Exploring Fine-Grained Audiovisual Categorization with the SSW60 Dataset

Exploring Fine-Grained Audiovisual Categorization with the SSW60 Dataset

Grant Van Horn[1⋆] Rui Qian[1⋆] Kimberly Wilber[2]
Hartwig Adam[2] Oisin Mac Aodha[3] Serge Belongie[4]

[1]Cornell University [2]Google
[3]University of Edinburgh [4]University of Copenhagen

Abstract. We present a new benchmark dataset, Sapsucker Woods 60 (SSW60), for advancing research on audiovisual fine-grained categorization. While our community has made great strides in fine-grained visual categorization on images, the counterparts in audio and video fine-grained categorization are relatively unexplored. To encourage advancements in this space, we have carefully constructed the SSW60 dataset to enable researchers to experiment with classifying the same set of categories in three different modalities: images, audio, and video. The dataset covers 60 species of birds and is comprised of images from existing datasets, and brand new, expert curated audio and video datasets. We thoroughly benchmark audiovisual classification performance and modality fusion experiments through the use of state-of-the-art transformer methods. Our findings show that performance of audiovisual fusion methods is better than using exclusively image or audio based methods for the task of video classification. We also present interesting modality transfer experiments, enabled by the unique construction of SSW60 to encompass three different modalities. We hope the SSW60 dataset and accompanying baselines spur research in this fascinating area.

Keywords: multi-modal learning, fine-grained, audio, video

1 Introduction

Image-based fine-grained visual categorization (FGVC) of natural world categories has seen impressive performance gains over the last decade of research. This progression has been fueled by both larger datasets and improved techniques for classification. For example, consider the domain of bird species classification. The popular CUB200 [82] dataset (covering 200 classes of birds, each with 30 train and 30 test images) has seen top-1 accuracy improve from 10.3% [82] to over 91.7% [30]. This dataset motivated the construction of the larger and better curated NABirds [78] dataset (covering 400 species of birds, each with 60 train and 60 test images), which subsequently gave rise to the larger iNaturalist competition datasets [2]. The latest dataset in this series has 1,486 species of birds, most with 300 training examples, and the winners of the 2021 iNaturalist

⋆ The first two authors contributed equally. `https://github.com/visipedia/ssw60`

Fig. 1: **Why audiovisual?** Left: American Crow and Common Raven are visually confusing but aurally distinguishable, as illustrated by the spectrograms. Right: Yellow Warbler and Chestnut-sided Warbler are aurally confusing but visually distinguishable. Individually, audio and visual modalities have both advantages and disadvantages. We present the Sapsucker Woods 60 dataset (SSW60), a new dataset to facilitate work in fine-grained audiovisual categorization.

competition [2] achieved 94% top-1 accuracy on these species (using geographic location information). The release of the CUB200 dataset was a catalyst for FGVC research, motivating the construction of improved datasets as well as providing the means to benchmark progress. But what about the challenge of fine-grained categorization (FGC) in modalities besides images?

Audio and video modalities receive less attention than images for the task of fine-grained categorization. What opportunities and challenges do these modalities present (see Fig. 1)? More importantly, which existing datasets allow us to study cross-modality performance and where do they fall short? Large-scale audiovisual datasets such as AudioSet [26] and VGGSound [14], provide a class hierarchy more akin to coarse grained categories as perceived by humans than fine-grained categories as typically used in the context of FGC (with classes such as "Chirp, tweet" and "Hoot" for bird vocalizations in AudioSet).

There are a few existing bird video datasets [24,67,91], each focused primarily on benchmarking the performance of video frame classification, as opposed to cross-modality or audiovisual analysis. The YouTube-Birds dataset [91] almost checks all the boxes, except upon close inspection we find multiple inconveniences, e.g. it consists of a collection of YouTube links, of which at least 7% are broken at the time of writing, it contains labelling errors (typical for fine-grained datasets curated by non-experts), and the videos are not trimmed to the content of interest and are thus long and unwieldy. Both VB100 [24] and IBC127 [67] sampled their videos from higher quality data sources, but they each lack the full complement of unpaired audio and image modalities that we require for exploring audiovisual categorization. Finally, none of the prior art show the utility of audiovisual fusion methods for FGC.

In this paper, we aim to fill this dataset gap and open up new avenues of research in FGC. Our new dataset, SSW60, spans 60 species of birds that all occur in a specific geographic location: Sapsucker Woods in Ithaca, New York, (unlike the random collection of species present in the existing video datasets [24,67,91]). SSW60 contains a new collection of expert curated ten-second video clips for each species, totaling 5,400 video clips. SSW60 also contains an "unpaired" expert cu-

rated set of ten-second audio recordings for the same set of species, totaling 3,861 audio recordings. Finally, we also collate image data for the same species from the existing expert curated NABirds dataset [78] and the citizen science collected iNat2021 dataset [79].

With this new dataset in hand, we perform a thorough investigation of audiovisual classification performance. Our baseline methods utilize state-of-the-art backbones trained on visual and audio modalities. We experiment with several different fusion methods to combine information from both modalities and make audiovisual informed classifications. These experiments reveal that audiovisual methods outperform their respective single modality counterparts, advancing the state of the art for fine-grained bird species classification. As SSW60 contains images and unpaired audio examples, we conduct additional experiments to investigate the utility of pretraining on these individual modalities prior to working with video. We identify several insights from these experiments, including the unexpected negative impact of pretraining on high quality images, and the high utility of pretraining on unpaired audio samples.

In summary, we make the following contributions: 1) A new fine-grained dataset that contains expert curated video and audio data for a shared set of object categories. 2) A detailed analysis of cross-modality learning in the context of fine-grained object categories, as well as benchmark results for fine-grained audiovisual categorization.

2 Related Work

2.1 Image, Audio, and Video Datasets

Fine-Grained Image Datasets. The most commonly used classification datasets in computer vision predominantly deal with coarse-grained object reasoning, e.g. [66,90,41,20,49,29]. In contrast, fine-grained datasets contain subordinate categories that can be much more challenging for non-expert human annotators to discriminate. There are many fine-grained datasets spanning a wide range of visual concepts including airplanes [54,81], automobiles [43,50,87,25], dogs [40,61,51], fashion [35], plants [58,59,45], food [11,34], and the natural world [80,79], to name a few.

Datasets featuring images of different species of birds have been particularly popular in the vision community [82,10,78,42]. As a taxonomic group they present an interesting set of challenges that make them well suited for benchmarking advances in vision. For example, their appearance can differ based on life stage or sex, their shape can vary significantly, and some species can be very challenging for even expert humans to tell apart. Inspired by this, we propose a new multi-modal bird dataset that contains data from three different modalities: images, audio, and video.

Fine-Grained Video Datasets. The most commonly used video action recognition datasets also tend to focus on coarse-grained concepts [44,72,37,12,38,18], with some emphasizing temporal reasoning [28,55,69]. Fewer fine-grained datasets

exist, but those that do cover concepts such as sports [48,63,70] and cars [91,5]. Most relevant to this work are the small number of existing video datasets containing birds [24,67,91], see Table 1 for an overview. IBC127 [67] contains 8,014 videos across 127 bird categories. In the paper, experiments are performed for bird (127 classes) and action (4 classes) classification from video. VB100 [24] contains 1,416 videos from 100 bird species and evaluates on the task of species classification from video. While a small number of audio files are also available, no experiments are actually performed using this data. Finally, YouTube-Birds [91] contains 18,350 videos spanning the same 200 classes represented in the CUB200 image dataset [82]. The exact same set of videos are also used in [32]. Experiments are performed on the task of bird classification from video, and they show that their approach gives a minor performance improvement compared to simple baselines[83] which do not use any temporal information. The YouTube-Birds data is provided as a list of YouTube video URLs, and at the time of writing only 17,031 videos are still publicly available.

While these existing fine-grained datasets are very related to our work, they stop short of performing any cross-modal experiments, and do not show the benefit of audiovisual fusion methods for fine-grained categorization. Further, the distribution of data in these datasets is highly skewed and the included species were obviously dictated by data availability from web scraping. The SSW60 dataset provides a nearly uniform data distribution for a set geo-spatially co-located species. See the supplementary material for additional details.

Fine-Grained Audio Datasets. There are numerous examples of human speech focused [23,64], coarse-grained audio classification [68,62,22,26], and binary sound event [53,73] datasets. However, in contrast to images, there are fewer established datasets for fine-grained audio classification. One task that is highly representative of a fine-grained audio challenge is that of species identification. As a result, there exists a number of audio datasets focused on species identification. Examples include bird [52,56,32,17,16] and bat [89,65] species classification. Like their image counterparts, these datasets can be challenging to collect and accurately annotate [8]. These annotation issues can also be compounded by factors such as background noise and low quality recordings. The audio recordings in the SSW60 dataset have been manually vetted by domain experts to ensure that the labels are reliable.

Audiovisual Datasets. In addition to visual content, video data can also contain rich and descriptive audio information. For some fine-grained concepts, this information can be highly complementary to the visual cues, see Fig. 1. Inspired by these types of relationships, the vision community has developed several benchmarks to facilitate the exploration of multi-modal reasoning. Several different approaches have been used to construct these types of datasets.

The most basic approach is to query video media websites with keywords of interest, with the assumption that relevant sound events will also be present. This is the approach taken by the Flickr-SoundNet dataset [7], which contains 2M video clips with audio downloaded from Flickr, and was queried using tags from YFCC100M [74]. An alternative approach is to use automatic filtering,

Table 1: Overview of existing bird datasets. °Only 17,031 videos are currently available online. ⋆Contains the same images as [82]. †Contains the same videos as [91]. ‡Only spectrogram images are available, no audio files are included.

dataset	classes	images	videos	audio
CUB200 [82]	200	11,788	-	-
NABirds [78]	555	48,562	-	-
VB100 [24]	100	-	1,416	502
IBC127 [67]	127	-	8,014	-
YouTube-Birds [91]	200	11,788⋆	18,350°	-
PKU FG-XMedia [32]	200	11,788⋆	18,350†	12,000‡
Ours	60	31,221	5,400	3,861

e.g. by making use of image or audio classification models. VGG-Sound [14] consists of 200k, ten-second video clips from 300 different audio classes. The object that emits each sound is visible in the video clip, however, each clip is only labeled with one class even though multiple audio-visual events can be present. ACAV100M [46] contains 100M ten-second clips and was constructed using an automatic curation pipeline that maximized the mutual information between the audio and visual channels. The final dataset construction approach is to manually annotate some or all of the data. Kinetics-Sounds [6] features 19k, ten-second, audio-visual clips covering 34 human orientated action classes. The videos are a subset of the Kinetics dataset [38], which were manually filtered to ensure the presence of the actions of interest. More detailed annotations include localizing sound events in time or space. AVE (Audio-Visual Event) [76] is a subset of the AudioSet dataset [26], and contains 4,143 ten-second videos covering 28 event categories with manually labeled temporal event boundaries. Each video contains at least one two-second long audio-visual event. The LLP dataset [75] contains 11,849 YouTube video clips with 25 event categories labeled. The goal of the dataset is audio-visual parsing, i.e. deciding whether an event is audible, visible, or both. Manual temporal event annotations are provided for a subset of the videos. Finally, [13] adds image bounding boxes to audible sound sources for 5k videos in VGG-Sound [14]. In addition, the community has been working on audiovisual datasets for violence detection[84], as well as VQA [88,47]

None of the above datasets explore the problem of fine-grained audiovisual reasoning. In this work, we make use of high quality image and audio classifiers in order to select video clips that are highly likely to contain the discriminative audiovisual events for a set of 60 bird species.

2.2 Multi-modal Learning

Audiovisual Fusion. There is large and growing literature on multi-modal fusion for audiovisual understanding. Early methods adopted straightforward early or score fusion strategies, e.g. [15]. Subsequent research applied modality-specific networks with learning-driven information combinations in mid or late stage fusion strategies. Representative methods include activation summations [39], lat-

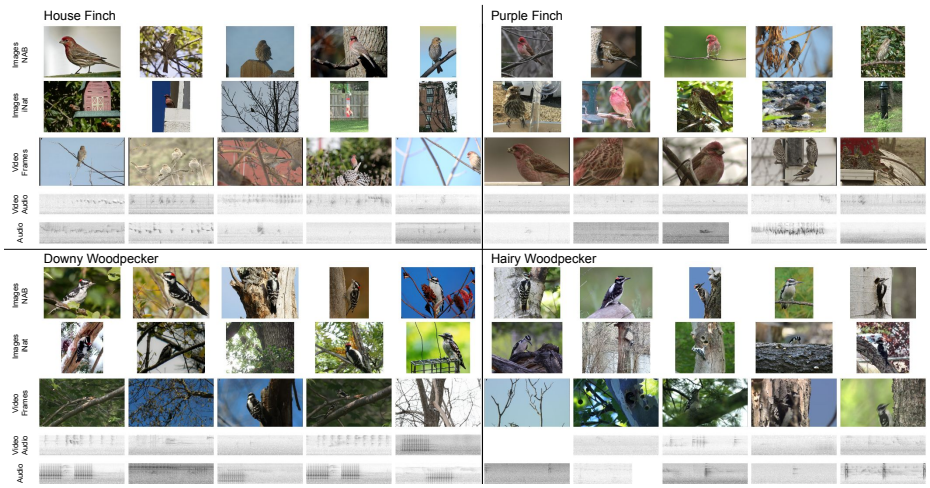


Fig. 2: Visual and audio examples (frames and spectrograms) for example bird species from the SSW60 dataset. Clockwise from top left: House Finch, Purple Finch, Hairy Woodpecker, and Downy Woodpecker. For each species, the five rows show modality samples from: (1) "Images NAB" - from NABirds [78]; (2) "Images iNat" - from iNaturalist2021 [79]; (3) "Video Frames" - center frames; (4) "Video Audio" - spectrogram that covers the three seconds of audio near the center frame; and (5) "Audio" - spectrogram generated from three seconds near the center of the file.

eral connections [85], attention based re-weighting [21], among others. A comprehensive review can be found in [9]. The recent success of adopting transformer architectures in the vision [19] and audio [27] communities empowered more advanced audiovisual fusion methods. A representative state-of-the-art work [57] carefully studied audiovisual fusion with transformers and we adopt this as our primary baseline.

Cross-Modal Analysis. [36] defined and measured the impact of several different domain shift factors in the context of training object detectors on video frames and images. These factors included the accuracy of the training bounding boxes, appearance diversity, image quality, and object size. They showed that these factors, in combination, are almost completely responsible for the performance difference as compared to training and testing on the same domain. Their conclusion was that if one wants to achieve the best performance they should train and test on the same domain. In this work, we analyse domain differences arising from depictions of the same concept (i.e. fine-grained bird categories) across different modalities.

3 SSW60 Dataset

In this section we describe the SSW60 dataset and the steps taken to construct it. The dataset is built around 60 species of birds that have a high propensity to be seen or heard on a live "feeder-cam" (i.e., a static camera monitoring a bird feeder) that is continuously recording in Ithaca, New York. These species, therefore, represent a realistic fine-grained challenge experienced by humans (unlike, e.g., CUB200 where the categories of birds come from all over the world). A model that can recognize these species and interpret their behaviors will be particularly relevant in assisting biologists with analyzing large collections of video footage from these cameras. We plan for future versions of this dataset to directly incorporate video from the live cams. For each of the 60 species we sampled data from three different modalities: videos (containing paired frames and audio), audio recordings, and images. Here videos are unique as they contain both visual and audio modalities, while the additional unpaired audio recordings and image datasets only consist of one modality respectively. See Fig. 2 for examples from the various modalities and Table 2 for per-modality statistics.

Video. The videos in SSW60 come from recordings archived at the Macaulay Library at the Cornell Lab of Ornithology [3]. These videos are contributed by professional and enthusiast videographers from the around the world, and can range in duration from a few seconds to multiple minutes. The camera view points are not fixed and can move in order to track the bird as it moves through the environment. Each video is associated with a particular "target species" that is known to be present in the video. For each video we isolated a ten-second clip where the task of species classification is particularly relevant. To accomplish this, we applied the following procedure: 1) For each of the 60 species of birds, we sampled all of their respective videos in the Macaulay Library. 2) We then used an image based bird detector and classifier (trained on the 30M+ images from the Macaulay Library) to identify the sections of video where the target species was present. This gave us candidate video sections to extract ten-second clips. 3) To further refine the candidate clips, we ran the Merlin Sound ID model [4], a high performing acoustic bird classification model, across the audio tracks of the candidate clips to determine if the target species was vocalizing. 4) For each video, we keep the clip with the highest likelihood of the target species vocalizing. 5) Finally, for each species, we select 90 video clips for the dataset, where each clip comes from a unique video.

We found that most videos in the Macaulay Library do not have complete metadata indicating the exact recording time and date. We therefore split the video files into train and test sets by splitting on the videographers. All of the videos from a particular videographer are either in the train split or the test split. We found that this tactic was necessary to prevent multiple, highly similar videos uploaded by the same videographer winding up in both the train and test sets (a problem found in existing datasets [24,67]). All videos are converted to a frame rate of 25FPS. This modality is referred to as "Video Frames" in the experiment section when considering only the frames of the videos, and is referred to as "Video Audio" when considering only the audio channel.

Table 2: Summary of the train/test split sizes for each modality in SSW60, along with information about the number of examples per class.

	source	total	min	max	median
Images NAB	[78]	5050, 5171	30, 31	221, 214	60, 60
Images iNat	[79]	18000, 3000	300, 50	300, 50	300, 50
Audio	ours	2597, 1264	28, 12	52, 30	45, 21
Video	ours	3462, 1938	38, 22	68, 52	59, 31

Note that some of the ten-second clips in SSW60 do not have the target species vocalizing; we do not treat this as a problem but view it as a challenge and inherent property of "in-the-wild" video. The process of a human uploading a video (as opposed to an audio recording) to the Macaulay Library, means that the videos in SSW60 will be biased towards visually relevant information for classification, as opposed to aurally relevant information. Using an acoustic classifier to find those sections of video with both visual and aurally relevant information helps mitigate this, but does not completely remove the visual bias.

Audio. All 60 bird species in SSW60 have unpaired audio recordings from the Macaulay Library. These recordings are unpaired in the sense that they do not have any associated visual data, i.e. no videos or images. Each audio recording is annotated with a particular "target species" that is known to be vocalizing in the file. However, it is not specified at what moment in time the target species is vocalizing, and recordings can be multiple minutes long. We sampled audio recordings for each of our 60 species and had an expert ornithologist provide temporal onset and offset annotations for the target species. We then trimmed the audio files to ten-second clips that contain the target species' vocalization. The result is an expert curated audio dataset for each of the 60 species in SSW60. The audio files are stored in WAV format at a sampling rate of 22.05kHz.

Audio is split between train and test sets by ensuring that audio files from the same recording session are placed in the same split. This prevents models from exploiting common background noise that might be heard across multiple recordings from the same location and time. This modality is referred to as "Audio" in the experiment section.

Images. Finally, we also preform experiments with images from two existing datasets: NABirds [78] and iNat2021 [79]. The 60 species in SSW60 conveniently overlap with the species in these existing datasets, and we incorporate all images available into SSW60 while maintaining the original train/test splits. For the NABirds dataset, we merged the respective "visual categories" that comprise each species. The images in NABirds are of particularly high quality, representing a best case scenario for visual classification (i.e. someone using high quality camera equipment to carefully compose a photograph for the goal of visual identification). The images in iNat2021 are more mixed in terms of quality and therefore represent a more difficult visual classification task. See Fig. 2 for sample images from both datasets. These modalities are referred to as "Images NAB" and "Images iNat" in the experiment section.

4 Methods

We are interested in exploring fine-grained categorization in two areas: **cross-modal analysis** and **audiovisual fusion**.

For **cross-modal analysis**, we assume a fixed backbone architecture that can be utilized for processing data from multiple modalities. For the audio modality, we convert the waveforms to spectrogram images. For videos, we adopt TSN [83] style methods using 2D image backbones to encode features and perform fusion on top of them. Our experimental procedure is straightforward: we train the backbone model using a particular training modality (see Sec 3 for the options) and then evaluate the performance on an evaluation modality directly. As we have the same species in each modality, the trained backbone can be used directly on the evaluation modality. However, there is a domain transfer problem to consider (i.e. moving from images to video frames), so we also evaluate the trained backbone by first fine-tuning the weights using the training split of the evaluation modality, and then evaluate on that evaluation modality. We use top-1 accuracy as the evaluation criteria for all experiments. Unlike existing bird video datasets [24,67,91], our evaluation splits are uniform for each species, which makes top-1 accuracy across examples an unbiased assessment of performance. All backbone models are trained using softmax cross-entropy.

In addition to cross-modality analysis, we study fine-grained **audiovisual fusion** using the paired "Video Frames" and "Video Audio" data in SSW60. We adopt a transformer-based backbone and experiment with mid-fusion through the state-of-the-art multimodal bottleneck fusion approach of [57], as well as late and score-fusion. Thanks to the image and audio recordings provided in SSW60, we are able to study the effect of different pretraining dataset choices (e.g. ImageNet, Images iNat, and Images NAB) on audiovisual fusion.

4.1 Implementation Details

Image Modality. We adopt the standard ImageNet [66] training paradigm. During training, we randomly crop and resize a square portion of the image to 224×224 pixels, followed by a random flip augmentation; during evaluation, the shorter edge of the image is resized to 256 pixels first, and then a center crop of 224×224 is extracted for classification. We perform evaluation using a CNN-based ResNet50 [31] and transformer-based ViT-B [19] for experiments on the image modality, as they are the most popular choices in the image recognition community. Both are initialized with ImageNet pretrained weights.

Audio Modality. For audio processing, we convert the audio waveforms into spectrogram images. Concretely, the raw audio signal is resampled to a rate of 16kHz. We then apply the short-time Fourier transform algorithm using a window size of 512 and a stride length of 128. The frequency values are then transformed using the "mel-scale" with 128 bins. Finally we convert the magnitude values to decibel units and normalize to generate the final spectrogram image. This image is duplicated three times to create the RGB input for the

network. For a 10-second long audio clip, the shape of the generated spectrogram image is approximately 128×1250, where 128 is the number of mel-scaled frequency bands and 1250 is the temporal span. During training, we utilize two augmentations to avoid overfitting: time cropping and frequency masking [60]. For time cropping, we randomly sample a window of length 400 time bins (spanning all 128 frequency bands) from the original spectrogram image (400 time bins corresponds to approximately 3 seconds of audio). For frequency masking, we randomly mask out 15 consecutive frequency bands. During evaluation, we densely sample five windows of length 400 time bins (spanning all 128 frequency bins) from the original spectrogram images using a stride of 150. We average the logits across the 5 windows to use as the final prediction. Earlier work [33] used VGG-style [71] backbones which we also compare to for completeness.

Video Frame Modality. We adopt the segment sampling strategy of TSN [83] where we first divide the video clip into eight uniform segments. During training we randomly sample one frame from each segment, while for evaluation we select the center frame of each segment. Each of the eight selected frames is passed through the 2D ResNet50 or ViT-B backbone for feature extraction using images of size 224×224 pixels. We then average the eight feature vectors to generate the final feature representation for the video clip. This global feature is then passed through a fully connected layer to produce a vector of logits. We initially conducted experiments with video-specific 3D convolution networks with dense frame sampling using S3D [86]. However we found that S3D (pretrained on Kinetics-400 [38]) performed worse than our TSN-style baselines (pretrained on ImageNet) for SSW60. Furthermore, using a 2D network like ResNet50 or ViT-B as the backbone provides the flexibility for easily studying feature transfer between video frames and images. We leave experimentation with more sophisticated video backbones for future work.

Audiovisual Fusion. We briefly recap the transformer architecture and then describe how we conduct audiovisual fusion experiments. Given an input image or audio spectrogram, it is first divided into non-overlapping patches. Each patch is projected to a token using a linear layer and a special learnable classification token is added. More details can be find in the original ViT paper [19]. After tokenization, the tokens are passed through a stack of transformer layers. We denote the input of the l-th layer as $\mathbf{z}^l$, which results in $\mathbf{z}^{l+1} = trans\_layer_l(\mathbf{z}^l)$. The computation inside $trans\_layer_l$ can be written as

$$\mathbf{y}^l = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}^l)) + \mathbf{z}^l, \tag{1}$$

$$\mathbf{z}^{l+1} = \mathrm{MLP}(\mathrm{LN}(\mathbf{y}^l)) + \mathbf{y}^l, \tag{2}$$

where LN denotes layer normalization, MSA denotes multi-head self-attention. In our audiovisual fusion, we use two identical $L = 12$ layer transformers to take the visual and audio input separately. The forward process for the visual modality is thus $\mathbf{z}_v^{l+1} = v\_trans\_layer_l(\mathbf{z}_v^l)$, and $\mathbf{z}_a^{l+1} = a\_trans\_layer_l(\mathbf{z}_a^l)$ for the audio modality.

For mid-fusion, we use the state-of-the-art multimodal bottleneck transformer [57]. Here a set of learnable tokens $\mathbf{z}_b$ are used as the fusion bottleneck

$$[\mathbf{z}_v^{l+1}||\hat{\mathbf{z}}_b] = v\_trans\_layer_l([\mathbf{z}_v^l||\mathbf{z}_b]), \tag{3}$$

$$[\mathbf{z}_a^{l+1}||\mathbf{z}_b] = a\_trans\_layer_l([\mathbf{z}_a^l||\hat{\mathbf{z}}_b]). \tag{4}$$

$[.||.]$ denotes the concatenation of tokens. In each layer, $\mathbf{z}_b$ first interacts with the visual tokens $\mathbf{z}_v^l$ and gets updated to $\hat{\mathbf{z}}_b$. Then $\hat{\mathbf{z}}_b$ interacts with the audio tokens $\mathbf{z}_a^l$ to finish the audio visual fusion. Following [57], we conduct this fusion in the last four layers of the transformer.

For late fusion, we concatenate the class tokens of both modalities after the last transformer block, written as $[\mathbf{z}_v^{L+1}[0]||\mathbf{z}_a^{L+1}[0]]$ and apply a linear classifier on top of it. For score fusion, we take the predictions from both modalities and use a weighted sum of the combined final predictions. In practice, we use a weight of 0.5 for both the visual and audio modalities.

5 Experiments

We first perform cross-modal experiments on the video and audio modalities separately, and then explore multi-modal fusion for audiovisual categorization.

Visual Modality Categorization. Here we benchmark the performance achieved on the Video Frames of SSW60. Table 3 (Left) shows the top-1 accuracy on the Video Frame test set when using a ResNet50 backbone trained on either the Images iNat, Images NAB, or the Video Frames training datasets. We split the results depending on whether we fine-tune (FT) the trained backbone on the Video Frames training dataset. Training directly on the Video Frames dataset achieves a top-1 accuracy of 54.92%. Interestingly, we see that evaluating the Images iNat model directly on the Video Frames achieves an even higher top-1 accuracy of 60.47%. This is further improved to 71.88% when fine-tuning on the Video Frames train split. We compare these numbers to those achieved by a model trained on the Images NAB dataset: 24.05% and 56.55%, top-1 accuracy respectively. The Images iNat dataset has more training samples than Images NAB, however, the images in the NABirds dataset are aesthetically higher quality (see Section 3). These results seem to indicate that performance on "in-the-wild" videos benefits more from "lower quality" training images.

Table 3: Top-1 accuracy on SSW60 Video-Frames using a ResNet50 backbone (left) and SSW60 Video-Audio using a ResNet18 backbone (right) when training on different datasets (columns). Results are presented with and without finetuning (FT) on the respective video modality.

Cross-Modal - Video Frames			
FT	Images iNat	Images NAB	Video Frames
	60.47	24.05	54.92
✓	71.88	56.55	-

Cross-Modal - Video Audio		
FT	Unpair Audio	Video Audio
	24.41	10.37
✓	15.33	-

Table 4: Comparison of audio backbones trained and tested on the unpaired audio modality in SSW60. All models are initialized from ImageNet pretrained weights. '↑384' indicates that a model is fine-tuned on ImageNet with a higher resolution [77] and 'AS' is further fine-trained on AudioSet [27] before use.

Backbone	VGG16	VGG19	ResNet18	ResNet50	ViT-B	ViT-B↑384	ViT-B↑384 AS
Top 1 Acc	52.1%	56.1%	59.01%	63.7%	66.8%	65.9%	67.4%

Audio Modality Categorization. We next benchmark the new unpaired audio dataset component of SSW60. Table 4 contains the results of these experiments. We trained and evaluated VGG16 and 19 [71], ResNet18 and 50 [31], and the transformer-based ViT-B [19] architectures. As expected, we see a progression of top-1 accuracy as we move from older architectures (52.1% for VGG16) to the latest architectures (66.8% for ViT-B). We attempted to push accuracy further by using a ViT-B model pretrained on a higher resolution image input (224 vs 384), but we actually see performance decrease to 65.9%. However, if we take this higher resolution model and add an additional pretraining step of training on AudioSet [26] then we achieve a top-1 accuracy of 67.4%.

We now benchmark the Video Audio component of SSW60. For these experiments we chose a ResNet18 backbone for convenience, but expect a more powerful backbone to be slightly more performant (see Table 4 and Table 5 (Direct Eval)). The obvious result is the low performance achieved when training exclusively with the Video Audio data, achieving a top-1 accuracy of just 10.37%. Directly using a model trained on the unpaired audio achieves 24.41%, a significant improvement. Interestingly, fine-tuning the unpaired audio model on the Video Audio training samples leads to a decrease in performance, down to 15.33%. This points to a recurring theme: video is biased to visual features (simply by the nature through which it was collected), and while it contains an audio channel, the ability to use the audio channel for classification appears to be difficult. We show in the next section however that it is possible to improve overall classification accuracy by incorporating audio.

Audiovisual Fine-Grained Categorization. In contrast to the rich literature of audiovisual fusion on coarse-grained video datasets, audiovisual fine-grained categorization (FGC) remains under-explored due to a lack of appropriate datasets. SSW60 fills this gap and allows us to conduct a comprehensive analysis that explores the impacts of various pretraining and fusion methods on audiovisual FGC. We follow the paradigm employed by Nagragni et al. [57] and use two uni-modal models to process the audio and visual modalities separately. We adopt the ViT-B [19] backbone for both modalities (see Section 4.1). We are interested in two research questions: 1) What is the effect of different fusion methods? and 2) What is the effect of different pretraining datasets? For fusion methods, we use the state-of-the-art MBT [57] as the mid-fusion algorithm, and compare to late and score fusion techniques. For pretraining datasets, we utilize ImageNet, Images NAB, and Images iNat for the visual modality, and ImageNet and unpaired audio recordings for the audio modality. By construction, once a

Table 5: Audiovisual fine-grained categorization results on SSW60 videos using ViT models. We split results into two different scenarios: evaluation on the modalities individually ("Direct Eval" and "No Fusion") and on both modalities together ("Fusion"). All numbers reflect top-1 accuracy. "Direct Eval" means we can conduct direct evaluation for modalities pretrained on datasets with the same 60 species. "No Fusion" means we take a pretrained network and fine-tune it on the respective modality from the SSW60 training videos. For the Mid and Late fusion algorithms in "Fusion", we initialize the model with pretrained weights from individual models trained on the "Pretrain" datasets. For Score fusion, we take the best individual model for each modality (considering both "Direct Eval" and "No Fusion" variants) and fuse their scores by a weighted sum.

Pretrain		Direct Eval		No Fusion		Fusion		
Visual	Audio	Vid Frames	Vid Audio	Vid Frames	Vid Audio	Mid	Late	Score
ImageNet	ImageNet	-	-	59.0%	14.3%	54.3%	**59.8**%	58.9%
ImageNet	Unpair Audio	-	28.3%	59.0%	30.4%	62.0%	62.5%	**63.5**%
Images NAB	Unpair Audio	60.0%	28.3%	64.4%	30.4%	67.5%	**68.4**%	68.2%
Images iNat	Unpair Audio	78.0%	28.3%	76.2%	30.4%	73.5%	78.3%	**80.6**%

backbone has been trained on Images NAB, Images iNat, or the unpaired audio, we are able to directly evaluate on the corresponding modality of the SSW60 video dataset, since all datasets share the same 60 species. Our results are summarized in Table 5. We also provide per-class analysis between uni-modal and audiovisual fusion performance in Fig. 3. Our best result on SSW60 (80.6% top-1 accuracy) comes from fusing the scores of a visual model pretrained on Images iNat (and **not** fine-tuned on the SSW60 video frames), and an audio model pretrained on the unpaired audio and further fine-tuned on the audio channels of the SSW60 videos.

We highlight three conclusions from our audiovisual fusion investigations. **First, the best result from audiovisual fusion is always better than training on each modality separately.** For each row in Table 5, the highest top-1 accuracy is always in the Fusion column, meaning that combining information from both modalities is always better than using a single modality. This finding aligns well with our motivation of audiovisual fusion in Fig. 1. **Second, there is no "best" fusion method**. In the four different pretraining configurations, we find that late fusion works best half the time, and score fusion works best in the other half. It is interesting that the state-of-the-art mid-fusion method does not work as well as the simpler methods. We leave this as an open question for the community to explore more advanced mid-fusion methods for audiovisual FGC. **Third, pretraining on external datasets can be *very* beneficial**. We observe a $\sim 20\%$ increase in top-1 performance when fine-tuning the ImageNet backbones on Images iNat and the unpaired audio (Fusion column, row 1 vs row 4 in Table 5).

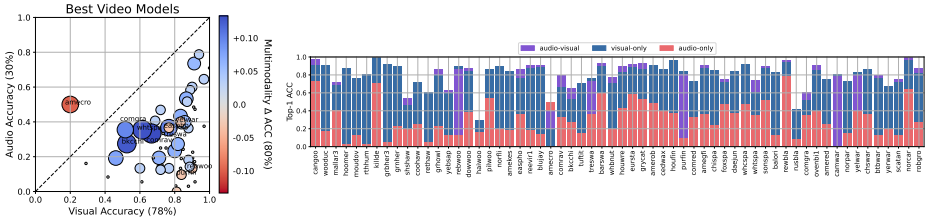


Fig. 3: Per-species audio and visual modality performance along with the resulting audiovisual performance after score fusion. These results correspond to the bottom right fusion model in Table 5. The size and color of the dots on the scatter plot indicate the resulting top-1 accuracy change (from the best uni-modal model) when fusing the predictions for audiovisual classification. Large blue dots correspond to better audiovisual accuracy. Large red dots correspond to worse audiovisual accuracy. Species with the largest positive and negative audiovisual changes have been labeled. The bars for each species in the bar plot are ordered by the modality performance. Purple bars on top reveal those species with improved audiovisual accuracy. 27 species improved, 27 species remained the same, and 6 species decreased after fusion.

6 Conclusion

We present SSW60, a new dataset for advancing fine-grained audiovisual categorization. This expert curated dataset provides researchers with the tools to explore categorization across three different modalities, enabling a comprehensive exploration of cross-modal and audiovisual fusion. Similar to how the CUB200 dataset paved the way for the larger, and better curated, NABirds and iNaturalist datasets, we envision SSW60 as a vital first step towards studying audiovisual fine-grained categorization. The availability of live "feeder-cam" video featuring the bird species in SSW60 also provides an interesting avenue for studying the deployment of trained models for real-time audiovisual categorization - an important problem for biodiversity monitoring. At its current size, SSW60 can also be used as an evaluation dataset for self-supervised audiovisual models. We envision SSW60 broadly benefiting the vision community by providing ample directions for future work on FGC and video analysis more generally.

Limitations. The size of the SSW60 dataset is a potential limitation for training models from scratch, which is why we used ImageNet pretrained models. ImageNet does contain ∼60 classes of birds, but all models started from an ImageNet pretrained backbone. The video and audio annotations in SSW60 are "weak" in the sense that they apply to the entire ten-second clip, as opposed to temporally localized annotations.

Acknowledgements. Serge Belongie is supported in part by the Pioneer Centre for AI, DNRF grant number P1. These investigations would not be possible without the help of the passionate birding community contributing their knowledge and data to the Macaulay Library; thank you!

References

1. iNaturalist, `www.inaturalist.org`, accessed Mar 7 2022
2. iNaturalist 2021 Challenge, `www.kaggle.com/c/inaturalist-2021`, accessed Mar 7 2022
3. Macaulay Library, `www.macaulaylibrary.org`, accessed Mar 7 2022
4. Merlin Sound ID, `merlin.allaboutbirds.org/sound-id`, accessed Mar 7 2022
5. Alsahafi, Y., Lemmond, D., Ventura, J., Boult, T.: Carvideos: A novel dataset for fine-grained car classification in videos. In: International Conference on Information Technology-New Generations (2019)
6. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV (2017)
7. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: NeurIPS (2016)
8. Baker, E., Vincent, S.: A deafening silence: a lack of data and reproducibility in published bioacoustics research? Biodiversity data journal (2019)
9. Bayoudh, K., Knani, R., Hamdaoui, F., Mtibaa, A.: A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. The Visual Computer (2021)
10. Berg, T., Liu, J., Lee, S.W., Alexander, M.L., Jacobs, D.W., Belhumeur, P.N.: Birdsnap: Large-scale fine-grained visual categorization of birds. In: CVPR (2014)
11. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: ECCV (2014)
12. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
13. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: CVPR (2021)
14. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020)
15. Chen, T., Rao, R.R.: Audio-visual integration in multimodal communication. Proceedings of the IEEE (1998)
16. Chronister, L., Rhinehart, T., Place, A., Kitzes, J.: An annotated set of audio recordings of eastern north american birds containing frequency, time, and species information. Ecology (2021)
17. Cramer, J., Lostanlen, V., Farnsworth, A., Salamon, J., Bello, J.P.: Chirping up the right tree: Incorporating biological taxonomies into deep bioacoustic classifiers. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020)
18. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018)
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
20. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010)
21. Fayek, H.M., Kumar, A.: Large scale audiovisual learning of sounds with weakly labeled data. arXiv:2006.01595 (2020)

22. Fonseca, E., Favory, X., Pons, J., Font, F., Serra, X.: Fsd50k: an open dataset of human-labeled sound events. arXiv:2010.00475 (2020)
23. Garofolo, J.S.: Timit acoustic phonetic continuous speech corpus. Linguistic Data Consortium (1993)
24. Ge, Z., McCool, C., Sanderson, C., Wang, P., Liu, L., Reid, I., Corke, P.: Exploiting temporal information for dcnn-based fine-grained object classification. In: International Conference on Digital Image Computing: Techniques and Applications (2016)
25. Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Fei-Fei, L.: Fine-grained car detection for visual census estimation. In: AAAI (2017)
26. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017)
27. Gong, Y., Chung, Y.A., Glass, J.: Ast: Audio spectrogram transformer. In: Interspeech (2021)
28. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: ICCV (2017)
29. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: CVPR (2019)
30. He, J., Chen, J.N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., Wang, C., Yuille, A.: Transfg: A transformer architecture for fine-grained recognition. arXiv:2103.07976 (2021)
31. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
32. He, X., Peng, Y., Xie, L.: A new benchmark and approach for fine-grained cross-media retrieval. In: International Conference on Multimedia (2019)
33. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: ICASSP (2017)
34. Hou, S., Feng, Y., Wang, Z.: Vegfru: A domain-specific dataset for fine-grained visual categorization. In: ICCV (2017)
35. Jia, M., Shi, M., Sirotenko, M., Cui, Y., Cardie, C., Hariharan, B., Adam, H., Belongie, S.: Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In: ECCV (2020)
36. Kalogeiton, V., Ferrari, V., Schmid, C.: Analysing domain shift factors between videos and images for object detection. PAMI (2016)
37. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
38. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv:1705.06950 (2017)
39. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: ICCV (2019)
40. Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L.: Novel dataset for fine-grained image categorization. In: First Workshop on Fine-Grained Visual Categorization (2011)

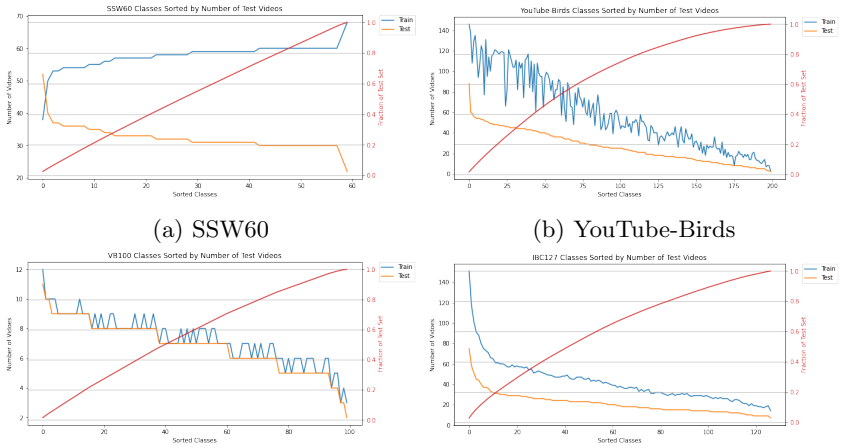
41. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Kamali, S., Malloci, M., Pont-Tuset, J., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., Murphy, K.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://storage.googleapis.com/openimages/web/index.html (2017)
42. Krause, J., Sapp, B., Howard, A., Zhou, H., Toshev, A., Duerig, T., Philbin, J., Fei-Fei, L.: The unreasonable effectiveness of noisy data for fine-grained recognition. In: ECCV (2016)
43. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: ICCV Workshops (2013)
44. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: ICCV (2011)
45. Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V.: Leafsnap: A computer vision system for automatic plant species identification. In: ECCV (2012)
46. Lee, S., Chung, J., Yu, Y., Kim, G., Breuel, T., Chechik, G., Song, Y.: Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In: ICCV (2021)
47. Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.R., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19108–19118 (2022)
48. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: ECCV (2018)
49. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
50. Lin, Y.L., Morariu, V.I., Hsu, W., Davis, L.S.: Jointly optimizing 3d model fitting and fine-grained classification. In: ECCV (2014)
51. Liu, J., Kanazawa, A., Jacobs, D., Belhumeur, P.: Dog breed classification using part localization. In: ECCV (2012)
52. Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., Bello, J.P.: Birdvox-full-night: A dataset and benchmark for avian flight call detection. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018)
53. Mac Aodha, O., Gibb, R., Barlow, K.E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G.R., Newson, S.E., et al.: Bat detective—deep learning tools for bat acoustic signal detection. PLoS computational biology (2018)
54. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv:1306.5151 (2013)
55. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al.: Moments in time dataset: one million videos for event understanding. PAMI (2019)
56. Morfi, V., Bas, Y., Pamuła, H., Glotin, H., Stowell, D.: Nips4bplus: a richly annotated birdsong audio dataset. PeerJ Computer Science (2019)
57. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. NeurIPS (2021)
58. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: CVPR (2006)
59. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics & Image Processing (2008)

60. Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V.: SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In: Interspeech (2019)
61. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: CVPR (2012)
62. Piczak, K.J.: Esc: Dataset for environmental sound classification. In: International Conference on Multimedia (2015)
63. Piergiovanni, A., Ryoo, M.S.: Fine-grained activity recognition in baseball videos. In: CVPR Workshops (2018)
64. Robinson, T., Fransen, J., Pye, D., Foote, J., Renals, S.: Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition. In: International Conference on Acoustics, Speech, and Signal Processing (1995)
65. Roemer, C., Julien, J.F., Bas, Y.: An automatic classifier of bat sonotypes around the world. Methods in Ecology and Evolution (2021)
66. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015)
67. Saito, T., Kanezaki, A., Harada, T.: Ibc127: Video dataset for fine-grained bird classification. In: International Conference on Multimedia and Expo (2016)
68. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: International Conference on Multimedia (2014)
69. Sevilla-Lara, L., Zha, S., Yan, Z., Goswami, V., Feiszli, M., Torresani, L.: Only time can tell: Discovering temporal data for temporal modeling. In: WACV (2021)
70. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: CVPR (2020)
71. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
72. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402 (2012)
73. Stowell, D., Wood, M.D., Pamuła, H., Stylianou, Y., Glotin, H.: Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. Methods in Ecology and Evolution (2019)
74. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM (2016)
75. Tian, Y., Li, D., Xu, C.: Unified multisensory perception: Weakly-supervised audio-visual video parsing. In: ECCV (2020)
76. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: ECCV (2018)
77. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021)
78. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: CVPR (2015)
79. Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., Mac Aodha, O.: Benchmarking representation learning for natural world image collections. In: CVPR (2021)
80. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: CVPR (2018)

81. Vedaldi, A., Mahendran, S., Tsogkas, S., Maji, S., Girshick, R., Kannala, J., Rahtu, E., Kokkinos, I., Blaschko, M., Weiss, D., et al.: Understanding objects in detail with fine-grained attributes. In: CVPR (2014)
82. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Technical Report, CNS-TR-2011-001 (2011)
83. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. PAMI (2018)
84. Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., Yang, Z.: Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: European conference on computer vision. pp. 322–339. Springer (2020)
85. Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C.: Audiovisual slowfast networks for video recognition. arXiv:2001.08740 (2020)
86. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 305–321 (2018)
87. Yang, L., Luo, P., Change Loy, C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: CVPR (2015)
88. Yun, H., Yu, Y., Yang, W., Lee, K., Kim, G.: Pano-avqa: Grounded audio-visual question answering on 360deg videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2031–2041 (2021)
89. Zamora-Gutierrez, V., Lopez-Gonzalez, C., MacSwiney Gonzalez, M.C., Fenton, B., Jones, G., Kalko, E.K., Puechmaille, S.J., Stathopoulos, V., Jones, K.E.: Acoustic identification of mexican bats based on taxonomic and ecological constraints on call design. Methods in Ecology and Evolution (2016)
90. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. PAMI (2017)
91. Zhu, C., Tan, X., Zhou, F., Liu, X., Yue, K., Ding, E., Ma, Y.: Fine-grained video categorization with redundancy reduction attention. In: ECCV (2018)

A Existing Bird Video Datasets

In this section we dive deeper into the existing bird video datasets [24,67,91] and discuss why they were not suitable for our investigations. See Table A1 and Table A2 for overview statistics comparing the different datasets. As mentioned in the main paper, *none* of these prior works explore cross modality or audiovisual fine-grained categorization. For reference when comparing the datasets, the distribution of train/test videos in the SSW60 dataset can be seen in Fig. A1a.



(a) SSW60 (b) YouTube-Birds

(c) VB100 (d) IBC127

Fig. A1: Train and test examples per species for various datasets. Note the (nearly) uniform train and test distributions for the SSW60 dataset compared to the other datasets.

A.1 YouTube-Birds

The YouTube-Birds dataset [91] is a collection of 18,350 videos that cover the same 200 categories as the CUB200 dataset [82]. The dataset is provided as a collection of YouTube links, with no information regarding which section of a video is relevant for classification (see Table A1 for statistics on the video duration). At the time of writing only 17,031 videos are still available (a link attrition rate of 7% over 3 years). The distribution of both the train and test videos per category is non-uniform, see Fig. A1b. In the benchmark experiments for this dataset, it is unclear whether the authors used top-1 accuracy averaged across all test videos ("micro") or if they first computed top-1 accuracy for each species and then averaged those values to get overall top-1 accuracy ("macro"). Given that the test distribution is non-uniform, "micro" accuracy would give a

very skewed sense of performance, since the 56 categories with the most train data have over 50% of the test videos.

Unlike websites organized around a particular fine-grained domain (like iNaturalist [1] or the Macaulay Library [3]), YouTube has no mechanisms to vouch for the reliability of tags or labels applied to videos (i.e. to confirm if the species labelled as being present are actually correct). Therefore the creators of YouTube-Birds had to query for videos using CUB200 category names (presumably searching the titles and descriptions for text matching the names) and then "used a crowd sourcing system to annotate the videos" [91]. No details are given describing the skill of the annotators, and it is well documented that crowd workers (e.g. those on Amazon Mechanical Turk) can provide noisy labels when annotating fine-grained data [78]. We therefore expect the error rate in YouTube birds to be at least has high as it is in the CUB200 dataset: 5% [78]. While conducting a thorough cleanup of YouTube-Birds is beyond the scope of this work, we did find particularly high error rates in those categories with few videos (e.g. only 1 / 5 videos were relevant for the "024.Red_faced_Cormorant" category, and 6 / 11 videos were relevant for the "151.Black_capped_Vireo" category).

The lack of a well defined 10 second clip also makes YouTube-Birds unwieldy for the task of classification. While some videos focused on a single individual, in others, the birds played a small role. For example, which species should a model focus on in this video: `www.youtube.com/watch?v=wiCr5Yqo5y0` - which is assigned to the '151.Black_capped_Vireo" category in the dataset? There are two different species, each in clear focus during different sections of the video, but neither are necessarily the focus of the video. In addition, large portions of the video consist of an interview with a human. The task is ambiguous for evaluation, and confusing for training. While narrowing a video down to a 10-second clip does not completely alleviate this problem, it does certainly help.

We chose not to use the YouTube-Birds dataset due to the challenges associated with downloading (potentially broken) YouTube links, the high probability of labeling errors, and the issue of untrimmed video clips. One final inconvenience of the YouTube birds dataset is that while the authors matched the categories of the CUB200 dataset, they used a different label assignment for their annotations. While just an inconvenience, it highlights that this dataset poses serious obstacles for effective analysis of cross modal performance. Our SSW60 dataset aims to alleviate many of the issues listed above, i.e. it will be distributed as a single download as opposed to a list of YouTube links, it has been curated by bird experts so the label quality is very high, and it contains 10-second video clips which focus on the bird of interest.

A.2 VB100

The VB100 dataset [24] is a collection of $1,416$ videos covering 100 bird species, with a non-uniform distribution of train and test images per species, see Fig.A1c. The authors do not provide information on the source of the videos, but upon visual inspection it is highly likely that most of these videos came from the Internet Bird Collection (IBC) website. The media on this website has since

Table A1: Video duration (in seconds) stats for existing bird video datasets. °18,350 videos originally.

dataset	classes	videos	Avg Dur	Med Dur	Min Dur	Max Dur
VB100 [24]	100	1,416	32.6	32.14	4.60	200.83
IBC127 [67]	127	8,014	31.2	28.66	3.00	266.72
YouTube-Birds [91]	200	17,031°	60.5	49.04	0.76	465.2
SSW60 (Ours)	60	5,400	9.7	9.96	2.20	9.96

Table A2: Train and test stats for existing bird video datasets, for each class. Means are rounded to the nearest tenth. °18,350 videos originally.

dataset	classes	videos	Total	Avg	Med	Min	Max
VB100 [24]	100	1,416	730, 686	7.3, 6.9	7, 7	3, 2	12, 11
IBC127 [67]	127	8,014	5343, 2671	42.1, 21.0	37, 19	14, 7	151, 75
YouTube-Birds [91]	200	17,031°	11735, 5296	58.7, 26.5	50, 25	3, 2	146, 88
SSW60 (Ours)	60	5,400	3462, 1938	57.7, 32.3	59, 31	38, 22	68, 52

been incorporated into the Macaulay Library[1]. One challenge of using media from IBC is that one has to be careful with how videos are separated into train and test splits. Many videos from IBC are actually shorter clips from a longer recording session or part of a longer original video. For example the VB100 videos corresponding to "American_Rock_Wren_00001.mp4"[2] and "American_Rock_Wren_00002.mp4"[3] are from the same recording session, but one is a test video and the other is a train video in the VB100 dataset. This leaks information across the train/test splits, providing an opportunity for models to 'cheat'. We aim to mitigate this from occurring in SSW60 by placing all the videos from a particular videographer into either the train or test split.

We chose not to use the VB100 due to its small size, random collection of species (see the IBC127 discussion below), and problems with the existing train/test splits. Also it should be noted that there are other minor issues with the dataset, e.g. the annotation files accompanying the dataset are incorrectly formatted, so that "Sandwitch_Tern" in the annotations files corresponds to the "Sandwich_Tern" directory of videos (note the typo).

A.3 IBC127

The IBC127 dataset [67] is a collection of 8,014 videos covering 127 species of birds. The videos in this dataset were originally downloaded from the Internet Bird Collection (IBC) website. As mentioned above, the media on this website has since been incorporated into the Macaulay Library. Similar to VB100, the IBC videos must be split into train and test splits carefully, so as to prevent

[1] www.macaulaylibrary.org/the-internet-bird-collection-the-macaulay-library/
[2] www.macaulaylibrary.org/asset/201760451
[3] www.macaulaylibrary.org/asset/201760441

leakage of information. In the paper the authors state that they "use 5,343 videos for learning and 2,671 videos for testing" [67], however these splits are not included with the dataset. It is unclear whether the authors attempted to maintain a uniform or non-uniform test set for each species. The dataset also does not provide user IDs for the videos, so we are unable to ensure that we create reliable train/test splits. We assume the authors used a non-uniform test split (because the numbers easily match those provided by the authors under this assumption), and generated the data in Table A2 for the IBC127 dataset by randomly creating a 2/1 train/test split for each species (to match the authors' 5,343 / 2,671 split).

Overall, IBC127 is actually a reasonable dataset to start from. It has an imbalanced data problem, and the train/test conundrum is a serious problem, but we could have invested time manually (or automatically) to review the videos. However, a big problem with IBC127 is the random collection of bird species that comprise the dataset (a problem that affects the VB100 dataset as well). These species were clearly chosen because they satisfied some data quantity threshold when the authors were downloading videos. As we are interested in image and audio modalities, each of which would have their own data collection requirements, we wanted to avoid a 'hodgepodge' of bird species. We built SSW60 around 60 species of birds that all occur in a specific geographic region. This makes the classification task realistic, and also means that progress on the dataset directly impacts the biologists working on these species. The live "feeder-cams" mentioned in the main body of the paper is a prime example of a real world use case for an audiovisual classifier built on SSW60.

We chose not to use the VB100 dataset due to its missing metadata for train/test set creation, skewed video distribution, and its random collection of species.

B Visual Cross-Modality Results

In Table A3 we provide detailed results for cross-modality experiments on the visual modalities of the SSW60 dataset. Results on rows 5, 8, 15, 18, and 22 are also presented in Table 3 of the main paper. For completeness, we present results for models that have either been pretrained on ImageNet or simply randomly initialized. These experiments also explore the linear classifier setting for both training and domain transfer evaluation settings. All datasets (regardless of source) use the same 60 categories. Each row is a different experiment. Simply put, each experiment consists of (1) choosing a training dataset, (2) training a model, (3) choosing an evaluation dataset, and (4) evaluating the trained model. These experiments explore various tactics for training the classifier and for handling the domain shift when shifting to different evaluation datasets. Columns:

- **Initialization:** specifies whether the ResNet-50 backbones starts from ImageNet weights or randomly intialized weights.

- **Pretrain dataset:** specifies the source of training data used for the experiment. This is either the NABirds dataset (NAB), the iNaturalist dataset (iNat'21), or the frames of the videos from the SSW60 video clips.
- **Pretrain modality:** specifies whether the ResNet-50 backbone was trained using images or video clips. See Section 4.1 in the main paper for details on how the different modalities are used to train the backbone.
- **Pretrain method:** specifies how the "Pretrain dataset" was used to train the ResNet-50 backbone. Options are: **Linear**: we leave the ResNet-50 backbone weights fixed and we train a linear classifier by extracting a feature vector for each train sample in the "Pretrain dataset". **Finetune**: we fine-tune the weights of the ResNet-50 backbone using the "Pretrain dataset."
- **Evaluation dataset:** specifies the source of evaluation data for measuring top-1 accuracy. This is either the NABirds dataset (NAB), the iNaturalist dataset (iNat'21), or the frames of the videos from the SSW60 video clips.
- **Evaluation modality:** specifies whether the trained model (either a linear classifier or a fine-tuned network) was evaluated using images or video clips. See Section 4.1 in the main paper for details on how the different modalities are used for evaluation.
- **Evaluation method:** specifies how we used the "Evaluation dataset" to evaluate the trained model. Options are: **Direct**: we directly evaluate on the test samples of the evaluation dataset. **Linear:** we train a linear classifier using the *training* samples from the evaluation dataset, and then evaluate on the test samples. **Finetune**: we fine-tune the weights of the ResNet-50 model on the *training* samples from the evaluation dataset, and then evaluate on the test samples.

C Audio Augmentations

We employ augmentations at training time for both the visual and audio modalities, see Section 4.1 in the main paper for descriptions. In Table A4 we provide results when we disable different augmentation types on the audio modality. The model is equivalent to the ViT-B backbone results in Table 4 of the main paper. We can see that the addition of augmentations improves performance.

D Video Clip Examples

In Figs. A2, A3, A4, and A5 we show frames sampled at 1Hz from randomly sampled videos from our SSW60 dataset.

Table A3: Full results for **visual** cross-modality experiments. For all experiments we use a ResNet-50 backbone. See Sec. B for a description of the experiment setup and column explanations.

#	Initialization	Pretrain dataset	Pretrain modality	Pretrain method	Evaluation dataset	Evaluation modality	Evaluation method	Top-1 acc. (%)
1	ImageNet	NAB	Image	Linear	NAB	Image	Direct	79.20
2	ImageNet	NAB	Image	Finetune	NAB	Image	Direct	90.31
3	Random	NAB	Image	Finetune	NAB	Image	Direct	59.56
4	ImageNet	NAB	Image	Linear	SSW60	Video	Direct	17.44
5	ImageNet	NAB	Image	Finetune	SSW60	Video	Direct	24.05
6	Random	NAB	Image	Finetune	SSW60	Video	Direct	3.41
7	ImageNet	NAB	Image	Finetune	SSW60	Video	Linear	46.54
8	ImageNet	NAB	Image	Finetune	SSW60	Video	Finetune	56.55
9	ImageNet	iNat'21	Image	Linear	NAB	Image	Direct	75.94
10	ImageNet	iNat'21	Image	Finetune	NAB	Image	Direct	91.67
11	ImageNet	iNat'21	Image	Linear	iNat'21	Image	Direct	53.40
12	ImageNet	iNat'21	Image	Finetune	iNat'21	Image	Direct	75.20
13	Random	iNat'21	Image	Finetune	iNat'21	Image	Direct	51.57
14	ImageNet	iNat'21	Image	Linear	SSW60	Video	Direct	37.87
15	ImageNet	iNat'21	Image	Finetune	SSW60	Video	Direct	60.47
16	Random	iNat'21	Image	Finetune	SSW60	Video	Direct	24.36
17	ImageNet	iNat'21	Image	Finetune	SSW60	Video	Linear	73.63
18	ImageNet	iNat'21	Image	Finetune	SSW60	Video	Finetune	71.88
19	Random	iNat'21	Image	Finetune	SSW60	Video	Linear	45.72
20	Random	iNat'21	Image	Finetune	SSW60	Video	Finetune	46.44
21	ImageNet	SSW60	Video	Linear	SSW60	Video	Direct	35.60
22	ImageNet	SSW60	Video	Finetune	SSW60	Video	Direct	54.92
23	Random	SSW60	Video	Finetune	SSW60	Video	Direct	10.06
24	ImageNet	SSW60	Video	Linear	NAB	Image	Direct	13.85
25	ImageNet	SSW60	Video	Finetune	NAB	Image	Direct	18.45
26	Random	SSW60	Video	Finetune	NAB	Image	Direct	1.59
27	ImageNet	SSW60	Video	Finetune	NAB	Image	Linear	8.97
28	ImageNet	SSW60	Video	Finetune	NAB	Image	Finetune	56.91
29	Random	SSW60	Video	Finetune	NAB	Image	Linear	8.41
30	Random	SSW60	Video	Finetune	NAB	Image	Finetune	58.67

Table A4: Audio augmentation ablations using a ViT-B backbone.

No augmentation	+ time crop	+ frequency mask
44.1	60.6	66.8

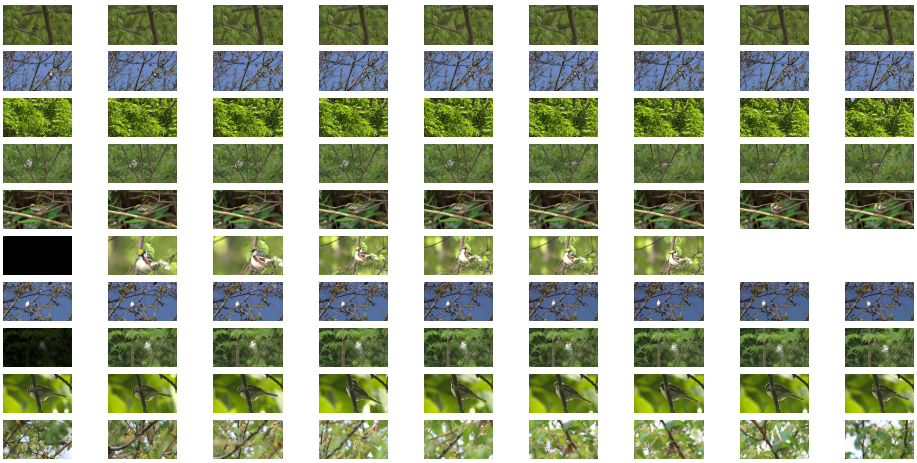


Fig. A2: 1Hz frames from Chestnut-sided Warbler videos from our SSW60 dataset.

Fig. A3: 1Hz frames from Northern Cardinal videos our SSW60 dataset.

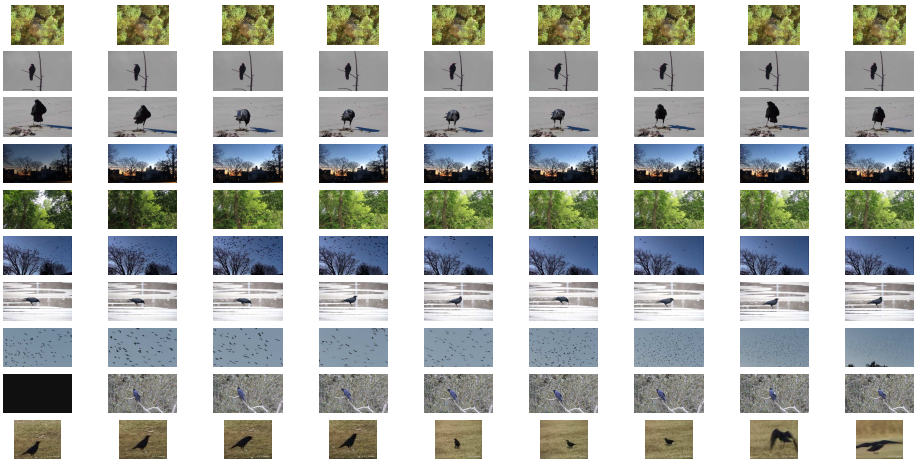


Fig. A4: 1Hz frames from American Crow videos our SSW60 dataset.

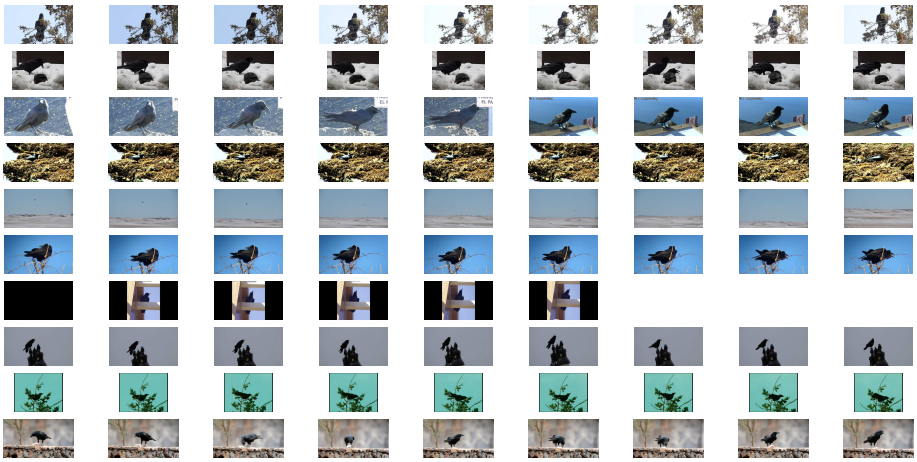


Fig. A5: 1Hz frames from Common Raven videos our SSW60 dataset.