OmniVec: Learning robust representations with cross modal sharing

Siddharth Srivastava, Gaurav Sharma
TensorTour Inc.
{siddharth, gaurav}@tensortour.com

Abstract

Majority of research in learning based methods has been towards designing and training networks for specific tasks. However, many of the learning based tasks, across modalities, share commonalities and could be potentially tackled in a joint framework. We present an approach in such direction, to learn multiple tasks, in multiple modalities, with a unified architecture. The proposed network is composed of task specific encoders, a common trunk in the middle, followed by task specific prediction heads. We first pre-train it by self-supervised masked training, followed by sequential training for the different tasks. We train the network on all major modalities, e.g. visual, audio, text and 3D, and report results on 22 diverse and challenging public benchmarks. We demonstrate empirically that, using a joint network to train across modalities leads to meaningful information sharing and this allows us to achieve state-of-the-art results on most of the benchmarks. We also show generalization of the trained network on cross-modal tasks as well as unseen datasets and tasks.

1. Introduction

Many applied machine learning methods aim to extract useful representations from data. However, a majority of such methods are modality and task specific. Building methods that can work with multiple modalities is a relatively recent research direction [25, 36, 37, 46, 62, 105]. Learning tasks together with a unified network can lead to regularization effects, as a large amounts of shared parameters are trained to perform varied tasks, and hence are more likely to extract meaningful representations from data without overfitting to one task or modality. It can also aid in utilizing available labelled data from different domains, hence potentially eliminating the cost and effort of labelling large amounts of data in a specific modality for a specific task. With the ability of sharing knowledge from multiple modalities (e.g. image, video, depth map and speech) from different domains (e.g. visual, acoustic, textual), the modality agnostic learning frameworks have been shown to provide better robustness [1, 28] to traditional unimodal networks. We contribute to that line of work, and develop a framework that can learn embeddings in a shared space from different modalities and also deliver high generalization performance. Specifically, we propose to learn embeddings from distinct modalities with modality specific encoders, and process them with a shared transformer backbone. The transformer backbone maps the input embeddings to a shared embeddings space. The network is then trained in an end-to-end manner.

Prior works towards generalized modality agnostic learning can be categorized into following three approaches. (i) Methods which directly take multiple heterogeneous modalities (image, 3D, audio) as input, without separate encoders for each modality, and directly learn representations from them [36, 37]. (ii) Methods that take representations from modality specific encoder as input and learn generalized modality specific embeddings using a common objective in the latent space [7], and, (iii) Methods which aim at sharing knowledge among different modalities by keeping either a common encoder [25] or separate encoders [1]. The first two approaches generally target modality agnostic input representation, which lend them capability to keep the network definition same for different modalities. However, such networks, in general, can be trained on one modality at a time, and hence do not facilitate cross modal knowledge sharing. On the other hand, the third approach facilitates jointly training networks on multiple modalities. Our work is closer to the third set of approaches. Specifically, similar to [7], the proposed method employs different encoder for each modality. Similar to [25] we share knowledge among modalities, and train on multiple modalities sequentially allowing embeddings to generalize across modalities. Unlike [25], we do not limit our method to a specific subset of modalities. and train on multiple modalities in a sequential manner. Further, we do not assume any correspondence between the training data i.e. paired training sets across modalities, which is different from previous works, e.g. [1], where correspondence in data among modalities is assumed.

Our proposed framework, OmniVec, consists of the following components: (i) a modality specific encoder, (ii) a shared backbone network, and (iii) task specific heads where tasks can be any machine learning task. The framework facilitates end-to-end training. In simple terms, OmniVec works as follows. For a given task and a modality we select a modality compatible encoder and an appropriate task head. We attach the encoder and task heads to, the beginning and end of the shared backbone network respectively. Then to train on another modality, we replace the encoder while keeping the backbone same. If the task is to be changed as well, we replace the task head. To further facilitate learning of better representations and cross-modal information sharing, we train the network numerous tasks. We borrow the motivation from earlier works where it has been shown that training networks on multiple related tasks can provide better generalization [103]. Similar improvements, in generalization, have been reported for multi-modal multi-task learning as well [16, 34, 59, 69]. However, we do not train in a traditional multi-task setting, where all tasks are available together and are trained for together. Instead, we train the network in a sequential manner, i.e. we train on different tasks, one after another.

Motivated by empirical observations and previous works indicating that robustness of multi-task mechanisms depends on the complexity of tasks selected for joint training [59, 71], we propose to group the tasks based on the extent of information exploited by the task across different modalities, e.g., a semantic segmentation task forces the network to embed more local information in the learned representation, as compared to a classification task [15]. In addition to grouping the tasks, we also construct training data by mixing samples from each modality for a particular task. We train the network by replacing modality encoder for each modality, while keeping the task heads and backbone network same. Based on earlier works indicating that self-supervised pretraining helps networks in better exploiting multiple modalities [16, 24], we pretrain the network with masked pretraining.

In summary, we make the following contributions. (i) We propose a novel method to learn embeddings from many modalities. The method has a common backbone to process the different modalities and perform different tasks. Specifically, we show that the proposed method works with RGB images and videos, depth images, point clouds, audio, speech and text data. (ii) We propose a novel training mechanism to allow learning using multiple tasks from both spatial (e.g. image, 3D point clouds, depth maps) and temporal (e.g. video, audio, speech, text) data. Owing to the common backbone of the method, and a synchronous training mechanism, the method shares knowledge between different modalities and tasks, resulting in improved performance and generalization. (iii) The proposed method allows for infusing cross domain information in the feature vectors, i.e. allowing embeddings from text data to be close to similar data in image domain. (iv) We propose an iterative training mechanism by mixing modalities and grouping tasks. Different from earlier works, we also propose to perform self supervised masked pretraining across visual as well as non visual modalities. (v) With exhaustive experiments on numerous popular benchmarks across, we show that the proposed framework achieves state-of-the-art results or performs close to the competing methods. (vi) We also study the generalization ability of the proposed framework by demonstrating the robust performance of the learned embeddings on unseen tasks. (vii) We conduct an extensive ablation study to demonstrate the impact of the design choices.

2. Related Works

In this section, we discuss similar works and various similar paradigms to our work. We begin with transformers, which are basis of our work, and then move to methods which work with multiple modalities. Among methods that work with multiple modalities, many of them work on utilizing the modalities simultaneously, while others propose networks which take the modalities as input, one at a time.

Transformers. Transformers were proposed originally for Natural Language Processing tasks [78]. The main contribution of this work was to demonstrate the effectiveness of multi-head attention in representing long-range correlation between words. Owing to the popularity of transformers in NLP tasks [49], attempts were made to extend it to vision tasks. Early work in this direction [14, 53, 93] involved utilizing features from convolutional neural networks. However, with vision transformers [19], transformers obtained an ability to process raw images and achieved performance competitive to CNNs. After that, transformers have dominated nearly all the vision related tasks [41]. As transformers have demonstrated robust performance across modalities, recent methods across various modalities use them to solve various tasks [21, 49, 63, 64, 91]

Multi-modal methods. Majority of the current multimodal methods use modality specific feature encoders [2, 38, 39, 62, 88] and are hence concerned with methods of feature fusion with their proposed architectures. In general the networks for different modalities differ from each other and can not be easily used together without architectural modifications. They also need to decide on when to fuse the features from various modalities, when to fine-tune, how to pre-train etc. [90]. Such problems inhibits extending networks such as transformers to be applied as a common backbone across multiple domains such as point clouds, audio and images.

Common network for multiple modalities. Recently, many methods have been proposed which learn from multiple modalities [7, 8, 25, 37]. Among the most popular, however recent, are methods that do not have separate encoders for each modality. Such methods generally transform the input raw data to a common input representation prior to generally being processed by a transformer network. Among them, the perceiver and similar methods [8,36,37] have tried to learn from multiple modalities together without separate encoders. Perceiver architecture works by cross-attention among a set of latent queries. Similarly, hierarchical perceiver [8] builds upon it proposes to group the input array while preserving the locality structure. On the other hand, methods such as data2vec [7] use modality specific encoders. Other methods such as Omnivore [25] have a common encoder. However, Omnivore is limited to only visual modalities (image, depth map, video). Then, methods such as VATT [1] have a common backbone for each text, image and audio. However, it processes each modality independently using a transformer. Such methods which learn from multiple modalities have been shown to provide better robustness [1, 28]. Our methods largely overlaps with the motivation of such methods, however, it differs from such methods in that earlier methods operate on training for one task or one modality at a time, while we learn by training on multiple modalities and multiple tasks while using a single common backbone architecture.

Multi-task learning. We have discussed many methods that attempt at learning from multiple inputs. As discussed in the previous section, recent years have seen many methods that work with multiple modalities. PerceiverIO [36] extends Perceiver [37] and enables learning multiple tasks using the same network architecture. While PerceiverIO can also learn multiple tasks at a time using a single architecture, generally multiple networks are used [102]. Many techniques [7, 16, 25, 34, 59] learn from multiple modalities and from their raw representation and apply to multiple tasks.

Multi-modal masked pretraining. Methods such as [52, 84, 92] use masked pre-training. Masked pretraining has shown to improve the performance of deep networks networks for various modalities and tasks [1,6,7,24,31,98] and motivated by such works we also use masked pre-training as a self supervised step leveraging large amounts of data available. However, different from earlier works, we perform masked pre-training on multiple modalities and multiple datasets on the same common backbone.

3. Approach

We now describe our framework for learning multiple tasks in multiple modalities with a common backbone network, allowing for cross modality knowledge sharing. The overview of the proposed framework is shown in Figure 1. The network comprises six building blocks, i.e. modality encoders, meta token block, projection block, transformer, vectorizer and task heads. We now explain each block in detail.

Modality	**Domain**	**Network**
Image	Visual	Vision Transformer (ViT) [19]
Depth maps	Visual	Vision Transformer (ViT) [19]
Video	Visual	Video Vision Transformer (ViViT) [4]
3D point clouds	Visual	Simple3D-former [82]
Audio	Auditory	Audio Spectrogram Transformer (AST) [27]
Text	Language	BERT [18]

Table 1. **Modality Encoders.** We select transformer based modality encoders for evaluating OmniVec framework

3.1. OmniVec Framework

Modality Encoder. The modality encoder takes as input, one modality at a time and extracts feature embedding for each of the modalities. In the proposed framework, the modality encoder can be a transformer, convolutional neural network or can directly use raw signals [1]. As we do not assume any specific structure for the modality encoder, the proposed framework allows incorporating any appropriate deep network as a modality encoder.

For current work, we use domain specific transformer based encoders for each of the modalities as shown in Table 1 followed by a common backbone network. It is worth noting that each of the networks in visual and auditory domain is based on Vision Transformer architecture i.e. image and depth directly use ViT, video (ViViT) differs from ViT in input tokenization that extends 2D patches to 3D (spatio-temporal mapping), audio (AST) transformers differ from ViT only in input representation i.e. uses log-mel spectrograms instead of images, Simple3D-former for point cloud uses a 2D ViT transformer as the base network with modified positional embeddings and tokenization approach. We use a standard BERT transformer for textual data. We train each of these models from scratch.

Meta Tokens. We extract meta tokens from the input modalities. This meta representation is a vector that encodes the type of modality (I), size of temporal dimension (T), height (H), width (W) in spatial dimension, number of channels (C) and length or number of tokens (L). In general, the meta tokens can also hold additional information to make the framework adapt to additional modalities. The value in each of these representation variables is conditioned on the type of modality e.g. non spatial data may have H, W only with the other non-spatial parameter set as a special token, denoting lack of such information.

Projection Layer. The projection layer inputs the intermediate representations from the modality encoder network and is conditioned on the meta tokens. It then converts the input representation to patches that are provided as input to the subsequent transformer network. We obtain n-dimensional vector for each patch by applying linear pro-

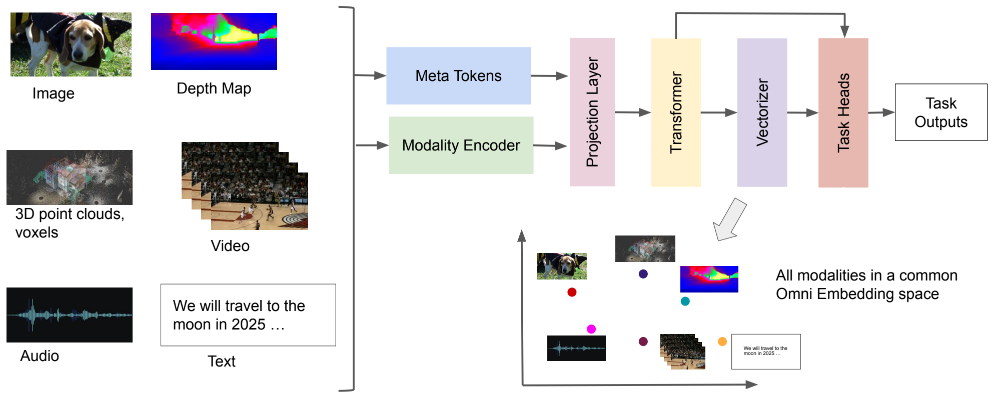

Figure 1. **OmniVec:** The proposed method takes data from one of the modalities and pass it through the modality encoder and combine it with the meta token and then pass through the projection layer to embedd the feature onto a common embedding space. Then it is passes through the common backbone of Transformer layers which is then vectorized by the vectorizer. Finally, the task heads are used for task specific outputs.

jection. Similar to ViT [19], this projection is applied with a learnable weight $W_{ip} \in \mathbb{R}^{t \cdot h \cdot w \cdot c \cdot l \times n}$ for each modality i. The meta tokens make the projection layer adaptable to varying number and dimensions of input patches and generate latent representations compatible with the subsequent transformer network. For instance, we represent RGB images as $I \in \mathbb{R}^{1 \times h \times w \times 3 \times 1}$ with t=1 frames and c=3 channels. Similarly, we represent video as $V \in \mathbb{R}^{t \times h \times w \times 3 \times 1}$ with t frames ($t > 1$) and $c = 3$ channels, depth as $D \in \mathbb{R}^{1 \times h \times w \times 4 \times 1}$ with $c = 4$ channels, point cloud as $P \in \mathbb{R}^{1 \times 1 \times 1 \times 3 \times l}$ with l points, audio as $A \in \mathbb{R}^{t \times h \times w \times c \times 1}$ with spectrogram input, and text as $L \in \mathbb{R}^{1 \times 1 \times 1 \times 1 \times l}$ with l tokens. Each patch $\mathbf{x}$ is processed independently and projected to an embedding $\mathbf{e}$ followed by a LayerNorm [5]

Transformer. The transformer network is the common part of the framework and is in effect a ‘bottleneck‘ block. While different modalities may arrive here through different encoders, they all have to pass through this transformer network. The transformer network inputs the patches generated by the projection layer and outputs features. While the OmniVec framework can use any standard transformer architecture, we use [18] as our backbone architecture. In our transformer network, the multi head attention involves standard self-attention [78], and GeLU [33] activation prior to the MLP layer.

Vectorizer The vectorizer layer takes patches from the transformer network as input, and outputs embeddings for the original data point. It outputs a single embedding $\mathbf{e} = f(\mathbf{X})$ for an input $\mathbf{X}$. We name the output embeddings of the vectorizer as Omni Embeddings, as these embeddings constitute knowledge from multiple tasks and modalities due to forward pass from the transformer block where cross modality and cross task information is infused.

For our implementation, we concatenate the output patches and pass them through a linear layer to obtain a d-dimensional embedding. At the time of training, we use the outcome of vectorizer as input to subsequent task heads. However, using the outcome of vectorizer as input to task heads is optional as the task head may also directly take input patches from the previous transformer bottleneck. Once the model has been trained, the output from vectorizer can be used for fine-tuning and evaluation on downstream tasks.

Task Heads The final parts of network, the task heads are $\sum T_{ih}$ independent networks which learn task h for every i^{th} modality. The task heads can generally be any computer vision, natural language processing or other modality specific task. We experiment with classification (image, video, audio, text), segmentation (image, point clouds) etc. We describe them in Section 4.

3.2. Training OmniVec Framework

We train the OmniVec Framework in two stages. First we perform masked pretraining. Then we fine tune the network on multiple modalities. Both these stages are described below.

Masked Pretraining. We pretrain the network with masked autoencoders [1, 24]. Specifically, for an input with N patches, we mask K patches, and feed non-masked patches and their positions to the encoder. For each modality, we use the encoder from Table 1 followed by our bottleneck transformer that outputs per patch embeddings i.e. we keep a shared bottleneck transformer encoder for each of the modalities. Similar to [1,32], the per patch embeddings are concatenated with K replicas of learnable mask tokens resulting in N embeddings. We add corresponding positional embeddings to each of the N embeddings, and pass to the decoder. We use the same masking strategy for modalities from visual and auditory domains. For textual data, we follow [66] and randomly permute the sentences [95] and use a small fraction f of tokens as predicted tokens, followed by utilizing 8:1:1 strategy of BERT [18] for constructing mask tokens. The training objective is to minimize the reconstruction error between the input and decoder outputs. For image, video, point clouds and audio spectrogram input, we minimize l_2 distance between the K predicted and target patches. For visual inputs, the input samples are normalized to zero mean and unit variance. For textual data, we use the permuted language modelling of XLNet [95] as

Method/Dataset	**Supp. Modalities**	**Cross-Modal sharing**	**Masked pretraining**	**Supp. Tasks**	**AudioSet (A+V.)**	**AudioSet (A)**	**SSv2**	**GLUE**	**ImageNet1K**	**Sun RGBD**	**ModelNet40**
Omni-MAE [24]	Image, Video	No	Yes	Class.	-	-	73.4	-	85.5	-	-
Perceiver [37]	Modality Agnostic	No	No	Class.	43.4	38.4	-	-	78.6	-	-
Heirarchical Perceiver [8]	Modality Agnostic	No	No	Class.	43.8	41.3	-	-	81.0	-	80.6
data2vec [7]	Modality Agnostic	No	Yes	Class.	-	34.5	-	82.9	86.6	-	-
Omnivore [25]	Image, Video, Depth map	Yes	No	Class.	-	-	71.4	-	84.0	65.4	-
VATT [1]	Image, Video, Audio, Text	Yes	Yes	Class.	-	39.4	-	-	-	-	-
Perceiver IO [36]	Modality Agnostic	No	No	Multiple	-	-	-	-	79.0	-	77.4
OmniVec (pretrained)	Image, Video, Audio, Text, Depth map, Point Clouds	Yes	Yes	Multiple	**48.6**	**44.7**	**80.1**	**84.3**	**88.6**	**71.4**	**83.6**

Table 2. **Comparison of OmniVec framework with similar methods that work on multiple modalities**. We compare OmniVec with masked pretraining with the best reported results from respective publications of the compared methods. Supp. Tasks and Supp. Modalities indicate Supported Tasks and Supported Modalities by respective networks. In Supported (Supp.) Tasks, Class. indicates classification.

the objective.

Training on multiple modalities and tasks. For training the network on multiple modalities and tasks, we introduce *modality mixing* and *task grouping*. We train our model using a collection of h tasks $\mathbf{T_{i,h}}$ for i^{th} modality. We group tasks into *simple* and *dense* tasks and refer to it as *task grouping*. We categorize the tasks into two categories namely, simple and dense based on the complexity of the dataset and outputs i.e. classification task predicts a single label for a given input, irrespective of the size of the input, therefore we refer it as a simple task. However, a segmentation or depth prediction task, requires each pixel to be predicted, and hence we refer it as a dense task. We detail each of the tasks, the datasets used to train them and their task grouping in Section 4.

As we do not assume any correspondence between data from various modalities, we propose mixing samples from all datasets for a particular task to share knowledge between various modalities. An alternative approach would be to construct mini-batches from each dataset separately. However, we found it performs poorly compared to mixing samples from modalities. We refer this strategy of constructing mini-batches as *modality mixing*. Specifically, for a particular task h belonging to a type of task t (simple, dense), for each modality i, we extract sample $s_{t,i,h}$ from the datasets.

After task grouping and modality mixing, we train the network in an end-to-end manner iteratively for simple and dense tasks. Specifically, we train the network for E epochs, we train the network for v_1 epochs with mini-batches from simple tasks and v_2 for dense tasks. We continue training the network in iterative manner i.e. switching between simple and dense tasks for E epochs.

4. Experiments

Masked pretraining. We do masked pretraining using the modality mixing as described in Section 3. We use AudioSet (audio) [23], Something-Something v2 (SSv2)(video) [30], English Wikipedia (text), ImageNet1K (image) [17], SUN RGB-D (depth maps) [67], ModelNet40 (3D point cloud) [87] for pretraining the network. As we perform autoencoder based pre-training, we do not group the tasks, and instead uniformly sample data from each of the datasets and modalities. Further, we randomly select patches for masking. For image, video and audio, we randomly mask 90% of the patches. For point cloud, we mask 80% of the patches, and for text we mask 95% of the patches. Further, we keep f = 5% of the tokens as predicted tokens (unlike 15% in [95]). We perform pretraining for 2000 epochs.

Modality Encoder. For modality specific encoders, we use the networks from Table 1. We use the same network configurations for these networks as in corresponding publications. We pretrain the model using masked pretraining as described in Section 3, followed by training on specific modalities as per task groups and modality mixing. For different tasks on a modality, we keep the modality encoder same, while changing the task heads with appropriate loss functions. We train modality encoders for $E = 900$ epochs with 2 consecutive epochs each for simple and dense task groups.

Datasets for training on multiple modalities and tasks. After masked pre-training, we fine tune the network on multiple tasks across modalities. The datasets and their corresponding task groups and modality are given in Table 7.

Task Heads. For classification tasks, we use standard classification head from ViT [19] while use [4] for video classification and [27] for audio classification. For image and point cloud segmentation tasks, we use the segmentation head from [61]. For text summarization, we use a 3-layered transformer.

We provide more implementation details in the supplementary material.

4.1. Results

Comparison of pretrained OmniVec with similar methods. Table 2 compares OmniVec model with masked pretraining to various similar methods. The table also indicates the modalities supported by various methods (Col.-Supp. Modalities), and that if the method supports sharing knowl-

Method/Dataset	iNaturalist 2018	Places 365
Omni-MAE [24]	78.1	59.4
Omnivore [25]	84.1	59.9
EfficientNet B8 [73]	81.3	58.6
MAE [32]	86.8	
MetaFormer [97]	87.5	60.7
InternImage [81]	92.6	61.2
OmniVec	**93.8**	**63.5**

Table 3. **iNaturalist-2018 and Places-365** top-1 accuracy.

Method/Dataset	Kinetics-400
Omnivore [25]	84.1
VATT [1]	82.1
Uniformerv2 [47]	90.0
InternVideo [83]	**91.1**
TubeViT [58]	90.9
OmniVec	**91.1**

Table 4. **Kinetics-400** top-1 accuracy.

Method/Dataset	Moments in Time
VATT [1]	41.1
Uniformer v2 [47]	47.8
CoCa [96]	47.4
CoCa-finetuned [96]	49.0
OmniVec	**49.8**

Table 5. **Moments in time** top-1 accuracy.

Method/Dataset	ESC50
AST [27]	85.7
EAT-M [22]	96.3
HTS-AT [10]	97.0
BEATs [55]	98.1
OmniVec	**98.4**

Table 6. **ESC50** top-1 accuracy.

Task	Dataset	Modality	Task Group
Image Recognition	iNaturalist-2018 [77]	Image	Simple
Scene Recognition	Places-365 [104]	Image	Dense
Video Action Recognition	Kinetics-400 [40]	Video	Simple
Video Action Recognition	Moments in Time [54]	Video	Dense
Audio Event Classification	ESC50 [57]	Audio	Simple
Point Cloud Segmentation	S3DIS [3]	Point Cloud	Dense
Text Summarization	DialogueSUM [13]	Text	Dense
Point Cloud Classification	ModelNet40-C [87]	Point Cloud	Simple

Table 7. **List of tasks and corresponding datasets for task group based training after masked pretraining**. We assign each task to a task group (simple, dense) based on complexity of the dataset and output.

edge between modalities (Col.-Cross-Modal sharing). Further, it also details the learning objectives by these methods. The table reports results on six benchmark datasets on seven tasks as AudioSet supports two tasks (audio only, and audio with video). These datasets are used to perform masked pretraining on the OmniVec model as described in Section 3. It can be observed that the proposed OmniVec model outperforms all the compared methods on all the datasets. It is important to note that, we do not fine tune on any of these datasets specifically while other methods, in general, fine tune the results, mostly using a linear layer with softmax classification. This demonstrates the robustness of the proposed model and its ability to learn generalized embeddings without task specific fine-tuning.

Comparison to state-of-the-art. For comparison with state of the art methods, we performed masked pretraining of OmniVec followed by training on multiple modalities and task groups as described in Section 3. We discuss the comparison on each modality below.

(i) Image Table 3 shows state of the art on image datasets. We compare with multi-modal methods (Omni-MAE, Omnivore) and specialized methods (MetaFormer, InternImage). We surpass the state of the art on iNaturalist with a top-1 accuracy of 93.8%, compared to InternImage's 92.6%. On Places-365, we beat all competitors, achieving 61.6% accuracy versus InternImage's 61.2%. Moreover, we best Omnivore by $\sim$ 7% on iNaturalist and $\sim$ 3% on Places-365. Our results either match or surpass modality-specific methods in image classification, and outperforming unified learning methods.

(ii) Video Table 4 and Table 5 show comparison against state of the art methods on Kinetics-400 and Moments in Time datasets.We observe that we outperform all the competing methods on Moments in Time dataset while perform same as the state of the art method InterVideo i.e. 91.9 top-1 accuracy.

(iii) Audio Table 6 highlights our comparison with top-performing methods on the ESC50 dataset. OmniVec outperforms competing methods, achieving an accuracy of 98.4%, significantly higher than the Audio Spectrogram Transformer (AST) at 85.7%. While most compared methods utilize supervised pretraining on AudioSet, we adopt masked pretraining without accessing labels. This suggests OmniVec's proficiency in learning from related tasks across different modalities, emphasizing its effectiveness in cross-modal knowledge transfer.

(iv) Point Cloud. Table 9 and Table 10 compare against state of the art methods on ModelNet40-C and S3DIS datasets respectively. On ModelNet40-C, we evaluate a classification task, while on S3DIS we evaluate semantic segmentation. On both the datasets, we outperform the competing method. This demonstrates that the proposed method is able to robust performance with the shared backbone network across tasks.

(v) Text Table 11 shows state of the art on DialogueSUM dataset for text summarization. OmniVec surpasses other methods in three out of four metrics and comes in second on the R-L metric. Despite utilizing significantly fewer datasets for text (only two) in comparison to visual tasks (ten datasets), OmniVec demonstrates strong performance. This suggests OmniVec's capacity to bridge the modality gap [48] across distinct domains in the latent space, even when the data distribution is skewed.

4.2. Ablations

Impact of task grouping and modality mixing. Table 8 shows the effect of task grouping and modality mixing. We evaluate four network variations: (i) OmniVec-1 without either of task grouping and modality mixing, (ii) OmniVec-2 with just task grouping, (iii) OmniVec-3 with only modality mixing, and (iv) OmniVec-4 combining both. OmniVec-1 uses masked pretraining on single datasets. OmniVec-

Method	Task Grouping	Modality Mixing	AudioSet (A+V.)	AudioSet (A)	SSv2	GLUE	ImageNet1K	Sun RGBD	ModelNet40
OmniVec-1 (baseline)	✗	✗	37.5	36.3	62.6	57.5	70.2	59.8	68.5
OmniVec-2	✓	✗	42.6	40.1	73.5	69.5	79.8	66.4	75.2
OmniVec-3	✗	✓	39.2	39.4	70.2	68.8	77.3	65.5	72.2
OmniVec-4	✓	✓	**48.6**	**44.7**	**80.1**	**84.3**	**88.6**	**71.4**	**83.6**

Table 8. **Impact of various training strategies on OmniVec.** We report results with and without each of task grouping and modality mixing. The results are reported with masked pretraining only. We observe that individually, both task grouping and modality mixing improve the results over the baseline method. However, there combination outperforms individual performance using these mechanisms.

Method/Dataset	Model Net40C
PointNet++ [60]	0.236
DGCN+PCM-R [100]	0.173
PCT + RSMIx [45]	0.173
PCT + PCM-R [72]	0.163
OmniVec	**0.156**

Table 9. **ModelNet40-C** Error Rate.

Method/Dataset	S3DIS
PointTransformer+CBL [74]	71.6
StratifiedTransformer [44]	72.0
PTv2 [86]	72.6
Swin3D [94]	74.5
OmniVec	**75.9**

Table 10. **Stanford Indoor Dataset** mIoU.

Method	R-1	R-2	R-L	B-S
CODS [85]	44.27	17.90	36.98	70.49
SICK [42]	46.2	20.39	**40.83**	71.32
OmniVec	**46.91**	**21.22**	40.19	**71.91**

Table 11. **DialogueSUM** text summarization ROGUE scores.

2 groups tasks by modality, OmniVec-3 mixes modalities randomly, and OmniVec-4 follows the settings from Section 3. Comparatively, OmniVec-1 lags behind the others. Both OmniVec-2 and OmniVec-3 outperform OmniVec-1 by around 30% to 45%, showing their efficacy. However, OmniVec-4, which combines both approaches, performs better, emphasizing the benefits of integrating tasks and modalities.

Influence of size of the modality encoder. We evaluated the impact of enlarging the base modality encoder to the scale of our suggested network, using modality-specific data. This change slightly improved performance. For example, on ImageNet1K, the top-1 accuracy went from 88.5% with the base ViT [19] to 89.1% with the augmented ViT having a similar parameter count, while OmniVec achieved 92.4%. These findings suggest that even with enhancements, the augmented base modality encoder still lags significantly behind OmniVec, highlighting OmniVec's advantage of leveraging information from multiple modalities.

Fine-tuning with the same datasets after masked pretraining and comparison to state-of-the-art. In Table 13, we show the results of fine-tuning the OmniVec-4 model on each of the datasets that was used for masked pretraining. As during masked pretraining, we use the standard train sets for each of these datasets for fine-tuning.

It can be observed from the results that OmniVec achieves better performance on each dataset than existing state of the art method. As we are using same backbone (OmniVec-4) for each of these datasets, it shows the robustness of the embeddings and the capacity of the network to adapt to different tasks and distribution of dataset.

4.3. Generalization Ability

Generalization on unseen datasets. We evaluate the performance of the learned embeddings on unseen datasets. Specifically, we show results on the tasks of fine grained image classification (Oxford-IIIT Pets [56]), Video Classification (UCF-101 [68], HMDB51 [43]), 3D point cloud classification (ScanObjectNN [76]), 3D point cloud segmentation (NYUv2 [65]) and text summarization (SamSum [26]). Our findings, tabulated in Table 12 [rows 1-6], demonstrates that even without fine-tuning, OmniVec surpasses most state-of-the-art methods. Further, while the pretrained OmniVec slightly underperformed on ScanObjectNN (92.1%) compared to PointGPT's 93.4%, when fine-tuned, OmniVec achieved 96.1% accuracy, outperforming PointGPT. This shows OmniVec's generalizability on datasets where it is exposed to analogous tasks.

Generalization on unseen tasks - Monocular Depth Prediction on KITTI Depth Prediction Benchmark. We fine tune the network for the task of depth prediction on KITTI Depth Prediction benchmark [75]. Our network has not seen such image to image style transfer tasks. The results on KITTI depth prediction benchmark are shown in Table 12 (row 7). We outperform the state of the method VA-DepthNet [50] i.e. 10.44 iRMSE on VA-DepthNet cf. 10.2 for OmniVec. As can be observed from Figure 2, the depth maps obtained by OmniVec are able to better capture the details near edges.

Cross-domain generalization. Following prior work [1], we evaluate on the task of zero-shot text-to-video retrieval. The results are reported in Table 12. On the YouCook2 dataset, our pretrained OmniVec surpasses the state of the art in zero-shot retrieval, achieving a Recall@10 of 64.2% compared to VideoCLIP's 63.1%. On MSR-VTT, when compared with SM [99], our fine-tuned OmniVec embeddings yield a Recall@10 of 89.4% against SM's 90.8%. With just pretraining, SM has a Recall@10 of 80%, slightly above our 78.6%. SM utilizes large-scale pretraining on internet scale data, while OmniVec uses much less data. Further, the second-best MSR-VTT method [11] achieves only

Dataset	Modality	Task	Metric	OmniVec (Pre.)	OmniVec (FT.)	SOTA
UCF-101	Video	Action Recognition	3-Fold Accuracy	98.7	**99.6**	**99.6** (VideoMAE V2-g [79])
HMDB51	Video	Action Recognition	3-Fold Accuracy	89.21	**91.6**	88.1 (VideoMAE V2-g [79])
Oxford-IIIT Pets	Image	Fine grained classification	Top-1 Accuracy	97.4	**99.2**	97.1 (EffNet-L2 [20])
ScanObjectNN	3D Point Cloud	Classification	Accuracy	92.1	**96.1**	93.4 (PointGPT [9])
NYU V2	RGBD	Semantic Segmentation	Mean IoU	58.6	**60.8**	56.9 (CMN [51])
SamSum	Text	Meeting Summarization	ROGUE(R-L)	51.2	**54.6**	50.88 (MoCa [101])
KITTI	RGB	Depth Prediction	iRMSE	-	**10.2**	10.4 (VA-DepthNet [50])
YouCook2	Video+Text	Zero Shot Text-to-Video Retrieval	Recall@10	64.2	**70.8**	63.1 (VideoCLIP [89])
MSR-VTT	Video+Text	Zero Shot Text-to-Video retrieval	Recall@10	78.6	89.4	80.0(Pre.)/**90.8**(FT)(SM [99])

Table 12. **Generalization performance of OmniVec** on *unseen datasets* (Oxford-IIIT Pets, UCF-101, HMDB51, ScanObjectNN, NYUv2 Seg, SamSum), *unseen tasks* (KITTI Depth Prediction) and *cross-domain* generalization (YouCook2, MSR-VTT). Pre. indicates network with pretraining only, FT indicates network finetuned on training set of respective datasets. See supplementary for more detailed results.

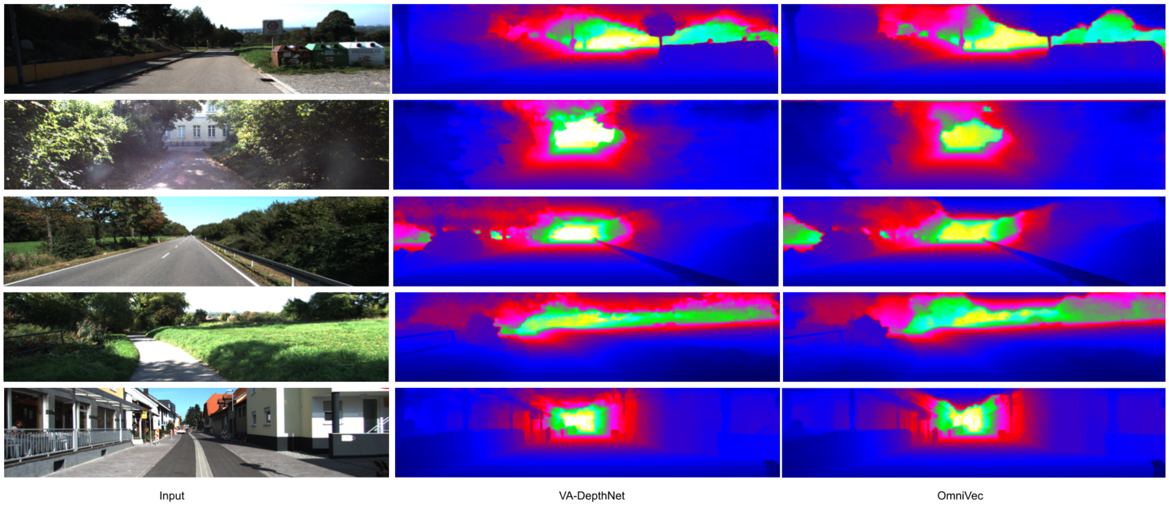

Figure 2. Qualitative results on test set of KITTI Depth Prediction. Ground truth is not available. For an RGB input image (left), the outputs from VA-DepthNet [50](middle) and OmniVec (right) are shown. See supplementary material for more qual. results.

Dataset	Metric	OmniVec	SOTA
AudioSet(A)	mAP	54.8	53.3 (MAViL [35])
AudioSet(A+V)	mAP	55.2	51.2 (CAV-MAE [29])
SSv2	Top-1 Acc	85.4	77.3 (MVD [80])
ImageNet1K	Top-1 Acc	92.4	91.1 (BASIC-L [12])
Sun RGBD	Top-1 Acc	74.6	67.2 (Omnivore [25])
ModelNet40	Overall Acc	96.6	95.4 (GeomGCNN [70])

Table 13. **Comparison with state of the art** after fine tuning on respective training sets.

73.9% Recall@10 (see supplementary), which is behind our pretrained OmniVec.

5. Conclusion and Limitations

Conclusion. We proposed OmniVec, a unified data and task agnostic learning framework with a single backbone. The main idea behind OmniVec is that modalities in different domains can aid learning process. Further, we also proposed a novel training mechanism by grouping tasks and constructing mini batches by mixing inter-modality datasets. With experiments on 22 datasets spanning across image, video, point cloud, depth, audio, text; we show that the proposed framework is highly generalizable along with being extremely robust. It can also generalize well to seen tasks with different data distribution as well as can adapt to unseen tasks effectively. We also studied the cross-domain knowledge sharing by evaluating a zero shot video-text retrieval task. We achieve state of the art or close to state of the art performance on all the evaluated datasets.

Limitations. OmniVec trains on unpaired multi-modal data, but paired data, though better, is expensive to obtain. The method employs multiple encoders per modality, increasing computational demands. Future research may address these computational challenges in unified networks.

Societal Impact. Modality agnostic techniques enhance realistic data cloning, risking misinformation and identity theft. These networks, syncing various modalities and using extensive internet data, amplify privacy, security, and bias concerns.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021. 1, 3, 4, 5, 6, 7

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

[3] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 6

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3, 5

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[6] Alan Baade, Puyuan Peng, and David Harwath. Mae-ast: Masked autoencoding audio spectrogram transformer. *arXiv preprint arXiv:2203.16691*, 2022. 3

[7] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 1, 3, 5

[8] Joao Carreira, Skanda Koppula, Daniel Zoran, Adria Recasens, Catalin Ionescu, Olivier Henaff, Evan Shelhamer, Relja Arandjelovic, Matt Botvinick, Oriol Vinyals, et al. Hierarchical perceiver. *arXiv preprint arXiv:2202.10890*, 2022. 3, 5

[9] Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. *arXiv preprint arXiv:2305.11487*, 2023. 8

[10] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2022. 6

[11] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *arXiv preprint arXiv:2305.18500*, 2023. 7

[12] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023. 8

[13] Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*, 2021. 6

[14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020. 2

[15] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2

[16] Yong Dai, Duyu Tang, Liangxin Liu, Minghuan Tan, Cong Zhou, Jingquan Wang, Zhangyin Feng, Fan Zhang, Xueyu Hu, and Shuming Shi. One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code. *arXiv preprint arXiv:2205.06126*, 2022. 2, 3

[17] 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. 5

[18] 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. 3, 4

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 4, 5, 7

[20] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 8

[21] Quentin Fournier, Gaétan Marceau Caron, and Daniel Aloise. A practical survey on faster and lighter transformers. *ACM Computing Surveys*, 2021. 2

[22] Avi Gazneli, Gadi Zimerman, Tal Ridnik, Gilad Sharir, and Asaf Noy. End-to-end audio strikes back: Boosting augmentations towards an efficient audio classification network. *arXiv preprint arXiv:2204.11479*, 2022. 6

[23] 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. 5

[24] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. *arXiv preprint arXiv:2206.08356*, 2022. 2, 3, 4, 5, 6

[25] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 1, 3, 5, 6, 8

[26] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*, 2019. 7

[27] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 3, 5, 6

[28] Yuan Gong, Alexander H Liu, Andrew Rouditchenko, and James Glass. Uavm: Towards unifying audio and visual models. *IEEE Signal Processing Letters*, 29:2437–2441, 2022. 1, 3

[29] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022. 8

[30] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 5

[31] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Mart´in-Mart´in, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022. 3

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 4, 6

[33] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[34] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021. 2, 3

[35] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Haoqi Fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, and Christoph Feichtenhofer. Mavil: Masked audio-video learners. *arXiv preprint arXiv:2212.08071*, 2022. 8

[36] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 1, 3, 5

[37] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 1, 3, 5

[38] Xingyu Jiang, Jiayi Ma, Guobao Xiao, Zhenfeng Shao, and Xiaojie Guo. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73:22–71, 2021. 2

[39] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017. 2

[40] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[41] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 2

[42] Seungone Kim, Se June Joo, Hyungjoo Chae, Chaehyeong Kim, Seung-won Hwang, and Jinyoung Yeo. Mind the gap! injecting commonsense knowledge for abstractive dialogue summarization. *arXiv preprint arXiv:2209.00930*, 2022. 7

[43] Hildegard Kuehne, Hueihan Jhuang, Est´ibaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 7

[44] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8500–8509, 2022. 7

[45] Dogyoon Lee, Jaeha Lee, Junhyeop Lee, Hyeongmin Lee, Minhyeok Lee, Sungmin Woo, and Sangyoun Lee. Regularization strategy for point cloud via rigidly mixed sample. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15900–15909, 2021. 7

[46] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, et al. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2691–2700, 2023. 1

[47] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 6

[48] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022. 6

[49] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 2022. 2

[50] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. Va-depthnet: A variational approach to single image depth prediction. *arXiv preprint arXiv:2302.06556*, 2023. 7, 8

[51] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838*, 2022. 8

[52] Jing Liu, Xinxin Zhu, Fei Liu, Longteng Guo, Zijia Zhao, Mingzhen Sun, Weining Wang, Hanqing Lu, Shiyu Zhou, Jiajun Zhang, et al. Opt: Omni-perception pre-trainer for cross-modal understanding and generation. *arXiv preprint arXiv:2107.00249*, 2021. 3

[53] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2

[54] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 6

[55] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*, 2019. 6

[56] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 7

[57] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. 6

[58] AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. Rethinking video vits: Sparse video tubes for joint image and video learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2214–2224, 2023. 6

[59] Subhojeet Pramanik, Priyanka Agrawal, and Aman Hussain. Omninet: A unified architecture for multi-modal multi-task learning. *arXiv preprint arXiv:1907.07804*, 2019. 2, 3

[60] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 7

[61] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 5

[62] Adrià Recasens, Jason Lin, João Carreira, Drew Jaegle, Luyu Wang, Jean-baptiste Alayrac, Pauline Luc, Antoine Miech, Lucas Smaira, Ross Hemsley, et al. Zorro: the masked multimodal transformer. *arXiv preprint arXiv:2301.09595*, 2023. 1, 2

[63] Javier Selva, Anders S Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B Moeslund, and Albert Clapés. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[64] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*, 2022. 2

[65] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV (5)*, 7576:746–760, 2012. 7

[66] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020. 4

[67] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 5

[68] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 7

[69] Siddharth Srivastava, Swati Bhugra, Vinay Kaushik, and Brejesh Lall. Hierarchical multi-task learning via task affinity groupings. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3289–3293. IEEE, 2023. 2

[70] Siddharth Srivastava and Gaurav Sharma. Exploiting local geometry for feature and graph construction for better 3d point cloud processing with graph neural networks. In *2021 IEEE INternational conference on robotics and automation (ICRA)*, pages 12903–12909. IEEE, 2021. 8

[71] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pages 9120–9132. PMLR, 2020. 2

[72] Jiachen Sun, Qingzhao Zhang, Bhavya Kailkhura, Zhiding Yu, Chaowei Xiao, and Z Morley Mao. Benchmarking robustness of 3d point cloud recognition against common corruptions. *arXiv preprint arXiv:2201.12296*, 2022. 7

[73] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6

[74] Liyao Tang, Yibing Zhan, Zhe Chen, Baosheng Yu, and Dacheng Tao. Contrastive boundary learning for point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8489–8499, 2022. 7

[75] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 7

[76] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1588–1597, 2019. 7

[77] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 6

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[79] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 8

[80] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. *arXiv preprint arXiv:2212.04500*, 2022. 8

[81] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023. 6

[82] Yi Wang, Zhiwen Fan, Tianlong Chen, Hehe Fan, and Zhangyang Wang. Can we solve 3d vision tasks starting from a 2d vision transformer? *arXiv preprint arXiv:2209.07026*, 2022. 3

[83] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 6

[84] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 3

[85] Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. Controllable abstractive dialogue summarization with sketch supervision. *arXiv preprint arXiv:2105.14064*, 2021. 7

[86] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 7

[87] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 5, 6

[88] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2

[89] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pretraining for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 8

[90] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *arXiv preprint arXiv:2206.06488*, 2022. 2

[91] Yifan Xu, Huapeng Wei, Minxuan Lin, Yingying Deng, Kekai Sheng, Mengdan Zhang, Fan Tang, Weiming Dong, Feiyue Huang, and Changsheng Xu. Transformers in computational visual media: A survey. *Computational Visual Media*, 8:33–62, 2022. 2

[92] Zhiqiang Yan, Xiang Li, Kun Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Multi-modal masked pre-training for monocular panoramic depth completion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 378–395. Springer, 2022. 3

[93] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020. 2

[94] Yu-Qi Yang, Yu-Xiao Guo, Jian-Yu Xiong, Yang Liu, Hao Pan, Peng-Shuai Wang, Xin Tong, and Baining Guo. Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. *arXiv preprint arXiv:2304.06906*, 2023. 7

[95] Z Yang, Z Dai, Y Yang, J Carbonell, RR Salakhutdinov, and XLNet Le QV. generalized autoregressive pretraining for language understanding; 2019. *Preprint at https://arxiv.org/abs/1906.08237 Accessed June*, 21, 2021. 4, 5

[96] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 6

[97] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 6

[98] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19313–19322, 2022. 3

[99] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 7, 8

[100] Jinlai Zhang, Lyujie Chen, Bo Ouyang, Binbin Liu, Jihong Zhu, Yujin Chen, Yanmei Meng, and Danfeng Wu. Pointcutmix: Regularization strategy for point cloud classification. *Neurocomputing*, 505:58–67, 2022. 7

[101] Xingxing Zhang, Yiran Liu, Xun Wang, Pengcheng He, Yang Yu, Si-Qing Chen, Wayne Xiong, and Furu Wei. Momentum calibration for text generation. *arXiv preprint arXiv:2212.04257*, 2022. 8

[102] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018. 3

[103] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021. 2

[104] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 6

[105] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815, 2022. 1