# A. Implementation details for Pretraining

We train using AdamW with a batch size of 4096 for each dataset, and use a cosine learning rate (LR) schedule with linear warm up and cool down phases for the first and last 10% of training, respectively. We train for 500 epochs with a peak LR of $2 \cdot 10^{-3}$ and a weight decay of $5 \cdot 10^{-2}$. Swin-T, Swin-S and Swin-L use a window size of $8 \times 7 \times 7$, whereas Swin-B uses a window size of 16×7×7. The models are trained with stochastic depth with a drop rate of 0.1 for Swin-T, 0.2 for Swin-S, and 0.3 for Swin-B, and Swin-L. We use exponential moving average (EMA) [73] with a decay of $10^{-4}$ and report the best results during training since EMA results peak before the end of training.

For IN1K and IN21K we use RandAugment [19], mixup [101], CutMix [98], label smoothing [85], and Random Erasing [104] with the same settings as used in [88], and color jittering of 0.4. For SUN RGB-D we clamp and normalize the disparity channel, drop the RGB channels with a probability of 0.5, and we also apply 0.5 Dropout [82] before the linear head when pre-training with ImageNet-21K. For Kinetics-400 we use mixup, CutMix and label smoothing, and Dropout of 0.5 before the linear head.

# B. Details on the Transfer Tasks

## B.1. Image Classification

We finetune all models on the downstream tasks for 100 epochs and optimize the models with mini-batch SGD. We use a half-wave cosine learning rate [54] and set the weight decay to zero. For all models, including the modality-specific models, we perform a grid search for the best learning rate in the range [5e-3, 1e-2, 2e-2, 4e-2, 8e-2, 1e-1, 2e-1, 3e-1, 4e-1, 5e-1, 6e-1] and drop path in [0.1, 0.3]. We use the strong augmentations from [88] for finetuning. For the evaluations in Tables 3 and 5, we follow [78] and resize the images to shortest side of 224px and evaluate the models on the center crop of $224 \times 224$. For higher resolution (384px) evaluations in Table 5, we similarly resize the images to shortest side of 384px and evaluate the models on the center crop of $384 \times 384$. We also increase the spatial window size for all the Swin models from 7 to 12.

## B.2. Video Classification

In Table 3, we finetune video models using hyperparameters as described in [52]. For Something Something-v2, we finetune for 60 epochs with AdamW optimizer. We use half-wave cosine learning rate with warmup. We start the learning rate from $10^{-6}$ and linearly warmup to a peak learning rate of $6 \cdot 10^{-3}$ over 5% of the training, and rest 95% we use half-wave cosine schedule to decay the learning rate back to $10^{-6}$. We train the classification head with this learning rate, and the backbone with $0.1\times$ the above learning rate.

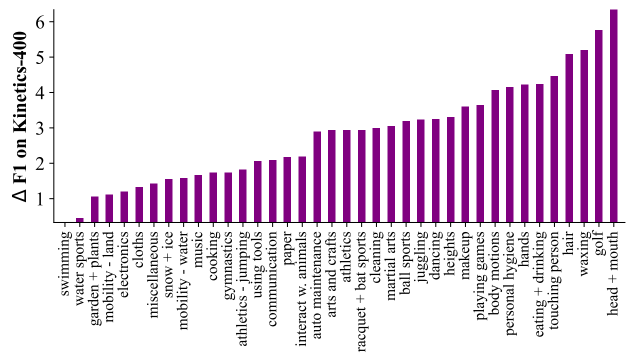


**Figure 6. Gain of OMNIVORE over baseline on Action recognition (per group).** We plot the gain in per-class F1-score on the K400 dataset for all the action groups defined in [13]. The baseline model is first pretrained on ImageNet-1K and then fine-tuned on K400 whereas OMNIVORE is trained jointly on ImageNet-1K, K400 and the single-view 3D SUN RGB-D dataset. OMNIVORE improves the performance for all the 38 groups.

Throughout we use a weight decay of 0.05. We use a batch size of $4 \times 64$ distributed over 64 32GB GPUs. For EPIC-Kitchens-100, we use similar hyperparamters with only difference being that we use a peak learning rate of $2 \cdot 10^{-3}$ and we train for 150 epochs. These settings provided better performance for the modality-specific baseline, and we use it for finetuning both the baseline and OMNIVORE models.

In terms of preprocessing, at train time we sample a 32 frame video clip at stride 2 from the full video using temporal segment sampling as in [52]. We scale the short side of the video to 256px, take a 224px random resized crop, followed by RandAugment and Random Erasing. At test time, we again sample a 32 frame clip with stride 2, scale the short side to 224px and take 3 spatial crops along the longer axis to get $224 \times 224$ crops. The final predictions are averaged over these crops.

For comparison to the state-of-the-art in Table 6, when finetuning OMNIVORE models trained with IN21K, we found slightly different hyperparameters to perform better. For Something Something-v2, we used peak learning rate of $1.2 \cdot 10^{-3}$ over 150 epochs. For EPIC-Kitchens-100, we used weight decay of 0.004, over 100 epochs, peak learning rate of $4 \cdot 10^{-4}$, with the same learning rate schedule for backbone and head. We also used cutmix augmentation and label smoothing. All other hyperparameters in both cases were as described earlier. We also use EMA with similar settings as used during pretraining.

## B.3. Single-view 3D Tasks

**NYU Scene classification.** We follow the setup from [33] for scene classification and use 10 classes derived from the original 19 classes. In Table 7 (classification) the best Swin

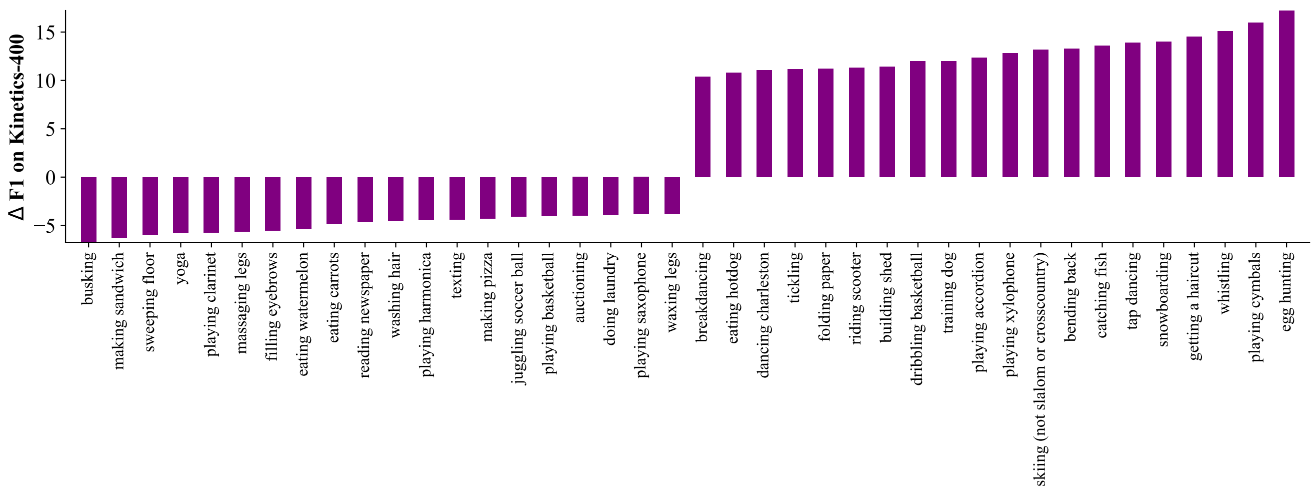


**Figure 7. Gain of OMNIVORE over baseline on Action Recognition (per class).** We plot the gain in per-class F1-score on the K400 dataset for the top twenty and bottom twenty classes. The baseline model is first pretrained on ImageNet-1K and then fine-tuned on K400 whereas OMNIVORE is trained jointly on ImageNet-1K, K400 and the single-view 3D SUN RGB-D dataset. OMNIVORE improves the F1 score on 308 out of the 400 total classes.

B and Swin L models were trained for 200 epochs with starting learning rate of $5 \times 10^{-3}$, weight decay of 0 for Swin B and $1 \times 10^{-4}$ for Swin L. All other hyperparameters were as described earlier.

**NYU RGBD Segmentation.** We follow the training and evaluation setup from [10]. We follow the Swin segmentation architecture which uses an UperNet [95] head with the Swin trunk. All models are finetuned with AdamW [53] with a weight decay of 0.01. The learning rate follows a Polynomial Decay (power 1) schedule and starts at 0.00006. We warmup the learning rate for 1500 iterations and train the model with a batchsize of 32. All the depth maps in NYU are converted into disparity maps by using the camera baseline and focal length of the Kinect sensor.

### B.4. $k$-NN experiments

**Extracting depth on ImageNet-1K.** We ran a monocular depth-prediction model [74] on the IN1K train set. We used the pretrained `dpt_large` model and followed the input image preprocessing steps as provided in [74].

**Classifying ImageNet-1K using different modalities.** For the experiments involving classification using different modalities, we extract features from the IN1K train set using the RGB, RGBD or just Depth (D) modalities, and on IN1K validation set using the RGB modality. We follow the $k$-NN protocol from [12] for evaluation and briefly describe it next. We extract the stage 3 [51] features and $L_2$ normalize them. For each validation feature as the query, we retrieve the nearest neighbors from the train set using euclidean distance, and take the top-$k$ closest matches. For each match we create a one-hot vector using its ground truth label, and scale it by $e^{s/\tau}$, where $s$ is the dot product between the feature of the matched image the query image, and $\tau$ is a temperature hyperparameter (set to 0.07). We compute an effective prediction for the query by summing the top-$k$ one-hot vectors. Similar processing is used for the visualizations in Figure 1 and Figure 4.

| | VideoSwin-B | OMNIVORE (Swin-B) |
|---|---|---|
| 3-split accuracy | 96.9 | **98.2** |

**Table 9. UCF-101.** As in Table 3, the VideoSwin model is inflated from IN1K and pre-trained on K400. OMNIVORE is pretrained with IN1K, K400 and SUN RGB-D. Both models are then finetuned and evaluated on UCF-101 for each split separately. Performance reported is averaged over the standard 3 splits.

## C. Other Results

**Results on UCF-101.** We also evaluate OMNIVORE on another popular (albeit smaller) video recognition benchmark, UCF-101 [81]. As shown in Table 9, OMNIVORE pretraining is effective for sports action recognition in UCF-101 as well. Note that the results shown are with RGB modality only; the state-of-the-art on these datasets often leverages additional features such as optical flow, dense trajectories (IDT) etc.

**Low-data regime fine-tuning.** We analyzed low-shot versions of the Places-365 benchmark (models from Table 3). As shown in Table 10, OMNIVORE outperforms the modality-specific baseline in the low-shot regime too.

| Method | Places-365 | | | |
|---|---|---|---|---|
| | 1% | 2% | 5% | 10% |
| OMNIVORE | **46.2** | **49.0** | **51.5** | **53.9** |
| Image-specific | 44.8 | 47.9 | 50.9 | 53.4 |

**Table 10. Low-shot finetuning.** Performance of finetuning OMNIVORE on low-shot versions of the Places-365 dataset.

**Per-class gains.** We present the gain of OMNIVORE over the VideoSwin baseline (§ 4.1 of the main paper) in Figs. 6 and 7.

## References


[1] Hassan Akbari, Linagzhe 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. *arXiv preprint arXiv:2104.11178*, 2021.

[2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2020.

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017.

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, 2018.

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. ViViT: A video vision transformer. In *ICCV*, 2021.

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

[7] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *arXiv preprint arXiv:2104.00650*, 2021.

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.

[9] Ali Caglayan, Nevrez Imamoglu, Ahmet Burak Can, and Ryosuke Nakamura. When cnns meet random rnns: Towards multi-level analysis for rgb-d object and scene recognition. *arXiv preprint arXiv:2004.12349*, 2020.

[10] Jinming Cao, Hanchao Leng, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *ICCV*, 2021.

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[13] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[14] Rich Caruana. Multitask learning. *Machine Learning*, 1997.

[15] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *CVPR*, 2016.

[16] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *ECCV*, 2020.

[17] 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 *ECCV*, 2020.

[18] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019.

[19] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR*, 2020.

[20] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *IJCV*, 2021.

[21] 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. In *ICLR*, 2021.

[22] Dapeng Du, Limin Wang, Huiling Wang, Kai Zhao, and Gangshan Wu. Translate-to-recognize networks for rgb-d scene recognition. In *CVPR*, 2019.

[23] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015.

[24] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021.

[25] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

[26] Kunihiko Fukushima. A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.*, 1980.

[27] Golnaz Ghiasi, Barret Zoph, Ekin D Cubuk, Quoc V Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *ICCV*, 2021.

[28] Rohit Girdhar, João Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, 2019.

[29] Rohit Girdhar and Kristen Grauman. Anticipative Video Transformer. In *ICCV*, 2021.

[30] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, 2014.

[31] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

[32] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018.

[33] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *CVPR*, 2013.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*,

2016.
[35] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. *arXiv preprint arXiv:2110.06915*, 2021.
[36] 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 *CVPR*, 2018.
[37] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *ICCV*, 2021.
[38] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. *ICML*, 2021.
[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.
[40] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021.
[41] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.
[42] 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.
[43] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. In *BMVC*, 2021.
[44] Iasonas Kokkinos. UberNet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, 2017.
[45] Stepan Komkov, Maksim Dzabraev, and Aleksandr Petiushko. Mutual modality learning for video action classification. *arXiv preprint arXiv:2011.02543*, 2020.
[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012.
[47] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *ICCV*, 2003.
[48] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.
[49] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.
[50] Yabei Li, Junge Zhang, Yanhua Cheng, Kaiqi Huang, and Tieniu Tan. Df$^2$net: Discriminative feature learning and fusion network for RGB-D indoor scene classification. In *AAAI*, 2018.
[51] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.
[52] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.
[53] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
[54] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.
[55] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.
[56] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.
[57] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020.
[58] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In *CVPR*, 2019.
[59] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020.
[60] Ishan Misra, Rohit Girdhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. In *ICCV*, 2021.
[61] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, 2016.
[62] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *CVPR*, 2021.
[63] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *CVPR*, 2021.
[64] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, 2021.
[65] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.
[66] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021.
[67] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018.
[68] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *CVPR*, 2021.
[69] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012.
[70] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018.
[71] Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transforma-

tions. *arXiv preprint arXiv:2003.04298*, 2020.

[72] Mandela Patrick, Dylan Campbell, Yuki M Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *NeurIPS*, 2021.

[73] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 1992.

[74] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020.

[75] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[76] Fadime Sener, Dibyadip Chatterjee, and Angela Yao. Technical report: Temporal aggregate representations. *arXiv preprint arXiv:2106.03152*, 2021.

[77] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.

[78] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *CVPR*, 2022.

[79] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015.

[80] Xinhang Song, Shuqiang Jiang, Bohan Wang, Chengpeng Chen, and Gongwei Chen. Image representations with spatial object-to-object relations for rgb-d scene recognition. *TIP*, 2020.

[81] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human action classes from videos in the wild. *CRCV-TR-12-01*, 2012.

[82] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[83] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[84] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[85] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[86] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[87] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. VIMPAC: Video pre-training via masked token prediction and contrastive learning. *arXiv preprint arXiv:2106.11250*, 2021.

[88] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

[89] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. In *NeurIPS*, 2019.

[90] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *CVPR*, 2015.

[91] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

[92] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[93] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021.

[94] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[95] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018.

[96] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *CVPR*, 2020.

[97] Yuchun Yue, Wujie Zhou, Jingsheng Lei, and Lu Yu. Two-stage cascaded decoder for semantic segmentation of rgb-d images. *IEEE Signal Processing Letters*, 2021.

[98] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[99] Amir Roshan Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018.

[100] Bowen Zhang, Jiahui Yu, Christopher Fifty, Wei Han, Andrew M. Dai, Ruoming Pang, and Fei Sha. Co-training transformer with videos and images improves action recognition. *arXiv preprint arXiv:2112.07175*, 2021.

[101] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[102] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014.

[103] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021.

[104] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020.

[105] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017.