Sound and Visual Representation Learning with Multiple Pretraining Tasks

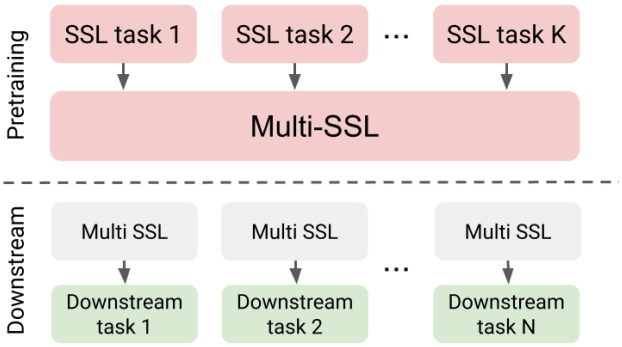
Arun Balajee Vasudevan[1], Dengxin Dai[2], Luc Van Gool[1,3]
ETH Zurich[1] MPI for Informatics[2] KU Leuven [3]
{arunv,vangool}@vision.ee.ethz.ch, ddai@mpi-inf.mpg.de

Abstract

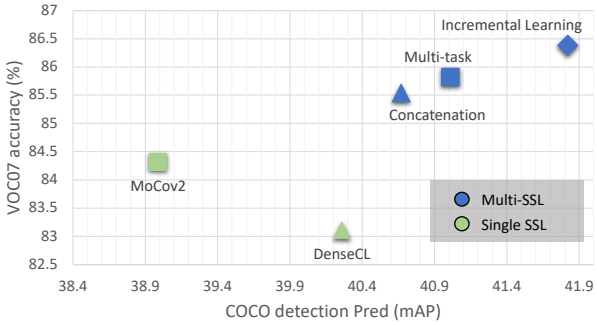
Different self-supervised tasks (SSL) reveal different features from the data. The learned feature representations can exhibit different performance for each downstream task. In this light, this work aims to combine Multiple SSL tasks (Multi-SSL) that generalizes well for all downstream tasks. Specifically, for this study, we investigate binaural sounds and image data in isolation. For binaural sounds, we propose three SSL tasks namely, spatial alignment, temporal synchronization of foreground objects and binaural sounds and temporal gap prediction. We investigate several approaches of Multi-SSL and give insights into the downstream task performance on video retrieval, spatial sound super resolution, and semantic prediction using OmniAudio dataset. Our experiments on binaural sound representations demonstrate that Multi-SSL via incremental learning (IL) of SSL tasks outperforms single SSL task models and fully supervised models in the downstream task performance. As a check of applicability on other modalities, we also formulate our Multi-SSL models for image representation learning and we use the recently proposed SSL tasks, MoCov2 and DenseCL. Here, Multi-SSL surpasses recent methods such as MoCov2, DenseCL and DetCo by 2.06%, 3.27% *and* 1.19% *on VOC07 classification and* +2.83, +1.56 *and* +1.61 *AP on COCO detection. Code will be made publicly available.*

1. Introduction

Self-supervised learning (SSL) is a popular paradigm to train deep networks on pretext tasks that readily extract supervision from the data. Typically, this allows models to learn data representations, which are then used for downstream tasks. The objectives for these SSL tasks are designed based on the corresponding downstream tasks. As a result, pretrained deep networks on SSL tasks yield good performance on that downstream task or its related tasks when finetuned. We note in the literature [16, 66] that these pretrained models are not generic enough to give a satisfactory performance on a diverse pool of downstream tasks. For instance, some pretext tasks [9, 28] on images focus on learning global image feature representation while few others [8, 43, 66, 60] focuses on local features. The former works well on downstream tasks like image retrieval or classification while the latter helps with dense prediction/labelling tasks. For instance, this is evident in Figure 1(b) where MoCov2 [28] performs good for classification while DenseCL [66] comes out better for object detection task.



(a) Pipeline of our approach



(b) Classification vs Object detection tradeoff

Figure 1: Self Supervised Learning (SSL) tasks are designed for specific downstream tasks. Our work demonstrates how a single model learns to combine Multiple SSL tasks (i.e., Multi-SSL) that generalizes well for all the downstream tasks. Figure (b) shows comparison of single SSL tasks: MoCov2 [28] and DenseCL [66] with our Multi-SSL models.

Be it visual, sound, or linguistic data, how well the data representations are learned determines the generalization of

a model. When the models are generic, feature representations from them perform satisfactorily on several diverse downstream tasks. In this light, our work tries to investigate how to train a self-supervised model using Multiple SSL tasks (Multi-SSL as in Figure 1(a)) that generalizes well.

In the last few years, a number of self-supervised approaches are proposed in the language, sound and vision research community, from natural language text corpus [14, 57, 58], images [43, 36], videos [1, 27, 38], and audios [24, 39]. In sound representation learning, a few prominent ones are audio-visual correspondence [4, 74], audio context prediction [62], and among others. Colorization [41], image inpainting [56], *etc.,* are among the vision tasks for image representations. To assess the model, a standard set of downstream tasks for their corresponding areas is picked, and the model is retrained and tested on several datasets.

With the aim of extracting well generalized sound and image representations, this paper explores ways of combining *Multiple SSL* tasks as shown in Figure 1(a). We call it Multi-SSL. While there have been a few works in this area on sound representation [59, 71, 68], language [65] or visual representation learning [16, 22], they only have only considered addressing this in the standard multi-tasking framework [16, 66]. This work, however, introduces Multi-SSL which investigates different design options to combine multiple SSL tasks and provide insights into the downstream tasks. Through this comprehensive analysis, we highlight how Multi-SSL models improve over strong baselines in the evaluation of different downstream tasks. In the paper, we experiment with binaural sound and image data representations.

For binaural sound representation learning, we propose a set of SSL tasks. Firstly, spatial alignment task is proposed for learning spatial features in sounds. The task leverages the correspondence between binaural sounds and the rich spatial cues present in 360° videos. Our second task is to learn temporal synchronization of moving objects in the scene and binaural sounds. We call this task as foreground alignment as it learns to align foreground objects and sounds. The third task of temporal gap prediction encourages the sound models to learn a sense of time gap between binaural sounds. For training the above tasks, we use the OmniAudio dataset [64] and evaluate the performance on three downstream tasks: a) video retrieval, b) auditory semantic prediction and c) spatial sound super resolution (S^3R).

In addition, we examine the performance of the proposed Multi-SSL approach to visual representation learning. For SSL tasks, we consider the recently proposed contrastive learning paradigms on representation learning. Following MoCoV2 [28], the first SSL task works with contrastive learning at the level of global image features. For the second task, we select dense contrastive learning [66] which focuses on the local features. We train the above SSL tasks on the ImageNet dataset [13] and then evaluate the performance on downstream tasks of image classification on Pascal VOC dataset [18] and object detection and instance segmentation on MS COCO dataset [45].

Furthermore, we propose different Multi-SSL methods such as Concatenation, Multi-task, ProgressiveNet, Incremental Learning (IL) and others, that are detailed in Section 4. Experiment results show that a) All the above Multi-SSL methods improve over single SSL tasks, b) and they also outperform supervised models, and finally c) IL approach performs the best among the Multi-SSL methods as in Figure 1(b). We note that these observations are consistent for both sound and vision domains.

Here is a summary of our contributions. (1) We propose different approaches to self-supervised learning (SSL) for binaural sound representation learning; (2) We introduce several approaches of Multi-SSL that learns to combine multiple SSL tasks; (3) We also train and evaluate our Multi-SSL approach for image representation learning.

2. Related Works

Self-supervised learning. Recently, SSL has become a key component to achieve good performance on downstream tasks predominantly with low-resource settings either in sounds [24, 39], natural language processing [40, 10] or computer vision [48, 36, 43]. Let us focus more on sound and image representation learning in this work. In vision research, many self-supervision tasks have been applied as a counter to ImageNet [13] pretraining. Early self supervised pretext tasks typically include image colorization [41, 73], orientation prediction [23], affine transform prediction [72], predicting contextual image patches [15], reordering image patches [6], counting visual primitives [51]. These pretext tasks typically predict some low-level image properties resulting in feature representations *i.e.,* covariant to image transformations. Recently, contrastive learning gained considerable traction in SSL [34, 5, 52, 31], which drives the concept of maximizing the similarity of a representation across views while minimizing its similarity with distracting negative samples [28, 9, 69]. Here, the positive pairs are usually created with multiple augmented views of the same image, while negative pairs are created from different images. However, there are few works [12, 26] that use just positive samples. Our work explores several of these contrastive learning based SSL such as [28, 66, 32, 60] and its variants for learning initial image representations. We pick MoCov2 [28] and DenseCL [66] in our work.

Audio-visual learning. Audio-visual data offers a variety of resources for knowledge transfer between different modalities [4, 7, 1]. Many works [64, 3] leverage the natural synchronization between vision and sound to learn representation of sounds and images without ground truth labels. This has been successfully used in various tasks such as

visually guided sound source separation [19] and sound localization [3], audio to visual generation [75], visual to audio generation[77, 54, 17], sound inpainting [76, 47] and sound classification [4]. Prior works are often implemented either by predicting audio-visual correspondences at the video level [2, 50], frame level and object level alignment [1]. Han *et al* [27] uses optical flow patterns in video to learn video representations. We inherit these flow patterns to find foreground objects and sound correspondence to learn sound representations. Morgado *et al* [49] learn representations by performing audio-visual spatial alignment of 360° video and spatial audio. In our spatial alignment SSL, we differ with them in learning binaural sounds representation. [62, 67] propose prediction of sense of time difference as SSL task, given two video/audio frames, to learn video/audio representations. We follow quite similar to the work of [62].

Multi-task self-supervised learning. While it has been shown extensively in supervised learning settings [63, 35], the literature on multi-tasking in SSL remains less explored. There are extensive studies on pretext tasks as we see in Section 2 for image or sound representation learning. In addition, a few SSL works [55, 59, 42] in computer vision and speech address combining multiple pretext SSL tasks in a multi-task setting. Successful pretext tasks such as Jigsaw [15], colourisation and rotation [23] have been combined successfully to improve downstream performance [37]. Wang *et al* [66] employs contrastive learning paradigm at image level and dense level features in multi-tasking settings. In our work, we explore more ways to combine these paradigms and explore how incremental learning helps in this context.

3. Self-Supervised Tasks

We use different set of pretext tasks for learning binaural sounds and image representations. We pick diverse set of SSL tasks aiming to extract diverse feature representations. These are more likely to span the space of features needed to understand general data content. Initially, let us see SSL pretext tasks for binaural sounds in Section 3.1, and then for visual representations in Section 3.2.

3.1. Binaural sounds

Given an audio-visual dataset with N raw video segments, e.g. $D = \{(a_1, v_1), (a_2, v_2), ..., (a_N, v_N)\}$, the objective for SSL task is to obtain a function $f(.)$ that is effectively used to generate sound representations for various downstream tasks. In our work, we formulate sound clips as spectrograms, which are effectively processed by convolutional neural networks(CNNs) as demonstrated by [20, 3]. Let us see the SSL pretext tasks.

Spatial alignment (denoted as $\mathcal{A}$). This pretext task learns to align 360° videos spatially with their corresponding binaural sounds. A straightforward way to implement audio-visual spatial alignment is to rotate the video randomly with R angle rotation with respect to sounds to create an artificial misalignment between them. And, later we learn to predict this rotation angle between the video and the sounds. In learning the spatial alignments of visual and sound contents, the network is encouraged to understand the scene composition (*i.e.,* where the different sources of sound are located), which results in better representations for downstream tasks. Closest work to this pretext task is [49] which employs contrastive learning setup to learn the spatial alignment between 360° videos and spatial sounds. In our work, we leverage the pairs of binaural sounds and 360° videos and frame the problem as an angle prediction between them. We divide 360° into 8 equal bins, each representing different orientations. We train the SSL model to predict the angular difference *i.e.,* rotation angle R. We employ cross entropy loss between the predicted and actual angular difference as $CE(h(v_{iR}, a_i), R)$, where $h(v_{iR}, a_t)$ is a prediction head followed by a softmax layer to predict the angle $\hat{R}$. v_{iR} and a_i refer to video features of rotated video segment v_i and sound features respectively. R is the groundtruth angular difference due to the rotation on 360° video segment v_i. CE denotes the cross entropy loss.

Foreground alignment (denoted as $\mathcal{B}$). Using unlabeled videos, our work aims to harness the natural synchronization of vision and sound to learn binaural sound representations. Audio-visual temporal synchronization (AVTS) [38, 53] distinguishes between a pair of audio and video clips belongs to the same timestamps (aligned) or from separate timestamps (misaligned) of the same video. We leverage this alignment to train our model in a contrastive learning setup.

The motion of objects and their sounds are closely related. In this light, we apply two approaches based on the fair assumption that the recorded sounds from the scene come from moving foreground objects alone. First, we extract the spatial masks of foreground objects outlined in [64]. Features of the masked foreground objects are learned to align with sounds similar to AVTS case. The results of this first part are in the supplementary material. Secondly, we use motion flow features. Based on our assumption, all the sound making objects in the scene are in motion. Indeed, this promotes self-supervision by aligning motion flow features with sound features. Our experiments with second approach are discussed in Section 5. Coming to the loss, we define it as: $-log\frac{exp(v_p \cdot a_i/\tau)}{exp(v_p \cdot a_i/\tau) + \sum_{v_n \in P_i} exp(v_n \cdot a_i/\tau)}$

where v_p and v_n are video feature vectors from aligned (positive) and misaligned (negative) video segments respectively, with respect to sound segment feature a_i, and τ is a temperature hyperparameter. P_i represents the set of misaligned video features for a_i.

Temporal gap prediction (denoted as $\mathcal{C}$). This pretext task consists of estimating the time difference between any two

Name	Train	Eval	Task	# Size
OmniAudio [64]	✓		SSL	64K
OmniAudio		✓	SP, VR, S^3R	64K
ImageNet [13]	✓		SSL	1.28M
Pascal VOC [18]		✓	Image Classification	16K
COCO [45]		✓	Detection, Inst Segm	118K

Table 1: Datasets used for Multi-SSL pretraining and for sound and visual downstream evaluation.

sound segments that are randomly sliced from a single longer sound clip. For the task, we frame a model that takes 2 sampled sound segments as inputs and learn to estimate the distance in time between them quite similar to [62]. Specifically, let us assume the length of sliced sound clip be T and original clip be T_{max}. We extract two sound slices a_i and a_j such that $\Delta = |t_i - t_j|$. Here, t_i and t_j are timestamps of a_i and a_j and Δ is sampled from a uniform distribution, $U(0, T_{max} - T)$. This temporal gap between sound slices is normalized as $\delta = \Delta/(T_{max} - T) \in [0, 1]$. It is important to note that there is no temporal order between the two slices. We concatenate the sound representations $[a_i, a_j]$ into a single vector and we feed this vector into a fully connected feed forward network with a single hidden layer of size 64 that produces the scalar output $\hat{\delta}$. We train the model end-to-end so as to minimize a huber loss $L_{gap}(\delta, \hat{\delta})$ between the ground-truth and the predicted temporal gap.

3.2. Images

A number of recent research on self supervised learning [52, 69, 5] has demonstrated the benefits of using a discriminative contrastive loss on data samples. Contrastive learning can drive a variety of pretext tasks and we choose a few of them, each carrying out different mechanisms.

MoCo. Momentum Contrast (MoCo) applies contrastive loss to features at the image level. MoCo [28, 11] shows that unsupervised learning can be superior to its ImageNet-supervised counterpart in image classification and detection tasks. MoCo uses two encoders, an encoder and a momentum encoder and the encoded representations are called queries and keys, respectively. MoCo trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss. We consider a query and a key as a positive pair if they originate from the same image and are two random views under random data augmentation, and otherwise as a negative sample pair. MoCo is designed for learning global representations, based on which the model is fine-tuned later for image classifications. Recent researches show that local features can also be extracted and compared using contrastive learning.

Dense contrastive learning. DenseCL [66] proposes a self-supervised learning framework to handle dense prediction/labelling tasks. DenseCL is primarily viewed as a dense pairwise contrastive learning as opposed to the global image representation learning. To begin with, a dense projection head is defined that takes the backbone features as inputs and then produces dense feature vectors. By producing a dense output format, we maintain spatial information unlike the existing global projection head that outputs a single, global feature vector for each image. Further, we determine the positive sample for each local feature vector by extracting the correspondence across views of the same image. We then construct an SSL loss function by extending the conventional InfoNCE loss [31] to a dense paradigm i.e., dense contrastive loss. We then perform contrastive learning densely using a fully convolutional network (FCN) [46], and the pretrained network is used to target dense prediction tasks.

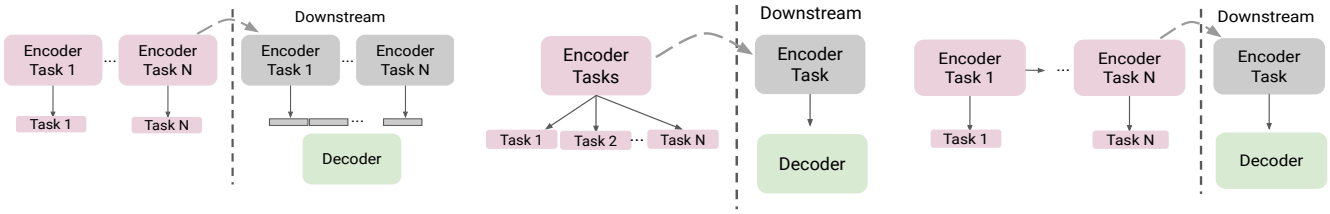
4. Multi-SSL pretraining

Several existing self-supervised learning (SSL) approaches choose a self-supervision objective based on the downstream task. The aim of our study is to determine whether we can combine Multiple SSL (Multi-SSL) tasks to simultaneously train a single encoder network. Furthermore, this encoder representations yield better downstream performance. Combining these tasks fairly in a multiple task learning objective is challenging and we discuss how we overcome this problem in subsequent sections. We call this approach as Multi-SSL. In our experiments, we investigate whether multiple SSL tasks extract general feature representations more effective than single task ones. Additionally, we examine which combination of SSL tasks are more beneficial and give a notable boost. In this section, we will discuss three prominent ways of combining multiple tasks we investigated and briefly summarize few other methods. Whenever possible, we follow the procedures established in the previous works, although in many cases modifications are necessary for our multiple task pretraining.

Let us assume we have K SSL tasks and N downstream tasks as shown in Figure 1. For single SSL task pretraining, we have different encoders Enc_k trained for each task to yield feature representations $\{f_1, f_2, ..., f_K\}$. Let us denote decoder for n_{th} downstream task D_n as d_n.

4.1. Concatenation

Firstly, let us consider a naive way of combining multiple tasks under Multi-SSL pretraining. Here, our approach is to concatenate the encoder features from K encoders. We train specific encoder networks Enc_k for specific SSL tasks separately. Each SSL task is designed to suit a particular downstream task D_n as we see in Section 3. Different features of encoder can complement each other. Here, we investigate the approach of concatenation of all SSL encoder features f_k to combine multiple SSL tasks. Further, these concatenated features $[f_1, f_2, ..., f_K]$ are frozen and passed



(a) Concatenation (b) Multi-tasking (c) Incremental learning

Figure 2: Different ways of combining multiple self-supervised approaches. Left side of each subfigure indicates the Multi-SSL pretraining and right side depicts the downstream task training and evaluation. Gray blocks denote frozen part.

to downstream task specific decoders d_n, where we just train the decoders as shown in Figure 2(a). The downstream task decoder d_n chooses the appropriate input features from concatenated features based on its task. However, this approach has its own limitation. In proportion to the number of SSL tasks added to *Concatenation*, the number of encoder trunks and concatenated features increase proportionally. In the next subsections, we can see how a single encoder trunk is learned for Multi-SSL approach.

4.2. Multi-task

Multi-task setting occurs when multiple tasks are combined with the goal of improving all tasks simultaneously through the sharing of common knowledge. Here, in our work, multiple SSL tasks (K tasks) are trained simultaneously to learn shared representations. As discussed in Section 3, we have three SSL tasks for learning sound representations and two for learning image representations. Inspired from the work of [16], we employ a base feature encoder trunk that are shared among all SSL tasks. The encoder features are further passed as input to the task-specific output heads as in Figure 2(b). These output heads are subjected to different SSL losses as discussed in Section 3. Here, we assign each SSL approach a separate task and train them jointly in a multi-task setting. To train the model, we use the weighted sum of loss from all K SSL tasks. Later, we use the encoder trunk of the trained model under multi-tasking for the downstream task training.

4.3. Incremental learning (IL)

Continual learning is another popular paradigm to learn new tasks one after the other, which we employ as another Multi-SSL approach for SSL tasks. Learning without forgetting [44] focuses on learning new tasks while preserving the performance of old tasks. Inspired from them, we learn multiple SSL tasks sequentially in an incremental manner. This means that we keep the same base encoder trunk model for all K tasks and attach task-specific output heads for each SSL task as shown in Figure 2(c). We learn the first SSL method using its task-specific layers and the responses are saved. Keeping the pretrained base trunk from the first task, we add task-specific layers of the second task and train it. During this training of the second task, we also retrain the first task's specific layers with their old responses. This way, we train the base trunk along with task specific layers of first and second task and backpropagate the combined loss from both tasks. In the same manner, we continue adding more SSL tasks to Multi-SSL IL to learn incrementally. Upon completion of all the tasks, we use the base encoder trunk of the trained model for the downstream tasks as shown in Figure 2(c).

We use the same dataset for all K SSL tasks in Multi-SSL, as shown in Table 1. Details are in Section 5. Hence, we store the output responses from task specific layers for all the tasks once the task is completed. Whenever we learn a new task with new task-specific layers, we use 2 kinds of losses. For the loss of the current task, groundtruth output from the current task is used anad secondly, for the loss of all the previous tasks, we use their stored output responses as groundtruth.

4.4. Other methods

As part of Multi-SSL, we also attempt other approaches. For *Euclidean dist* and *Contrastive dist* of Table 2, we learn separate models for each SSL task. Each data point in the dataset is passed through the learned model and feature representations are extracted for each SSL task and are stored. Then, we learn a new base encoder trunk model which is learnt to output feature representation using their corresponding losses. We apply L2 loss between base trunk features and stored latent representations of K SSL tasks to train the base model for Euclidean dist. In the case of *Contrastive dist:*, contrastive loss is applied which pulls together base trunk features and stored SSL representations of positive pairs while pushing apart latent representations of misaligned data points. Further, we investigate ProgressiveNet [61] as in Table 2 which is another continual learning approach. For single SSL tasking, we adopt baseline2 approach in [61] and later we follow the same work to add more SSL tasks.

Table 2: Single-SSL in (a) and Multi-SSL methods in (b). All methods are trained and evaluated on OminAudio dataset.

SSL	Downstream tasks		
	SP↑	S^3R↓	VR↑
Sup	26.82	0.2085	-
$\mathcal{A}$	15.32	**0.2105**	9.13
$\mathcal{B}$	**24.33**	0.2501	**27.35**
$\mathcal{C}$	16.85	0.2931	20.44

(a) Sound representations from 3 SSL tasks are evaluated on 3 downstream tasks. $\mathcal{A}$:Spatial alignment, $\mathcal{B}$: Foreground alignment, $\mathcal{C}$: Temporal gap prediction.

Multi-SSL Methods	Semantic prediction↑		Video Retrieval ↑		S^3R ↓	
	$\mathcal{B}$+$\mathcal{C}$	$\mathcal{B}$+$\mathcal{C}$+$\mathcal{A}$	$\mathcal{B}$+$\mathcal{C}$	$\mathcal{B}$+$\mathcal{C}$+$\mathcal{A}$	$\mathcal{B}$+$\mathcal{C}$	$\mathcal{B}$+$\mathcal{C}$+$\mathcal{A}$
Euclidean dist	25.38	25.82	27.64	27.45	0.2607	0.2188
Contrastive dist	25.59	27.39	27.96	27.95	0.2589	0.2145
ProgressNet [61]	30.38	32.45	29.06	29.68	0.2397	0.2035
Concatenate	26.37	30.14	28.31	28.09	0.2495	0.2101
Multi-task	27.28	31.21	28.94	29.52	0.2411	0.2066
IL	**32.76**	**34.05**	**29.72**	**30.32**	**0.2378**	**0.1988**

(b) Different methods of Multi-SSL approaches are evaluated on two-task $\mathcal{B}$+$\mathcal{C}$ and three-task $\mathcal{B}$+$\mathcal{C}$+$\mathcal{A}$ combination.

5. Experiments

In this section, we elaborate the experiments on transferability of single SSL and Multi-SSL models to different downstream tasks. The details of sound and image downstream tasks are in the Section 5.1 and Section 5.2 respectively and we discuss about the results of single SSL in Section 5.3 and Multi-SSL in Section 5.4. In addition, we conduct extensive ablation studies on combination of SSL tasks in Multi-SSL in Section 5.5.

5.1. Sound Downstream tasks

We apply three SSL tasks to extract initial sound representations, as discussed in Section 3.1. We use OmniAudio dataset [64] for all SSL tasks. Later, for the evaluation of SSL and Multi-SSL models, we consider three downstream tasks as described below, which tests the diversity and generalizability of sound representations.

Settings. We use OmniAudio dataset [64] for training and testing of downstream tasks. For all the experiments, we collect training and testing samples of 2-second video segments and a pair of binaural sound channels. We preprocess sound samples following techniques from [20, 64]. More details are added in the supplementary material.

Video retrieval (VR) is a common downstream task that aims to retrieve relevant videos based on a given sound clip. Following standard practices, we extract the sound representations from pretrained models on OmniAudio dataset [64], and measure the top-1 accuracies of retrieving the video segment, obtained for a single sound segment.

Semantic prediction (SP) [64] is a downstream task for binaural sounds which deals with the prediction of semantics of sound-making objects as pixel-level labelling task given the binaural sounds. We use the base encoder trunk model of sound network pretrained on SSL approaches and attach a decoder and train the whole model to predict the semantic segmentation masks of 5 classes- bus, car, tram, motorcycle and trucks.

Spatial sound super resolution (S^3R) [64] is a downstream task which aims to increase the directional resolution of sounds. This can be another testbed for binaural sounds. Here again, we attach a specific decoder to the encoder trunk pretrained on SSL approaches. Later, the model is trained and evaluated for S^3R task.

5.2. Visual Downstream tasks

We train two kinds of SSL approaches for visual representations as discussed in Section 3.2. We use ImageNet [13] dataset with 1.28M images for training SSL tasks. For the evaluation of these SSL tasks and Multi-SSL models, we investigate on three downstream tasks. We use Pascal VOC [18] on image classification and MS COCO datasets [45] on object detection and instance segmentation.

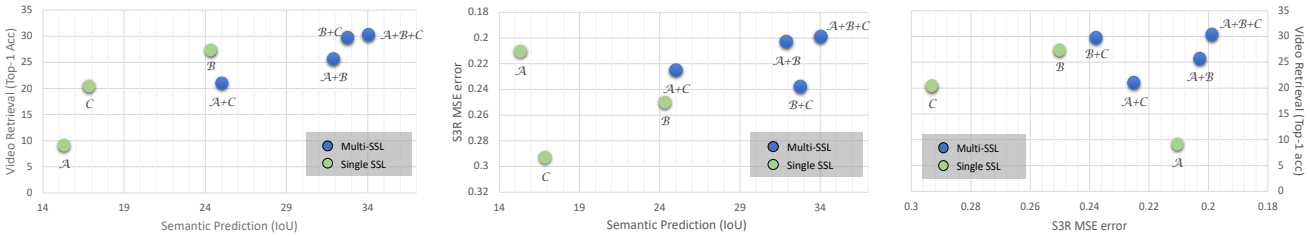
Settings For image representation learning on ImageNet [13], we follow the settings from [11, 28]. A ResNet50 [30] is adopted as the backbone. The global projection head in MoCov2 [28] and dense projection head in DenseCL [66] have a output of 512D feature vector and dense 512D feature vectors respectively. We adopt SGD as the optimizer and set its weight decay and momentum to 0.0001 and 0.9. We train for 200 epochs. More details are in the supplementary material. Let us now move to the standard image downstream tasks.

Image classification. We investigate results of image classification on Pascal VOC dataset [18]. We follow [25] and train linear SVMs using the feature representations extracted from the frozen encoder pretrained on SSL tasks. Finally, we evaluate on the VOC07 top-1 accuracy as in Table 4.

Detection and instance segmentation We evaluate our pretrained SSL and Multi-SSL models on object detection and instance segmentation. We train a Mask R-CNN detector [29] using pretrained FPN-backbone on COCO `train2017` split and evaluate on COCO `val2017` split. Table 4 reports the results of object detection and instance segmentation results on COCO dataset.

Table 3: Ablation studies on different combination of SSL tasks are provided. SSL tasks are $\mathcal{A}$: Spatial alignment, $\mathcal{B}$: Foreground alignment, $\mathcal{C}$: Temporal gap prediction. Different methods of Multi-SSL approaches like Concatenation, Multi-task and Incremental Learning and for evaluation downstream task of semantic prediction, video retrieval and S^{3}R is considered. Arrows indicate whether higher or lower is better.

Self-supervised tasks			Semantic prediction ↑			Video Retrieval ↑			S^{3}R ↓		
$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	Concat	Multi-task	IL	Concat	Multi-task	IL	Concat	Multi-task	IL
✓	✓		29.55	29.23	31.89	24.47	25.10	25.65	0.2165	0.2073	0.2029
✓		✓	23.57	24.73	25.02	19.03	20.22	21.00	0.2347	0.2308	0.2251
	✓	✓	26.37	27.28	32.76	**28.31**	28.94	29.72	0.2495	0.2411	0.2378
✓	✓	✓	**30.14**	**31.21**	**34.05**	28.09	**29.52**	**30.32**	**0.2101**	**0.2066**	**0.1988**



(a) Semantic Pred vs VR (b) Semantic Pred vs S^{3}R (c) S^{3}R vs VR

Figure 3: This presents the trade-off between semantic prediction, video retrieval accuracy and S^{3}R performance of our single SSL tasks (green) and Multi-SSL approaches (blue). All the methods are trained and evaluated on OmniAudio dataset.

5.3. Learning with single SSL

We tabulate the downstream task results of single SSL tasks for sound and image representation learning in Table 2(a) and 4 respectively.

Good performers in Single SSL tasks. In Table 2(a), we see that SSL task $\mathcal{A}$ performs best in S^{3}R task compared to SSL tasks $\mathcal{B}$ and $\mathcal{C}$. For instance, $\mathcal{A}$ gets S^{3}R error of 0.2105 which is 0.039 and 0.092 lower than other single SSL counterparts. This can be because aligning binaural sounds with spatial video cues in $\mathcal{A}$, help in learning representations variant to spatial directions. These representations further help in learning the directional resolution of sounds. Coming to other tasks, $\mathcal{B}$ outperforms other two tasks ($\mathcal{A}$ and $\mathcal{C}$) in semantic prediction and video retrieval performance as in Table 2(a). This is because, learning to align the foreground objects and binaural sounds may have allowed the sound representations to capture the semantics and spatial cues of objects. Coming to visual part, we see that DenseCL performs good in object detection and instance segmentation performance with COCO, compared to MoCov2 as in Table 4 while the latter performs better in image classification on VOC07. This has clear explanation in previous work [66] that MoCov2 captures global image features while DenseCL captures the local ones.

Supervised models vs single SSL. As we note in Table 2(a), supervised pretrained models pretrained on AudioSet [33, 21], perform better than single SSL tasks. For *eg.,* $\mathcal{B}$ gets semantic prediction mean IoU of 24.33 while $\mathcal{A}$ has an S^{3}R error of 0.2105 and these are less than the supervised counterparts with 26.82 and 0.2085 respectively. Contrasting to this, visual SSL of contrastive learning approaches MoCov2 and DenseCL in Table 4, surpass supervised models that are pretrained on ImageNet [13], by 0.2% in VOC07 classification and 1.34 on AP and 0.82 on AP^{mk}.

5.4. Learning with Multi-SSL

Different approaches of Multi-SSL models are tabulated for sound and image representations in Table 2(b) and Table 4 respectively on their respective downstream tasks.

Different Multi-SSL methods. *Concatenation* and *Multi-task* approach with SSL tasks $\mathcal{B}$ and $\mathcal{C}$, as discussed in Section 3.1, achieve 26.37 and 27.28 respectively. We further adopt the work of ProgressiveNet [61] where we use baseline2 approach of [61] which shows promising performance of 30.38 on tasks $\mathcal{B}$ & $\mathcal{C}$ as in Table 2. Then, we introduce incremental learning (IL) that achieves 32.78 mean IoU with $\mathcal{B}$+$\mathcal{C}$. Further, we see that IL with $\mathcal{B}$+ $\mathcal{C}$+$\mathcal{A}$ scores mean IoU of 34.05 that outperforms all other methods with a huge margin. For video retrieval and S^{3}R tasks too, we see that IL approach performs better than other Multi-SSL models. Coming to image downstream tasks, IL approach for image representations performs better than *Multi-task* and *Concatenation*, with an improvement of 0.81 AP and

Table 4: Evaluation of image classification on VOC07, Object detection and instance segmentation on COCO dataset using pretrained Multi-SSL models from ImageNet for 200 epochs, having ResNet50 as the trunk. For downstream, we use Mask R-CNN detector (FPN-backbone).

Downstream ⟶	VOC07	COCO detection			COCO instance segm		
Pretrain Tasks	Acc	AP	AP_{50}	AP_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Supervised	84.12	38.92	59.55	42.83	35.40	56.60	38.14
DetCo [70]	85.19	40.21	61.11	43.84	36.36	58.12	38.97
MoCov2(M) [28]	84.32	38.99	59.78	42.57	35.64	56.62	38.02
DenseCL(D) [66]	83.11	40.26	59.92	44.35	36.22	57.61	38.78
MTL(M+D) [66]	85.82	41.01	60.96	44.66	36.41	57.84	39.15
Concat(M+D)	85.54	40.67	60.59	43.98	36.61	58.18	39.20
IL(M+D)	**86.38**	**41.82**	**62.02**	**45.01**	**37.21**	**59.10**	**39.93**

1.15 AP respectively on COCO detection in Table 4. We notice similar trend with VOC07 accuracy and COCO instance segmentation.

Multi-SSL improves over single SSL task. We see that IL of $\mathcal{B}$+$\mathcal{C}$+$\mathcal{A}$ clearly outperforms best performing single SSL tasks with **+9.72** mean IoU and **+2.97%** video retrieval top-1 accuracy higher than $\mathcal{B}$ and **0.01** S^3R error lower than $\mathcal{A}$ as in Table 2. Further, we plot single SSL and Multi-SSL IL approach as a tradeoff between different downstream task performance in Figure 3. We notice that $\mathcal{B}$+$\mathcal{C}$+$\mathcal{A}$ outbeats single SSL tasks and two task combination in the tradeoff between all downstream tasks. Further we notice in Table 4 that IL approach has significant advantages over MoCov2 and other recent methods *e.g.* DenseCL [66] and DetCo [70] by **+2.83**, **+1.56** and **+1.61** AP on COCO detection. On instance segmentation, IL outperforms MoCov2 and DenseCL by +1.57 and +0.99 on AP^{mk}. On image classification, IL is also 2.06% and 3.27% higher than MoCov2 and DenseCL on VOC07 accuracy. In Figure 1, we see that IL achieves the best performance trade-off on both classification and detection unlike SSL tasks.

Multi-SSL outperforms supervised counterparts. We notice that Multi-SSL models improve over supervised models both in sound and image downstream tasks. In Table 2(a,b), Multi-SSL IL method significantly outperforms over supervised models, especially +7.23 on mean IoU and -0.01 on S^3R error. Coming to Table 4, we see that IL(M+D) is higher than ImageNet supervised models by 2.26% in VOC07 accuracy, 2.57 on AP_{50} and 1.81 on AP^{mk}.

5.5. Ablation studies

More SSL tasks in Multi-SSL better the performance: We perform ablation studies on different combination of SSL tasks in Multi-SSL for sound representations in Table 3. We showcase the performance of different Multi-SSL methods, *i.e., Multi-task, Concatenation* and IL using 3 downstream tasks. In Table 3, Concatenation of $\mathcal{A}$+$\mathcal{B}$ get a mean IoU of 29.55 and it boosts to 30.14 when $\mathcal{C}$ is added. This can be observed with $\mathcal{B}$+$\mathcal{C}$ and $\mathcal{A}$+$\mathcal{C}$. With *Multi-task* approach, we see that joint training of all 3 SSL tasks achieves 31.21 mean IoU bettering all other combinations of tasks. Coming to IL approach, $\mathcal{B}$+$\mathcal{C}$ scores 32.76 mean IoU which is higher than the single task $\mathcal{B}$ and $\mathcal{C}$. Finally, when $\mathcal{A}$ is added to IL model, performance improves to 34.05 in mean IoU which outbeats two tasks combination and single task performance as in Table 3. We see the same trend with video retrieval top-1 accuracy and S^3R error measures as well. This shows that adding more SSL tasks to Multi-SSL approaches improve the downstream performance.

Incremental learning approach for general features: From Figure 1 and 3, we note that single SSL task display good performance with one or two downstream tasks for which they are designed but not on all the tasks. For instance, in Table 2(a), Task $\mathcal{A}$ performs best in S^3R task but poor in other tasks.Task $\mathcal{B}$ performs good in video retrieval and semantic prediction while $\mathcal{C}$ shows mediocre performance in all three downstream tasks. In contrast, we see in Table 3 that Multi-SSL performs well in all three downstream tasks. Figure 3 more clearly distinguishes the performances of single task and Multi-SSL (IL), displaying the tradeoff between downstream task performance. We note that Multi-SSL (IL) outperforms on both the tasks in all subfigures of Figure 3. We see the same trend in Figure 1 showing its generalizability. In addition, Multi-SSL IL shows advantages over *Multi-task* and *Concatenation* here. Overall, above experiments indicate that Multi-SSL approach extract more generic features that perform well in all the downstream tasks.

5.6. Limitations and future works

Our paper investigates a handful of approaches of Multi-SSL. We strongly believe that this work will open up more directions to combine SSL tasks and this can raise SSL benchmarks. Further, in our work, the tasks are learned sequentially in Multi-SSL models and this results in more training time than single SSL tasks. For Multi-SSL models, we limit to 3 SSL tasks during the experiments. In future, it would be interesting to bring in more tasks and examine affinity mapping between combination of SSL tasks and downstream tasks.

6. Conclusion

This work proposes different approaches of Multi-SSL that learn to combine multiple SSL tasks. Experiments on OmniAudio dataset show that Multi-SSL via incremental learning outperforms all single SSL tasks and supervised models on the downstream tasks of semantic prediction, video retrieval, and spatial sound super resolution. We see similar trends of Multi-SSL on image representation learning using MoCov2 and DenseCL as SSL tasks. It demonstrates state-of-the-art performance on VOC classification, COCO detection and instance segmentation.

Acknowledgement: This work is funded by Toyota Motor Europe via the research project TRACE-Zurich.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020. 2, 3

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 3

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 2, 3

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29:892–900, 2016. 2, 3

[5] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019. 2, 4

[6] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 2

[7] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2940–2949, 2016. 2

[8] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 2

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 4, 6

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 4, 6, 7

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[15] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2, 3

[16] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017. 1, 2, 5

[17] Ariel Ephrat and Shmuel Peleg. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5095–5099. IEEE, 2017. 3

[18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 4, 6

[19] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 3

[20] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 324–333, 2019. 3, 6

[21] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 7

[22] Golnaz Ghiasi, Barret Zoph, Ekin D Cubuk, Quoc V Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8856–8865, 2021. 2

[23] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2, 3

[24] Ritwik Giri, Srikanth V Tenneti, Fangzhou Cheng, Karim Helwani, Umut Isik, and Arvindh Krishnaswamy. Self-supervised classification for detecting anomalous sounds. In *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pages 46–50, 2020. 2

[25] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6391–6400, 2019. 6

[26] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 2

[27] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*, 2020. 2, 3

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2, 4, 6, 8

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[31] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020. 2, 4

[32] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. *arXiv preprint arXiv:2103.10957*, 2021. 2

[33] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 7

[34] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 2

[35] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Köring, Suman Saha, and Luc Van Gool. Three ways to improve semantic segmentation with self-supervised depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[36] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[37] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Learning image representations by completing damaged jigsaw puzzles. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 793–802. IEEE, 2018. 3

[38] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018. 2, 3

[39] Volodymyr Kuleshov, S Zayd Enam, and Stefano Ermon. Audio super resolution using neural networks. *arXiv preprint arXiv:1708.00853*, 2017. 2

[40] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. 2

[41] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017. 2

[42] Wonhee Lee, Joonil Na, and Gunhee Kim. Multi-task self-supervised object detection via recycling of bounding box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4984–4993, 2019. 3

[43] Xiaoni Li, Yu Zhou, Yifei Zhang, Aoting Zhang, Wei Wang, Ning Jiang, Haiying Wu, and Weiping Wang. Dense semantic contrast for self-supervised visual representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1368–1376, 2021. 1, 2

[44] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 5

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 4, 6

[46] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4

[47] Andrés Marafioti, Nathanaël Perraudin, Nicki Holighaus, and Piotr Majdak. A context encoder for audio inpainting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2362–2372, 2019. 3

[48] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2

[49] Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. *arXiv preprint arXiv:2011.01819*, 2020. 3

[50] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 3

[51] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5898–5906, 2017. 2

[52] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 4

[53] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 3

[54] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016. 3

[55] Santiago Pascual, Mirco Ravanelli, Joan Serra, Antonio Bonafonte, and Yoshua Bengio. Learning problem-agnostic speech representations from multiple self-supervised tasks. *arXiv preprint arXiv:1904.03416*, 2019. 3

[56] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[57] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018. 2

[58] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 2

[59] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993. IEEE, 2020. 2, 3

[60] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1153, 2021. 1, 2

[61] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 5, 6, 7

[62] Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, and Dominik Roblek. Self-supervised audio representation learning for mobile devices. *arXiv preprint arXiv:1905.11796*, 2019. 2, 3, 4

[63] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[64] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Semantic object prediction and spatial sound super-resolution with binaural sounds. In *European Conference on Computer Vision*, pages 638–655. Springer, 2020. 2, 3, 4, 6

[65] Shaolei Wang, Wangxiang Che, Qi Liu, Pengda Qin, Ting Liu, and William Yang Wang. Multi-task self-supervised learning for disfluency detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9193–9200, 2020. 2

[66] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 1, 2, 3, 4, 6, 7, 8

[67] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018. 3

[68] Ho-Hsiang Wu, Chieh-Chi Kao, Qingming Tang, Ming Sun, Brian McFee, Juan Pablo Bello, and Chao Wang. Multi-task self-supervised pre-training for music classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 556–560. IEEE, 2021. 2

[69] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2, 4

[70] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021. 8

[71] Salah Zaiem, Titouan Parcollet, and Slim Essid. Pretext tasks selection for multitask self-supervised speech representation learning. *arXiv preprint arXiv:2107.00594*, 2021. 2

[72] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2547–2555, 2019. 2

[73] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[74] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 2

[75] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306, 2019. 3

[76] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 283–292, 2019. 3

[77] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3550–3558, 2018. 3