turk for reliability estimation of labels provided by CLIP for movie shots. As shown in Fig 8, we provide annotators with top-5 CLIP labels for each shot. If none of the scene labels appear relevant for the given movie shot, the annotators choose the **Not relevant** option. The remaining shots (N=1883) with valid visual scene labels are considered as part of evaluation set. We also show some samples from the evaluation set along with the associated visual scene labels in Fig 9.

## 4. Experiments

### 4.1. Visual scene recognition - Movies

We implement the end-to-end 3D convolutional and video transformer models using mmaction2 [5] repository. Our experiments are carried out using 4 NVIDIA T4 gpus. The model configurations listed as follows:

#### 4.1.1 3d convolutional models

##### i3D:

For i3D[3] we use a Resnet50 [10] backbone with the following settings:

- **Optimizer:** *SGD: lr=1e-3, decay=1e-4, momentum=0.9, gradient clipping (maximum norm=40)*
- **Learning scheduler:** *policy: step, steps: 20, 40*
- **Batch size, Epochs:** *32,38*
- **Clip length:** *32 frames(frame interval=2, number of clips=1)*
- **Video resizing:** *Resize with shorter side as 256.*
- **Input Resizing:** *Resize to 224 x 224*
- **Training-augmentation:** *Multi-scale crop(Size: 224 x 224, scales=(1,0.8)), Flip(ratio=0.5)*
- **Normalization:** *Mean(123.675,116.28,103.53), Std(58.395,57.12,57.375)*
- **Testing setting:** *Num clips=10, ThreeCrop (crop size=256), Average=prob*

##### SlowFast:

For SlowFast[8] we use a Resnet50[10] backbone with the following settings:

- **Optimizer:** *SGD: lr=0.1, decay=1e-4, momentum=0.9, gradient clipping (maximum norm=40)*
- **Learning scheduler:** *policy: CosineAnnealing, warmup: linear, warmup<sub>iters</sub> : 34*
- **Batch size:** *32*
- **Clip length:** *32 frames (frame interval =2, number of clips =1)*
- **Video resizing:** *Resize with shorter side as 256.*
- **Input Resizing:** *Resize to 224 x 224*
- **Training-augmentation:** *RandomResizedCrop, Flip (ratio=0.5)*
- **Normalization:** *Mean(123.675,116.28,103.53), Std(58.395,57.12,57.375)*

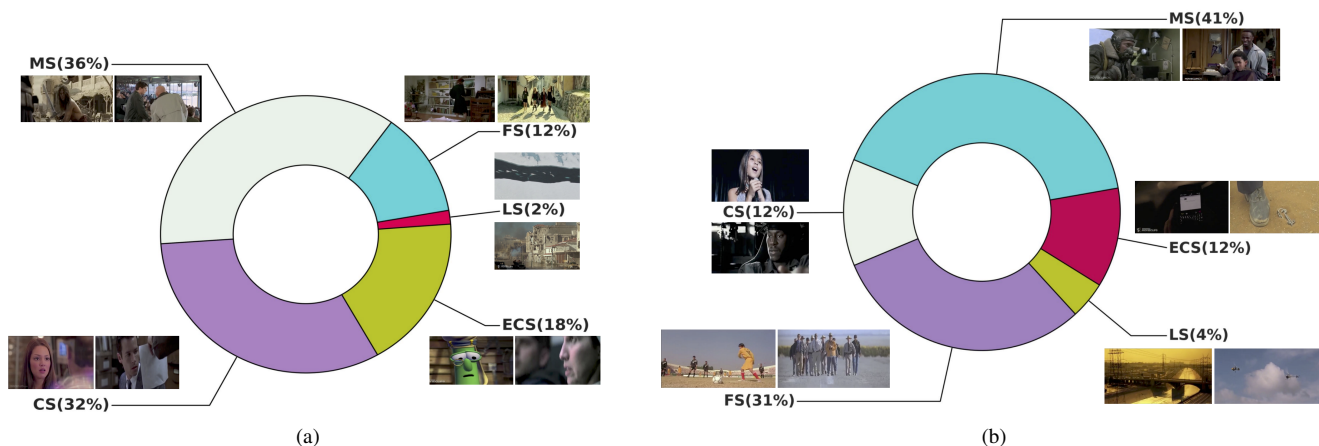



Figure 7. (a) Distribution of predicted scale labels for shots (N=745059) in MovieCLIP dataset having top-1 CLIPSceneScore  $\leq 0.2$ (b) Distribution of predicted scale labels for shots (N=107420) in MovieCLIP dataset having top-1 CLIPSceneScore  $\geq 0.4$  and top-k CLIPSceneScore  $\geq 0.1$  ( $k=2, 3, 4, 5$ ). Scale labels include : ECS: Extreme Close-up shot, CS: Close-up shot, MS: Medium shot, LS: Long shot, FS: Full shot.

## Verification experiment for movie scene understanding

Please watch the following snippet from movies and mark appropriate choices.



• Automatically generated tags for movie backgrounds are provided for the movie snippets.  
 • Select all the background tags(out of 5 provided tags) that apply.  
 • If none of the background tags are relevant, then select the option "Not relevant"

Following are the automatic tags for the snippet. Mark all that apply.

- cafe
- restaurant
- lounge
- penthouse
- salon
- Not relevant

Figure 8. Schematic design of the mturk experiment used for human verification of visual scenes.

- **Testing setting:** *Num clips=10, ThreeCrop (crop size=256), Average=prob*

**R(2+1)D:**

For R(2+1)D[17] we use a Resnet34[10] backbone with the following settings:

- **Optimizer:** Adam: *lr=1e-4, betas=(0.9, 0.999), eps=1e-08, gradient clipping (maximum norm=40)*

- **Learning scheduler:** *policy: step, steps: 20, 40*

- **Batch size, Epochs:** *16, 15*

- **Clip length:** *8 frames (frame interval=8, number of clips=1)*

- **Video resizing:** *Resize with shorter side as 256.*

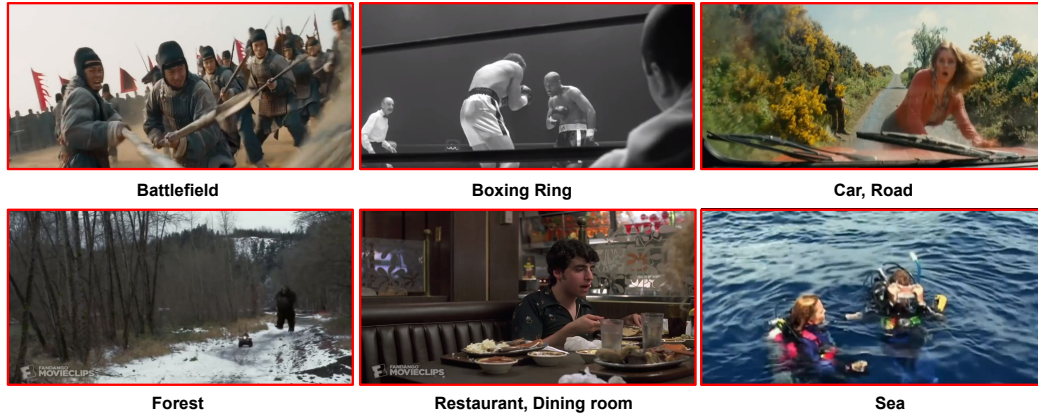


Figure 9. Example of samples in the evaluation set along with visual scene labels obtained after human verification experiment

- **Input Resizing:** *Resize to 224 x 224*
- **Training-augmentation:** *RandomResizedCrop, Flip (ratio=0.5)*
- **Normalization:** *Mean(123.675,116.28,103.53), Std(58.395,57.12,57.375)*
- **Testing setting:** *Num clips=10, ThreeCrop (crop size=256), Average=prob*

#### 4.2. Video transformer models

For video transformer models, we use TimeSformer [2] and Video Swin Transformer [13].

##### TimeSformer:

For TimeSformer [2], we use the divided space time attention configuration with the following details:

- **Patch configuration:** *16 (patch size), 768 (patch embeddings)*
- **Number of Layers and Heads:** *12 (Heads), 12(Layers)*
- **Optimizer:** *SGD: lr=5e-3, betas=(0.9, 0.999), nesterov=True, weight decay=1e-4, gradient clipping (maximum norm=40)*
- **Learning scheduler:** *policy: step, steps: 5, 10*
- **Batch size, Epochs:** *8, 15*
- **Clip length:** *8 frames (frame interval=32, number of clips=1)*
- **Training-augmentation:** *RandomRescale(256,320), RandomCrop(224), Flip(ratio=0.5)*
- **Normalization:** *Mean(127.5,127.5,127.5), Std(127.5,127.5,127.5)*
- **Testing setting:** *Num clips=1, ThreeCrop (crop size=224), Average=prob*

##### Video Swin Transformer

For Video Swin Transformer [13], we use the Swin-B configuration with the following model details:

- **Patch configuration:** *[2,4,4], 128 (patch embeddings)*
- **Depth and Head configuration:** *[4, 8, 16, 32](Heads), [2, 2, 18, 2](Depths)*
- **Window size:** *[8,7,7]*
- **Optimizer:** *AdamW: lr=1e-4, betas=(0.9, 0.999), weight decay=0.05, nesterov=True, gradient clipping (maximum norm=40)*
- **Learning scheduler:** *policy: Cosine Annealing, warmup: linear, warmup\_iters : 2.5*
- **Batch size, Epochs:** *32, 9*
- **Clip length:** *32 frames (frame interval=2, number of clips=1)*
- **Video resizing:** *Resize with shorter side as 256.*
- **Input Resizing:** *Resize to 224 x 224*
- **Training-augmentation:** *RandomResizedCrop, Flip(ratio=0.5)*
- **Normalization:** *Mean(123.675,116.28,103.53), Std(58.395,57.12,57.375)*
- **Testing setting:** *Num clips=4, ThreeCrop (crop size=224), Average=prob*

#### 4.3. Downstream tasks

##### 4.3.1 Visual scene recognition - web videos

We train a three layer fully connected architecture called  $M_{scene}$  with 1024 dimensional features as input for visual scene recognition in HVU [6] dataset. The architecture details of  $M_{scene}$  is shown in Fig 10 (a).



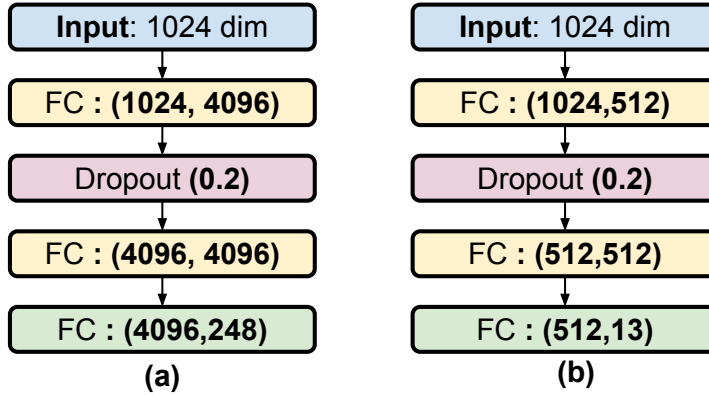


Figure 10. (a) Architecture details of  $M_{scene}$  fully connected model used for multi-label scene recognition in HVU[6] dataset. (b) Architecture details of  $M_{trailer}$  fully connected model used for genre classification in Moviescope[4] dataset

### 4.3.2 Multi label genre classification - Movie trailers:

We train a three layer fully connected architecture called  $M_{trailer}$  with 1024 dimensional features as input for multi-label genre classification in Moviescope [4] dataset. The architecture details of  $M_{trailer}$  is shown in Fig 10 (b).

For both the models  $M_{scene}$  and  $M_{trailer}$ , we use the following settings of hyperparameters:

- **Batch size:** 256
- **Optimizer:** AdamW [14] (learning rate =  $1e-4$ , Step decay (gamma=0.85) as scheduler after 5 and 50 epochs)

## 5. Results

### 5.1. Impact of MovieCLIP pretraining

We show the impact of MovieCLIP pretraining on visual scene recognition (HVU dataset) by considering the following two settings of fully connected model  $M_{scene}$ :

- $M_{scene}$ :  $M_{scene}$  model having inputs as 1024 dimensional features from best-performing Swin-B [13] model trained on MovieCLIP.
- $M_{scene}(Kin)$ :  $M_{scene}$  model having 1024 dimensional features from Swin-B [13] model pretrained on Kinetics400 [12].

From Fig 11, we can see that for certain classes present in our taxonomy like *tunnel*, *restaurant*, *apartment*, *attic* and *concert hall*,  $M_{scene}$  performs better when compared to  $M_{scene}(Kin)$ . Similar trends can be seen for HVU scene classes that are part of broader scene classes in our taxonomy like *riverbed* (part of *river*), *mountain pass* (part of *mountain*) and *track* (part of *race track*).

## References

[1] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings, 2020.

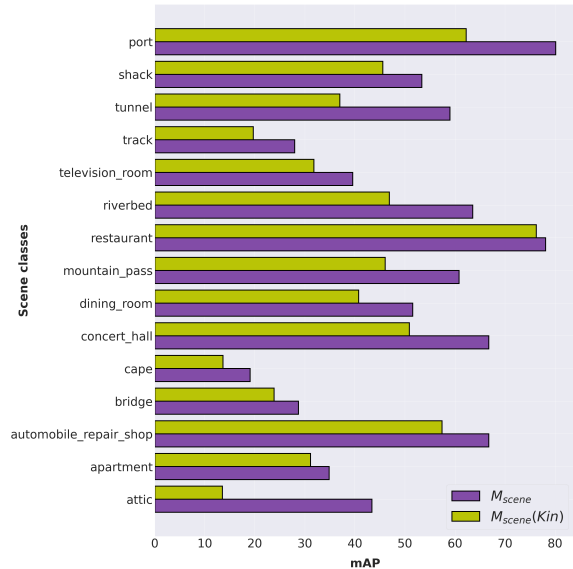


Figure 11. Sample scene classes in HVU where  $M_{scene}$  performs better in comparison to  $M_{scene}(Kin)$

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *ArXiv*, abs/2102.05095, 2021.

[3] J. Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. pages 4724–4733, 07 2017.

[4] Paola Cascante-Bonilla, Kalpathy Sitaraman, Mengjia Luo, and Vicente Ordonez. Moviescope: Large-scale analysis of movies using multiple modalities. *ArXiv*, abs/1908.03180, 2019.

[5] MMAction2 Contributors. Openmmlab’s next generation video understanding toolbox and benchmark. <https://github.com/open-mmlab/mmdetection>, 2020.

[6] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large

- Scale Holistic Video Understanding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 593–610, Cham, 2020. Springer International Publishing.
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.
- [8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019.
- [9] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
- [11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [12] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.
- [13] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *ArXiv*, abs/2106.13230, 2021.
- [14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.
- [15] Anyi Rao, Jiaze Wang, Linning Xu, Xuekun Jiang, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A unified framework for shot type classification based on subject centric lens. In *The European Conference on Computer Vision (ECCV)*, 2020.
- [16] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.
- [17] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.
- [18] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.