Multimodal Multi-camera Timecode Synchronization via Audio-visual Embeddings

Machine Learning for Visual Understanding (2025)

5 조 : 양효걸 민동준 권흥찬 박준형 조원진

Abstract

We propose a multimodal synchronization framework for multi-camera videos in real-world broadcast environments. Leveraging pretrained audio-visual feature extractors from Synchformer, we extract segment-level embeddings from each view and estimate alignment via cross-moda similarity. Empirical results on three datasets—Replay, CASTLE 2024, and a real-world broadcast dataset (Womenera)—demonstrate that audio features significantly outperform video features in discriminative power, with a similarity gap of 0.19 vs. 0.001. Our classifier extension, Multi-Synchformer, achieves low-latency offset prediction by reusing pretrained components with minimal adaptation. We further propose a confidence-weighted soft voting mechanism that improves alignment accuracy by up to 25% over baseline methods. Our approach offers a scalable, accurate, and practical solution for automatic multi-camera synchronization in unconstrained environments.

1. Introduction

Synchronizing multi-camera video is essential in broadcasting, yet remains challenging in the absence of shared time references such as genlock or clapperboards. Traditional unimodal alignment methods, especially audio-only, are often brittle under noisy or partial data conditions. While prior work has explored audio-visual synchronization, most focus on within-video tasks such as lip-sync, overlooking the more complex problem of aligning independently recorded multi-camera footage.We address this gap by proposing a segment-based alignment framework built on Synchformer's pretrained audio-visual extractors. Our method measures cross-moda similarity between segments from master and node streams, using cosine distance and L1/L2 metrics to determine optimal alignment lags. We further introduce Multi-Synchformer, a lightweight offset classifier repurposing Synchformer's second-stage module with minor input modifications.To validate our approach, we evaluate performance on three datasets—Replay, CASTLE 2024, and a real-world public broadcast set—under both synthetic and real misalignment. Results show that audio modality offers superior alignment cues, and that our soft voting ensemble reduces prediction errors by up to 25%. These findings establish a robust baseline for multimodal multi-camera synchronization in complex production settings.

Our contributions are summarized as follows:

Problem formulation and data curation.We formalize the task of multi-camera synchronization as a supervised offset classification problem and propose a general-purpose data processing pipeline applicable to a wide range of existing multi-view datasets. Furthermore, we release a novel in-the-wildbroadcast dataset that reflects the complexities and variability encountered in real-world production environments.

Multimodal synchronization model. We extend prior unimodal approaches by jointly utilizing audio and visual cues. Leveraging Segment-AVCLIP representations, we propose a unified multimodal synchronization framework that significantly enhances alignment accuracy across diverse camera viewpoints.

Cross view ensemble evaluation. We propose a confidence-weighted ensemble strategy that integrates complementary view pairs across cameras. The resulting soft votingmechanism produces more stable and robust alignment estimates compared to single-pair evaluations, making it well suited for practical deployment in multi-camera synchronization workflows.

2. Related work

Multi-camera video synchronization remains a core challenge in video processing, especially when shared time references such as genlock or clapperboards are unavailable. Traditional methods rely on visual cues (e.g., scene changes, motion trajectories) or audio cues (e.g., waveform peaks, onset detection), but perform poorly under noise or viewpoint shifts. To overcome these limitations, Casanovas et al. [1] proposed a multimodal

approach based on detecting co-occurring audio-visual events, demonstrating clear gains over unimodal baselines in both fixed and mobile setups.

Recent advances in audio-visual representation learning have enabled self-supervised models to infer synchronization from embeddings. Chung et al. [2] and Owens & Efros [3] introduced contrastive training schemes to distinguish aligned vs. misaligned audio-visual pairs. Follow-up work, such as PerfectMatch [4], incorporated spatio-temporal attention [5] and improved cross-moda alignment. Transformer-based approaches [6] further extended this to large-scale in-the-wild datasets like VGGSound.

To address sparse alignment scenarios, Iashin et al. [7] proposed learnable segment selectors, which were integrated into Synchformer [8], a state-of-the-art model combining segment-level contrastive pretraining with a lightweight offset predictor. While these methods achieved success in within-video tasks, most prior work focuses on lip-sync or single-stream alignment.

In contrast, our work targets the underexplored problem of synchronizing independently recorded multi-camera streams. We propose a scalable framework that reuses pretrained AV feature extractors and introduces a confidence-weighted ensemble strategy for robust offset prediction. Our findings also highlight that traditional similarity metrics (e.g., cosine) may underestimate video modality performance, which can be recovered by alternative metrics such as L2 distance.

In addition, we benchmark our method against AE2 [15], an object-centric alignment model for egocentric-exocentric video pairs, and the ICASSP 2023 stereo camera synchronization method [16], confirming the need for explicit multimodal integration strategies.

3. Method

3.1 Feature Extraction for Multi-camera Synchronization

For multi-camera synchronization, the ability to effectively capture audio-visual information from specific time segments within videos is crucial. In this study, we utilized the pre-trained audio and visual feature extractors from Synchformer[8], which has achieved state-of-the-art performance in audio-visual synchronization research.

Synchformer[8] was specifically developed for in-the-wild environments where synchronization cues are sparse, and employs an efficient two-stage learning approach that separates feature extractor and synchronization module training. The core of this approach is the Segment-level Audio-visual CLIP (Segment-AVCLIP) pre-training. In this stage, audio and visual streams are divided into short time segments, and feature extractors (Fa: AST[9], Fv: Motionformer[10]) are trained to extract high-quality, identifiable features through segment-level contrastive learning based on CLIP[11] methodology. Audio-visual segment pairs from the same time period are learned to be close, while other pairs are pushed apart, enabling the feature extractors to effectively represent the relationship between audio and visual elements within segments. These pre-trained feature extractors have demonstrated excellent adaptability not only for synchronization tasks but also for other audio-visual related subtasks.

3.2. Multi-camera Synchronization Framework

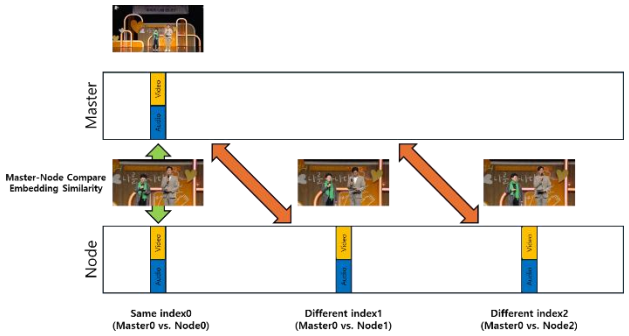

Fig. 1 Experimental Framework for Comparing Audio-visual Embeddings from Multiple Cameras

For multi-camera synchronization, the ability to effectively capture audio-visual information from specific time segments within videos is crucial. In this study, we utilized the pre-trained audio and visual feature extractors from Synchformer[8], which has achieved state-of-the-art performance in audio-visual synchronization research.

Synchformer[8] was specifically developed for in-the-wild environments where synchronization cues are sparse, and employs an efficient two-stage learning approach that separates feature extractor and synchronization module training. The core of this approach is the Segment-level Audio-visual CLIP (Segment-AVCLIP) pre-training. In this stage, audio and visual streams are divided into short time segments, and feature extractors (Fa: AST[9], Fv: Motionformer[10]) are trained to extract high-quality, identifiable features through segment-level contrastive learning based on CLIP[11] methodology. Audio-visual segment pairs from the same time period are learned to be close, while other pairs are pushed apart, enabling the feature extractors to effectively represent the relationship between audio and visual elements within segments. These pre-trained feature extractors have demonstrated excellent adaptability not only for synchronization tasks but also for other audio-visual related subtasks.

To verify temporal alignment across views, we periodically sample fixed-length audio-visual segments, treat the longest, scene-covering stream as the **Master** camera, and offset each **Node** segment by an integer lag τ

relative to the master (Fig. 1). Let $E_m \in R^d$ and $E_{n,\tau} \in R^d$ denote the Segment-AVCLIP embeddings of the master segment and the node segment shifted by τ frames, respectively. Their cosine similarity is

$$\mathrm{Sim}_{cos}(E_m, E_{n,\tau}) = \frac{E_m^{\top} E_{n,\tau}}{|E_m| \cdot |E_{n,\tau}|}$$

which ranges from -1 to 1; higher values indicate stronger audio-visual correspondence. For completeness we additionally compute the L1 and L2 distances

$$d_1(E_m, E_{n,\tau}) = |E_m - E_{n,\tau}|_1$$
$$d_2(E_m, E_{n,\tau}) = |E_m - E_{n,\tau}|_2$$

where lower values signify greater similarity. Because all segments are captured under identical scene and acoustic conditions, this triplet of metrics—cosine similarity, d_1, and $d_2$offers a robust, objective basis for determining the lag τ^* that maximizes correspondence, thereby confirming that disparate cameras are indeed recording the same subject at the same moment.

3.3 Multi Camera Offset Classification: Multi-Synchformer

In 3.1 we showed that L2 distances between AudioSet-pre-trained audio-visual embeddings reliably separated segments drawn from the same time-stamp from those drawn elsewhere on the timeline. This result suggested training an explicit classifier for multi-camera synchronization. However, the time and GPU budget required to construct a new data set and to train a large model from scratch proved prohibitive. Instead, we repurpose the **second-stage synchronization module of Synchformer** by modifying only its inputs, thereby retaining all pre-trained weights.

The original synchronization head accepts a sequence of video tokens V and their co-occurring audio tokens A, projects them to $d_{\text{model}} = 768$, and processes the concatenated stream with a 3-layer, 8-head Transformer encoder. The task is formulated as 21-way classification, covering discrete offsets $\{-2.0, -1.8, \ldots, 1.8, 2.0\}$ s in 0.2-s steps. The $[CLS]$ token produced by the encoder is passed through a two-layer MLP followed by a soft-max to yield the posterior over offsets.

For the multi-camera setting we form a cross view pair by combining the master camera's video tokens V^{M} with the node camera's audio tokens A_{τ}^{N}, where $\tau \in \{0.1, 0.2, \ldots, 2.0\}$ s is an artificially imposed lag. Because Segment-AVCLIP ensures that video embeddings of the same moment remain similar across viewpoints, we hypothesise that the pre-trained classifier can still infer τ from the audio–video mismatch. Training examples are generated by sampling time-aligned segments from the master stream, pairing them with node segments shifted by a random τ, and labelling each pair with the corresponding offset class. This strategy enables low-cost adaptation of Synchformer to the multi-camera synchronization problem while preserving its audio-visual reasoning capabilities.

4. Experiments

In this chapter, we evaluated the cross-scene transfer effectiveness of Synchformer's pretrained feature extractors by assessing multimodal and multi-camera synchronization performance across three datasets. We used the Replay dataset[12] and CASTLE 2024 dataset[13] as our base datasets, and additionally utilized a custom-built public broadcast dataset. This public broadcast dataset, based on "womenera" data, includes actual broadcast footage captured from various angles in a studio environment, showing full stage views and character close-ups. We employed cosine similarity as the primary metric to evaluate how well each modality distinguishes between same index and different index features.

4.1. Datasets and Experiment Setup

Datasets for Multi-View AV Synchronization: The community has developed several datasets to facilitate research in multi-camera multimodal analysis. The **Replay** dataset , introduced by Shapovalov et al[12], provides 68 multi-view video scenes (each ~5 minutes, captured by 12 cameras) with spatial audio recorded by an array of microphones. All sensors in Replay are temporally synchronized and calibrated, making it a valuable resource for studying multi-view alignment and cross-moda understanding. More recently, the **CASTLE** dataset. (Rossetto et al[13].) offers an unprecedented scale of egocentric and exocentric video: 15 time-aligned camera streams (10 first-person wearable cameras and 5 static cameras) recorded over four days, totaling over 600 hours of 50 FPS UHD video with audio. CASTLE's combination of first-person and third-person footage presents rich opportunities for multimodal synchronization research in real-world settings. Our work uses these datasets (along with a new internal broadcast-video dataset) as testbeds by taking their synchronized videos and verifying that our method can recover the known alignments. Notably, since these datasets come pre-synchronized by design, they allow us to simulate misalignment scenarios and will enable quantitative benchmarking of alignment accuracy in future work.

In this study, we utilized two publicly available datasets to validate the cross-scene transfer effectiveness of Synchformer's pretrained feature extractor. The Replay dataset[12] contains temporally synchronized videos captured from various viewpoints using fixed cameras and

action cameras, all sharing a common audio file. The CASTLE 2024 dataset[13] consists of recordings from cameras installed at fixed angles in multiple rooms of a residential environment, with independent video capture and audio recording.

For our experiments, we used the original 1 minute 15 seconds videos from DSLR-1, DSLR-2, and GOPRO1 cameras in the Replay dataset. From the CASTLE dataset's 60-minute recordings, we selected Kitchen, LivingRoom1, and LivingRoom2 videos, segmenting them into 120-second clips for use as master-node video pairs. Additionally, we employed Real-world multi-camera footage from a public broadcast as a supplementary evaluation dataset. This studio-recorded content consists of two camera angles: one capturing the entire stage and another providing close-up shots of the performers' upper bodies. The total broadcast duration was 95 minutes, which we segmented into 4-minute intervals for evaluation purposes.

All three datasets provide simultaneous multi-camera recordings of identical subjects with audio-video information, making them suitable for multimodal and multi-camera embedding research

4.2 Embedding Distance Metric Results

Table. 1 All Dataset Embedding Comparision by Index

Dataset	Metric	Modality	Same	Diff
Replay	Cosine Similarity	Video	0.99995	0.99981
		Audio	1	0.68755
		AV	0.99995	0.93446
	L2 dist	Video	5.53	6.08
		Audio	0.00	143.21
		AV	5.24	143.15
Castle	Cosine Similarity	Video	0.99999	0.99998
		Audio	0.86562	0.82815
		AV	0.96779	0.96156
	L2 dist	Video	7.04	6.82
		Audio	102.39	118.45
		AV	103.11	119.07
Womenera	Cosine Similarity	Video	0.99805	0.99894
		Audio	0.9501	0.74179
		AV	0.9875	0.93816
	L2 dist	Video	125.02	185.32
		Audio	1604.37	3121.12
		AV	1751.43	3250.08

Table. 1 summarizes the modality-specific embedding similarity comparison across the Replay, Castle, and Womenera datasets, using two metrics: cosine similarity and L2 distance. When analyzing with cosine similarity, the Audio modality generally demonstrated the highest discriminative power, while the Video modality showed minimal differences. However, when evaluating with L2 distance, a different pattern emerges. The L2 distance metric demonstrates that all modalities, including Video, can distinguish between same index and different index pairs, as indicated by consistently lower distances for 'Same' pairs compared to 'Diff' pairs across the datasets. The perfect matching in the Replay dataset's Audio modality (Cos=1.0, L2=0.00) is due to its design with shared audio files.

In the real-world Womenera broadcast dataset, analysis via cosine similarity shows the Video modality struggling, with nearly identical values for both same index (0.998) and different index (0.999) pairs, making scene differentiation difficult. In stark contrast, the L2 distance analysis reveals a different outcome. The Video modality shows a clear separation, with an average distance of 125.02 for same index pairs and 185.32 for different index pairs. This confirms that video embeddings can indeed differentiate scenes when an appropriate distance metric is used. While the Audio and AV modalities also demonstrate even stronger discriminative power with L2 distance, the key finding is that L2 distance unlocks the potential of video-based differentiation where cosine similarity fails.

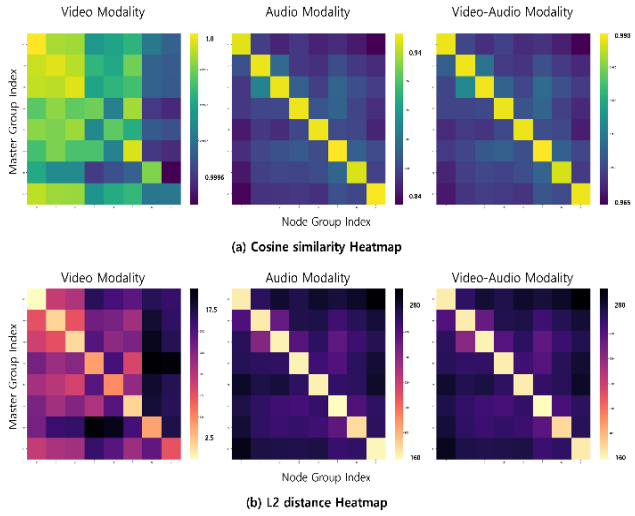

Fig. 2 Master-Node Heatmap by distance method

Heatmap visualizations of similarity matrices (Fig. 3) reveal consistent patterns across both cosine similarity and L2 distance metrics. The cosine similarity heatmaps (Fig. 3a) show that the Video modality displays uniformly high values (0.9996-1.0), making group distinction nearly impossible due to the narrow value range. The Audio modality demonstrates the most distinct block diagonal structure with broadly distributed similarity values (0.84-0.94), clearly differentiating between same and different temporal indices. The Audio-Video (AV) fusion exhibits a moderate range of values (0.965-0.990) with somewhat observable block structure, though less pronounced than Audio alone.

The L2 distance heatmaps (Fig. 3b) provide more encouraging results for Video modality discrimination.

Video modality shows a discernible block diagonal pattern with same index distances (2.5-10) clearly lower than different index distances (10-17.5), indicating improved discriminative capability under L2 distance metric. Audio modality exhibits the most pronounced block diagonal structure where same index distances remain minimal (dark regions) while different index distances are substantially higher (160-280 range), demonstrating the strongest temporal synchronization cues. The AV fusion maintains a similar pattern to Audio (160-280 range) with well-defined contrast between same index and different index distances.

Our results reveal that metric choice significantly impacts modality effectiveness for multi-camera scene synchronization. While cosine similarity favors Audio modality exclusively, L2 distance demonstrates that Video modality also possesses discriminative power, though Audio remains superior. The L2 distance metric appears more suitable for capturing the temporal dynamics necessary for synchronization tasks, as it better reveals the discriminative capabilities across all modalities.

4.3 Multi-camera Offset Classification

Evaluation protocol

Offset prediction was assessed on the Replay and Womenera datasets with four inference strategies. ConfSelect feeds both cross view pairs—Master-Video → Node-Audio (MV–NA) and Node-Video → Master-Audio (NV–MA)—into Multi-Synchformer and selects the offset whose soft-max confidence is higher; the predicted value is converted to its absolute magnitude because all injected lags are positive. Soft Voting (our primary scheme) averages the absolute offsets of the two pairs, weighted by their respective confidences. $Cross_{MV-NA}$ and $Cross_{NV-MA}$ evaluate each pair in isolation, exposing the model's viewpoint sensitivity. Finally, $Same_{MV-MA}$ applies the synchronization head to a single stream with an artificially inserted offset, probing the ceiling performance when no cross view variation exists. All methods were scored with mean-absolute-error (MAE) and mean-squared-error (MSE) averaged over the four lags {0.1, 0.5, 1.0, 2.0 s}.

Results

Table. 2 Performance Metrics for Evaluation Methods on Replay and Womenera Datasets

Dataset	Method	MAE	MSE
Replay	ConfSelect	0.765200	0.870140
	SoftVoting	0.688300	0.752080
	$Cross_{MV-NA}$	0.781560	0.893840
	$Cross_{NV-MA}$	0.769060	0.873800
	$Same_{MV-MA}$	0.781560	0.893840
Womenera	ConfSelect	0.788667	0.923667
	SoftVoting	0.665733	0.695167
	$Cross_{MV-NA}$	0.716333	0.782567
	$Cross_{NV-MA}$	0.773300	0.896133
	$Same_{MV-MA}$	0.775200	0.901500

As summarised in Table 2, Soft Voting achieved the lowest errors on both datasets. On Replay it reduced MAE from 0.7652 to 0.6883 (-10 %) and MSE from 0.8701 to 0.7521 (-14 %) relative to ConfSelect. On Womenera the gains were larger, lowering MAE to 0.6657 (-16 %) and MSE to 0.6952 (-25 %).

The single-pair evaluations ($CrossModal_{MV-NA}$ and $CrossModal_{NV-MA}$) lagged behind Soft Voting on both metrics, and the same-source baseline registered the highest errors, confirming that cross view information is essential for accurate alignment.

The comparison reveals that ConfSelect benefits from confidence ranking, yet Soft Voting further stabilises predictions by blending the two views, thereby mitigating outliers. Taken together, the results establish Soft Voting as the most reliable strategy for fine-grained multi-camera synchronization across heterogeneous viewpoints.

4.4 Comparison Method

For comparison, we re-implemented AE2 [15], an object-centric network that aligns egocentric–exocentric video pairs by minimizing Dynamic Time Warping (DTW) cost. Replay and Womenera datasets were partitioned 70%/20%/10% into train, validation and test splits, and processed into 10-second clips. Two evaluation settings were used: a +1 second offset and perfect synchrony. All preprocessing followed the procedure described in Section 4. AE2 was trained for 50 epochs with the Adam optimizer (learning rate 1×10^{-4}, weight decay 1×10^{-5}), hidden dimension 256, and DTW cost as both loss and metric. Training was performed on NVIDIA RTX 4080 and A4000 GPUs.

Table. 3 Performance of AE2 on Multi Camera Datasets

Offset(sec)	Replay	Womenera
1	0.11203	0.16623
0	0.10762	0.10850

Table 3 reports the mean DTW costs. On Replay, the score decreased from 0.11203 under the 1-second offset to 0.10762 when no offset was applied; on Womenera, it fell from 0.16623 to 0.10850. While the cost improves as temporal misalignment is reduced, the change is modest, indicating that AE2 provides a coarse yet consistent visual baseline for subsequent multimodal experiments.

As a second comparison, we reproduce the method from [16]. The system consists of a Matching-Frames (MF) network that learns an object-centric distance metric followed by a Delay-Estimation (DE) network that regresses the frame lag. Using the Womenera4, Womenera5 pair, we extract the first 24,000 frames at 224×224 resolution and compute Farneback optical flow for the 23,999 inter-frame intervals.

The MF network is trained with Triplet-Euclidean loss (margin = 0.5) where the anchor is a frame from Womenera4, the positive is its temporally aligned counterpart in Womenera5, and the negative is a frame shifted by −10…−1 or 1…20 frames; the split is 80%/20% and the batch size is 64. After convergence, the MF weights are frozen and 40,000 clip pairs of length 20 frames are generated with relative shifts of −19…20 frames. Each pair is converted to a 20×20 distance matrix, flattened, and fed to the DE network, which is trained (80%/20%, batch size 32) to predict the integer offset.

The MF stage attains a training loss of 0.2047 and a validation loss of 0.2424 (initially 0.5), confirming effective metric learning. The DE stage reports 48.9% training accuracy and 44.2% validation accuracy with losses of 1.87/2.13 and a mean absolute error of 0.032 frames. Precision remains high (0.86/0.78) while recall is lower (0.35/0.32).

5. Conclusion and Future Work

This study presents a multimodal synchronization framework for aligning multi-camera video streams in real-world broadcasting environments. Through extensive experiments across three datasets—Replay, CASTLE 2024, and Womenera—we empirically confirm that the audio modality provides significantly stronger synchronization cues than the visual modality. Specifically, the similarity gap between same- and different-timestamp segments for audio (0.19) was approximately 190× larger than that for video (0.001), highlighting the superior discriminative power of audio features for scene alignment. Although simple audio-visual concatenation yielded moderate gains (0.08), it did not outperform the audio-only approach, suggesting that naïve fusion methods remain suboptimal.

We hypothesize that this disparity arises from (1) the limited visual variance present in our dataset, and/or (2) the representational limits of current visual embedding extractors. To verify these hypotheses, we plan to conduct further experiments with diverse datasets, varying scene structures, segment granularities, and fusion strategies.

In addition, we introduced a lightweight offset classifier, Multi-Synchformer, which repurposes Synchformer's second-stage synchronization module with minimal modifications. Among four evaluated inference schemes, our proposed Soft Votingstrategy—which fuses predictions across view pairs using confidence-weighted averaging—consistently outperformed all baselines. Compared to ConfSelect, Soft Voting reduced alignment errors by up to 14% on Replay and 25% on Womenera, demonstrating the importance of cross view integration for fine-grained synchronization.

We further benchmarked our approach against re-implementations of AE2 [15] and the stereo-video synchronization model from [16]. While AE2 produced modest DTW cost improvements under reduced misalignment, the method from [16] achieved stable metric learning but exhibited limited recall and coarse offset regression performance. These comparisons reaffirm the need for synchronization frameworks that explicitly leverage modality-specific cues and structured fusion.

In future work, we aim to enhance our system through the following directions:
(1) modality reliability-aware weighted fusion,
(2) improved visual feature encoders sensitive to temporal variation,
(3) timecode search optimization via embedding quantization and dimensionality reduction,
(4) dedicated offset prediction modules for timecode localization, and
(5) deployment of a unified synchronization architecture robust to diverse production scenarios.

Together, these extensions will contribute to more accurate, efficient, and scalable multi-camera synchronization pipelines suited for real-world broadcasting and video production workflows.

6. References

[1] Llagostera Casanovas, A., & Cavallaro, A. (2015). Audio-visual events for multi-camera synchronization. Multimedia Tools and Applications, 74(4), 1317-1340.

[2] Chung, J. S., & Zisserman, A. (2016). Out of time: automated lip sync in-the-wild. In Proceedings of the ACCV Workshop on Multi-view Lip-reading, 2016.

[3] Owens, A., & Efros, A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In European Conference on Computer Vision, 2018.

[4] Chung, J. S., et al. (2019). PerfectMatch: Audio-visual synchronization for in-the-wild videos. In Proceedings of the International Conference on Computer Vision, 2019.

[5] Khosravan, N., Ardeshir, S., & Puri, R. (2019). On attention modules for audio-visual synchronization. In CVPR Workshop on Sight and Sound, 2019.

[6] Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., & Zisserman, A. (2021). Audio-visual synchronisation in-the-wild. In British Machine Vision Conference, 2021.
[7] Iashin, V., Xie, W., Rahtu, E., & Zisserman, A. (2022). Sparse in space and time: Audio-visual synchronisation with trainable selectors. In British Machine Vision Conference, 2022.
[8] Iashin, V., Xie, W., Rahtu, E., & Zisserman, A. (2024, April). Synchformer: Efficient synchronization from sparse cues. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (pp. 5325-5329). IEEE.
[9] Gong, Y., Chung, Y., & Glass, J. (2021). AST: Audio Spectrogram Transformer. In Interspeech.
[10] Mandela, P., Campbell, D., Asano, Y., Misra, I., Metze, F., Feichtenhofer, C., Vedaldi, A., & Henriques, J. F. (2021). Keeping your eye on the ball: Trajectory attention in video transformers. In Advances in Neural Information Processing Systems.
[11] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning.
[12] Shapovalov, R., Kleiman, Y., Rocco, I., Novotny, D., Vedaldi, A., Chen, C., ..., & Neverova, N. (2023). Replay: Multi-modal multi-view acted videos for casual holography. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20338-20348).
[13] Rossetto, L., Bailer, W., Dang-Nguyen, D. T., Healy, G., Jónsson, B. Þ., Kongmeesub, O., ..., & Gurrin, C. (2025). The CASTLE 2024 Dataset: Advancing the Art of Multimodal Understanding. arXiv preprint arXiv:2503.17116.
[14] Shrestha, P., Barbieri, M., Weda, H., & Sekulovski, D. (2010). Synchronization of multiple camera videos using audio-visual features. IEEE Transactions on Multimedia, 12, 79–92.
[15] Xue, Z. S., & Grauman, K. (2023). Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. Advances in Neural Information Processing Systems, 36, 53688-53710.
[16] Boizard, N., El Haddad, K., Ravet, T., Cresson, F., & Dutoit, T. (2023). Deep learning-based stereo camera multi-video synchronization. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.