AUDIO-VISUAL SYNCHRONISATION IN THE WILD

Honglie Chen
hchen@robots.ox.ac.uk
Weidi Xie
weidi@robots.ox.ac.uk
Triantafyllos Afouras
afourast@robots.ox.ac.uk
Arsha Nagrani
arsha@robots.ox.ac.uk
Andrea Vedaldi
Vedaldi@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford
Oxford, UK

Abstract

In this paper, we consider the problem of audio-visual synchronisation applied to videos 'in-the-wild' (*i.e.* of general classes beyond speech). As a new task, we identify and curate a test set with high audio-visual correlation, namely VGG-Sound Sync. We compare a number of transformer-based architectural variants specifically designed to model audio and visual signals of arbitrary length, while significantly reducing memory requirements during training. We further conduct an in-depth analysis on the curated dataset and define an evaluation metric for open domain audio-visual synchronisation. We apply our method on standard lip reading speech benchmarks, LRS2 and LRS3, with ablations on various aspects. Finally, we set the first benchmark for general audio-visual synchronisation with over 160 diverse classes in the new VGG-Sound Sync video dataset. In all cases, our proposed model outperforms the previous state-of-the-art by a significant margin. Project page: `https://www.robots.ox.ac.uk/~vgg/research/avs`

1 Introduction

In videos, the audio and visual streams are often strongly correlated, presenting effective signals for self-supervised representation learning [7, 47]. A useful task in this area is audio-visual synchronisation, and several studies have shown promising results even without requiring any manual supervision [4, 17, 22]. However, these works study this problem extensively on only one class – human speech – where even a slight offset is easily discernable.

In this paper, rather than focusing on a specialised domain, *e.g.* human speech [4, 16, 17, 22], or videos with periodic sounds such as the tennis shots in a match [27], we aim to explore audio-visual synchronisation on general videos in the wild (characterized by more than 160 sound classes). Solving this task would be extremely useful for a number of applications including video conferencing, television broadcasts and video editing, which are largely done by 'off-line' measurements or heavy manual processing [24, 51, 53].

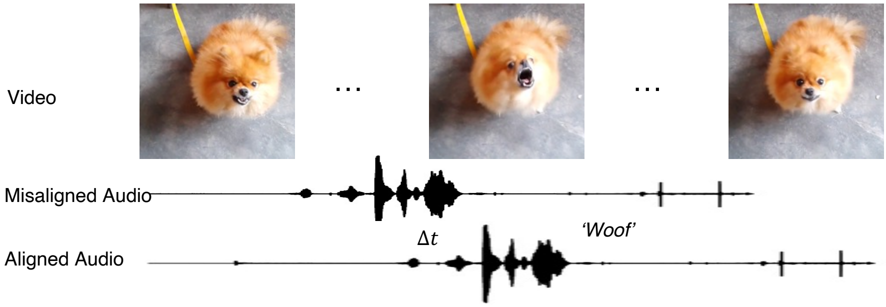


Figure 1: **Audio-visual synchronisation in the wild**. The goal of this work is to develop an audio-visual synchronisation method that performs well on general videos in-the-wild. Unlike speech videos, highly correlated audio and visual events for general classes may occur briefly in the video (e.g. the bark of the dog in the centre of this clip). A short clip sampled randomly from the video might miss this fleeting moment; with longer input videos this becomes less probable. With this in mind, we propose a Transformer based architecture that can operate on long sequences, and is able to perform audio-visual synchronisation on videos of 160 general sound classes.

There are several challenges in automatic audio-visual synchronisation for general classes. First, unlike the task of synchronising speech [2, 19, 21, 44], which contains audio-visual evidence from the lips most of the time, videos from general classes may contain uniform sounds (*e.g.* airplane engine sound, electric trimmer), ambient sounds (*e.g.* wind, water, crowds, traffic), or small object sound sources (*e.g.* players in an orchestra, birds), which make synchronisation extremely challenging or even impossible; Second, for categories with strong audio-visual evidence, localising such signals can also be difficult, for example: temporally, 'dog barking' may happen instantaneously, as shown in Figure 1, and spatially, unlike in speech synchronisation where visual cues are largely localised to lip motions, in general videos the entire frame must be processed to accommodate different object classes; Third, due to the aforementioned challenges, it is unclear how to evaluate the synchronisation in general classes.

In order to address these issues, first, we curate a new benchmark for general audio-visual synchronisation called VGG-Sound Sync using a subset of VGG-Sound [13]. Specifically, this is built by selecting classes and video clips that potentially have audio-visual correlation, and removing those classes and video clips that don't, *e.g.* uniform, ambient sound; Second, compared with previous works, we use substantially longer input video sequences, so that the chance of having a synchronised audio and video event in the input increases. We explore several variants of Transformer-based architectures that can elegantly deal with these long sequences of variable lengths, and that use self-attention to implicitly pick out the relevant parts in both space and time. Finally, we conduct a thorough study on the VGG-Sound Sync test set, estimating the chance of audio-visual synchronisation for different clip lengths, and also define a set of metrics for evaluation.

Concretely, in this paper, we consider the problem of audio-visual synchronisation applied to 'in-the-wild' videos, *i.e.* general classes beyond speech. We make the following contributions: (i) we identify and curate a subset of general classes from VGG-Sound, namely VGG-Sound Sync, with potentially high audio-visual correlation; (ii) we introduce a set of transformer-based architectures for audio-visual synchronisation, which can exploit the spatial-temporal correlations between audio and visual streams, such models can train and predict on variable length video sequences; (iii) we conduct an analysis on the VGG-Sound Sync test set, and define an evaluation metric for audio-visual syncrhonisation on these videos; (iv) we achieve state-of-the art synchronisation performance on standard lip reading speech benchmarks, LRS2, LRS3; and more importantly, set the first benchmark for audio-visual synchronisation in general (non-speech) classes.

2 Related Work

Audio-visual synchronisation. Early works studied audio-visual synchronisation in talking faces [35, 52] using handcrafted features and statistical models. [16] developed a model for synchronizing lip movements to audio speech, based on a dual-encoder architecture trained with contrasting learning. Follow-up works improved this pipeline by moving to noise-contrastive objectives [22], or directly inferring the audio-visual offset conditional on cross-similarity patterns [38]. Lip synchronisation is an important component for pipelines used for various visual speech related tasks, such as lipreading [2, 18], active speaker detection [16] and sign language recognition [5]. Although these works demonstrate strong synchronisation performance, they are limited in terms of deployment as they are applicable only on videos that include speech. Our method generalizes to broader sound source classes and conditions, while also outperforming these works in the speech domain. Other closely related works have investigated lip-syncing [34], i.e. the temporal alignment of video and speech clips from different sources, speech-conditioned face animation [20, 58], and audio-visual dubbing [49, 62]. Audio-visual synchronisation has been also used as a pre-text task for learning general visual and audio representations [4, 15, 40, 47, 48]. [37] investigate the use of attention for audio-visual synchronisation on speech data. [27] train models to detect synchronisation errors based on mismatch of event detection between the audio and visual stream. [12] propose a method for synchronising audio-visual recordings of the same events from different cameras. Unlike the works above which use simple concatenation between audio and visual features, we employ encoder-based and decoder-based Transformers to implicitly match the relevant parts.

Audio-visual learning. Our work is more broadly related to various works on audio-visual learning, including audio-visual event detection [43, 54], sound-source localization [4, 8, 28, 50, 61], representation learning [6, 9, 45], audiovisual fusion [36, 46, 60] and sound source separation [32, 56, 64]. More recently, [65] proposed to leverage temporal motion information to separate musical instrument sound. [30] further improved the sound separation models with explicit keypoint-based representations. Another line of work explored audio synthesis using visual input: [29] utilized body keypoints to synthesize music from a silent video, and [39] synthesized piano music from overhead views of the hands. [31] converted monaural audio into binaural audio by injecting visual spatial information.

Transformers. Transformers [57] were originally introduced for NLP tasks, in particular machine translation where they showed improvement over recurrent-based encoder-decoder architectures. Since then they have been widely applied to a great range of problems, including speech recognition [33], language modelling [23, 25], object detection [11, 63]. Recent works have even extended their use to visual feature extraction, replacing CNNs, for classification [26], semantic segmentation [26, 59] and video representation learning [10]. In the multi-modal domain, [42, 55] explored unimodal and cross-modal temporal contexts simultaneously to detect audio-visual events, and [41] alleviated the high memory requirement of a vanilla Transformer by sharing the weights across layers and modalities. Audio-visual fusion using transformers has also been explored by new architectures such as Perceiver [36] and MBT [46].

3 Method

In this section, we describe our proposed method, which we call Audio-Visual Synchronisation with Transformers (AVST). Our goal is to detect audio-visual synchronisation without the use of any manual annotation. Similar to prior work [4, 16, 47], we first use CNN encoders to extract visual and audio representations from unlabelled video (described in section 3.1.1). In section 3.1.2, we introduce three variants of our Transformer-based module that can jointly process visual and audio features, and discuss the pros and cons for each architecture. Finally in section 3.2, we describe the contrastive learning objective used to train the model. An overview of our architecture can be found in Figure 2.

3.1 Architecture

3.1.1 Audio and visual representations

The proposed model has two input streams, one ingesting a short video clip $v_i \in \mathbb{R}^{3 \times T \times H_v \times W_v}$ consisting of T visual frames and the other taking in an audio spectrogram $a_j \in \mathbb{R}^{1 \times H_a \times W_a}$, where i, j index the source of each modality (e.g. when $i = j$ the visual and audio signals come from the same video and are temporally aligned). We compute representations for each modality using functions $f(\cdot;\theta_1)$ and $g(\cdot;\theta_2)$, which in this case are instantiated using CNN encoders:

$$V_i = f(v_i;\theta_1), \quad V_i \in \mathbb{R}^{c \times t_v \times h \times w} \tag{1}$$

$$A_j = g(a_j;\theta_2), \quad A_i \in \mathbb{R}^{c \times t_a} \tag{2}$$

Both representations V_i and A_j have the same number of channels c, which allows us to jointly model the input video and audio with cross-modal attention.

3.1.2 Synchronisation module

The visual and audio representations are formulated into a sequence of tokens, and passed through a Transformer [57] consisting of N layers. We introduce three variants of AVST, each one with a slightly different design choice for modelling audio-visual information.

Encoder variant (AVST$_{\text{enc}}$). The most straightforward step is to simply treat the dense visual features as a sequence of "visual tokens". To that end, the visual features are flattened over the spatial dimensions and concatenated to the audio features after also prepending a learnable `class` token ([CLS]), inspired by the BERT model [25]. In order for the model to distinguish the signals from the two modalities and maintain spatio-temporal positional information (as all subsequent Transformer layers are permutation invariant), three types of encodings are also added to the audio and visual features: modality encodings $E_m \in \mathbb{R}^{c \times 2}$, that indicate the type of feature (i.e. audio or visual); temporal encodings $E_{t_{\{v,a\}}} \in \mathbb{R}^{c \times t_{\{v,a\}}}$ and spatial encodings $E_s \in \mathbb{R}^{c \times h \times w}$, that keep track of absolute positions for the tokens:

$$\overline{V_i} = \text{FLATTEN}(V_i) + E_m + E_{t_v} + E_s, \tag{3}$$

$$\overline{A_j} = A_j + E_m + E_{t_a}, \tag{4}$$

$$Z_{ij} = [[\text{CLS}]; \overline{V_i}; \overline{A_j}] \tag{5}$$

where $[;]$ denotes a concatenation operation. The output result, $Z_{ij} \in \mathbb{R}^{c \times (1 + hwt_v + t_a)}$, is then fed into a Transformer Encoder [57], that is composed of a stack of Multihead Self-Attention (MSA), and feed forward networks (FFNs). This module allows the tokens from both modalities to directly interact with each other through the self-attention operations:

$$Y_{ij} = \text{TRANSFORMER-ENCODER}(Z_{ij}). \tag{6}$$

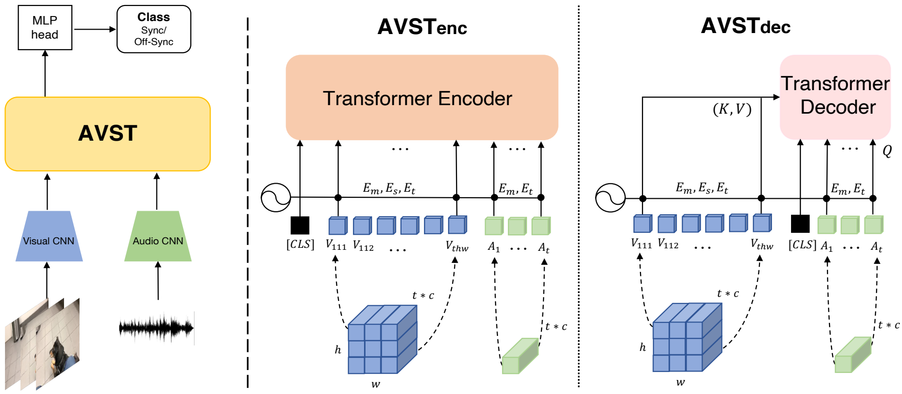


Figure 2: **The AVST model architecture and variants.** We use AVST to jointly model visual and audio representations computed from backbone CNNs, with an MLP head to predict audio-visual synchronisation (left). On the right, we show two variants of the AVST transformer backbone, the Encoder (AVST$_{\text{enc}}$) and Decoder variant (AVST$_{\text{dec}}$). AVST$_{\text{enc}}$ uses self-attention for all audio and visual features, whereas in AVST$_{\text{dec}}$ the visual information is kept fixed, and the audio latents are used to QUERY the visual information which forms the KEY, VALUE pairs.

Max-pooled encoder variant (AVST$_{\text{enc-mp}}$). Naively feeding all visual features densely into the Transformer is computationally expensive, with a quadratic cost, $\mathcal{O}((hwt_v + t_a)^2)$, which significantly limits the scalability to longer video sequences. Thus, although the architecture is powerful, it heavily limits the audio-visual samples that can be processed in each batch, which in turn limits the number of negatives that can be used for training, resulting in sub-optimal performance, as we will show in the experiments (section 4).

Rather than taking dense visual features as input, we propose a cheaper alternative, which consists of a simple Global Max Pooling (GMP) operation spatially on each frame. This reduces the length of the sequence that is input to the Transformer from $(hwt_v + t_a + 1)$ to $(t_v + t_a + 1)$; and thereby significantly lowers the memory footprint of the MSA module. To obtain AVST$_{\text{enc-mp}}$ we simply replace the flattening operator in Equation 4 with GMP:

$$\overline{V_i} = \mathrm{GMP}(V_i) + E_m + E_{t_v} + E_s \tag{7}$$

Decoder variant (AVST$_{\text{dec}}$). Using the visual feature from max-pooling is computationally efficient, however, it also removes spatial information in the visual representations, impairing the ability of audio features to probe fine-grained visual information, which may be required for certain general object categories.

To resolve the aforementioned challenge, we consider an alternative architecture that uses a Transformer decoder [57], as shown in Figure 2 (right), where dense visual features are kept fixed without self-attention and passed as the KEY and VALUE inputs to every decoder layer, and audio features (concatenated along a [CLS] token, similarly to AVST$_{\text{enc}}$) are passed as the QUERY inputs:

$$\textsc{Query} = \textsc{concat}([\mathrm{CLS}], \overline{A}_j),\ \textsc{Key} = \textsc{Value} = \overline{V_i} \tag{8}$$

$$Y_{ij} = \textsc{Transformer-Decoder}(\textsc{Query},\ \textsc{Key},\ \textsc{Value}) \tag{9}$$

3.1.3 Output head

For all variants, we only use the first token (Y_{ij}^1), of the output of the final encoder (or decoder) layer, corresponding to the [CLS] position in the input sequence. This functions as an aggregate representation of the whole output sequence and is fed to $h(\cdot;\theta_3)$, which we implement as an MLP head. The output is a synchronisation score that indicates to what degree the inputs v_i and a_i are in sync, $s_{ij} = h(Y_{ij}^1; \theta_3)$.

3.2 Training objectives

Given mini-batches $\mathcal{B} = \{(v_1, a_1), (v_2, a_2), ..., (v_k, a_k)\}$ of temporally aligned audio-visual pairs, the goal is to jointly optimize the entire pipeline in an end-to-end manner, so that the prediction scores for synchronised pairs (v_i, a_i) are maximised, while the scores of out-of-sync pairs (v_i, a_j) are minimised. Training proceeds by minimising the commonly used InfoNCE loss, defined as:

$$\mathcal{L} = -\frac{1}{k}\sum_{i=1}^{k}\left[\log\frac{\exp(s_{ii})}{\sum_j \exp(s_{ij})}\right]$$

Discussion. Unlike previous works, which simply compute either the Euclidean distance or the cosine similarity between the audio and visual representations obtained from separate CNN streams to predict synchronisation, we use a Transformer model that jointly models the relationship between the audio and visual streams using attention over multiple layers. This is useful for attending to longer input sequences, where informative audio and video may only be localised in a short sub-sequence of the video.

4 Experiments

In the following sections, we describe the datasets, evaluation protocol and experimental details to thoroughly assess our method.

4.1 Datasets

Audio-visual speech datasets: We conduct experiments on two public audio-visual speech datasets, namely, LRS2 [2, 19, 21] and LRS3 [1], which have been created from British television footage and TED talks from YouTube respectively. Both datasets are distributed as short video clips of tightly cropped face tracks around the active speaker's head. Since LRS3 is based on public YouTube videos, we also extract full-frame versions of the same clips for all splits ("pretrain", "trainval" and "test"). To distinguish between these two versions of LRS3, we refer to them as "cropped" and "full-frame" respectively. Note that for LRS2, only the "cropped" version is available.

General sound dataset: Here, we construct a new benchmark called VGG-Sound Sync using a subset of VGG-Sound [13], which was recently released with over 200k clips, and each clip is labelled as one of the 300 different sound categories. This dataset is conveniently audio-visual, in the sense that the object that emits sound is likely to be visible in the corresponding video clip. In the next section, we will detail the curation process.

4.2 Evaluation protocol

Depending on the downstream benchmarks, we consider two different evaluation protocols.

4.2.1 Audio-visual synchronisation on speech

For LRS2 and LRS3, we follow previous works and use an input of 5 frames, extracted at 25fps. During testing, the synchronisation scores were computed densely between each 5-frame video feature and all audio features in ±15 frame range. Synchronisation was then determined to be correct if the lip-sync error was not detectable to a human, *i.e.* the maximum score between two streams is within ±1 frame ($\pm 0.04s$) from the ground truth [22].

4.2.2 Audio-visual synchronisation on general classes

Compared to speech videos with audio-visual cues (the lip motion and speech) spanning almost the entire clip, evaluating synchronisation on general videos potentially incurs two challenges: (1) videos with only ambient or uniform sound, *e.g.* wind, wave, engine sound, are unlikely to have any cues that can be used for synchronisation; (2) the audio-visual cues for synchronisation are sometimes instantaneous, *e.g.* a dog barking may only last for less than 1s. In the following, we describe the evaluation benchmark and how it was constructed.

Seconds	1s	2s	3s	4s	5s	6s
Audio-visual evident	50%	56%	57%	62%	60%	59%
Uniform/ambient sound	30%	34%	35%	31%	35%	38%
No sound/object	20%	10%	8%	7%	5%	3%

Table 1: Categorisation of video clips as duration of video varies.

Categorising video clips. Here, we analyse the statistics of videos in the VGG-Sound test set, by categorising each video clip into three classes, namely, audio-visual evident, uniform / ambient sound, missing sound / visual object. Specifically, we randomly sample 1200 video clips, where each clip is of different lengths between 1s and 6s for manual verification. As shown in Table 1, the following phenomenon can be observed: *First*, the proportion of clips with uniform or ambient sound remains roughly constant, as this error is caused by all the video clips of particular sound categories; *Second*, as expected, with the increase of temporal lengths, the chance of having audio-visual cues for synchronisation increases. Notably, when clips are over 2s, the error rate drops to around 10%.

At this stage, we curate a subset of VGG-Sound by filtering the sound categories to remove ones that are potentially dominated by uniform / ambient sound, resulting in a test set of over 160 classes, 95k training videos and 5k testing videos (each of them lasts 10 seconds).

Verifying synchronisation of YouTube videos. In this section, we conduct manual verification to serve two purposes: *first*, as the video clips in VGG-Sound are all sourced from YouTube, their audio-visual alignments are not always guaranteed, we aim to understand the chance of these videos being audio-visual synchronised, at least from the perspective of an ordinary human observer; *second*, we aim to understand the human tolerance, by that we mean, how much temporal misalignment is noticeable for human observers. In a practical evaluation, offsets smaller than such tolerance should be ignored or considered as correct. In detail, we randomly sample 500 example videos from our test set with 25fps, and create 1000 audio-visual pairs, with each lasting 5s. The temporal offsets between both streams vary from $[-0.8, +0.8]$ second, for example, for one visual clip sampled at time t, its paired audio signal can be centered at any time between $t - 0.8, t + 0.8$, we feed these pairs to human observers and ask a binary question: *is the given audio-visual pair synchronised ?* Please check the detailed statistics on proportions of videos considered to be syncd by a manual observer in our ArXiv version.

Summary. To evaluate the synchronisation for videos of general classes, we curate a test set from VGG-Sound, namely VGG-Sound Sync, with ambient, uniform sound categories removed. We only include audio-visual pairs of length between 2 - 6 second, that have a sufficiently high chance of containing informative cues for synchronisation. During evaluation, we decode the videos with 25fps, and construct audio or visual input by taking every

5th frame, note that, this has the same effect as using input decoded from 5fps. The synchronisation scores are computed for all audio-visual pairs with $[-15,-14,\ldots,+14,+15]$ frame gaps. Considering the challenging nature for audio-visual synchronisation in natural videos, synchronisation is determined to be correct if the synchronisation error is not detectable by a human, i.e. the maximum score between two streams is within ±5 frames ($\pm 0.2s$) from the ground truth. We refer the reader to our ArXiv version for the study on human tolerance on synchronisation in general videos.

4.3 Implementation details

Training curriculum. Following prior work [4, 40], we train our models in two stages: in the first stage, we construct the mini-batches by sampling audio-visual clips from different videos, this provides easy (correspondence) negatives that helps the training converge. In the second stage, all the clips in a mini-batch are sampled from the same video, which provides harder (synchronisation) negatives.

Training hyper-parameters. On a P40 GPU with 24GB memory, we train AVST_{enc} with a batch-size of 4 (due to memory restrictions), for $\text{AVST}_{\text{enc-mp}}$ and AVST_{dec}, we use a batch-size of 16 and 12 respectively, thereby allowing more negatives per batch.

Architectural Details. Unless otherwise specified, our Transformer encoder consists of 3 layers, 4 attention heads and a hidden unit dimension of 512. Typically $H = W = 224$ and $h = w = 14$. We refer the readers to the ArXiv version for more details.

4.4 Results on speech datasets

We first report experimental results on LRS2 and LRS3, and perform a number of ablations on different architectural design choices. We also analyse the model's robustness on cases, where the visual or audio signal is partially unavailable.

Architectures comparison. To compare our proposed architecture variants and assess their trade-offs, we train and evaluate them on the "full-frame" version of LRS3 and show results of all three Transformer variants in Table 2. Due to the memory restrictions, we can only train AVST_{enc} with a fixed length of 5 frames, whereas for the other two architectures, training is done with variable sequence length and larger batch size (see section 4.3). We observe a large gap ($6\% - 7\%$) between the performance of AVST_{enc} and the other two variants, which indicates that AVST_{enc} suffers from the reduced number of negatives. We also note that AVST_{dec} can localise sound sources because it preserves spatial information, but shows slightly worse performance than $\text{AVST}_{\text{enc-mp}}$ on speech datasets. We conclude that $\text{AVST}_{\text{enc-mp}}$ is a light-weight solution that offers the best performance when the sounding objects (*e.g.* lips) are clear and unique, which need little fine-grained spatial information.

Comparison to the state-of-the-art. We compare our method to previous work on "full-frame" LRS3 in the top half of Table 2. We show a significant improvement compared to the AVobjects baseline (16% gain) on short input (5 frames) reaching up to an almost saturated 98.6% accuracy with 15 frames. In the bottom half of Table 2, we further summarise our results for experiments on the "cropped" LRS2 dataset. Here too, we observe that our method greatly outperforms both the SyncNet[17] and PM [22] baselines, and achieves almost perfect accuracy with 15 frames of input during test time.

Since $\text{AVST}_{\text{enc-mp}}$ shows superior performance on speech datasets using a light-weight architecture, we conduct the rest of the analysis on speech data using $\text{AVST}_{\text{enc-mp}}$. In addition, in order to compare with SyncNet and PM, we use the same fixed length of 5 frames

				Clip Length in frames (seconds)					
Model	# Params.	Var.	Dataset	5(0.2s)	7(0.28s)	9(0.36s)	11(0.44s)	13(0.52s)	15(0.6s)
AVobjects [4]	69.4M	✗	LRS3	61.8	72.0	79.7	85.4	89.5	91.8
AVST$_{\text{enc}}$	42.6M	✗	LRS3	70.2	77.1	83.3	88.4	92.0	94.4
AVST$_{\text{dec}}$	44.5M	✓	LRS3	75.7	86.4	89.4	94.0	95.1	96.9
AVST$_{\text{enc-mp}}$	42.4M	✓	LRS3	**77.3**	**88.0**	**93.3**	**96.4**	**97.8**	**98.6**
SyncNet [17]	13.6M	✗	LRS2	75.8	82.3	87.6	91.8	94.5	96.1
PM [22]	13.6M	✗	LRS2	88.1	93.8	96.4	97.9	98.7	99.1
AVST$_{\text{enc-mp}}$	42.4M	✓	LRS2	**91.9**	**97.0**	**98.8**	**99.6**	**99.8**	**99.9**

Table 2: **Architecture comparison on LRS3 and LRS2.** We use the 'full-frame' dataset. 'Var': whether models are trained and tested using variable length inputs. '5-15' refers to the number of input frames to corresponding models.

	Clip Length (frames)					
# Layers	5	7	9	11	13	15
1	89.1	94.0	96.8	98.4	99.1	99.4
2	91.6	95.4	97.6	98.8	99.1	99.6
3	92.0	95.5	97.7	98.8	99.3	99.6

Table 3: **Ablation on Transformer depth (LRS2).** Performance increases with depth.

	Clip Length (frames)					
Mask	5	7	9	11	13	15
Audio	73.1	85.3	92.6	96.1	98.0	99.2
Visual	76.5	87.3	93.4	96.9	98.2	99.3
Both	71.7	84.0	91.2	95.6	97.7	99.1

Table 4: **Robustness test on LRS2.** 1 frame is masked during train and test.

during training and testing.

Number of Transformer Layers. We ablate the Transformer depth on the LRS2 dataset in Table 3. As more layers are added, the performance consistently improves, achieving the best performance with 3 layers. This confirms the effectiveness of self-attention in jointly modelling audio and visual information.

Robustness test. To mimic real-world scenarios, where sound sources and their corresponding sound might not appear together at every frame, we further conduct experiments to assess the robustness of our model on the LRS2 dataset by randomly masking input audio or video frames. We mask 1 frame for each or both modalities. As can be seen in Table 4, we find that for short inputs this causes a significant performance drop, however with longer inputs, we achieve comparable results to the non-masked case in Table 2.

4.5 Results on general sound classes

In this section, we report audio-visual synchronisation results on the VGG-Sound Sync dataset consisting of videos with general sound classes, and compare with several strong baselines. Results are provided in Table 5. First, while comparing with the recent AVobjects [3] method, both of our models show superior results on all input lengths, this is because (1) we trained on variable input lengths, where longer samples contain richer audio-visual evidence; and (2) the use of Transformer based architectures (AVST$_{\text{enc}}$ and AVST$_{\text{dec}}$) can implicitly discover the important temporal parts in long sequences. Second, in contrast to the results in speech datasets (Table 2), we note that AVST$_{\text{dec}}$ has higher accuracy than AVST$_{\text{enc}}$ on general videos. The reason is that general videos contain complex visual scenes and, compared to other variants, AVST$_{\text{dec}}$ can extract fine-grained spatial information in such situation by explicitly computing the attention between image regions and the audio sequence, therefore showing better performance. Finally, we analyse the performance for each class of VGG-Sound Sync dataset in Figure 3, and find that the performance is highly class dependant, with the best class ('child singing') achieving 75.7%, and most highly per-

	Clip Length in frames (seconds)				
Method	10(2s)	15(3s)	20(4s)	25(5s)	30(6s)
AVobjects [4]	37.2	42.6	45.1	47.3	49.4
AVST$_{\text{enc-mp}}$	39.0	44.1	46.8	49.7	51.8
AVST$_{\text{dec}}$	40.1	45.6	48.2	50.9	52.9

Table 5: **Audio-visual synchronisation results on VGG-Sound Sync**. We show results for sequences of $10-30$ frames. Our model outperforms the state of the art AVObjects [4] on sequence lengths $\geqslant 10$.

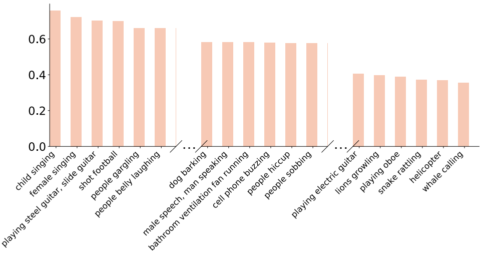
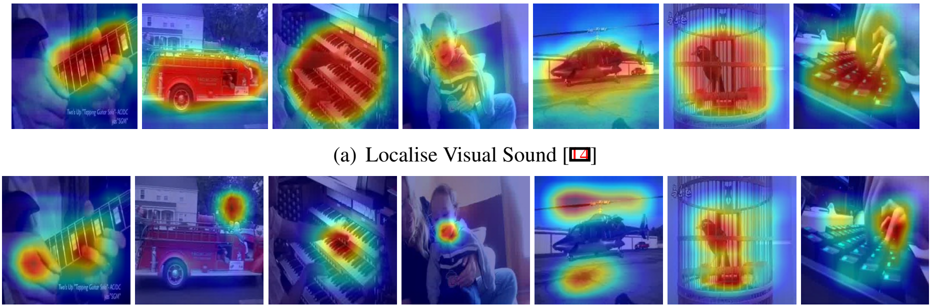


Figure 3: **Per-class accuracy on VGG-Sound Sync**.


(a) Localise Visual Sound [14]

(b) AVST$_{\text{dec}}$

Figure 4: **Attention heatmaps on VGG-Sound Sync.** We compare the heatmaps that we obtain with the AVST$_{\text{dec}}$ model to the state-of-the-art method for sound source localization [14]. It is interesting to note that while [14] highlights discriminative parts of the objects that are generally associated with the sound and are therefore *sufficient to identify it* – *i.e*. the entire musical instrument, firetruck and helicopter – our method focuses on the parts that exhibit some motion – *i.e*. the player's hands, the firetruck siren and the helicopter's rotor – that *modify or create sound* and are necessary to solve the much more challenging synchronisation task.

forming classes containing strong audio-visual correlations, *e.g*. 'female singing','playing steel guitar', etc.

4.5.1 Visualisation of attention heatmaps

We visualise the attention heatmaps of our model on samples from VGG-Sound Sync in Figure 4 (refer to the ArXiv version for more qualitative results). For general object classes in VGG-Sound Sync, the model manages to pick up on interesting sources of motion that produce sound. When comparing the heatmaps produced by AVST$_{\text{dec}}$ to the current *state-of-the-art* sound source localization method [14], we notice that our method attends to regions that *create or modify* sound, *e.g*. hands, lips, helicopter rotor, etc, while [14] tends to localise the entire object.

5 Conclusion

We revisit the problem of audio-visual synchronisation in human speech and introduce a new general class audio-visual synchronisation benchmark called VGG-Sound Sync. We experiment with different variants of Transformer-based architectures, analyse several critical components, and conduct thorough ablation studies to validate their necessity. Consequently, our proposed architecture sets new *state-of-the-art* results on LRS2 and LRS3, and provides baselines for general sound audio-visual synchronisation.

Acknowledgements

This work is supported by the UK EPSRC CDT in Autonomous Intelligent Machines and Systems, the Oxford-Google DeepMind Graduate Scholarship, the Google PhD Fellowship, the EPSRC Programme Grant VisualAI EP/T028572/1, and a Royal Society Research Professorship.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. In *arXiv preprint arXiv:1809.00496*, 2018.

[2] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE PAMI*, 2019.

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *Proc. ICASSP*, 2020.

[4] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proc. ECCV*, 2020.

[5] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *Proc. ECCV*, 2020.

[6] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *NeurIPS*, 2019.

[7] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proc. ECCV*, 2018.

[8] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proc. CVPR*, 2016.

[9] Yuki M. Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*, 2020.

[10] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proc. ICML*, 2021.

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *arXiv preprint arXiv:2005.12872*, 2020.

[12] Anna Casanovas and Andrea Cavallaro. Audio-visual events for multi-camera synchronization. *Multimedia Tools and Applications*, 74:1317–1340, 2014.

[13] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vgg-sound: A large-scale audio-visual dataset. In *Proc. ICASSP*, 2020.

[14] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proc. CVPR*, 2021.

[15] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proc. ACMM*, 2020.

[16] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Proc. ACCV*, 2016.

[17] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Proc. ACCV Workshop*, 2016.

[18] Joon Son Chung and Andrew Zisserman. Signs in time: Encoding human motion as a temporal image. In *Proc. ECCV Workshop*, 2016.

[19] Joon Son Chung and Andrew Zisserman. Lip reading in profile. In *Proc. BMVC.*, 2017.

[20] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *Proc. BMVC.*, 2017.

[21] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proc. CVPR*, 2017.

[22] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *Proc. ICASSP*, 2019.

[23] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *ACL*, 2019.

[24] Vansh Dassani, Jon Bird, and Dave Cliff. Automated composition of picture-synched music soundtracks for movies. In *Proc. CVMP*, 2019.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, 2019.

[26] 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. In *Proc. ICLR*, 2021.

[27] Joshua P. Ebeneze, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Zongyi Liu. Detection of audio-video synchronization errors via event detection. In *Proc. ICASSP*, 2021.

[28] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proc. ICCV*, 2019.

[29] Chuang Gan, Deng Huang, Peihao Chen, Joshua B. Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Proc. ECCV*, 2020.

[30] Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proc. CVPR*, 2020.

[31] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *Proc. CVPR*, 2019.

[32] Ruohan Gao, Rogério Schmidt Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proc. ECCV*, 2018.

[33] Anmol Gulati, Chung-Cheng Chiu, James Qin, Jiahui Yu, Niki Parmar, Ruoming Pang, Shibo Wang, Wei Han, Yonghui Wu, Yu Zhang, and Zhengdong Zhang. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. INTERSPEECH*, 2020.

[34] Tavi Halperin, A. Ephrat, and Shmuel Peleg. Dynamic temporal alignment of speech to lips. In *Proc. ICASSP*, 2019.

[35] John R. Hershey and Javier R. Movellan. Audio-vision: Locating sounds via audio-visual synchrony. In *NeurIPS*, 1999.

[36] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*, 2021.

[37] Naji Khosravan, Shervin Ardeshir, and Rohit Puri. On attention modules for audio-visual synchronization. In *Proc. CVPR Workshop*, 2019.

[38] You Jin Kim, Hee-Soo Heo, Soo-Whan Chung, and Bong-Jin Lee. End-to-end lip synchronisation based on pattern classification. In *Proc. SLT Workshop*, pages 598–605, 2021.

[39] Sophia Koepke, Olivia Wiles, Yael Moses, and Andrew Zisserman. Sight to sound: An end-to-end approach for visual piano transcription. In *Proc. ICASSP*, 2020.

[40] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018.

[41] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. In *Proc. ICLR*, 2021.

[42] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proc. ACCV*, 2020.

[43] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *Proc. ICASSP*, 2019.

[44] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A large-scale speaker identification dataset. In *Proc. INTERSPEECH*, 2017.

[45] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. *ECCV*, 2018.

[46] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *NeurIPS*, 2021.

[47] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proc. ECCV*, 2018.

[48] Mandela Patrick, Yuki M. Asano, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv*, abs/2003.04298, 2020.

[49] Renukananda Prajwal, Kondajji, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C V Jawahar. Towards automatic face-to-face translation. In *Proc. ACMM*, 2019.

[50] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proc. ECCV*, 2020.

[51] Prarthana Shrestha, Mauro Barbieri, Hans Weda, and Dragan Sekulovski. Synchronization of multiple camera videos using audio-visual features. *IEEE Transactions on Multimedia*, 2010.

[52] Malcolm Slaney, Michele Covell, and Facesync Is. Facesync:a linear operator for measuring synchronization of video facial images and audio tracks. In *NeurIPS*, 2000.

[53] Nicolas Staelens, Jonas De Meulenaere, Lizzy Bleumers, Glenn Van Wallendael, Jan De Cock, Koen Geeraert, Nick Vercammen, Wendy Van den Broeck, Brecht Vermeulen, Rik Van de Walle, et al. Assessing the importance of audio/video synchronization for simultaneous translation of video sequences. *Multimedia systems*, 18(6):445–457, 2012.

[54] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proc. ECCV*, 2018.

[55] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proc. ECCV*, 2020.

[56] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Dan Ellis, and John R. Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In *Proc. ICLR*, 2021.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[58] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *IJCV*, 2019.

[59] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Péter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv*, abs/2006.03677, 2020.

[60] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[61] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *Proc. ACMM*, 2020.

[62] Yi Yang, Brendan Shillingford, Yannis Assael, Miaosen Wang, Wendi Liu, Yutian Chen, Yu Zhang, Eren Sezener, Luis C Cobo, Misha Denil, et al. Large-scale multilingual audio visual dubbing. *arXiv preprint arXiv:2011.03530*, 2020.

[63] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: Improving end-to-end object detector with dense prior. *arXiv preprint arXiv*, abs/2104.01318, 2021.

[64] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proc. ECCV*, 2018.

[65] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proc. ICCV*, 2019.