OmniVec2 - A Novel Transformer based Network for Large Scale Multimodal and Multitask Learning

Siddharth Srivastava, Gaurav Sharma
Typeface
{siddharth.srivastava, gaurav}@typeface.ai

Abstract

We present a novel multimodal multitask network and associated training algorithm. The method is capable of ingesting data from approximately 12 different modalities namely image, video, audio, text, depth, point cloud, time series, tabular, graph, X-ray, infrared, IMU, and hyperspectral. The proposed approach utilizes modality specialized tokenizers, a shared transformer architecture, and cross-attention mechanisms to project the data from different modalities into a unified embedding space. It addresses multimodal and multitask scenarios by incorporating modality-specific task heads for different tasks in respective modalities. We propose a novel pretraining strategy with iterative modality switching to initialize the network, and a training algorithm which trades off fully joint training over all modalities, with training on pairs of modalities at a time. We provide comprehensive evaluation across 25 datasets from 12 modalities and show state of the art performances, demonstrating the effectiveness of the proposed architecture, pretraining strategy and adapted multitask training.

1. Introduction

Extracting meaningful representations from data is a central task in machine learning. Majority of the approaches proposed are usually specialized for specific modalities and tasks. The development of methods capable of handling multiple modalities, in a holistic way, has been an active topic of research recently [21, 34, 35, 45, 63, 105]. Multi task learning has a large body of literature [10], but has been traditionally limited to tasks from single modality. Learning a unified network that trains shared parameters across diverse tasks in different modalities, like image, video, depth maps, audio, has been shown to be more robust and give better generalization and reduce overfitting to a single task or modality [1, 24] cf. unimodal networks. Such joint learning also enables more efficient use of available labeled data across various modalities, potentially reducing the need for extensive labeling in specific modalities for particular tasks.

In the present work, we extend such line of research and propose a multimodal multitask method which learns embeddings in a shared space across different modalities and then employs task specific sub-networks for solving specific tasks in specific modalities. The method utilizes a common transformer based bottleneck block to map the input to embeddings in a shared space, thus incorporating knowledge from multiple tasks associated with different respective modalities. This structure leads to learning of very robust representations informed and regularized by all tasks and modalities together. The embeddings are then used by the task heads to make required predictions.

Previous research in generalized multimodal learning falls into three main categories. First, there are methods that process multiple heterogeneous modalities such as images, 3D, and audio, directly without using separate encoders for each modality, learning representations directly from these inputs [34, 35]. Second, some approaches use modality specific encoders and then learn generalized embeddings, for data from each modality, based on a unified objective in the latent space [5]. Third, there are methods focused on knowledge sharing across different modalities, employing either a single common encoder [21] or distinct encoders for each modality [1]. Our work aligns more closely with the third type of approaches, while incorporating elements from the first. We employ modality specific tokenizers and encoders, and have a bottleneck shared transformer backbone. Tokenization is tailored to each modality, drawing inspiration from the Uni-Perceiver model but with key modifications detailed in Sec. 3. After tokenization, transformer based network is used to obtain initial representations for the modalities which are passed through fully connected layers and then fused together with cross attention module. The fused representation then passes through the transformer backbone. The features from the transformer are then individually fused with original modality features using cross attention and are in turn fed to the modality specific task head.

Two-phase masked pretraining!

The training procedure involves a dual-stage masked pretraining and a full task based loss optimization. The first stage of masked pretraining is the standard unsupervised masked pre-training with one modality at a time. The second state masked pretraining involves masked pretraining with pairs of modalities at a time, employing a two stream setup as shown in Fig. 1. In this stage two modalities are used together, tokens are randomly masked and the full network is used to predict the masked tokens using the unmasked tokens for both modalities together. This allows for knowledge sharing across all modalities as the training proceeds by randomly sampling training batches from two modalities from all modalities. The final training step is then training for multiple tasks for different modalities. This is done similar to the second stage of masked pretraining, i.e. pairs of modalities are sampled, and a pair of tasks are sampled, one from each modality. Training batches are then constructed, half each from the two modality-task pairs. These are then used to optimize standard losses corresponding to the tasks, e.g. cross entropy for classification and ℓ_2 loss for pixelwise prediction. The pretraining and final task training using pairs of modalities is the key component of the training strategy, that enables the cross modal knowledge sharing across all modalities together, which we discuss more in the following.

In summary, the contributions of the work are as follows. (i) We propose a multimodal multitask network based on transformer architectures with modality specific tokenizers, shared backbone, and task specific heads. (ii) We provide comprehensive empirical results on 25 benchmark datasets over 12 distinct modalities *i.e.* text, image, point cloud, audio and video along with applications to X-Ray, infrared, hyperspectral, IMU, graph, tabular, and time-series data. The method achieves better or close to state of the art performances on these datasets. (iii) We propose a novel multimodal pretraining approach that alternates between a pair of modalities to enable crossmodal knowledge sharing. (iv) We propose a multimodal and multitask supervised training approach to leverage knowledge sharing between modalities for robust learning, simplifying the complex processes proposed in previous works on modality integration, e.g. [45, 94].

2. Related Works

In this section, we discuss similar works and various similar paradigms to our work.

Multi-modal methods. Contemporary multi-modal methods predominantly employ modality-specific feature encoders [2, 36, 37, 63, 85], focusing on fusion techniques within their architectural designs. These networks usually vary across modalities, necessitating architectural modifications for combined usage. They must address challenges related to feature fusion timing, fine-tuning, and pre-training *etc*. [87]. Such complexities restrict the adaptability of universal frameworks like transformers for diverse domains, including point clouds, audio, and images.

Common network for multiple modalities. A growing body of research aims to learn from multiple modalities without modality-specific encoders [5, 7, 21, 35]. Notably, architectures like the perceiver [7, 34, 35] employ cross-attention among latent queries to process multiple modalities together. The hierarchical perceiver [7] expands on this by structuring the input while maintaining locality. Other approaches, such as data2vec [5], use modality-specific encoders. Omnivore [21], with a common encoder, is limited to visual modalities only. Contrarily, VATT [1] employs a unified transformer backbone but processes each modality independently. These multi-modal methods have demonstrated enhanced robustness [1, 24].

Multi-task learning. As explored in the preceding section, there has been a surge in methods that process multiple modalities. PerceiverIO[34] extends the capabilities of Perceiver [35] to facilitate learning multiple tasks with a singular network architecture. Although PerceiverIO is capable of multitasking, often separate networks are employed [98]. Various techniques [5, 11, 21, 32, 59] learn from raw representations of multiple modalities and are applicable to numerous tasks.

Multi-modal masked pretraining. Approaches such as [50, 79, 88] implement masked pre-training. This technique has proven beneficial for improving the performance of deep networks across different modalities and tasks[1, 4, 5, 20, 28, 95].

Comparison to similar works. We draw motivations from UniPerceiver [105], MetaFormer [16] and OmniVec [68]. Unlike UniPerceiver line of methods, we do not use a unified task head definition, while similar to it we use task specific task heads. This allows our method to learn more robust and leverage fine details from each task depending upon the complexity of the tasks, which is important as each modality has distinct definition of complexity. For ex., in vision task, classification is a relatively simpler task as compared to segmentation, as segmentation tasks enforces networks to learn pixel level attention and learning better neighbourhood relationships [27, 69]. Further, MetaFormer uses unified tokenizers, and instead, we utilize modality specific tokenizers. Our experiments indicate that modality specific tokenizers perform better than MetaFormer's unified tokenizer when training on multiple modalities. Further, OmniVec uses separate encoders for each modaity, that makes the network heavy and computationally expensive. In contrast, we use modality specific tokenizers with a shared backbone. Additionally, unlike other works, we train on multiple modalities in a multi task manner, allowing the network to learn from multiple modalities with varying task complexities simultaneously.

3. Approach

Overview. We are interested in multimodal multitask learning. Say we have modalities indexed by $m \in [1, M]$, and each modality has T tasks indexed by $t \in [1, T]$. Note that here we assume same number of tasks for all modalities for notational convenience, in practice different modalities would have different number of tasks. Examples of modality and their tasks could be classification into categories for point cloud modality, and dense pixel wise segmentation in image modality. We are interested in jointly learning classifiers $\phi_{mt}(\cdot|\theta_{mt})$ which take inputs x_m from modality m and make predictions for task t, with θ_{mt} being the respective parameters. We assume that the learning is to happen by loss minimization where $\ell_{mt}(\cdot)$ denotes the loss for task t on modality m. Examples of such losses are cross entropy loss for classification tasks, and ℓ_2 loss for dense image prediction tasks such as image segmentation. We would like to solve the following optimization.

$$\Theta^* = \min_{\Theta} \sum_{m,t} \ell_{mt}(\mathcal{T}_{mt}), \tag{1}$$

where $\Theta = \{\theta_{mt}|m, t\}$ are the parameters of all the predictors, and $\mathcal{T}_{mt}$ is the training set provided for task t of modality m. This is the extension of multiple task learning to multiple modalities as well.

We present a network and associated unsupervised pretraining and supervised training algorithm for the above task of multimodal multitask learning. The network consists of $M \times T$ modality specific tokenizers, followed by common feature transformation and feature fusion networks built with transformers, with cross attention modules in between, denoted by $f(\cdot), g(\cdot)$ in Fig. 1. The final part of the network are $M \times T$ task specific prediction heads, denoted by $h_{mt}(\cdot)$ for task t on modality m, which provide the final outputs for the tasks. At inference the prediction function is the composition of the three functions, i.e. $\phi(x) = h_{mt} \circ g \circ f(x)$ where x is the tokenized form of the input. While training, we sample a pair of modalities from all the available modalities, and then sample one task each for the sampled modalities. We then construct training batch, half from each sampled task. Once the tokenization is done, the features x_i, x_j are passed into the first feature transformation subnetwork to obtain $f(x_i), f(x_j)$. These are then passed through the cross attention module to fuse them together. The fused features are then input to the second part of the network, i.e. $g(\cdot)$. The output $\hat{x}_{ij} = g \circ \mathcal{A}(f(x_i), f(x_j))$, where $\mathcal{A}(\cdot)$ is the cross attention function, is then again fused with the respective input features x_i, x_j. These features, i.e. $\mathcal{A}(\hat{x}_{ij}, x_i), \mathcal{A}(\hat{x}_{ij}, x_j)$ are then fed to the task predictors h_{iq} and h_{jr}, to obtain the final predictions for task q, r on modalities i, j respectively. The sum of losses $\ell_{iq} + \ell_{jr}$ are then minimized for the current batch by backpropagation. Thus the learning proceeds by optimizing pairs of losses at a time, to stochastically minimize the sum over all the losses.

Along with the supervised multimodal joint training explained above, the learning also consists of two stages of unsupervised masked pretraining with the first stage being unimodal and the second stage being multimodal pretraining, to achieve knowledge sharing between tasks and modalities leading to regularized and robust predictors. We now present each of the components and the full training algorithm in detail.

3.1. Network components

We now go through the network components sequentially from input to output.

Tokenizers. Each modality is tokenized using a modality specific tokenizer. The tokenizers are similar to those used in Uni-Perceiver [45], however, instead of attaching an embedding to the tokens, we provide transformer with one type of modality at a time. Further, Uni-Perceiver utilizes a combination of tokens from multiple modalities passed to a single transformer. This limits the Uni-perceiver to a limited set of modalities, i.e. text, image and videos. However, our method does not suffer from any such limitation. The details of specific tokenizers for the different modalities are provided in Supplementary.

Feature transformation network. Once the features are tokenized, they are then passed through a transformer network. While the method can utilize any transformer backbone, in the current implementation we use a transformer based on BERT [13]. Here, the multi head attention involves standard self-attention [76], and GeLU [30] activation prior to the MLP layer. The output from the transformer network is passed to a fully connected neural network with three fully connected layers with ReLU activation. This transformer network along with the fully connected layers is denoted a $f(\cdot)$ in Fig. 1. The network could be used without the fully connected layers—we added the fully connected layers to reduce the dimensions of the features so that the computational complexity of the remaining part of the network could be reduced.

Mixing features with cross attention. When training, we fuse the features from the two transformer streams, corresponding to two modalities, with cross attention module. The output fused features are then passed to another transformer network, denoted a $g(\cdot)$ in Fig. 1. The architecture of the transformer network is same as the transformers used in feature transformation network.

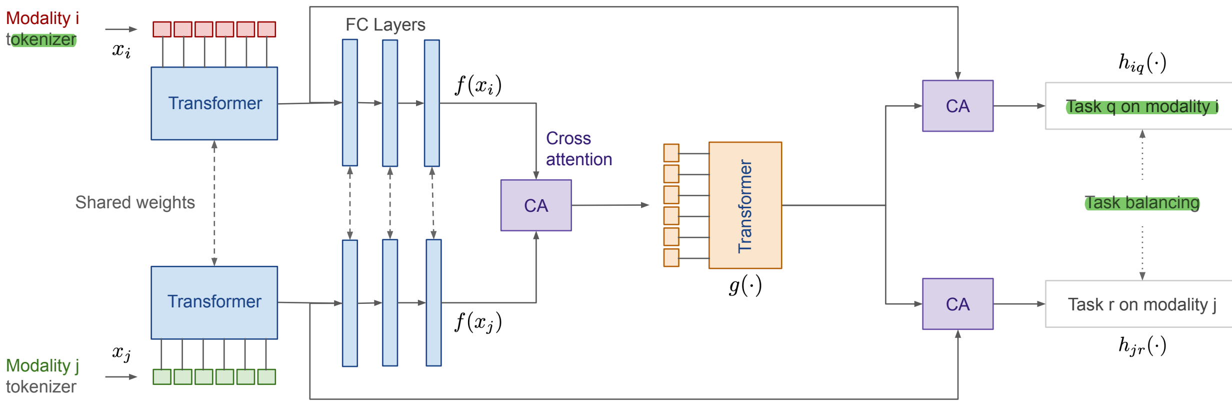

Figure 1. **Overview of the proposed method.** The proposed method consists of three parts, the feature transformation network $f(\cdot)$ which consists of a transformer followed by fully connected layers to reduce feature dimensions, another transformer $g(\cdot)$ and finally the task prediction heads $h_{mt}(\cdot)$ for task t on modality m. The input data is tokenized with corresponding modality specific tokenizer. While training, pairs of modalities are used and the features are fused between the two modalities using cross attention layers, in a two stream configuration as shown here. While making prediction, the network is a single stream with cross attention layers removed, and the output is $h_{mt} \circ g \circ f(x)$ where x is the output of the corresponding modality specific tokenizer.

Modality and task specific heads. The part of the network are the modality and task specific heads, denoted a $h_{mt}(\cdot)$ in Fig. 1. These task heads take as input, features from respective modality streams fused with features from the above network, fused with cross attention module. The task heads consist of a vanilla ViT-Tiny networks [82].

3.2. Training

The training is done in three steps: (i) masked pretraining iterating over modalities but doing masked prediction with one modality at a time, (ii) multimodal masked pretraining where two modalities are simultaneously used to do masked prediction for each, and (iii) finally supervised task based training.

Stage 1 masked pretraining. The first step in training is self supervised pretraining of the transformer in the feature transformation network. We follow earlier works [1, 20, 68] and add a decoder for predicting masked tokens. Specifically, for an input modality with P patches, we randomly mask P_m patches, and feed non-masked patches and their positions to an encoder network attached in addition to the feature transformer. Further, we iterate between modalities while keeping the transformer network common, so that it learns to work with all modalities. Once this stage is complete we discard the decoder added, and keep only the encoder transformer.

Stage 2 masked pretraining. We engage the full network, except the task specific prediction heads. We take two inputs from two different modalities and pass them through the network till just before the task prediction heads. Instead of task prediction heads we add decoders to predict the masked tokens for respective input modalities. This process involves decoding the modalities in parallel, utilizing the outputs from the cross-attention modules and the modality-specific feature vectors. This alternating approach is key to achieving effective multimodal masked pretraining. Here also, we randomly mask tokens for both the modalities. Task balancing is not employed in this pretraining stage. Such a multi task multi modality approach allows us to utilize unpaired data across modalities. As in stage 1 pretraining, once this stage of training is finished, we discard the decoders added and keep the trained network f, g.

3.2.1 Multimodal multitask supervised training

In the final part of the training, we train for two tasks at a time from two different modalities. This lets up stochastically minimize the loss function in Eq. 1, but minimizing sum of two losses at a time instead of minimizing the sum of all of them. When we use two modalities, we use the network as shown in Fig. 1 in a two stream configuration. With the two modality features being fused together in the middle, passed through a transformer $g(\cdot)$ and then fused back with themselves, before finally being input to the task prediction heads. Such fusion of the the features from two modalities leads to knowledge sharing between the tasks of different modalies and makes the learning robust and regularized.

Given the varying complexities of these task pairs, as underscored in previous research [17], we found it essential to balance the complexity of tasks in a multitask learning setting. Hence, the we train while employing standard task balancing techniques. We adjust the loss magnitude for each task based on its convergence rate. As our ablation studies will demonstrate, this approach allows for random pairing of modalities, in contrast to the need for selecting specific pairs as suggested in prior works [45, 68, 94, 105]. We give details of such task balancing in the Supplementary material.

3.2.2 Masked pretraining for different modalities

We use the best practices when pretraining with different modalities, following existing works. We use image, video, text, audio and 3D point clouds modalities for masked pretraining. We employ a consistent masking approach across visual and auditory modalities. We follow [65] for textual data, utilizing random sentence permutation [90]. We designate a fraction f of tokens for prediction, following the 8:1:1 token masking ratio of BERT [13]. Our primary goal is to reduce the discrepancy between the input and the outputs of the decoder. For inputs such as images, videos, point clouds, and audio spectrograms, we aim to minimize the ℓ_2 distance between the predicted and actual target patches. Normalization to zero mean and unit variance is applied to visual inputs. For textual data, we utilize the permuted language modeling objective of XLNet [90].

3.2.3 Inference

When doing prediction, the network is used as a single stream without the cross attention layers in Fig. 1. The input data is tokenized with the tokenizer for its modality, passed through the feature transformation network $f(\cdot)$ followed by the second transformer $g(\cdot)$, and finally input to the task prediction head $h_{mt}(\cdot)$, i.e. the full forward pass is $h_{mt} \circ g \circ f(x)$ where x is the output of the tokenizer.

4. Experimental results

Masked pretraining. We use AudioSet (audio) [19], Something-Something v2 (SSv2) (video) [25], English Wikipedia (text), ImageNet1K (image) [12], SUN RGB-D (depth maps) [66], ModelNet40 (3D point cloud) [84] for pretraining the network. For Stage 1 of masked pretraining (Sec. 3.2), we use the samples from the training set of the respective datasets. For Stage 2 of masked pretraining, we randomly select two modalities, and sample data from them to pretrain the full network. Further, we randomly mask patches. For image, video and audio, we mask 95% of the patches. For point cloud and text, we mask 90% and 95% of the patches respectively. We perform pretraining for 3000 epochs. We use fraction f as 5%.

Downstream tasks. We train the model on downstream tasks and report results. The datasets used for single modality methods are iNaturalist-2018 [75] (Image Recognition), Places-365 [100] (Scene Recognition), Kinetics-400 [38] (Video Action Recognition), Moments in Time [53] (Video Action Recognition), ESC50 [57] (Audio Event Classification), S3DIS [3] (3D point cloud segmentation), DialogueSUM [9] (Text summarization).

Adaptation on unseen datasets. To assess our method's adaptability to datasets not seen at training, we report comparisons with image classification on Oxford-IIIT Pets [56], action recognition in videos using UCF-101 [67] and HMDB51 [41], 3D point cloud classification on ScanObjectNN [74], point cloud segmentation with NYU v2 seg [64], text summarization using the SamSum dataset [22]. As the number of classes and labels differ in each dataset as compared to the datasets used during pretraining, we randomly sample 10% data from each of the training set. Further, we extract the embeddings using the pretrained network, and train two fully connected layers with task specific loss functions. This allows us to demonstrate the ability of the proposed method to generate embeddings which can generalize across datasets.

Cross domain generalization. We follow prior work [1] and evaluate on video-text retrieval on two benchmark datasets *i.e.* YouCook2 [104], and MSR-VTT [86], for multiple modalities.

Adaptation on unseen modalities. We also evaluate our method on unseen modalities. Specifically, we evaluate our method on the following (i) X-Ray scan, and hyperspectral data recognition, where we utilize the RegDB [54], Chest X-Ray [62], and Indian Pine datasets[1]. (ii) Time-series forecasting, where our experiments are based on the ETTh1 [103], Traffic[2], Weather[3], and Exchange datasets [42]. (iii) Graph understanding through the PCQM4M-LSC dataset [33], which comprises 4.4 million organic molecules with quantum-mechanical properties, focusing on predicting molecular properties with applications in drug discovery and material science. (iv)Tabular analysis, where we engage with the adult and bank marketing datasets from the UCI repository[4], (v) IMU recognition, where we conduct experiments on IMU sensor classification using the Ego4D dataset [26], assessing the capability to understand inertial motion systems. We follow [16] for the train test splits and evaluation metrics on these datasets. Further, we use modality specific tokenizers and follow similar network settings as for generalization on unseen datasets.

We provide more details on the tokenizers used for each modality, description of task heads, and formulations of loss functions in the supplementary material.

[1] https://github.com/danfenghong/IEEE_TGRS_SpectralFormer/blob/main/data/IndianPine.mat
[2] https://pems.dot.ca.gov/
[3] https://www.bgc-jena.mpg.de/wetter/
[4] http://archive.ics.uci.edu/ml/

Method/Dataset	iN2018	P365
Omni-MAE [20]	78.1	59.4
Omnivore [21]	84.1	59.9
EfficientNet B8[71]	81.3	58.6
MAE[29]	86.8	
MetaFormer [94]	87.5	60.7
InternImage[77]	92.6	61.2
OmniVec [68]	93.8	63.5
Ours	**94.6**	**65.1**

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

Method	K400
Omnivore [21]	84.1
VATT [1]	82.1
Uniformerv2 [46]	90.0
InternVideo[78]	91.1
TubeViT[58]	90.9
OmniVec[68]	91.1
Ours	**93.6**

Table 2. **Kinetics-400** top-1 accuracy.

Method	MIT
VATT [1]	41.1
Uniformer v2[46]	47.8
CoCa[93]	47.4
CoCa-finetuned[93]	49.0
OmniVec[68]	49.8
Ours	**53.1**

Table 3. **Moments in time** top-1 accuracy.

Method	ESC50
AST [23]	85.7
EAT-M[18]	96.3
HTS-AT[8]	97.0
BEATs[55]	98.1
OmniVec[68]	98.4
Ours	**99.1**

Table 4. **ESC50** top-1 accuracy.

Method	MN40C
PointNet++[60]	0.236
DGCN+PCM-R[97]	0.173
PCT + RSMIx[44]	0.173
PCT + PCM-R[70]	0.163
OmniVec[68]	0.156
Ours	**0.142**

Table 5. **ModelNet40-C** Error Rate.

Method	S3DIS
PointTransformer+CBL[72]	71.6
StratifiedTransformer[43]	72.0
PTv2[83]	72.6
Swin3D[89]	74.5
OmniVec[68]	75.9
Ours	**77.1**

Table 6. **Stanford Indoor Dataset** mIoU.

Method	R-1	R-2	R-L	B-S
CODS[81]	44.27	17.90	36.98	70.49
SICK[39]	46.2	20.39	40.83	71.32
OmniVec[68]	46.91	21.22	40.19	71.91
Ours	**47.6**	**22.1**	**41.4**	**72.8**

Table 7. **DialogueSUM** text summarization ROGUE scores.

4.1. Comparison with state of the art methods

We performed masked pretraining followed by training on multiple modalities and task groups as described in Section 3 for comparing with existing methods. We discuss the comparison on each modality below.

Image. Table 1 shows state of the art on iNaturalist 2018 and Places 365 datasets. On the iNaturalist 2018 dataset, our method achieves a top-1 accuracy of 94.6%, surpassing notable contenders such as OmniVec (93.8%), MetaFormer (87.5%), and MAE (86.8%). This superior accuracy demonstrates capability of the proposed method in accurately recognizing a diverse range of natural species. In the context of the Places 365 dataset, our method achieves an accuracy of 65.1%, notably outperforming OmniVec (63.5%), and significantly surpassing MetaFormer's 60.7% and Omnivore's 59.9%. The substantial margin of improvement, particularly in the challenging and variable environment of Places 365, underscores the robustness and adaptability of the proposed architecture. We also conduct experiments on ImageNet [12] (classification), MSCOCO [49] (object detection), and ADE-20K [101] (semantic segmentation) datasets (detailed table is in supplementary). 89.3% (accuracy) on ImageNet, 60.1 (AP) on MSCOCO and an mIoU of 58.5 on ADE-20K.

Video. Table 2 and Table 3 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 both the datasets achieving top-1 accuracy of 93.6% and 53.1% respectively.

Audio. Table 4 shows our comparison with top-performing methods on the ESC50 dataset. We outperform competing methods, achieving an accuracy of 99.1%, significantly higher than the Audio Spectrogram Transformer (AST) at 85.7%, and OmniVec at 98.4%.

Point Cloud. Table 5 and Table 6 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 methods. On ModelNet-C, we achieve an error rate of 0.142, which is notably lower than the rates observed in other contemporary methods. This is particularly evident when compared against methods like OmniVec, which recorded an error rate of 0.156, and PCT + PCM-R, with an error rate of 0.163. On S3DIS, we achieve an mIoU of 77.1, which is the highest among all the methods evaluated c.f. 75.9 of OmniVec, and 74.5 of Swin3D. This demonstrates that the proposed method is able to obtain a robust performance with the shared backbone network across tasks.

Text. Table 7 shows state of the art on DialogueSUM dataset for text summarization. Our method surpasses other methods in all the metrics. Despite utilizing significantly fewer datasets for text in comparison to visual tasks , our method demonstrates strong performance. This suggests proposed method's capacity to bridge the modality gap [48] across distinct domains in the latent space, even when the data distribution is skewed.

Table 9 illustrates the experimental results on the GLUE benchmark for text understanding tasks, comparing various

Dataset	Modality	Task	Metric	Ours	SOTA	Ref.
UCF-101	Video	Action Recognition	3-Fold Accuracy	<u>99.1</u>	**99.6**	OmniVec [68]
HMDB51	Video	Action Recognition	3-Fold Accuracy	**92.1**	<u>91.6</u>	OmniVec [68]
Oxford-IIIT Pets	Image	Fine grained classification	Top-1 Accuracy	**99.6**	<u>99.2</u>	OmniVec [68]
ScanObjectNN	3D Point Cloud	Classification	Accuracy	**97.2**	<u>96.1</u>	OmniVec [68]
NYU V2	RGBD	Semantic Segmentation	Mean IoU	**63.6**	<u>60.8</u>	OmniVec [68]
SamSum	Text	Meeting Summarization	ROGUE(R-L)	**55.4**	<u>54.6</u>	OmniVec [68]
YouCook2	Video+Text	Zero Shot Text-to-Video Retrieval	Recall@10	<u>69.9</u>	64.2(Pre) / **70.8**(FT)	OmniVec [68]
MSR-VTT	Video+Text	Zero Shot Text-to-Video retrieval	Recall@10	<u>85.8</u>	80.0(Pre) / **90.8**(FT)	SM [96]

Table 8. Adaptation on *unseen datasets.* (Oxford-IIIT Pets, UCF-101, HMDB51, ScanObjectNN, NYUv2 Seg, SamSum), and *cross-domain* generalization (YouCook2, MSR-VTT). See supplementary for more detailed results.

Method	GLUE Benchmark				
	SST-2 Sentiment	MRPC Paraphrase	QQP Duplication	MNLI Inference	QNLI Answering
BiLSTM+ELMo+Attn	90.4	84.9	64.8	76.4	79.8
OpenAI GPT [61]	91.3	82.3	70.3	82.1	87.4
$\text{BERT}_{\text{BASE}}$ [13]	88.0	<u>88.9</u>	71.2	84.6	<u>90.5</u>
$\text{RoBERTa}_{\text{BASE}}$ [52]	**96.0**	**90.0**	**84.0**	84.0	**92.0**
ChatGPT	92.0	66.0	78.0	**89.3**	84.0
Meta-Transformer-B16_{T} [16]	81.3	81.8	78.0	70.0	60.3
Ours	<u>95.6</u>	85.8	<u>82.2</u>	<u>87.9</u>	84.2

Table 9. **Text understanding on the GLUE benchmark.** We compare existing advanced methods from paraphrasing, sentiment, duplication, inference, and answering tasks.

state-of-the-art methods such as BERT [13], RoBERTa [52], and ChatGPT. The comparison centers on paraphrasing, sentiment, duplication, inference, and answering tasks. We achieve second best performance on three out of five tasks demonstrating its capability to perform reasoning and adaptability to natural language tasks.

Comparison on pretraining datasets. We fine tune our pretrained network on the respective training sets with related task heads. We obtain an mAP of 55.8 and 56.4 on AudioSet(A) and AudioSet(A+V) respectively. Further, on SSv2, ImageNet-1K, SUN-RGBD, and ModelNet we achieve top-1 accuracies of 86.1%, 93.6%, 75.9% and 97.1% respectively. We outperform the competing state of the art methods on these datasets(detailed results are in supplementary).

4.2. Adaptation on unseen datasets

In Table 8 (rows 1-6), we observe that our method performs close to SoTA on all the datasets. Specifically, except on UCF-101, we outperform the SoTA (OmniVec) on all the datasets. We observe that on NYUv2, we obtain a performance improvement of 3%, while on an average perform better by approx 1% on other datasets. It must be noted that we freeze the base embeddings, and unlike other methods do not fine tune the full network, and use simpler task head for analysis on these datasets.

4.3. Cross domain generalization

Table 8 (rows 7,8) demonstrates results using our pretrained network on various tasks. On the YouCook2 dataset, our pretrained network surpasses the state of the art in zero-shot retrieval, achieving a Recall@10 of 69.9% compared to OmniVec's 64.2% on pretrained network. Interestingly, we are very close to the full fine tuned OmniVec *i.e.* 70.8. This demonstrates that our method is able to leverage the cross domain information better potentially due to multi task pretraining while OmniVec sequentially trains on one modality at a time. On MSR-VTT, when compared with SM [96], our fine-tuned method has a Recall@10 of 89.4% cf. SM's 80.0% (pretrained). It must be noted that SM uses internet-scale data while our method utilizes significantly less data.

4.4. Adaptation on Unseen Modalities

Infrared, Hyperspectral, and X-Ray data. Table 10a presents the performance comparison on the RegDB dataset [54] for infrared image recognition. Our method achieves state of the art performance *i.e.* R@1 of 86.21 c.f. 83.86 of MSCLNet, and mAP of 84.24 c.f. 78.57 of SMCL. This demonstrates that our method can transfer knowledge across unseen modalities. Specifically, we significantly outperform Meta-Transformer, which pretrains on similar modalities as ours. This could be potentially due to separate tokenizers for each modality allowing better integration with the transformer encoder as compared to a common tokenizer in meta-transformer.

In addition, Table 10b presents the performance on the Indian Pine dataset for hyperspectral image recognition. We achieve an overall accuracy of 90.6%, which is better than the SpectralFormer (81.76%) and significantly better than Meta-Transformer(67.6%). For X-Ray images (table in supplementary), our method achieves an accuracy of 98.1%, significantly outperforming competing methods.

Graph and IMU Data. We show results in Table 11. We achieve performance close to the state of the art methods *i.e.* validate MAE of 0.1397 c.f. 0.1234 of Graphormer. It is important to note that our method was not designed for graphical data, while competing methods are designed to exploit graphical data. Meta-Transformer, which is a unified learning mechanism like ours, significantly lies behind with 0.8863 MAE cf. 0.1397 of ours.

Time series forecasting. We achieve an MSE of 0.399,

Method	R@1 (%)	mAP (%)
AGW [91]	70.49	65.90
SMCL [80]	83.05	78.57
MSCLNet [99]	83.86	78.31
Meta-Transformer-B16$_F$ [16]	73.50	65.19
Ours	**86.21**	**84.24**

(a) SYSU-MM01 (infrared)

Method	OA (%)	AA (%)
ViT [14]	71.86	78.97
SpectralFormer [31] (Pixel)	78.55	84.68
SpectralFormer [31](Patch)	81.76	87.81
RPNet-RF [73]	90.23	
HyLITE [102]	89.80	
TC-GAN [6]	87.47	
Meta-Transformer-B16$_F$ [16]	67.62	78.09
Ours	**90.6**	**89.3**

(b) Indian Pine (hyperspectral)

Table 10. **Infrared and hyperspectral classification**. Metrics are Rank-1 (R@1), mean Average Precision (mAP), Overall Accuracy (OA), Average Accuracy (AA).

Method	train MAE	val MAE
Graph Transformer [15]	0.0944	0.1400
Graph Transformer-Wide [15]	0.0955	0.1408
Graphormer$_{SMALL}$ [92]	0.0778	**0.1264**
Graphormer [92]	**0.0582**	**0.1234**
Meta-Transformer-B16$_F$ [16]	0.8034	0.8863
Ours	0.0594	0.1397

Table 11. **Graph data understanding** . MAE on PCQM4M-LSC dataset.

0.601, 0.210, 0.330 on ETTh1, Traffic, Weather and Exchange datasets respectively, outperforming all the competing methods such as Pyraformer [51], Informer [103], LogTrans [47], Meta-former [94] and Reformer [40]. The detailed results are in supplementary.

Tabular Data. We achieve an accuracy of 88.1 and 92.3 on Adult and Bank Marketing datasets respectively, outperforming the competing methods (details in supplementary). Our method has never seen tabular data or structured textual information demonstrating its generalization ability to adapt to unseen patterns within data while providing better performance than competing methods.

4.5. Ablations

We study the impact of various components of the network in Table 12 on image (iNaturalist), video (Kinetics-400) and audio (ESC50) modalities. Specifically, we study the impact of pretraining with a single modality only, using the full pretraining mechanism, and then fine tuning on the respective training set. We also study the impact of modality specific tokenizers compared to unified tokenizers of MetaFormer [94], and impact of utilizing multiple task heads as compared to unified task head design of UniPerceiver-v2 [45]. For unimodal pretraining (Table 12-row 1), we train the network on a single modality following Step 1 of Masked pretraining (see Sec. 3.2). We use corresponding modality for each dataset *i.e.* for iNaturalist, we pretrain on ImageNet1K, for K400, we pretrain on SSv2 and for ESC50, we pretrain on AudioSet. For multimodal multitask pretraining (Table 12-row 2), we pretrain using the full pretraining discussed previously. For fine tuning, we utilize the respective train sets.

Impact of unimodal vs. multimodal pretraining We can observe that multimodal multitask pretraining using our approach (row 5) provides a significant improvement in comparison to unimodal pretraining (row 1). Specifically, it outperforms unimodal pretraining by $\sim 16\%$ on iNaturalist and K400 datasets while is better by $\sim 8\%$ on ESC50. This demonstrates that the network is able to leverage the information from multiple modalities.

Modality	Tokenizer	Task Head	iN2018	K400	ESC50
Single	Modality	Autoencoder	74.2	78.6	82.4
Single	Unified [16]	Autoencoder	74.1	78.3	82.1
Multiple	Unified [16]	Unified [45]	80.1	81.8	82.7
Multiple	Modality	Unified [45]	85.4	84.8	86.8
Multiple	Unified [16]	Task specific	86.1	85.2	87.0
Multiple	Modality	Task specific	90.3	88.4	92.4

Table 12. **Ablation experiments.** We vary the different components of the network to study the impact (Sec 4.5). Metric reported is top-1 accuracy.

Impact of modality specific tokenizer vs. unified tokenizer. We observe that the performance of unified tokenizer (row 3) lags behind that of a modality specific tokenizer (row 4) by an average of $\sim 5\%$ across all the tasks, while keeping unified heads. Similarly, while keeping task specific heads, and modality specific tokenizer (row 6) vs unified tokenizer (row 5), we observe an average performance gap of 4% in favour of modality specific tokenizer.

Multiple task heads vs unified task head. Comparing row 4 and row 6, we see that the the task specific heads contribute to an increase (average 3.5%) in performance while keeping a modality specific tokenizer.

5. Conclusion

We presented a novel multimodal multitask network and associated training algorithm. The proposed method utilizes modality specific tokenizers and then uses shared transformers based backbone feeding to task specific heads. The traning proceeds in three stages, (i) masked pretraining with one modality at a time, (ii) masked pretraining with pairs of modalities together, and (iii) supervised traning for tasks with pairs of modalities together. The pairwise pretraining and supervised training allows for knowledge sharing between tasks and modalities and leads to a robust and regularized network. We showed empirical results on 25 challenging benchmark datasets over 12 modalities obtaining better or close to existing state of the art results. The method can incorporate arbitrary number of modalities, with only the tokenizer and task heads being modality specific.

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, 2, 4, 5, 6

[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. 5

[4] Alan Baade, Puyuan Peng, and David Harwath. Mae-ast: Masked autoencoding audio spectrogram transformer. *arXiv preprint arXiv:2203.16691*, 2022. 2

[5] 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, 2

[6] Jing Bai, Jiawei Lu, Zhu Xiao, Zheng Chen, and Licheng Jiao. Generative adversarial networks based on transformer encoder and convolution block for hyperspectral image classification. *Remote Sensing*, 14(14):3426, 2022. 8

[7] 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. 2

[8] 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

[9] Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*, 2021. 5

[10] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020. 1

[11] 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

[12] 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, 6

[13] 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, 5, 7

[14] 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. 8

[15] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021. 8

[16] Zhang et al. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023. 2, 5, 7, 8

[17] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021. 4

[18] 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

[19] 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

[20] 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, 4, 6

[21] 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, 2, 6

[22] 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. 5

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

[24] 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, 2

[25] 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

[26] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson

Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 5

[27] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 270–287, 2018. 2

[28] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martin, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022. 2

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, 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. 6

[30] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[31] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021. 8

[32] 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

[33] Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. Ogb-lsc: A large-scale challenge for machine learning on graphs. *arXiv preprint arXiv:2103.09430*, 2021. 5

[34] 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, 2

[35] 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, 2

[36] 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

[37] 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

[38] 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. 5

[39] 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. 6

[40] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020. 8

[41] Hildegard Kuehne, Hueihan Jhuang, Estibaliz 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. 5

[42] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 95–104, 2018. 5

[43] 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. 6

[44] 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. 6

[45] 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, 2, 3, 5, 8

[46] 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

[47] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *NeurIPS*, 2019. 8

[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] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[50] 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. 2

[51] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-

complexity pyramidal attention for long-range time series modeling and forecasting. In *ICLR*, 2021. 8

[52] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 7

[53] 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. 5

[54] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 5, 7

[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. 5

[57] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. 5

[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

[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. 6

[61] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 7

[62] Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, et al. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020. 5

[63] Adria Recasens, Jason Lin, Joao 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

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

[65] 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. 5

[66] 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

[67] 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. 5

[68] Siddharth Srivastava and Gaurav Sharma. Omnivec: Learning robust representations with cross modal sharing. *arXiv preprint arXiv:2311.05709*, 2023. 2, 4, 5, 6, 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] 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. 6

[71] 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

[72] 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. 6

[73] Denis Uchaev and Dmitry Uchaev. Small sample hyperspectral image classification based on the random patches network and recursive filtering. *Sensors*, 23(5):2499, 2023. 8

[74] 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. 5

[75] 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. 5

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[77] 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

[78] 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

[79] 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. 2

[80] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *ICCV*, pages 225–234, 2021. 8

[81] 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. 6

[82] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision*, pages 68–85. Springer, 2022. 4

[83] 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. 6

[84] 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

[85] 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

[86] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 5

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

[88] 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. 2

[89] 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. 6

[90] 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. 5

[91] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020. 8

[92] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 8

[93] 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

[94] 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. 2, 5, 6, 8

[95] 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. 2

[96] 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

[97] 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. 6

[98] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018. 2

[99] Yiyuan Zhang, Sanyuan Zhao, Yuhao Kang, and Jianbing Shen. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 462–479. Springer, 2022. 8

[100] 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. 5

[101] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 6

[102] Fangqin Zhou, Mert Kilickaya, and Joaquin Vanschoren. Locality-aware hyperspectral classification. *arXiv preprint arXiv:2309.01561*, 2023. 8

[103] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, 2021. 5, 8

[104] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 5

[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, 2, 5