Supplementary Material

OmniVec: Learning robust representations with cross modal sharing

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

1. Ablation on increasing number of parameters of base encoders

The details on influence of increasing the number of parameters termed as Modified encoder, of base modality encoders is provided in Table 1. Our observations are as follows:

OmniVec's Performance: OmniVec (FT), which is OmniVec(Pre.) after fine-tuning, consistently outperforms the other methods across all datasets. This suggests that fine-tuning OmniVec is beneficial and leads to superior performance.

Base vs Modified Encoder: The Modified Encoder generally performs better than the Base Encoder. While, the degree of improvement varies across datasets such as on datasets like Sun RGBD, we notice a substantial improvement of 5.1 percentage points, others like ImageNet1K and AudioSet(A) show relatively minor improvements. However, this relative improvement is significantly lower as compared to that obtained with OmniVec(Pre.) or OmniVec(FT). This suggests that the modifications may be especially beneficial for certain types of data or tasks, while training on multiple modalities provides consistent improvement across all modalities and tasks. This also indicates robustness and versatility achieved by OmniVec.

2. More Implementation Details

In addition to the datasets used for masked pretraining and training on multiple modalities, we also report results on additional datasets including both seen and unseen tasks. We use standard train/test split for each of the datasets for training and evaluating OmniVec i.e. masked pretraining, training on multiple tasks, modalities and generalization.

For demonstrating the generalization on unseen datasets, we compare the results against state-of-the-art methods on Oxford-IIIT Pets (image classification) [11], UCF-101 [15], HMDB51 (video action recognition) [7], ScanObjectNN (3D point cloud classification) [17], NYU v2 seg (point cloud segmentation) [14] and SamSum (text summarization) [5]. We evaluate the method on unseen task on KITTI depth prediction [16]. We obtain results on standard test sets for each of the tasks.

We do not fine-tune the base OmniVec network on any of these tasks and term it as OmniVec(Pre.) throughout the main manuscript (unless specified explicitly otherwise). The input to the network is the respective modality (text, image, point cloud, audio etc.). It is encoded with the respective encoders for these modalities as described in Table 1 (main manuscript) irrespective of the task.

Segmentation and Summarization. For segmentation and summarization, we use the same networks as described in Section 4-Task Heads (main manuscript). For reporting results with OmniVec(Pre.), we do not fine tune either encoder or decoder for evaluation on these tasks.

Classification. For classification/recognition tasks, as the classes differ from our training classes, following earlier works, we replace the Task Heads with a network consisting of two fully connected layers and a softmax classifier. We train these two layers by extracting OmniEmbeddings using the pretrained encoders of Table 1 (main manuscript) and the backbone Transformer network. We term it as OmniVec(Pre.) and we do not fine tune the backbone or the respective encoders to report results on it. For reporting results with fine-tuning (OmniVec(FT)), we use the pretrained OmniVec and fine-tune the network end-to-end on the respective training sets.

Depth Prediction. We use convolution decoder from [13] with our common transformer backbone. As the decoder works on patch wise output from the transformer encoder, we do not use a linear layer to reduce the features. We fine-tune the network in an end-to-end manner.

3. Detailed comparison with SoTA

Video Classification on UCF-101. Table 2 shows results on UCF-101 dataset for action recognition on 3-fold accuracy.

Video Classification on HMDB51 Table 3 shows comparison of state of the art method on HMDB51 dataset.

Dataset	Metric	Modality Encoder	Base Encoder	Modified Encoder	OmniVec (Pre.)	OmniVec (FT)
AudioSet(A)	mAP	AST	48.5	49.4	44.7	54.8
AudioSet(A+V)	mAP	AST	-	-	48.6	55.2
SSv2	Top-1 Accuracy	ViViT	65.4	68.6	80.1	85.4
ImageNet1K	Top-1 Accuracy	ViT	88.5	89.1	88.6	92.4
Sun RGBD	Top-1 Accuracy	Simple3D-former	57.3	62.4	71.4	74.6

Table 1. **Impact of increasing backbone size of base modality encoders.** All the base modality encoders above are based on ViT architecture. We increase the number of parameters equivalent to our OmniVec-4 model, by replicating the number of layers.

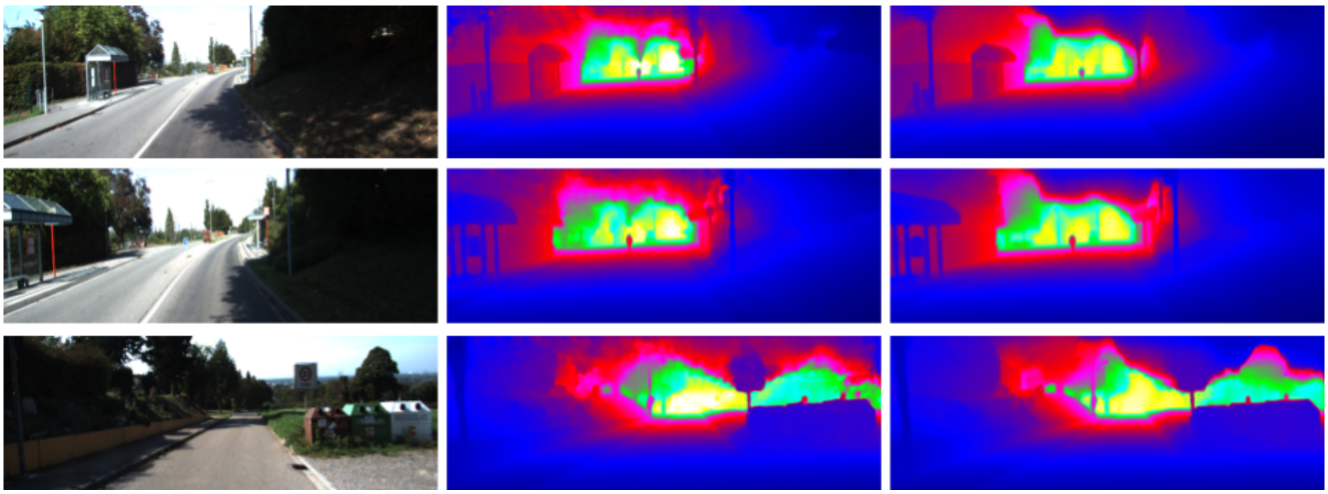

Figure 1. More qualitative results on Monocular depth prediction on KITTI test set. (From left to right) Input Image, Depth image generated using VA-DepthNet, Depth image generated using OmniVec. It can be observed that OmniVec predicts sharper depth around far away objects and on boundaries.

3D Point Cloud Classification on ScanObjectNN. Table 5 compares OmniVec against state of the art methods.

Semantic Segmentation on NYU v2 seg. Table 6 shows result on NYU v2 segmentation against state of the art methods.

Text summarization on SamSum. Table 7 compares our method against state of the art methods. An important observation here is that our method has not been specifically trained for solving text related tasks nor the OmniVec(Pre.) network has been fine tuned for this dataset. This shows that the learning mechanism is able to generalize across domains as well.

4. Qualitative Results.

We present additional qualitative findings in Figure 1. When compared to the previous benchmark, VA-DepthNet, our suggested approach demonstrates superior depth perception at boundaries and distant objects. Notably, our method offers enhanced depth discernment for objects such as the bus shelter (highlighted in the top and middle images) as well as houses (depicted in the bottom image).

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. 3

[2] 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. 3

[3] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 3

[4] 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. 3

[5] 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. 1

[6] Shreyank N Gowda, Marcus Rohrbach, and Laura Sevilla-Lara. Smart frame selection for action recognition. In *Pro-*

Method	UCF-101
VATT [1]	87.6
Omnivore [4]	98.2
Text4Vis [20]	98.2
SMART [6]	98.6
VideoMAE V2-g [18]	**99.6**
OmniVec(Pre.)	98.71
OmniVec(FT)	**99.6**

Table 2. **UCF-101 Action Recognition**. Metric is 3-fold accuracy.

Method	HMDB51
VATT [1]	66.4
DEEP-HAL [19]	87.56
VideoMAE V2-g [18]	88.10
OmniVec(Pre)	89.21
OmniVec(FT)	**91.6**

Table 3. **Comparison to state-of-the-art methods on HMDB51 dataset for Action Recognition**. Metric is 3-split accuracy.

Method	Pets (top-1)	Pets (top-5)
Omnivore [4]	95.1	99.1
IELT [21]	95.28	-
DINOv2 [10]	96.70	-
EffNet-L2 [3]	97.10	-
OmniVec(Pre.)	97.36	99.3
OmniVec(FT)	**99.2**	**99.7**

Table 4. **Comparison to state-of-the-art methods on Fine grained image classification on Oxford-IIIT Pets dataset**. The metrics are top-1 and top-5 accuracy.

Method	Scan Object NN
PointConT [9]	90.3
ReCon [12]	91.3
ULIP-2 [22]	91.5
PointGPT [2]	93.4
OmniVec(Pre.)	92.10
OmniVec(FT.)	**96.10**

Table 5. **Comparison to state-of-the-art methods on ScanObjectNN for 3D point cloud classification**. Metric is Overall Accuracy.

Method	NYUv2
Omnivore [4]	56.8
CMN [8]	56.9
OmniVec(Pre.)	58.6
OmniVec(FT)	**60.8**

Table 6. **Comparison to state-of-the-art methods on NYU v2 for semantic segmentation**. Metric is mean IoU. Note that the network has not been fine-tuned on this dataset nor any additional network has been used.

Method	R-1	R-2	R-L
Pegasus [24]	54.37	29.88	45.89
MoCa [23]	55.13	**30.57**	50.88
OmniVec(Pre.)	54.81	30.10	51.21
OmniVec(FT)	**58.81**	**31.1**	**53.4**

Table 7. **SamSum dataset for meeting summarization**. Metric are ROGUE scores. Note that the network has not been fine-tuned on this dataset nor any additional network has been used.

ceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 1451–1459, 2021. 3

[7] Hildegard Kuehne, Hueihan Jhuang, Est´ıbaliz 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. 1

[8] 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. 3

[9] Yahui Liu, Bin Tian, Yisheng Lv, Lingxi Li, and Feiyue Wang. Point cloud classification using content-based transformer via clustering in feature space. *arXiv preprint arXiv:2303.04599*, 2023. 3

[10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[11] 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. 1

[12] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. *arXiv preprint arXiv:2302.02318*, 2023. 3

[13] 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. 1

[14] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV (5)*, 7576:746–760, 2012. 1

[15] 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. 1

[16] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 1

[17] 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. 1

[18] 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. 3

[19] Lei Wang and Piotr Koniusz. Self-supervising action recognition by statistical moment and subspace descriptors. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4324–4333, 2021. 3

[20] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. *Proceedings of the AAAI, Washington, DC, USA*, pages 7–8, 2023. 3

[21] Qin Xu, Jiahui Wang, Bo Jiang, and Bin Luo. Fine-grained visual classification via internal ensemble learning transformer. *IEEE Transactions on Multimedia*, 2023. 3

[22] Le Xue, Ning Yu, Shu Zhang, Junnan Li, Roberto Mart´in-Mart´in, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. *arXiv preprint arXiv:2305.08275*, 2023. 3

[23] 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. 3

[24] Yao Zhao, Misha Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. Calibrating sequence likelihood improves conditional language generation. *arXiv preprint arXiv:2210.00045*, 2022. 3