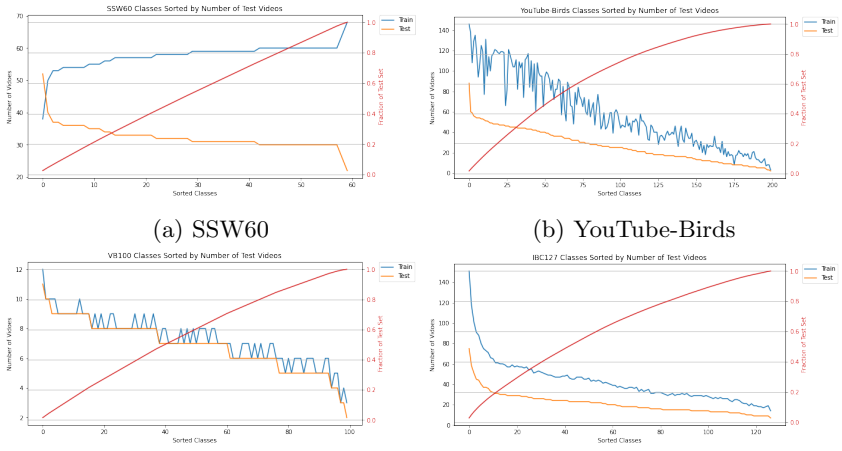
Exploring Fine-Grained Audiovisual Categorization with the SSW60 Dataset - Supplementary Material

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

A Existing Bird Video Datasets

In this section we dive deeper into the existing bird video datasets [3, 4, 7] 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 [7] is a collection of 18,350 videos that cover the same 200 categories as the CUB200 dataset [6]. 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 [2]), 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" [7]. 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 [5]. We therefore expect the error rate in YouTube birds to be at least has high as it is in the CUB200 dataset: 5% [5]. 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.

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 [3]	100	1,416	32.6	32.14	4.60	200.83
IBC127 [4]	127	8,014	31.2	28.66	3.00	266.72
YouTube-Birds [7]	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 [3]	100	1,416	730, 686	7.3, 6.9	7, 7	3, 2	12, 11
IBC127 [4]	127	8,014	5343, 2671	42.1, 21.0	37, 19	14, 7	151, 75
YouTube-Birds [7]	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

A.2 VB100

The VB100 dataset [3] 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 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

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

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 [4] 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 leakage of information. In the paper the authors state that they "use 5,343 videos for learning and 2,671 videos for testing" [4], 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 finetune 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.

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

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

No augmentation	+ time crop	+ frequency mask
44.1	60.6	66.8

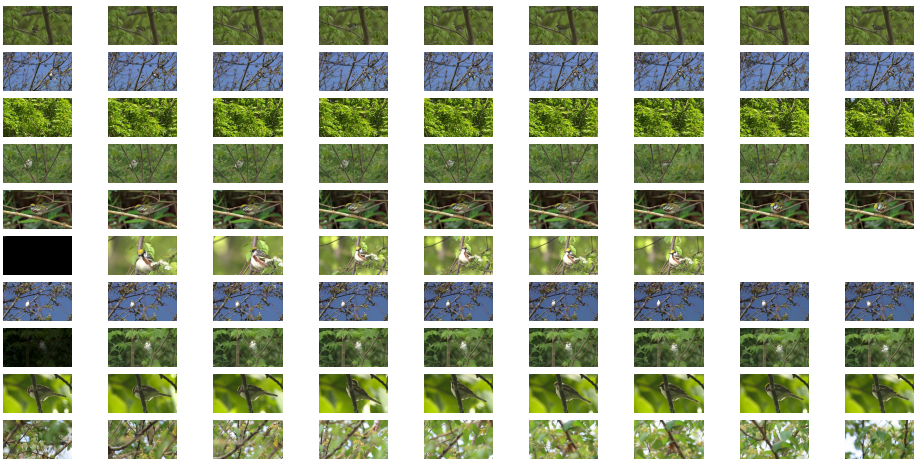


Fig. A2: 1Hz frames from Chestnut-sided Warbler videos from our SSW60 dataset.

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

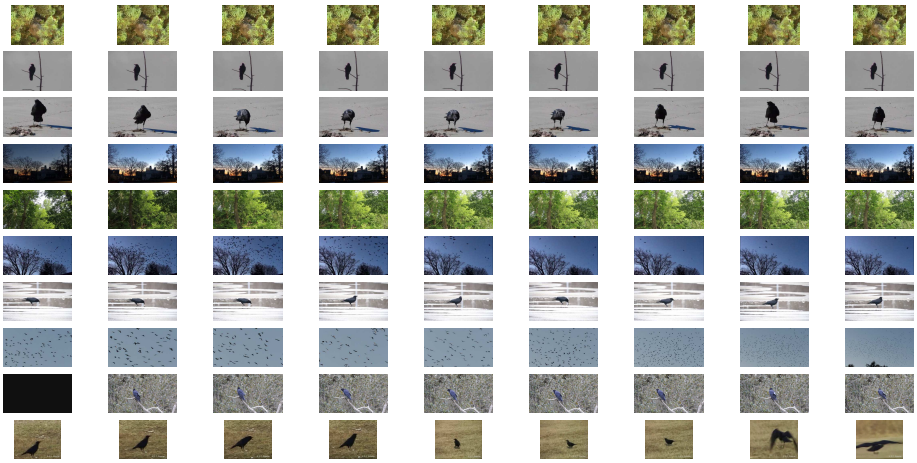


Fig. A4: 1Hz frames from American Crow videos our SSW60 dataset.

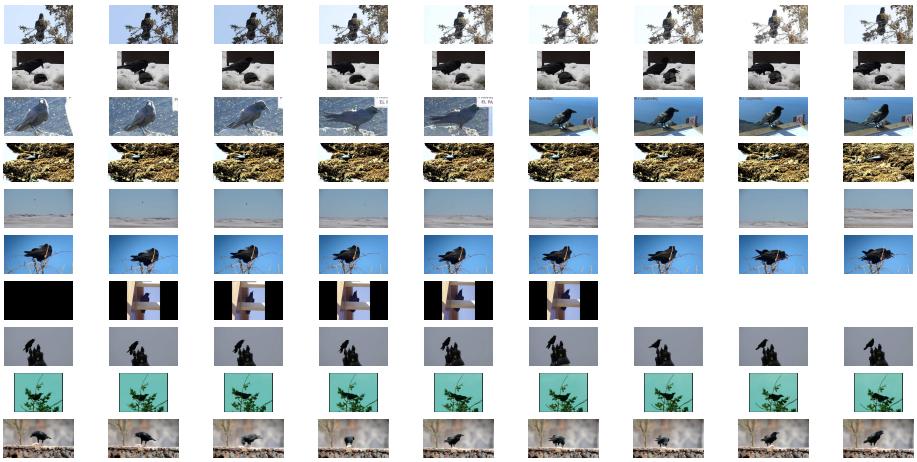


Fig. A5: 1Hz frames from Common Raven videos our SSW60 dataset.

References

1. iNaturalist, www.inaturalist.org, accessed Mar 7 2022
2. Macaulay Library, www.macaulaylibrary.org, accessed Mar 7 2022
3. 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)
4. Saito, T., Kanezaki, A., Harada, T.: Ibc127: Video dataset for fine-grained bird classification. In: International Conference on Multimedia and Expo (2016)
5. 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)
6. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Technical Report, CNS-TR-2011-001 (2011)
7. 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)