All that Sound

Kyuyeon Kim   Hyeongyeol Ryu   Yeonjae Kim

Abstract

Deep learning has recently brought great opportunities to obtain meaningful information from visual and audio data. While some previous works show good results on representing contextual information of audio and video, many have experimented with quite limited classes – mostly musical instruments. In this paper, we propose a better approach dealing with the lack of generalization by applying much more comprehensive datasets and also improve the performances using data augmentation on visual input and spectrogram. We replicate AVE-Net (audiovisual embedding network) and AVOL-Net (sound localization network), using L^3-Net and multisensory network as our baseline. We construct three different dataset – *AudioSet-Instruments, AudioSet-Animals,* and *AVE-Dataset* to confirm the scalability of network on other domains. The results show that AVE-Net and AVOL-Net can be applicable on any domain – *all that sound* – while preserving the performance on cross-modal retrieval and sound localization.

1. Introduction

If humans hear something, they can easily express it by drawing or writing. Similarly, humans can also point out an object that sound in the video. Then, is it possible to do these tasks with artificial systems? Cross-modal learning suggests an answer; to learn relevant information from multiple modalities.

Among variety of fields, one strategy broadly-used is cross-modal retrieval. Cross-modal retrieval refers to the tasks that retrieve related data from different modalities (Wen et al., 2019).

Recently, with the rapid growth of online media (e.g., YouTube) (Surís et al., 2018), audio-visual retrieval has been deeply discussed. Prior works, for example, use a video dataset in a supervised manner (Aytar et al., 2017; Gupta et al., 2016; Owens et al., 2016) or Teacher-Student supervision (Ramaswamy & Das, 2020).

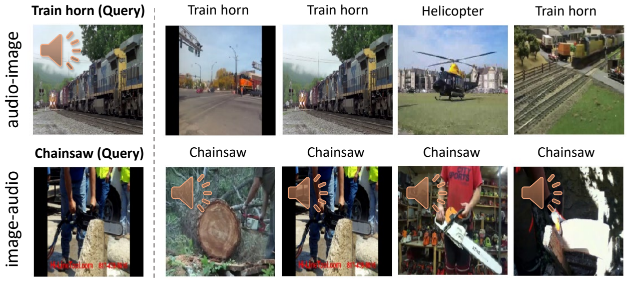


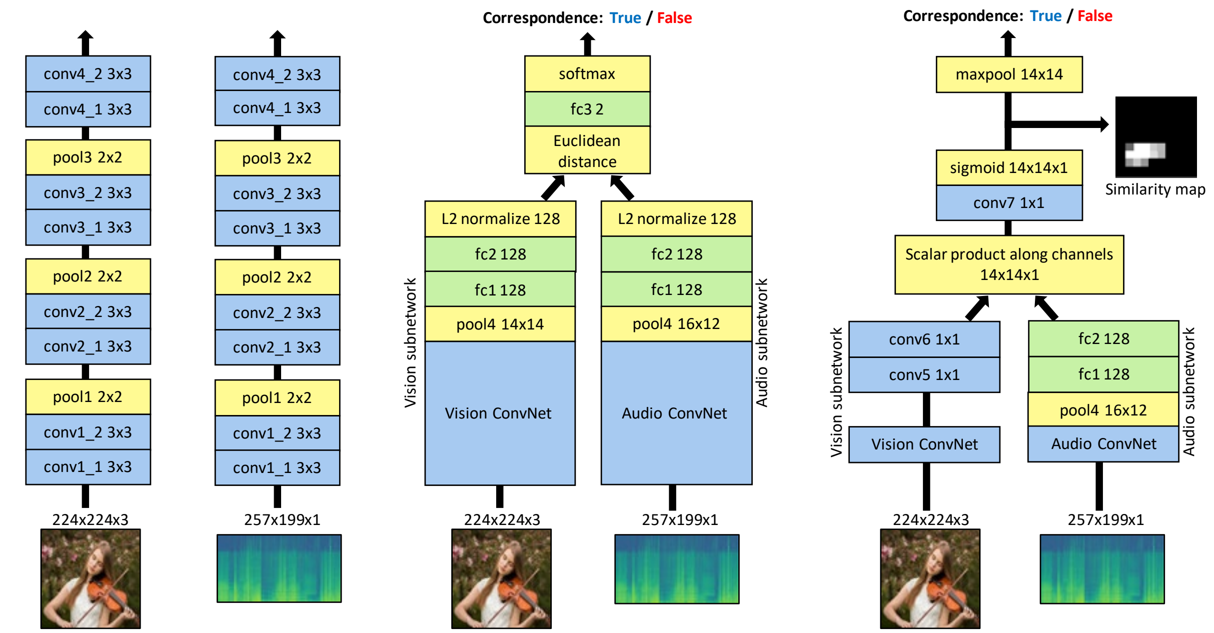
Figure 1. **Audio-image and image-audio retrieval.** When an audio or image is queried (visualized images with audio icon are just for easier understanding, not used as queries), most relevant top-4 are retrieved.

Yet, generating labeled datasets requires human labor a lot. Because of this problem, some approaches work in a self-supervised manner. They first create true and false samples utilizing the correspondence between video frames and audios, then train the networks to learn the audio-visual matching. (Owens & Efros, 2018b) proposes an idea that uses multiple video frames and raw sound data as input via 3D-convolutions to determine each embedding of them without any labeled data. Likewise, (Arandjelović & Zisserman, 2017) fuses the image and audio features by concatenating to extract both of their information. Lastly, (Arandjelovic & Zisserman, 2018) approaches this problem as (Arandjelović & Zisserman, 2017) did but also modifies network architectures to calculate Euclidean distance, so that the network can be more aware of distance between embeddings.

Among these papers of self-supervised audiovisual learning, we choose (Arandjelovic & Zisserman, 2018) as our target to replicate, since it shows state-of-the-art performances and exhibits availability of future extensions. However, this work only considers limited classes of objects, especially musical instruments, which cannot be generalized in wild, real-world objects. Therefore, we aim to solve this skewness by applying more various object classes while preserving the performance. Our approach mainly contributes that we show suggested networks are applicable on any domain of objects.

2. Approach

We aim to generalize the task by extending the domain of videos while maintaining the performance. To prove this, we first replicate our target paper from scratch, including main networks: AVE-Net and AVOL-Net. We first follow up authors' work with *AudioSet-Instruments*, the same dataset



(a) Vision ConvNet (b) Audio ConvNet (c) AVE-Net (d) AVOL-Net

Figure 2. **AVE-Net and AVOL-Net**

mentioned in the paper. We then check if they are scalable on another domain using *AudioSet-Animals*, our newly built dataset. We finally examine the models' ability to generalize with *AVE-Dataset*. As models are easily overfitted, we figure out to apply data augmentation on image and spectrogram for improvement. We evaluate the results on three tasks: audio-visual correspondence (AVC task), cross-modal retrieval, and sound localization. Detailed explanations are introduced in the next section.

3. Data and Experiments

3.1. Models

We decide to replicate two models: AVE-Net and AVOL-Net (Arandjelovic & Zisserman, 2018), utilizing L^3-Net (Arandjelovic & Zisserman, 2017) and the multisensory network (Owens & Efros, 2018a) as our baseline. As the structure of L^3-Net is not so different from AVE-Net, we also implement and train L^3-Net to set up a fair baseline. Details of each network are followed below.

3.1.1. Targeting Models

Our target paper (Arandjelovic & Zisserman, 2018) includes two models: AVE-Net (Audio-Visual Embedding Network) and AVOL-Net (Audio-Visual Object Localization Network), each targeting for different tasks. The overall pipeline of AVE-Net and AVOL-Net is shown in Figure 2.

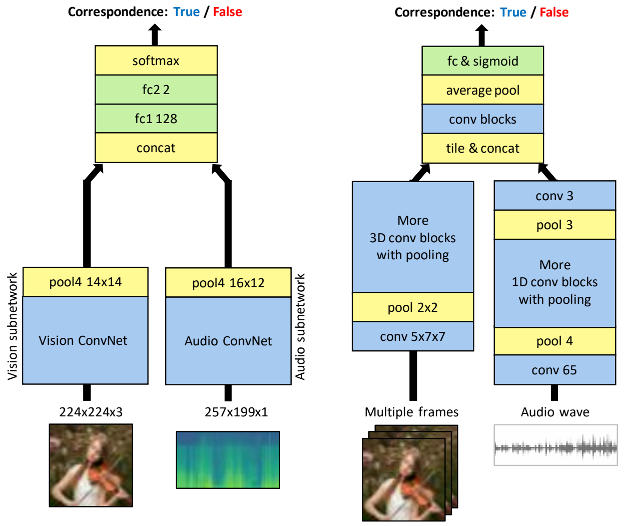
AVE-Net aims to represent the image and audio into the embedding vector, which is frequently-used approach of audiovisual representation learning (Zhu et al., 2020). AVE-Net employs AVC task for predicting whether given visual and audio data matches. Euclidean distance of image and audio embedding is passed to tiny fully connected layer, which produce the correspondence prediction of two inputs.

AVOL-Net is similar to AVE-Net, except that it produces similarity map for sound localization. Vision subnetwork in AVOL-Net produces 128-channel 14×14 feature map, rather than a single vector. A channel-wise product of visual feature map and audio embedding is utilized as a similarity map after activation.

3.1.2. Baseline Models

The first baseline model called L^3-Net is the previous version of AVE-Net (Arandjelovic & Zisserman, 2017). Our replication includes reproducing L^3-Net as we should compare the performance with AVE-Net under the same condition. Like AVE-Net, L^3-Net also focuses on extracting both visual and audio features. While most of the building blocks are identical, The significant difference comes from the feature fusion strategy: concatenation of audio and visual embedding. Thus, L^3-Net is less likely to be aware of Euclidean distance-based alignment.

The second baseline is a multisensory network (Owens & Efros, 2018a). Training multisensory network also utilizes AVC task as a proxy task. However, the definition of false correspondence is slightly different from that of AVE-Net



(a) L^3-Net

(b) Multisensory-Net

Figure 3. **Baseline models**

described in Section 3.3. For example, the negative sample is generated by pairing a given video frame with synthetically shifted audio. So the misalignment arises in the temporal feature, not in semantic details. Also, the multisensory network requires multiple video frames as input. It tries to concentrate on spatio-temporal features by applying 3D convolution on visual data. Lastly, it treats audio data as 1D-vector input, instead of being converted into a spectrogram. The multisensory network is far from obtaining high-quality audiovisual embedding, so there is no numerical measurement of feature embedding quality. Thus, we use only its accuracy on the AVC task to compare its performance with other works.

3.2. Dataset

Our target paper (Arandjelovic & Zisserman, 2018) used *AudioSet* (Gemmeke et al., 2017) as a base dataset. In this paper, *AudioSet* was filtered into *AudioSet-Instruments*, which has videos only containing events of human voices, playing musical instruments, and using tools. We suggest that *AudioSet-Instruments* is insufficient to generalize even when AVE-Net and AVOL-Net successfully work on targeting tasks since the categories are limited. To prove that AVE-Net and AVOL-Net comprehensively work on any sound events, we use additional categories: *AudioSet-Animals* containing animal-related events and *AVE-Dataset* with extensive audio-visual categories. We emphasize that using these additional datasets is not a target to improve the quantitative performance of given models. Rather, the purpose of using extra data is to show that the models can learn semantic concepts from any given audiovisual events.

AudioSet consists of more than 2 million videos that are annotated with 632 classes and structured with hierarchical ontology. (Arandjelovic & Zisserman, 2018) filters *AudioSet* into *AudioSet-Instruments* to make the dataset more manageable for their purposes, yielding 110 audio-visual classes. Due to the constraints of resources, we choose 50 out of110 classes and use the subset of all videos containing at least one of those classes. Thus, our reproduced *AudioSet-Instrument* has 60k of videos divided into 48k, 6k, and 6k for the train, validation, and test splits.

AudioSet-Animals is similar to *AudioSet-Instrument* as it is the subset of *AudioSet*. One difference is that *AudioSet-Animals* is the collection of videos, all annotated with ‘Animal’ tag. The number of videos in *AudioSet-Animals* is 35k and we split them into 80%-10%-10% proportions for the train, validation, and test set.

AVE-Dataset (Tian et al., 2018) is also the subset of *AudioSet*, but not constrained to domain-specific categories. *AVE-Dataset* consists of videos with 28 categories sufficient to generalize the real-world audio-visual events. It contains not only musical instruments, human voices, and animals but also any objects/events such as vehicles and cooking. Unlike *AudioSet-Instruments* and *AudioSet-Animals*, videos in *AVE-Dataset* are not skewed and well-balanced. In other words, the source of the sound event always exists in videos from *AVE-Dataset*, and the distribution of videos along categories is relatively even. We use 3,085 video clips in *AVE-Dataset* and split them into same proportion as *AudioSet-Animals*.

We re-emphasize that *AudioSet-Animals* and *AVE-Dataset* are not the supplementary data to boost the performance compared to using only *AudioSet-Instruments*, but instead used to insist that AVE-Net and AVOL-Net can learn the features from any objects in the wild. We first examine their scalability on animal category with *AudioSet-Animals* and then measure the ability to generalize on any objects with *AVE-Dataset*.

3.3. Data Processing

Datasets used in our experiments consist of 10-second video clips from YouTube. As AVE-Net and AVOL-Net require a single video frame with a 1-second spectrogram, we process the video clips into images and spectrograms which can be directly fed into the network.

We first slice a 10-second video clip to have 1-second intervals. For each 1-second slice, we extract a video frame at the mid-point with corresponding 1-second audio. Next, we resample the sound at 48 kHz and convert it into a log-spectrogram, using a 0.01-second window length with half-window overlapping. Each pair consequently has the image from the mid-point of sliced video with its corresponding audio spectrogram. Not all videos have a 10-second du-

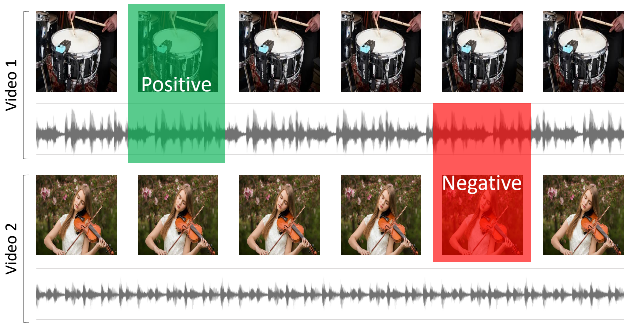


Figure 4. **How labels can be generated from video clips in self-supervised manner?** We obtain positive samples by pairing a video frame and audio from the same timestamp. Negative samples are made by pairing a image and audio from a different video.

ration. Hence we produce nine pairs for each video clip. The shape of the input frame is resized to 224×224 with an RGB channel, and the spectrogram has 257×199 shapes having a single channel. Unlike the original work, we do not treat the spectrogram as a gray-scale image but directly use the log-scaled value. We use mean and standard deviation from *ImageNet* to normalize input frames, and normalize spectrograms to have unit variance before feeding them into the network.

To train the networks in a self-supervised manner, we need to label image-audio pairs by defining a pretext task. The video itself can offer the binary classification pretext task, which comes from the correspondence. Here, the correspondence means whether the frame and the audio come s from the same video. For example, the correspondence between a dog image and dog-barking audio is likely to be true, while a dog image with a cat-meowing sound may have false correspondence. All image-audio pairs from the previous processing step have true correspondence since we retrieve the frame and the spectrogram from the same interval.

We obtain positive image-audio pairs automatically from video processing steps. In contrast, samples of negative correspondence can be produced by pairing the frame and the audio from a different video. In other words, with the video frame from a particular video slice, we pick the spectrogram from exactly different video and pair them as a negative sample. As summarized in Figure 4, generating positive and negative pairs does not require any human labor, which makes self-supervised learning able to be applied to training the suggested models. The proportions of positive and negative samples in training, validation, and test set are all 50%-50%. In the case of pairs whose frame and audio come from the same video, but different times are not produced as inputs; they are not the interest of the AVC task in our project.

3.4. Implementation

For each epoch, only a single positive and negative pair are made from each video in training time. A specific interval is randomly chosen for each video to make a positive pair. A negative pair is made by selecting a frame from an arbitrary interval of a given video and pairing it to a random audio interval from another video.

We first train the networks in vanilla form without any improvement strategy. We then try out two improvement strategies: data augmentation and initializing the last tiny fully connected layer (`fc3` in Figure 2). Data augmentation is applied to input frames – random cropping, horizontal flipping, and jittering on brightness and saturation. Without image augmentation, input frames are just resized to fit into the input layer. When augmenting images, we first resize the image into 256×256 and then randomly crop it into 224×224. In terms of weight initialization, AVE-Net is inevitably subject to weight change of the last fully connected layer.

The network is trained with the same hyperparameters in the paper, except for the learning rate and batch size. The learning rate is not mentioned in the paper, so we find our learning rate of 5×10^{-5}. The authors use a batch size of 2,048 by using 16 GPUs in parallel. However, because of resource limitations, we use a batch size of 64. Optimizer and regularization method is similar to the original; Adam optimizer with weight decay by 10^{-5}. The model, which shows the minimum validation loss, is chosen for performance evaluation.

Source codes of our project are available on the link below[1].

4. Results

4.1. AVC Task

The AVC task is a common task given to all networks described in Section 3.1. Although this task is utilized as a proxy task to help network achieve significant goals, its performance is still remarkable to check if the network succeeds to grab semantic information from the audio-visual data.

As shown in Table 2, using *AudioSet-Instruments*, we obtain accuracy of 72.8% on AVC task using AVE-Net without any improvement strategy. After applying the standard augmentation method on both image and spectrogram, the accuracy of AVE-Net is boosted to 75.1%. As it is obvious that data augmentation helps boost performance, we decide to apply

[1] https://bit.ly/2Z3SHMj

Table 1. **Cross-modal and intra-modal retrieval performance of AVE-Net and L^3-Net on *AudioSet-Instruments* and *AVE-Dataset*.** Top 2 rows show nDCG score of each network mentioned in the paper. Models on the bottom rows with bold text are the reproduced version of networks we implement.

	AudioSet-Instruments (nDCG@30)				*AVE-Dataset* (nDCG@5)			
Model	img-aud	aud-img	img-img	aud-aud	img-aud	aud-img	img-img	aud-aud
AVE-Net	**.561**	**.587**	**.604**	**.665**	-	-	-	-
L^3-Net	.418	.385	.567	.653	-	-	-	-
AVE-Net	.731	.728	.760	.780	.551	.539	.642	.677
AVE-Net + Aug.	**.743**	**.757**	**.772**	**.792**	**.572**	**.554**	.683	**.738**
L^3-Net + Aug.	.627	.611	.755	.781	.413	.401	**.687**	.711

Table 2. **Accuracy on AVC task of models trained on different kinds of dataset.** Top 4 rows show the accuracy of each network mentioned in the paper. Models on the bottom rows with bold text are the reproduced version of networks we implement.

Model	Dataset		
	AudioSet Instruments	*AudioSet Animals*	*AVE Dataset*
AVE-Net	81.9	-	-
AVOL-Net	81.9	-	-
L^3-Net	80.8	-	-
Multisensory	59.9	-	-
AVE-Net	72.8	66.9	66.4
AVE-Net + Aug.	75.1	68.5	64.8
AVOL-Net + Aug.	73.4	67.1	67.9
L^3-Net + Aug.	**77.7**	**70.6**	**71.1**

data augmentation on training AVOL-Net and L^3-Net. In the meanwhile, our version of networks cannot follow the accuracy suggested in the paper. We note that we use only 20% of the training data due to the resource limitation. Also, smaller batch size and unknown learning rates account for this result, leading the networks to the sub-optimal point.

We try to show improvement by training networks on the broader domain of objects. We first examine their extensibility with *AudioSet-Animals* and plunge into a much general area using *AVE-Dataset*. The performance of networks on other datasets is relatively low. However, the accuracy is much better than just random guessing. Here, the critical point is that models can obtain audio-visual information from any object. Although the networks have shown lower performance on other categories of objects, we note that the AVC task is just proxy task to let the network know how to make good embeddings. Moreover, L^3-Net has unexpectedly shown the best accuracy. However, we again stress that the AVC task is just an auxiliary task. The performance of networks on real task – cross and intra-modal retrieval – will be shown in the following section.

4.2. Cross-modal Retrieval

Retrieval task is a useful measurement to evaluate how well representations are aligned. Cross-modal retrieval aims to retrieve items that have different modalities with a query (e.g., retrieving audio related to the queried image). Intra-modal has the opposite goal (e.g., retrieving image related to a given image). To quantify the quality of retrieved items, we use the same metric in the paper – nDCG. As *AudioSet* offers a hierarchical ontology tree, we can measure the tree distance between the classes. We give higher relevance scores on retrieved items with lower tree distance to queried items. In the case of AVE-Net, we use 128-length L2 normalized embeddings from each subnetwork to measure the similarity between items. For L^3-Net, we use features after `pool4` layer in Figure 3. Euclidean distance between embeddings from two different items is computed to measure the similarity. The multisensory network is not included as it is not purposed for embedding the audiovisual data.

As shown in Table 1, the evaluation starts by measuring nDCG@30 on *AudioSet-Instruments*. It seems that our implementation surpasses the result suggested in the paper. Recall that we use the subset of classes and a small proportion of dataset due to the resource constraint, it may not be the fair comparison. However, aspects between the result of the target paper and ours are very similar, which may implicitly convey that we have trained successfully. AVE-Net with data augmentation-applied – one of our improvement methods – shows slightly better retrieval quality than the plain version. A reproduced version of L^3-Net shows relatively low quality on cross-modal retrieval. Because L^3-Net simply concatenates the visual feature with an audio descriptor, it is unaware of the Euclidean distance feature between representations having different modality.

As our improvement, we check the scalability of AVE-Net on a more comprehensive domain with *AVE-Dataset*. As the number of test samples is only 308, including various kinds of classes, we have found that evaluation on five retrieved items is sufficient instead of 30. Comparing to authors' result on *AudioSet-Instruments*, our result with *AVE-Dataset*

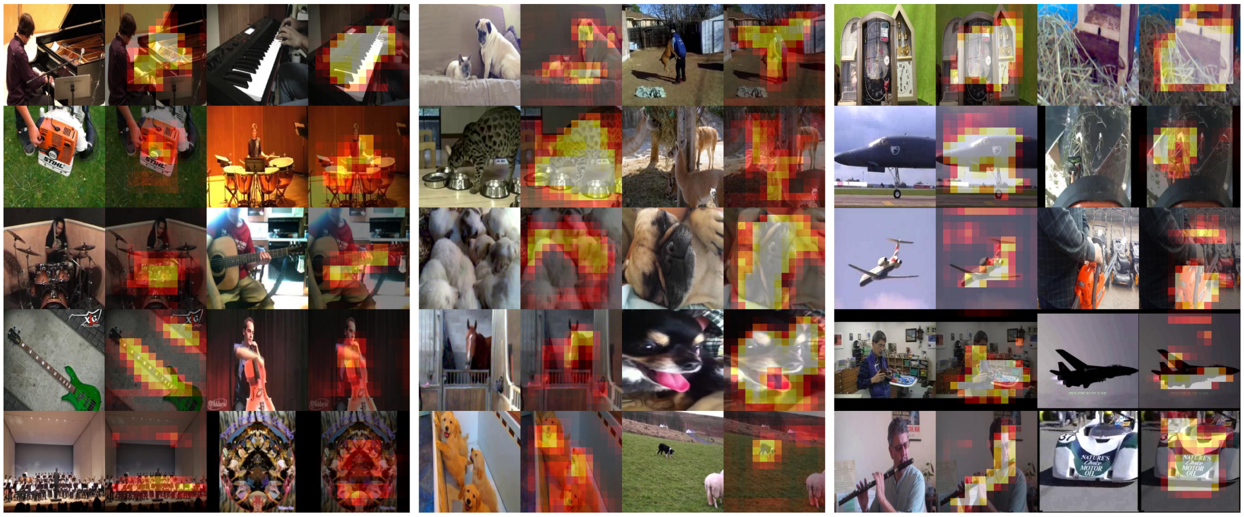

(a) *AudioSet-Instruments* (b) *AudioSet-Animals* (c) *AVE-dataset*

Figure 5. **Qualitative results of sound localization on different datasets.**

shows reasonable retrieval quality. Although L^3-Net works comparably on intra-modal retrieval, AVE-Net easily beats L^3-Net in cross-modal retrieval tasks.

4.3. Sound Localization

To obtain the attention map that highlights objects that sound, we visualize the similarity map of AVOL-Net depicted in Figure 2. We scale the similarity map of 14×14 resolution by 16 times so that the image can be completely overlapped with the similarity map.

We focus on giving a qualitative result of sound localization in Figure 5, as there is no standardized way to quantify the localization quality. The author of replicating paper gives a quantified result, using their test set by drawing a bounding box, but not open in public. In the case of a multisensory network, it also shows only qualitative results.

4.4. Embedding visualization with t-SNE

We visualize image and audio embeddings with t-SNE, a useful tool to check whether embeddings are well-aligned. We utilize AVE-Net trained on *AudioSet-Instruments* to obtain audiovisual embeddings. As shown in Figure 6, embeddings with various kinds of classes are well-clustered, which means model has successfully learned semantic concepts from variety of sound and object.

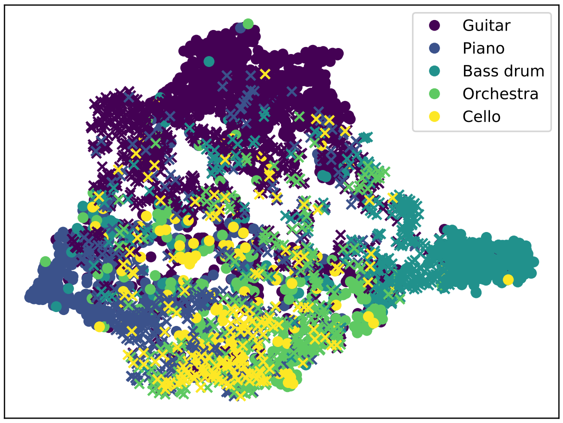


Figure 6. **Visualizing embeddings with t-SNE.** Dots with ● represent image embeddings, and embeddings with × are from audios.

5. Discussions

We propose an idea that can efficiently make use of any objects that sound for cross-modal retrieval and sound localization. Our approach tries to solve a question: *Is suggested audiovisual models able to generalize their tasks on any objects?* To answer this question, we first collect more various datasets not only limited to musical instruments, but also including humans, animals, and whatever existing around. Then, we mainly train two neural networks: AVE-net and AVOL-net for showing whether our model produces good audio and visual embeddings. According to the results of the evaluation with nDCG and sound localization, both of our networks perform well even with more classes. We can see that the performances are comparable to existing works, in spite of using many more classes of sound and objects.

While our strategy gives a proof of generalization, there is a room for improvement. One possible suggestion could be a change in model architectures, and this is left for the future work.

References

Arandjelovic, R. and Zisserman, A. Look, listen and learn. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Arandjelovic, R. and Zisserman, A. Objects that sound. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Arandjelović, R. and Zisserman, A. Look, listen and learn, 2017.

Aytar, Y., Vondrick, C., and Torralba, A. See, hear, and read: Deep aligned representations, 2017.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

Gupta, S., Hoffman, J., and Malik, J. Cross modal distillation for supervision transfer. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2827–2836, 2016.

Owens, A. and Efros, A. A. Audio-visual scene analysis with self-supervised multisensory features. In *The European Conference on Computer Vision (ECCV)*, September 2018a.

Owens, A. and Efros, A. A. Audio-visual scene analysis with self-supervised multisensory features, 2018b.

Owens, A., Wu, J., McDermott, J. H., Freeman, W. T., and Torralba, A. Ambient sound provides supervision for visual learning. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *Computer Vision – ECCV 2016*, pp. 801–816, Cham, 2016. Springer International Publishing.

Ramaswamy, J. and Das, S. See the sound, hear the pixels. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

Surís, D., Duarte, A., Salvador, A., Torres, J., and i Nieto, X. G. Cross-modal embeddings for video and audio retrieval, 2018.

Tian, Y., Shi, J., Li, B., Duan, Z., and Xu, C. Audio-visual event localization in unconstrained videos. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Wen, X., Han, Z., Yin, X., and Liu, Y.-S. Adversarial cross-modal retrieval via learning and transferring single-modal similarities, 2019.

Zhu, H., Luo, M., Wang, R., Zheng, A., and He, R. Deep audio-visual learning: A survey, 2020.