360+*x*: A Panoptic Multi-modal Scene Understanding Dataset (Supplementary)

Hao Chen Yuqi Hou Chenyuan Qu Irene Testini Xiaohan Hong Jianbo Jiao

The Machine Intelligence + *x* Group, University of Birmingham, UK

Project page: https://x360dataset.github.io/

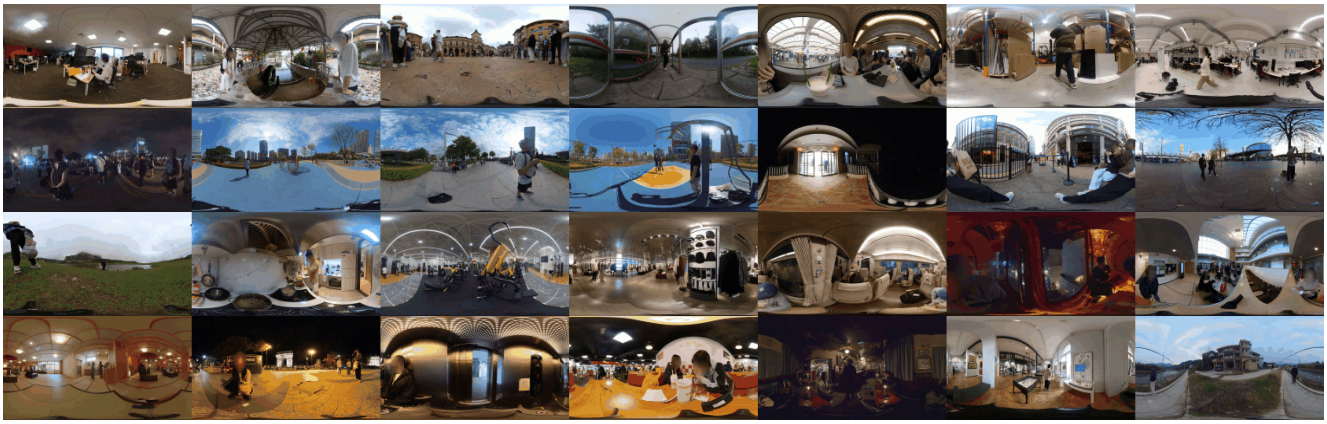


Figure 1. Example 360° panoramics videos from all 28 scene categories.

Introduction

This document provides supplementary materials for the main paper. Figure 1 offers a glimpse of all 28 scene categories of our *360+x* Dataset. Specifically, section 1 describes the data organisation in detail, while section 2 explains the procedure used to select the scene labels and the temporal segmentation labels. More statistics of the proposed dataset are presented in section 3. The ethical use of the dataset and the author's statement are discussed in section 4. Self-supervised methods and modality feature fusion methods employed in our work are introduced in section 5 and section 6, respectively. Additional experimental results are presented in section 7, and more samples from the dataset are shown in section 8. The social impact of the proposed dataset and the limitations of this work are analysed in section 9 and section 10, respectively. Potential future work is discussed in section 11.

Author statement. The authors acknowledge that they are fully responsible for any potential violations of rights, ethical issues, or legal disputes related to their work. The authors further confirm that they have obtained all necessary permissions and licenses for the data used in the research.

1. *360+x* Dataset Organisation

For each data instance, we provide a comprehensive set of views, including:

- 360° panoramic view
- Third-person front view
- Egocentric binocular view
- Egocentric monocular view

For each view, we offer a variety of data modalities and the original file, allowing for a more comprehensive understanding of the scene, which is structured as follows:

- Video
- Multi-channel audio
- Directional binaural delay
- Temporal segments label

Along with the data instance, we also provide accompanying metadata including scene category labels, textual scene descriptions, weather conditions, capture time, and GPS information. This provides an opportunity for exploring a comprehensive understanding of the scene from various angles.

Accessibility. Large-scale data collection can present challenges for researchers due to limitations in hardware resources such as storage and computing power. To address this, we offer a three-step solution:

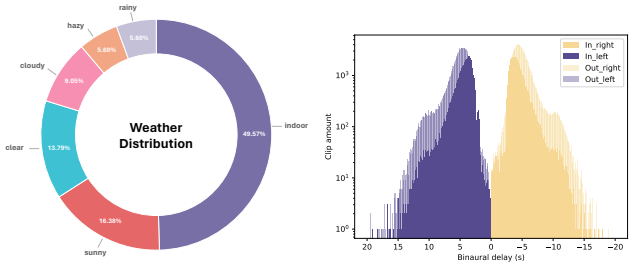
- Partitioned data: We provide standardised mini-sets of data for quick overviews and initial experimentation, allowing researchers to explore the dataset without being overwhelmed by its size.
- Reduced-resolution: We offer reduced-resolution versions of our extracted frame-by-frame images, which can be used to speed up exploration of the data in the early stages of research. The original high-resolution images are also available for those who require them.
- Pre-computed features: We provide pre-computed features such as video and audio features, which have been extracted using the methods described in the main paper. These features offer a convenient and efficient way for researchers to access and analyse the data without having to perform extensive processing.

2. Selection of Scene Label and Temporal Segmentation Label

The scene labels in *360+x* dataset aim to represent common real-world environments and activities people routinely experience in daily life. During data collection, we strived to capture diverse scenarios across different locations that resemble natural experiences. The categories emerged organically from the range of spaces and events we were able to access and record.

For the classification of database, it is generally based on scenes [14, 18, 19, 19] or action behaviours [4, 6, 8, 10]. However, considering that scene locations and activities often overlap, for example, '*speaking*' can occur in '*dining & food outlets*' or '*indoor residential spaces*', and even in the same location '*campus*' may have various actions such as '*walking*' and '*speaking*'. Our multi-modal data set is based on video recordings of natural behaviours in natural scenes. Each video contains rich naturally occurring behavioural information and scene information, to annotate the video more completely and efficiently, we divide the scene and behavioural actions into two layers of labels: scene labels and temporal segmentation labels.

Scene labels are based on the place where the scene occurs. We learn from the places dataset [18], which extracts 401 scenes based on wordnet [7]. However, those scenes are not all common in daily life scenes, such as '*archaeological excavation*', '*server room*', *etc*. The division is also more detailed, such as 'indoor residential spaces' can have multiple categories: '*bedroom*', '*living room*', '*dining room*', '*attic*', *etc*. Therefore, in order to more accurately fit daily life, we put these 401 directories into the large language model [9] for classification and summary, and then through manual screening, we finally obtained 28 categories including indoor and outdoor. After the scene categories were confirmed, we collected several videos for each category, considering a balanced contribution of weather, lightness and captured locations.



(a). Overall weather distribution (b). Overall binaural delay

Figure 2. Additional dataset indoor/outdoor statistics.

Temporal segmentation labels are the behavioural activities that occur in the scene. We obtained the time segmentation tags of the *360+x* database based on the activity level standard of ActivityNet [4] and combined them with the actual activities in the collected videos. Then we sampled about 50 videos from each directory and performed label pre-annotation. After about two rounds of pre-annotation, we analysed the differences between labels and the length of timeline coverage of each annotation, and then generated a temporal segmentation labels dictionary. To capture the diversity and granularity of activities within each category, we defined a total of 38 action instance labels covering specific actions and behaviours. Finally, we selected three professional annotators to annotate all the videos in the database according to the dictionary.

3. Additional Dataset Statistics

Beyond the action and scene categories mentioned in the main paper section 3.3, we also include weather tags. As illustrated in Figure 2(a), we collected data from both outdoor and indoor environments. For those *purely* indoor scenes that cannot tell any weather conditions, we label them as '*indoor*' tag, while for outdoor scenes or some indoor scenes that can tell the weather, we further categorise them into '*sunny*', '*clear*', '*cloudy*', '*hazy*' and '*rainy*'. Figure 2(b) represents the balanced clip histogram distribution of binaural delay in both indoor and outdoor environment.

4. Privacy and Ethics

We acknowledge data collectors have ethical obligations and standards to uphold when conducting data collection efforts. While specifics vary per site, three common obligations and guidelines have been followed:

1. Compliance with legal terms and consortium conditions of use, specifically for research purposes only.

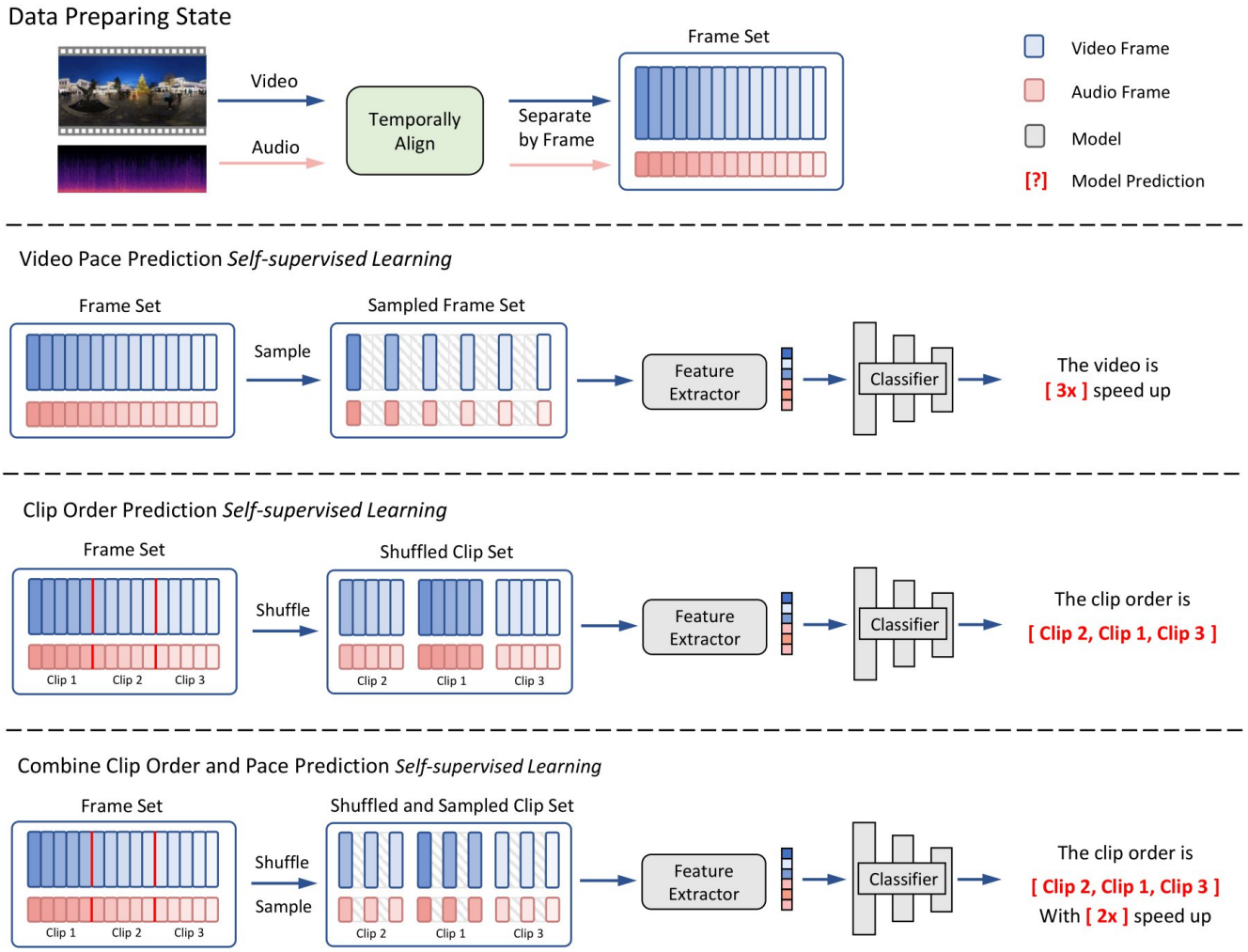


Figure 3. Elucidation of the self-supervised learning (SSL) techniques employed in our study: within SSL, audio is treated in tandem with video frames. To illustrate, when the video speed is augmented by a factor of 2, the audio sample rate is attenuated by 2 (thus speeding it up) to maintain synchronisation. Correspondingly, if the sequence of video clips is rearranged, the audio clips undergo a commensurate reshuffling. The processing of ITD data mirrors this approach used for audio data.

2. Protection of participant confidentiality and privacy.
3. Avoidance of sensitive areas to prevent any potential breaches of confidentiality.

Sensitive information processing. To protect the privacy of individuals, we use an automated face-blurring tool, *Deface*[1], to redact personally identifiable information (PII) from the videos. *Deface* employs the CenterFace [16] face detection model to identify facial regions in frames, then applies Gaussian blurring to mask each detected face.

While completely removing faces could maximise privacy, blurred faces retain some visual information and context. The blurring parameters were tuned to balance privacy protection and data utility based on established practices [2]. All videos were manually reviewed post-redaction to catch any errors or missings detection.

Despite our efforts to maintain efficiency and consistency, certain limitations exist. Factors such as occlusion, lighting, and face angle can affect face detection accuracy, and the blurring strength may be too weak or too strong in some instances. Additionally, our process does not address other forms of personally identifiable information like voices and text. While not perfect, our approach does reduce the privacy risk compared to fully visible faces, and allows the altered data to remain valuable for research purposes.

[1] https://github.com/ORB-HD/deface

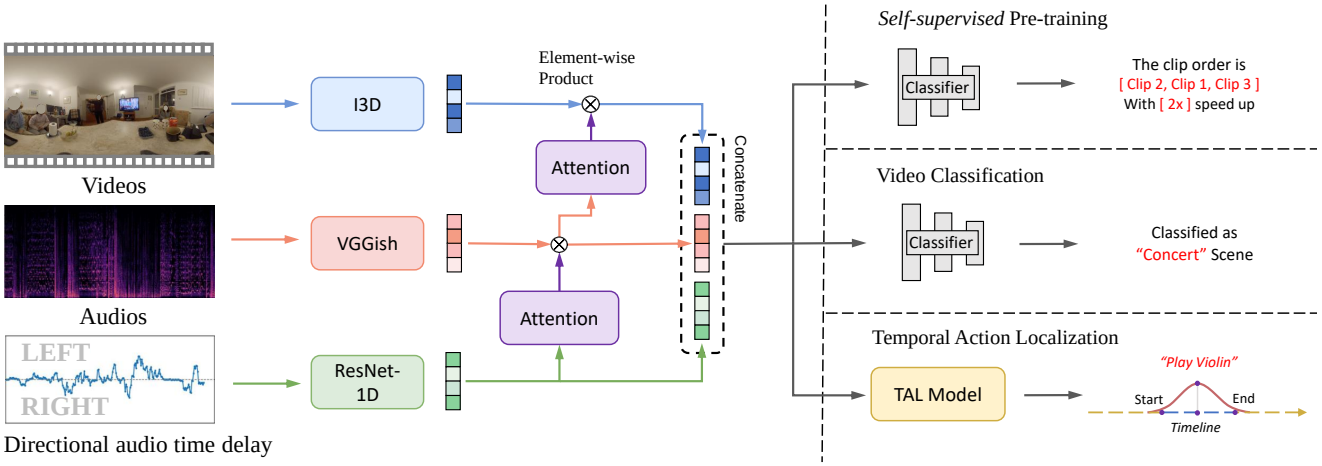


Figure 4. Illustration of Modality Fusion: The features from video, audio, and ITD are extracted utilizing I3D, VGGish, and ResNet-1D correspondingly. Subsequently, these features are concatenated for each sub-task.

5. Explain of Self-supervised Learning

In this study, we utilise self-supervised learning (SSL) techniques proposed in video pace (VP) prediction [12] and clip order (CO) shuffle prediction [15] to pre-train models for enhanced feature learning and subsequent task performance. These two methods are originally tailored for video data, and involve using speed perturbation or clip order permutation on the visual content.

However, our dataset provides more modalities beyond merely video. To fully leverage the power of self-supervised learning, we extend these methods to incorporate more modalities (*i.e*. audios and direction binaural delay). Figure 3 depicts how SSL methods can be applied to both video and audio modalities, while ensuring synchronisation between them. For example, if the video playback speed is altered (*e.g*. $\times 2$), the corresponding audio sample rate is changed accordingly (*i.e*. $\times 0.5$) to maintain synchronisation. Similarly, when the sequence order of video clips is shuffled, the order of audio clips is also rearranged identically to preserve alignment. The direction binaural delay data, which contains spatial audio information, undergoes similar synchronised transformations during SSL pre-training as the audio data. By treating all three modalities (*i.e*. video, audio, and direction binaural delay) jointly and applying transformations consistently across them, we enable cross-modal coordination and representation learning.

It is noteworthy that the VP and CO primarily focus on leveraging temporal information as training guidance, applying it either globally (pace) or locally (clip) to offer distinct interventions to this temporal data. By combining these interventions, there is potential to enhance the model's capability to capture global and local temporal dependencies simultaneously. This integration, depicted in Figure 3, is delineated as 'combine clip order and pace prediction' or varied pace clip order (VP+PO) shuffle. This integration is highlighted in our experiments detailed in the main paper Tables 5 and 6, where noted benefits become evident.

In summary, a core aspect of our self-supervised multimodal learning approach is ensuring aligned cross-modal augmentations and fusing representations across video, audio, and spatial audio domains. This provides a strong foundation for the multi-modal benchmarks in our work.

6. Explain of Modalities Fusion

Simply concatenating the modalities without proper fusion can lead to a reduction in the benefits of multi-modal learning, as pointed out in [13]. Therefore, instead of solely concatenating modality features, we leverage a hierarchical attention mechanism for multi-modality integration as depicted in Figure 4. To simplify the illustration, we use V - video, A - audio, and D - direction binaural delay data as simplified symbols representing each modality.

In nature of multi-modality, the direction binaural delay data contains spatial audio information, and audios can indicate the rich movement region to the videos. We design the hierarchical attention with D as an attention query to direct focused attention towards A. Afterwards, A is also leveraged as a query to attentively interact with V. The experimental supports for selecting A as the attention medium is also presented in section 7.2. This hierarchical design enables the encapsulation of directional and spatial information into audio and video modalities, creating a synergistic representation of the underlying data that integrates the features across modalities.

7. More Experiment Results

7.1. Temporal Action Localisation

As a supplement to section 4.3 in the main paper, we expand the experiments to variations of views, as detailed in Table 1. The results therein show a trend consistent with those observed in Table 2 in the main paper, indicating that the utilisation of multiple views contributes positively.

Table 1. TAL results for different views using TriDet, with extractors being *I3D* pretrained on *360+x*. The lines with a grey background were reported in the main paper.

Selected View	V				V+A				V+A+D			
	mAP @0.5	mAP @0.75	mAP @0.95	Avg.	mAP @0.5	mAP @0.75	mAP @0.95	Avg.	mAP @0.5	mAP @0.75	mAP @0.95	Avg.
Egocentric Only	12.5	9.8	4.3	8.9 (±0.0)	16.2	12.3	4.6	11.0 (±0.0)	16.9	12.7	4.7	11.4 (±0.0)
Front Only	19.7	14.4	5.2	13.1 (+4.2)	24.5	17.6	6.1	16.1 (+5.1)	25.6	18.0	6.2	16.6 (+5.2)
360°Only	21.1	15.3	5.5	14.0 (+5.1)	26.4	18.5	6.9	17.3 (+6.3)	27.1	18.7	7.0	17.6 (+6.2)
360°+ Egocentric	21.4	15.8	5.7	14.3 (+5.4)	27.3	19.2	7.2	17.9 (+6.9)	27.8	19.6	7.2	18.2 (+6.8)
360°+ Front	24.2	16.8	6.1	15.7 (+6.8)	28.1	20.3	7.3	18.6 (+7.6)	28.2	20.8	7.3	18.8 (+7.4)
360°+ Front + Ego	**24.6**	**17.1**	**6.3**	**16.0** (+7.1)	**28.2**	**20.6**	**7.3**	**18.7** (+7.7)	**28.8**	**21.0**	**7.4**	**19.1** (+7.7)

7.2. Cross-modality Retrieval

As we mentioned in the main paper section 4.4, we are embarking on a series of retrieval tasks that traverse the audio, video and directional time delay modalities. This section provides more experimental results on Query-to-Audio and Query-to-Directional information results.

Q-to-Audio retrieval results. Table 2 illustrates the retrieval results for the retrieving audios. In this table, the notation *V+D* represents a set of video and directional binaural features that are trained independently. Additionally, the superscript * indicates that these features are collaboratively trained rather than being treated separately.

The query *V+D* exhibits superior audio retrieval performance, surpassing the use of videos alone. Additionally, the suppression of *(V+D)** suggests that the modalities V and D are not directly related, which forms the foundation for designing our hierarchical attention mechanism that employs audio modality as the attention medium.

Table 2. Q-to-Audio retrieval results. The superscript* indicates modalities are co-trained. Recall reported with rank in $\{1, 5, 10\}$.

Query Modality	R1 (%)	R5 (%)	R10 (%)
V	54.17 (±0.00)	68.32 (±0.00)	80.72 (±0.00)
V + D	**66.36** (+12.19)	**76.78** (+8.46)	**88.59** (+7.87)
(V + D)*	59.21 (+5.04)	72.65 (+4.33)	86.84 (+6.21)

Q-to-Directional feature retrieval results. Table 3 illustrates the retrieval results for the Query modality retrieve directional features. In this table, the notation *V+A* represents video and audio, respectively. The query *(V+A)** exhibits better directional feature retrieval performance than other queries. The effective retrieval results across modalities demonstrate the high quality and compliance with the modalities of the *360+x* dataset.

Table 3. Q-to-Directional binaural delay retrieval results. The superscript* indicates modalities are co-trained. Recall reported with rank in $\{1, 5, 10\}$.

Query Modality	R1 (%)	R5 (%)	R10 (%)
V	6.02 (±0.00)	17.64 (±0.00)	25.93 (±0.00)
V + A	54.15 (+48.13)	76.10 (+58.46)	90.32 (+64.39)
(V + A)*	**67.26** (+61.24)	**89.47** (+71.83)	**94.26** (+68.33)

7.3. Modality Fusion

We also explored alternative modality fusion approaches, such as direct concatenation of modalities, concatenation followed by a linear layer, concatenation followed by self-attention, and varied hierarchical structures of hierarchical attention. The performance of these fusion methods on Temporal Action Localisation is systematically compared and presented in Table 4, suggesting the effectiveness of our presented hierarchical attention approach.

Table 4. TAL with TriDet, *I3D* pretrained on *360+x*, under the setting 360+Egocentric+F and V+A+D. X→Y: X as the query and Y as the key-value pair in the attention mechanism.

Feature Fusion	mAP@0.5	mAP@0.75	mAP@0.95	Avg.
Concatenation	19.2 (±0.0)	14.6 (±0.0)	5.3 (±0.0)	13.0 (±0.0)
Concat + Linear Layer	21.2 (+2.0)	15.1 (+0.5)	5.5 (+0.2)	13.9 (+0.9)
Concat + Self-Attention	26.9 (+7.7)	18.9 (+4.3)	6.8 (+1.5)	17.5 (+4.5)
D→V + Concat A	17.8 (−1.4)	13.8 (−0.8)	5.2 (−0.1)	12.3 (−0.8)
D→A + Concat V	24.6 (+5.4)	17.2 (+2.6)	6.2 (+0.9)	16.0 (+3.0)
A→D + Concat V	20.5 (+1.3)	14.9 (+0.3)	5.7 (+0.4)	13.7 (+0.7)
A→V + Concat D	28.3 (+9.1)	20.6 (+6.0)	7.3 (+2.0)	18.7 (+5.7)
Hierarchical Attention, D→A, A→V	**28.8** (+9.6)	**21.0** (+6.4)	**7.4** (+2.1)	**19.1** (+6.0)

7.4. Migration of the Dataset Pre-training Model

Regarding the integration with the EPIC-Kitchens [3] dataset, we follow the experiment setup in [17] and deploy the SlowFast architecture [5] for feature extraction. The outcomes of the experimentation, centred around the *verb* and *noun* sub-tasks within the EPIC-Kitchens dataset, are concisely displayed in Table 5 and Table 6. These tables provide a comprehensive overview of mean average precision (mAP) scores across a spectrum of IoU thresholds, spanning from 0.1 to 0.5.

In accordance with the EPIC-Kitchens [3], which offers a large amount of monocular egocentric data, we solely employ monocular egocentric information from the *360+x* for this section, thereby ensuring a consistent and reliable basis for experimental analysis. Examining Table 5 and Table 6, the *360+x* dataset extractor does not perform as well as the EPIC-Kitchens model when trained only with EPIC-Kitchens. This is likely due to the fact that the EPIC-Kitchens model is better suited for the EPIC-Kitchens dataset. However, pre-training with the *360+x* dataset followed by fine-tuning on EPIC-Kitchens [3] results in enhanced performance when compared with training solely on the EPIC-Kitchens dataset. This observation suggests that despite the disparate data formats inherent in the two

Table 5. The test outcomes for the ***verb*** sub-task within the EPIC-Kitchens dataset [3]. We utilise the ego-centric monocular modality for training as the sole source of feature extraction. PT: pre-train, FT: Fine-tune.

Feature Extractor	mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP0.7	Avg.
EPIC-Kitchens dataset [1]	28.6 (±0.0)	27.4 (±0.0)	26.1 (±0.0)	24.2 (±0.0)	20.8 (±0.0)	25.4 (±0.0)
360+x Dataset	28.1 (−0.5)	27.1 (−0.3)	25.9 (−0.2)	24.3 (+0.1)	21.2 (+0.4)	25.3 (−0.1)
360+x (PT), Epic-Kitchens (FT)	**28.8** (+0.2)	**27.8** (+0.4)	**26.5** (+0.4)	**24.9** (+0.7)	**21.7** (+0.9)	**25.9** (+0.5)

Table 6. The test outcomes for the ***noun*** sub-task within the EPIC-Kitchens dataset [3]. We utilise the ego-centric monocular modality for training as the sole source of feature extraction.

Feature Extractor	mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP0.7	Avg.
EPIC-Kitchens dataset [1]	27.4 (±0.0)	26.3 (±0.0)	24.6 (±0.0)	22.2 (±0.0)	18.3 (±0.0)	23.8 (±0.0)
360+x Dataset	26.9 (−0.5)	26.0 (−0.3)	24.4 (−0.2)	22.3 (+0.1)	18.6 (+0.3)	23.7 (−0.1)
360+x (PT), Epic-Kitchens (FT)	**27.9** (+0.5)	**26.9** (+0.6)	**25.4** (+0.8)	**23.2** (+1.0)	**19.3** (+1.0)	**24.5** (+0.7)

datasets, pre-training on the 360+x dataset holds the potential to contribute to improved performance within the EPIC-Kitchens context [3].

7.5. Transformer-Based Backbone

We used I3D as our backbone as it was widely adopted in video understanding tasks in the literature. However, we further explore *more contemporary Transformer-based* models as our backbone, *e.g.* VideoMAE [11], pretrained on the Kinetics dataset, akin to the I3D model setting in the main paper. Table 7 reports the performance on temporal action localisation using VideoMAE. Compared to the results in Table 3 in the main paper (*i.e.* the greyed line I3D in Table 7), Transformer shows better performance. Additionally, this experiment further validates the impact/benefits of various views and modalities.

Table 7. TAL using TriDet with extractors being *Transformer-based* model pretrained on *kinetics*. The greyed line was reported in the main paper using *I3D* extractor, for reference.

Selected View	V				V+A			
	mAP @0.5	mAP @0.75	mAP @0.95	Avg.	mAP @0.5	mAP @0.75	mAP @0.95	Avg.
360°Only, with I3D	16.7 (+0.0)	10.1 (+0.0)	4.8 (+0.0)	10.5 (+0.0)	23.6 (+0.0)	17.2 (+0.0)	6.4 (+0.0)	15.7 (+0.0)
360°Only	17.1 (+0.4)	13.4 (+3.3)	5.2 (+0.4)	11.9 (+1.4)	25.9 (+2.3)	18.5 (+1.3)	6.1 (−0.3)	16.8 (+1.1)
360°+ Egocentric	16.9 (+0.2)	13.1 (+3.0)	5.0 (+0.2)	11.7 (+1.1)	26.4 (+2.8)	19.0 (+1.8)	6.2 (−0.2)	17.2 (+1.5)
360°+ Front	19.5 (+2.8)	16.3 (+6.2)	5.6 (+0.8)	13.8 (+3.3)	27.6 (+4.0)	21.2 (+4.0)	6.5 (+0.1)	18.4 (+2.7)
360°+ Front + Ego	**19.2** (+2.5)	**15.8** (+5.7)	**5.4** (+0.6)	**13.5** (+2.9)	**27.8** (+4.2)	**21.7** (+4.5)	**6.6** (+0.2)	**18.7** (+3.0)

8. More Data Examples

Here we provide additional examples of the data (Figures 5 ∼ 32) to show a better understanding of the content and quality of the *360+x* Dataset.

9. Social Impact

Our contribution has the potential to positively impact *scene understanding* through multi-modality learning. The proposed *360+x* Dataset provides the research community with a multi-view perspective with rich modalities for *scene understanding* accompanied by rigorous privacy and ethics standards. Additionally, it offers a diversity and density of activities and reproducible benchmarks for technical advances in scene understanding and beyond.

We acknowledge that large-scale data collection with inadequate oversight could raise privacy and ethical concerns. Therefore, we intend to hinder potential negative applications by making *360+x* data available only for users who sign a license agreement with the statement enumerating the allowable uses of the data.

10. Limitations

Our dataset aims to encompass various aspects of daily life to reflect the real world, yet we acknowledge that it still possesses certain biases and cannot fully represent all aspects of the real world. Despite our efforts to collect massive everyday videos from geographically and demographically diverse sources, the current 28 scenes and 15 cities are still far from complete coverage of the full spectrum of everyday life. Furthermore, while we have included footage from rural and field locations, the majority of the videos remain concentrated in urban or college town areas, resulting in a biased representation of reality.

Another limitation pertains to the potential for biases and noise in our data collection procedures. The unscripted nature of video capturing can introduce inconsistency noise since collectors might choose scenes based on their personal interests, leading to an incomplete or biased depiction of daily experiences. Additionally, the video capturing results are also susceptible to the location of the recorder, which may introduce geometrical bias.

Finally, there remains the potential for temporal labelling bias. While we have taken steps to minimise bias through multiple annotator merging, there still exists the possibility of variations in interpretations of the scene or temporal activities due to individual differences in knowledge backgrounds and natural language use. This can result in subtle

yet potentially significant biases in the language-based narrations and action boards.

11. Future Work

The *360+x* dataset is a collaborative project aimed at driving forward the development of foundational AI research in the realm of panoramic multi-modal machine perception and scene understanding. We actively seek and encourage global collaborations with researchers and participants from diverse and underrepresented regions, as their contributions are critical for capturing the richness and diversity of daily life activities. Therefore, we have developed our data collection and annotation methods to be comprehensive and transparent, allowing researchers from diverse backgrounds to participate in expanding the diversity and quality of the dataset.

In addition to the current benchmarks, we plan to expand the scope of our dataset to encompass other video-audio scene understanding tasks such as audio-visual diarization, scene querying, pre/post conditions, and forecasting, which will further advance the state-of-the-art techniques in this field. However, our current dataset is lacking in spatial-temporal localisation of objects, actions, and audio sources, which we are currently working to address through the augmentation of our labelling process. Although we have made significant progress, the substantial annotation workload has postponed the completion of this task. Spatial annotations will be included in a future update.

To ensure the long-term utility of the dataset, we commit to providing regular updates and maintenance. This includes verifying and correcting any issues related to data accessibility and integrity, as well as expanding the dataset with new content to maintain its relevance with the latest advancements and challenges in academia and industry.

References

[1] Dima Aldamen, Davide Moltisanti, Evangelos Kazakos, Hazel Doughty, Jonathan Munro, William Price, Michael Wray, Tobias Perrett, and Jian Ma. Epic-kitchens-100, 2020. 6

[2] Victoria Cheng, Vinith M Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 149–160, 2021. 3

[3] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 5, 6

[4] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 2

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019. 5

[6] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 2

[7] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[8] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 2

[9] OpenAI. Chatgpt. https://openai.com/chatgpt, 2023. Version 4.0. 2

[10] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[11] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 6

[12] Jiangliu Wang, Jianbo Jiao, and Yunhui Liu. Self-supervised video representation learning by pace prediction. In *European Conference on Computer Vision*, 2020. 4

[13] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020. 4

[14] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 2

[15] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[16] Yuanyuan Xu, Wan Yan, Genke Yang, Jiliang Luo, Tao Li, and Jianan He. Centerface: joint face detection and alignment using face as point. *Scientific Programming*, 2020:1–8, 2020. 3

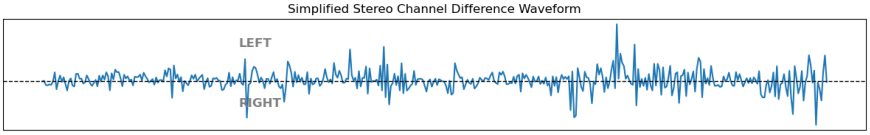
[17] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 5

[18] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 2

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



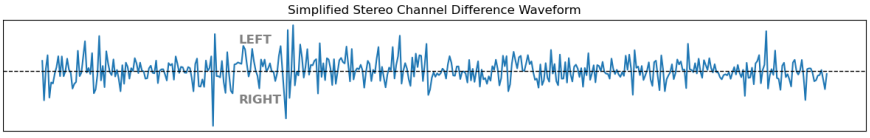
(d). Stereo Waveform Difference Figure

Figure 5. Frame examples in the category of Agriculture & Rural

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



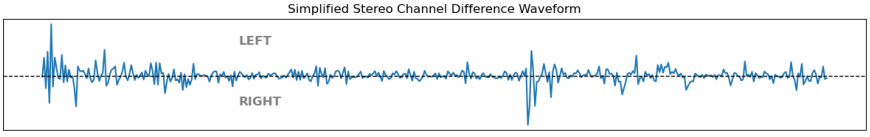
(d). Stereo Waveform Difference Figure

Figure 6. Frame examples in the category of Artistic Spaces

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



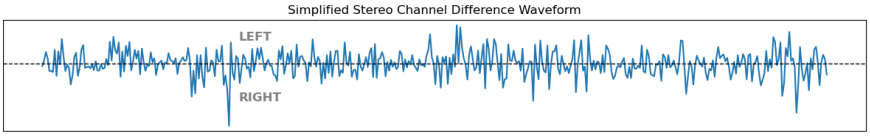
(d). Stereo Waveform Difference Figure

Figure 7. Frame examples in the category of Bars & Nightlife

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



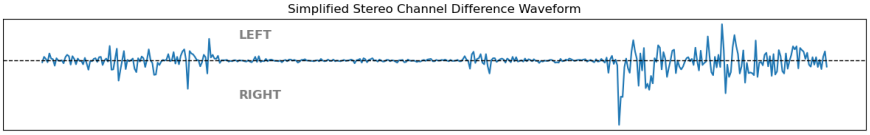
(d). Stereo Waveform Difference Figure

Figure 8. Frame examples in the category of Dining & Food Outlets

(a). 360° panoramic video frame example

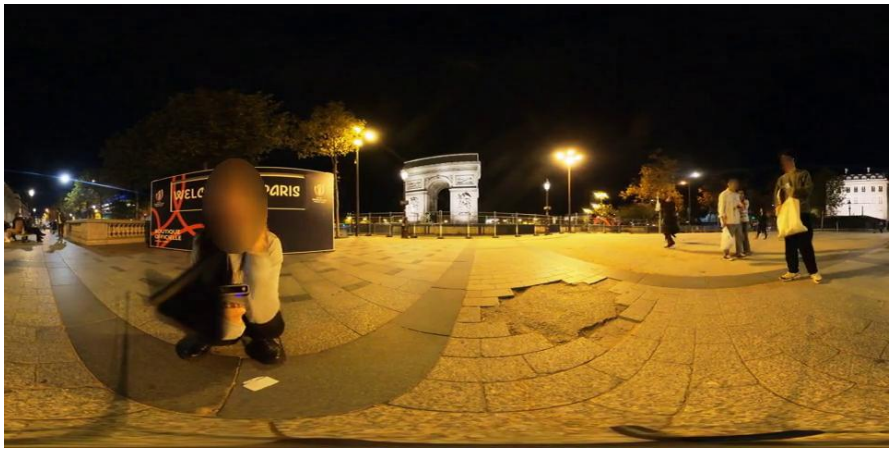
(b). Third-person front view video frame example

(c). Binocular video frame example



(d). Stereo Waveform Difference Figure

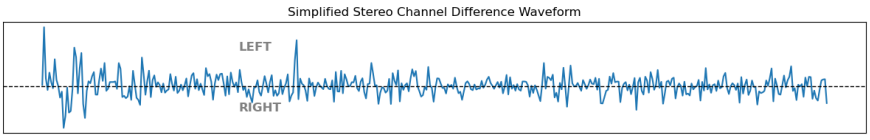
Figure 9. Frame examples in the category of Elevators & Escalators



(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



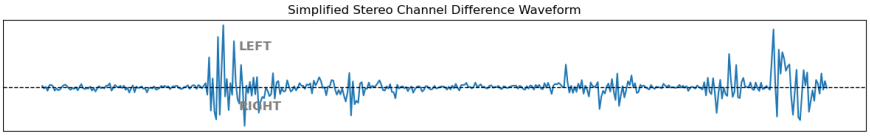
(d). Stereo Waveform Difference Figure

Figure 10. Frame examples in the category of Historic & Religious Sites

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

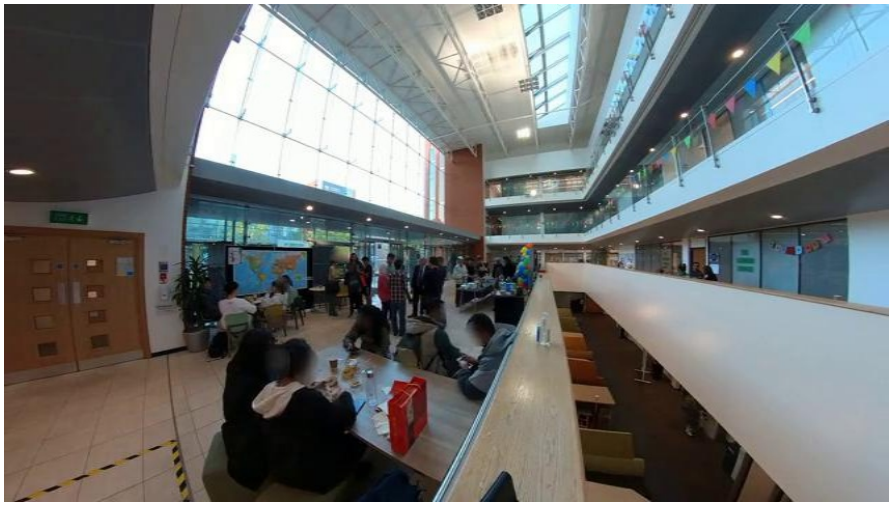
(c). Binocular video frame example



(d). Stereo Waveform Difference Figure

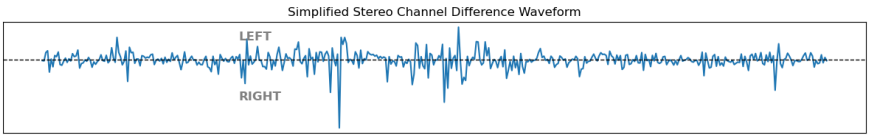
Figure 11. Frame examples in the category of Hotel & Temporary Stay

(a). 360° panoramic video frame example



(b). Third-person front view video frame example

(c). Binocular video frame example



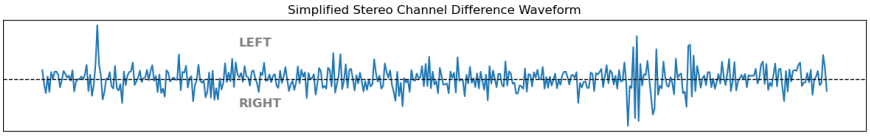
(d). Stereo Waveform Difference Figure

Figure 12. Frame examples in the category of Indoor Educational Spaces

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



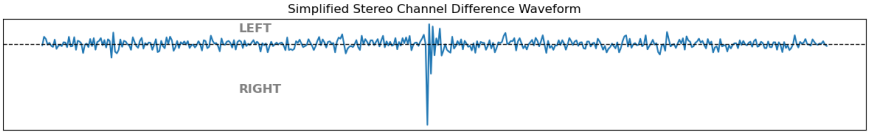
(d). Stereo Waveform Difference Figure

Figure 13. Frame examples in the category of Indoor Entertainment Venues

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



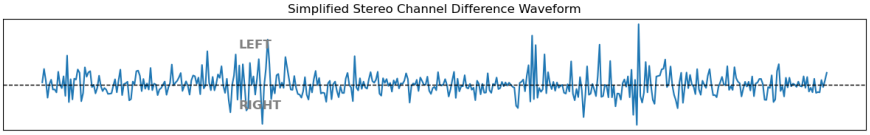
(d). Stereo Waveform Difference Figure

Figure 14. Frame examples in the category of Indoor Residential Spaces

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

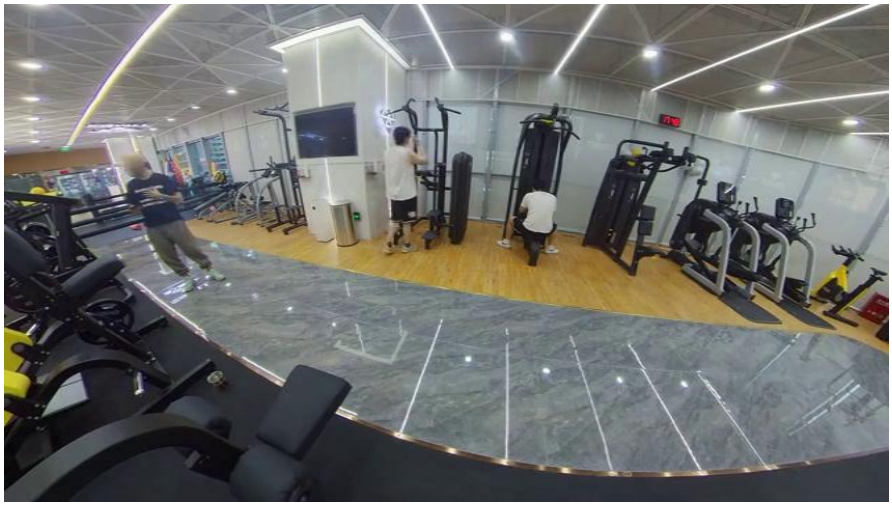
(c). Binocular video frame example



(d). Stereo Waveform Difference Figure

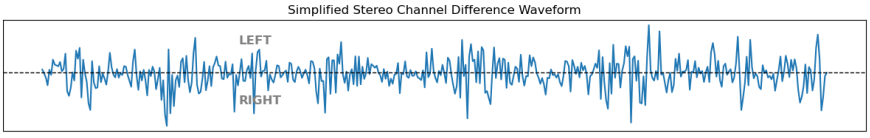
Figure 15. Frame examples in the category of Indoor Shops & Retail & Commercial

(a). 360°panoramic video frame example



(b). Third-person front view video frame example

(c). Binocular video frame example



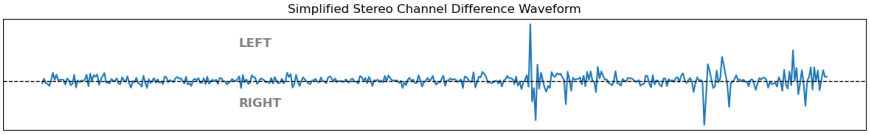
(d). Stereo Waveform Difference Figure

Figure 16. Frame examples in the category of Indoor Sports Venues

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



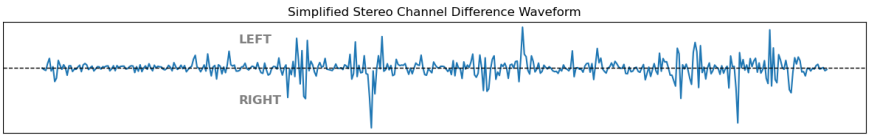
(d). Stereo Waveform Difference Figure

Figure 17. Frame examples in the category of Kitchen

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



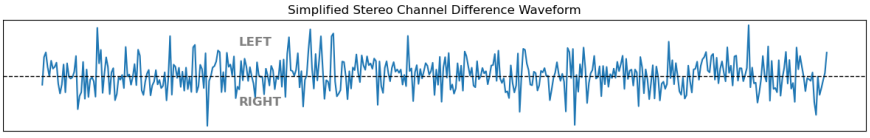
(d). Stereo Waveform Difference Figure

Figure 18. Frame examples in the category of Nature

(a). 360° panoramic video frame example

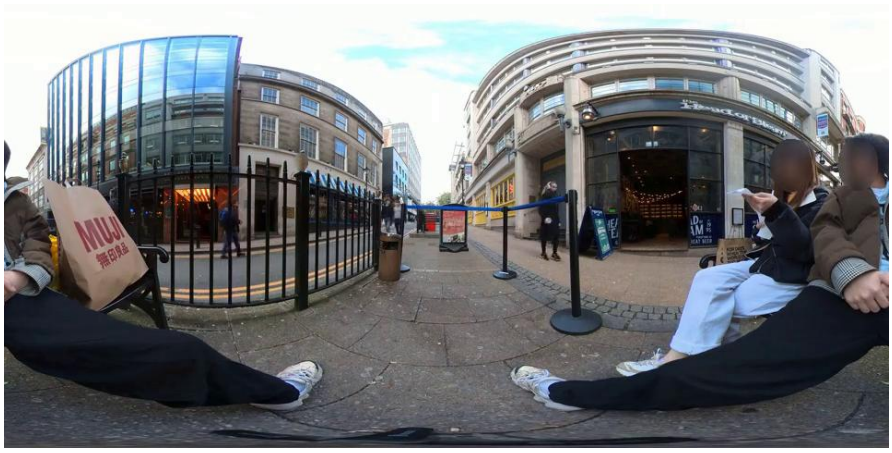
(b). Third-person front view video frame example

(c). Binocular video frame example



(d). Stereo Waveform Difference Figure

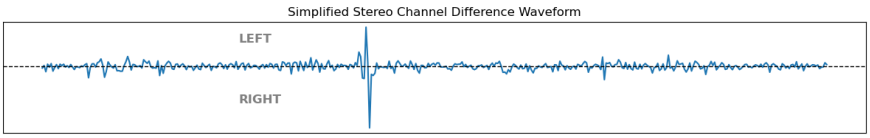
Figure 19. Frame examples in the category of Open Public Spaces



(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



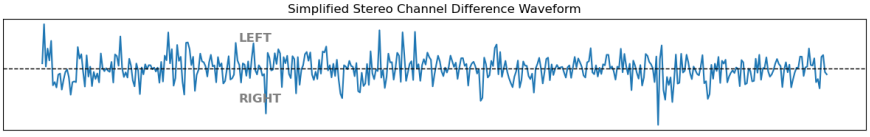
(d). Stereo Waveform Difference Figure

Figure 20. Frame examples in the category of Outdoor Commercial & Markets Outside

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



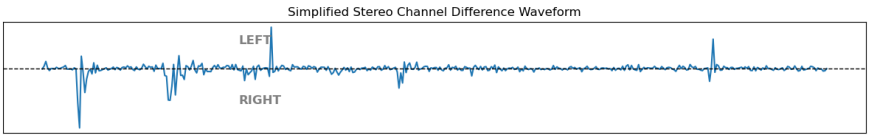
(d). Stereo Waveform Difference Figure

Figure 21. Frame examples in the category of Outdoor Residences & Living

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



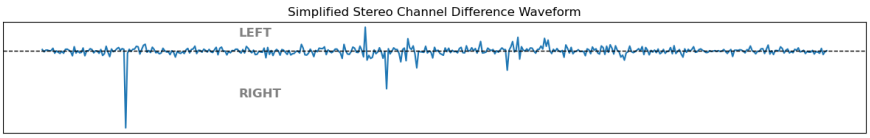
(d). Stereo Waveform Difference Figure

Figure 22. Frame examples in the category of Outdoor Sports & Athletic Fields

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



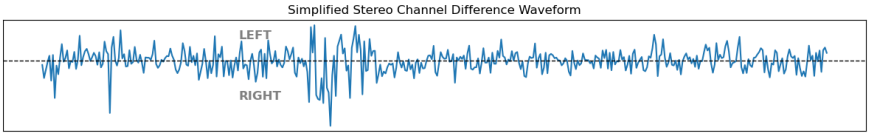
(d). Stereo Waveform Difference Figure

Figure 23. Frame examples in the category of Outdoor Transportation

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



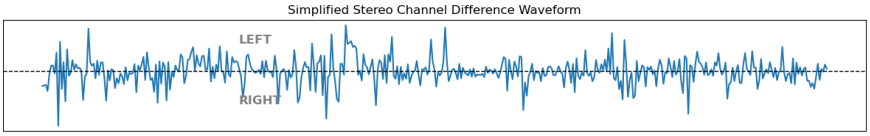
(d). Stereo Waveform Difference Figure

Figure 24. Frame examples in the category of Parks & Recreational Areas

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



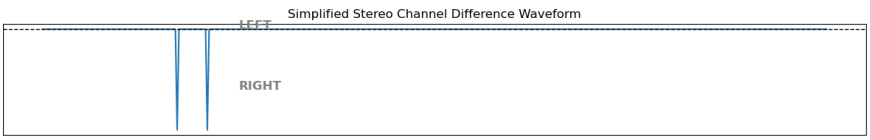
(d). Stereo Waveform Difference Figure

Figure 25. Frame examples in the category of Public Gathering & Conference Spaces

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

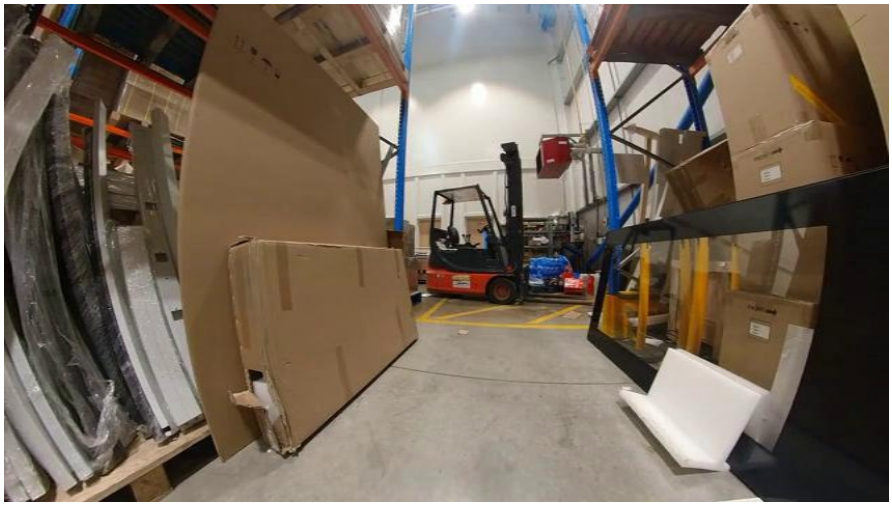
(c). Binocular video frame example



(d). Stereo Waveform Difference Figure

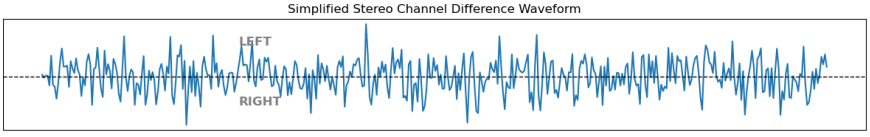
Figure 26. Frame examples in the category of Scientific Interior Space

(a). 360° panoramic video frame example



(b). Third-person front view video frame example

(c). Binocular video frame example



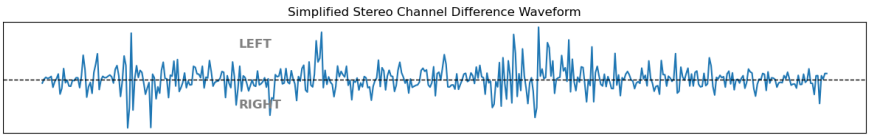
(d). Stereo Waveform Difference Figure

Figure 27. Frame examples in the category of Storage & Utility

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



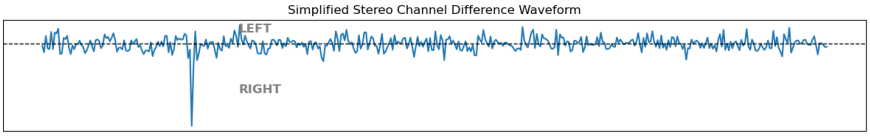
(d). Stereo Waveform Difference Figure

Figure 28. Frame examples in the category of Transportation Interiors

(a). 360°panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



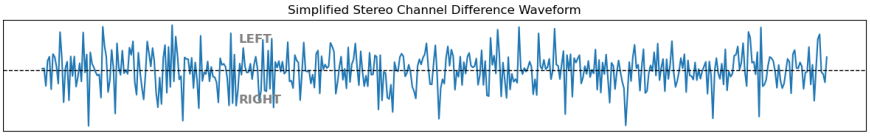
(d). Stereo Waveform Difference Figure

Figure 29. Frame examples in the category of Transportation Stops

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



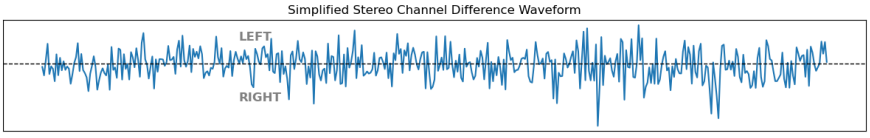
(d). Stereo Waveform Difference Figure

Figure 30. Frame examples in the category of Urban Constructions & street

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



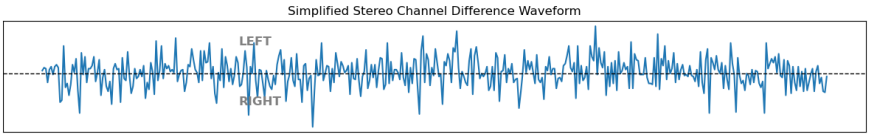
(d). Stereo Waveform Difference Figure

Figure 31. Frame examples in the category of Waterfronts & Water Bodies

(a). 360° panoramic video frame example

(b). Third-person front view video frame example

(c). Binocular video frame example



(d). Stereo Waveform Difference Figure

Figure 32. Frame examples in the category of Workspaces

[19] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2