OmniMAE: Single Model Masked Pretraining on Images and Videos

Rohit Girdhar* Alaaeldin El-Nouby* Mannat Singh*
Kalyan Vasudev Alwala* Armand Joulin Ishan Misra*
FAIR, Meta AI

https://github.com/facebookresearch/omnivore

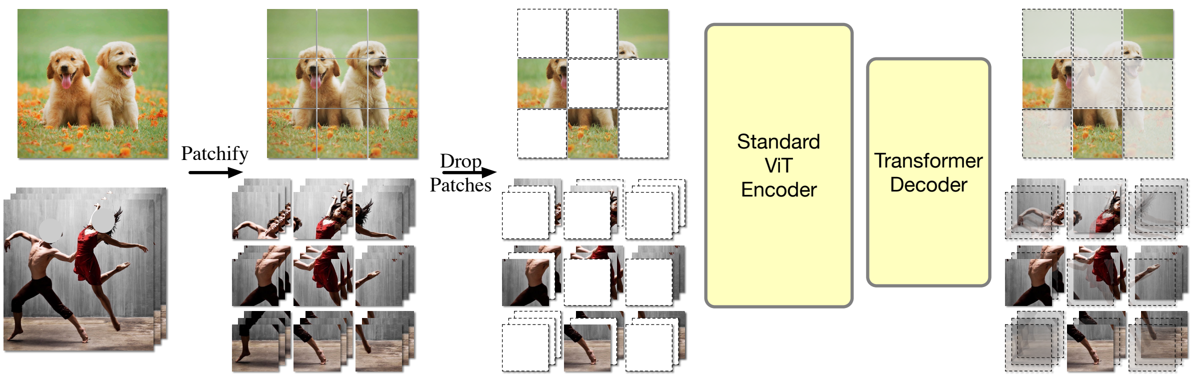


Figure 1. **OmniMAE** is a single model for images and videos that is trained using masked autoencoding [40]. We use a plain Vision Transformer [24] architecture but with spatio-temporal patches as input. At training, we 'patchify' the visual input (images or videos), and feed the encoder only a subset of the patches. The decoder reconstructs the pixels for the missing patches using the encoder's output. The encoder-decoder model is trained using a pixel reconstruction loss. After training, our single plain Transformer encoder performs competitively compared to specialized architectures on downstream image and video recognition tasks.

Abstract

Transformer-based architectures have become competitive across a variety of visual domains, most notably images and videos. While prior work studies these modalities in isolation, having a common architecture suggests that one can train a single unified model for multiple visual modalities. Prior attempts at unified modeling typically use architectures tailored for vision tasks, or obtain worse performance compared to single modality models. In this work, we show that masked autoencoding can be used to train a simple Vision Transformer on images and videos, without requiring any labeled data. This single model learns visual representations that are comparable to or better than single-modality representations on both image and video benchmarks, while using a much simpler architecture. Furthermore, this model can be learned by dropping 90% of the image and 95% of the video patches, enabling extremely fast training of huge model architectures. In particular, we show that our single ViT-Huge model can be finetuned to achieve 86.6% on ImageNet and 75.5% on the challenging Something Something-v2 video benchmark, setting a new state-of-the-art.

1. Introduction

The Transformer architecture [78] is rapidly becoming competitive across the different visual modalities in Computer Vision, from images [24, 27, 55, 77], to 3D [57, 60, 89] and videos [2, 9, 27, 31, 32, 56]. This convergence toward a unified architecture naturally suggests that we should be able to train a single model that works across different visual modalities. However, recent attempts to train unified models either lead to worse performance compared to single modality models [53], or require the use of an alternative architecture [33], namely the Swin Transformer [55], with inductive biases tailored towards vision tasks. While specialized Transformer architectures for vision [27, 55, 56, 81] can offer better performance for visual modalities, they lose the generality and flexibility of the vanilla Transformer, making it harder to later model different domains like text, speech, 3D *etc*. in multi-modal architectures.

In this work, we train a single vanilla Transformer that works for both images and videos, as illustrated in Figure 1.

*Equal technical contribution.

To this end, we leverage the findings of several recent works on the use of the masked pretraining [23] to greatly improve the training and performance of Transformers in the domain of images [6, 40, 80, 84], videos [76, 80, 82] or across text, audio and images [5]. We show that this masked pretraining is a viable strategy to pretrain a unified 'omnivorous' Transformer across visual modalities. In particular, we consider the Masked Auto-Encoding (**MAE**) approach [40] to train an **Omni**vorous visual encoder [33]. The resulting **OmniMAE** model learns from all the modalities with the same objective function and does not require any supervision.

Using a masked pretraining objective has several advantages over supervised objectives [33, 53] or discriminative self-supervised objectives [13, 17, 41]. First, as opposed to supervised objectives, a general-purpose unsupervised loss does not require any human labeling effort. As a result, it is robust to biases introduced by a predefined set of labels [35]. Moreover, it does not require a multi-head architecture to incorporate supervision from each of the label spaces corresponding to each modality, which is hard to maintain and scale with new modalities. Second, although discriminative self-supervised methods produce superior frozen features compared to reconstruction objectives, they are non trivial to scale in model and data size [36]. Our masked pretraining objective is simple, efficient to train, and scales to different visual modalities with minimal changes.

Contributions. (1) We show that the simple Vision Transformer architecture (ViT) [24] originally designed for images can naturally be applied on videos, and videos and images jointly. OmniMAE is a single ViT-based model for videos and images that outperforms architectures and models specifically designed for either modality. **(2)** Prior and concurrent work design self-supervised methods and architectures for either image or video and we find that these models do not transfer well across modalities. OmniMAE is the first single self-supervised model that achieves good performance on both modalities. **(3)** We show that our joint training using both images and videos enables us to use much higher masking ratios than any prior work for training MAE. Since ViT can processes only the non-masked input, we train OmniMAE models with only 10% of image and 5% of video patches. This enables us to train large (650M parameter) models with a $\sim 7\times$ and $\sim 11\times$ reduction in compute and memory on images and videos. **(4)** Finally, we propose improvements to the MAE training. We show that repeating samples in a mini-batch reduces dataloading (and thus training) time without loss in final transfer performance. Sample replication is particularly useful for masked pretraining as the unmasked patches are different across sample replicas. We also show that using a shallow shared decoder for videos and images leads to better performance while reducing the number of parameters by $2-4\times$.

2. Related Work

Our work builds upon research in self-supervised learning, masked pretraining and unified modeling in computer vision.
Self-supervised learning. In recent years, self-supervised approaches have been dominated by joint embedding methods which can rely on different objectives including contrastive [17,39,41,51,61,65,75,83], non-contrastive [7,18,26, 87], clustering [3, 11, 12, 85] or self-distillation [5, 13, 38, 92]. Such methods are trained to learn invariance to a set of predefined transformations which results in image descriptors with a strong linear probing and KNN performance. However, such methods can be challenging to scale since they can suffer from instabilities [19]. Additionally, the strongest performance is typically obtained with the help of augmentations like multi-crop [12, 13, 92] which can be hard to apply at scale due their compute and memory overhead.
Masked pretraining. We build upon masked prediction methods where the representation is learned by predicting masked parts of the input. Such methods have recently gained popularity given their immense success in NLP. In particular, BERT [23] showed that masked language modeling by predicting a subset of masked words is a powerful pre-training objective and leads to impressive finetuning performance on various downstream tasks. In computer vision, input reconstruction methods have a rich history with non-linear PCA [49], sparse reconstruction [64], autoencoders [10, 30, 42, 50], RBMs [71] *etc*. Masked prediction methods can be viewed as a special case of denoising autoencoders [30, 79] where the input 'noise' is a masking function. An example of such a method that uses masking as noise is context encoders [68]. With the recent and rapid rise of Vision Transformers [24], masked prediction was revisited by multiple efforts. SiT [4] replaces variable sized patches in the image with random noise and trains a ViT model for reconstruction, among other objectives. BEiT [6] moves a step closer to BERT with replacing full patches with mask tokens and training a ViT encoder to predict the discrete visual words of masked patches using a cross-entropy loss. Masked prediction has also shown impressive performance for specialized vision transformer architectures such as MViT [27, 52, 82] and Swin transformer [54, 55, 84]. SimMIM [84] and MaskedFeat [82] predict pixel values and HOG features of the masked patches using a Swin-v2 [54] and MViT-v2 [52] backbones respectively. Finally, SplitMask [25] studied the interesting properties of masked prediction methods in terms of high sample efficiency.

Of particular interest to OmniMAE, masked autoencoders (MAE) [40] demonstrated impressive scaling properties by utilizing a patch dropping strategy for masked patches accompanied with a high masking ratio of 75%. Under this setting, the encoder only process a small subset of the image patches followed by a relatively low capacity decoder which reconstructs the image pixels. This property is even more crucial for video representation learning given the large number of patches, and a concurrent work, ST-MAE [76], shows that MAE pretraining with an even higher masking ratio of 90% works well and obtains strong finetuning performance on downstream video recognition benchmarks. Notably, the efficiency gained by patch dropping is specific to vanilla ViTs, and multiscale architectures such as MViT and Swin are unable to benefit due to their design. Hence, we use the simple ViT as our architecture and show that it can be trained efficiently and jointly for images and videos using extremely high masking ratios (90-95% on both modalities), and yet perform competitively to specialized architectures like MViT [27] and Swin [55].

Unified modeling and multi-modal learning. Multi-modal learning in computer vision has a long history that includes training using images and text [15, 34, 47, 58, 59], video and optical flow [29, 72], and video and audio [1, 62, 63, 66]. The majority of such methods rely on training a separate backbone for each modality as well as the availability of alignment across modalities. More recently, Omnivore [33] was proposed for joint training of multiple modalities like images, videos and single-view 3D, attaining a strong performance in each of the modality specific benchmarks with a single shared trunk. PolyViT [53] co-trains a shared transformer encoder using images, videos and audio data and provides a competitive performance on various downstream tasks for each modality. The aforementioned methods differ compared to OmniMAE in that they are trained with supervised learning and require human annotations. BEVT [80] tackles BERT pre-training for videos and proposes that jointly pre-training using static images improves the finetuning performance of video recognition benchmarks. Unlike OmniMAE, BEVT uses the specialized Swin transformer architecture with separate decoder heads for images and videos. Thus, BEVT cannot drop patches while training which can limit its scalability. Furthermore, it relies on a tokenizer which must be trained apriori, and the tokenizer training itself can affect the model's performance.

3. OmniMAE

Our goal is to pretrain a single unified model for images and videos. Rather than use specialized architectures tailored for a visual modality, we build upon the vanilla Vision Transformer (ViT) [24] architecture that has limited inductive biases for vision. For pretraining, we extend the simple self-supervised masked auto-encoding (MAE) approach [40]. The original architecture and pretraining method are tested only on images, and we show simple design decisions for a unified model.

3.1. Training OmniMAE jointly on images and videos

We illustrate our method in Figure 1. For pretraining, we use an encoder-decoder architecture where the encoder only operates on a 'non-masked' subset of the input. The decoder predicts the pixel values for the entire input, *i.e.*, masked and non-masked pixels. The model is trained to minimize the reconstruction error for the masked (unseen) part of the input. After pretraining, we evaluate the encoder by transfer learning (the decoder is discarded). Next, we describe the pretraining details.

Images and videos as spatio-temporal patches. The input image or video can be represented as a 4D tensor of shape $T \times H \times W \times 3$ where T is the temporal dimension and H, W are the spatial dimensions, and 3 represents the color channels. We treat images as being single-frame videos with $T = 1$. The input is split into N spatio-temporal patches, each of size $t \times h \times w \times 3$ [33].

Omnivorous visual encoder. We use an omnivorous [33] visual encoder that processes both images and video using the same parameters. The encoder operates on the N spatio-temporal patches from the images and videos. The encoder can naturally handle variable number N of patches from images and videos as it uses the Transformer architecture [78]. The encoder shares the same parameters for both image and video inputs and processes them in the same way to output per-patch embeddings.

Pretraining. We convert the input into N spatio-temporal patches and randomly mask M of them. Following [40], we feed the non-masked $N - M$ patches (and their positions) to the encoder to produce per-patch embeddings. The encoder output is concatenated with M copies of a learnable 'mask token' to obtain N embeddings. We add the appropriate positional encoding to the N embeddings and feed them into the decoder, which outputs $N \times t \times h \times w \times 3$ pixel values.

Loss function and optimization. We minimize the reconstruction error between the decoder predictions and the input pixel values. The input pixel values are normalized to zero mean and unit variance [46] to get the targets for the loss. We minimize the ℓ_2 distance between the predictions and targets over the M masked patches. At each point in the training, we sample a mini-batch of either images or video and compute the loss using the decoder predictions. We study the effect of different mini-batch construction strategies on the overall performance and speed of training in § 4.2.

Dataset sample replication. Since we operate on a small subset of the input patches ($M \ll N$), we find that our data reading speed is unable to keep up with the number of data points we can theoretically process in a single optimization step. This issue is even more pronounced for videos, where we spend significant compute to read and decode a full video, only to discard $> 90\%$ of it. Inspired from the success of RepeatAugment, which shows that replicating samples

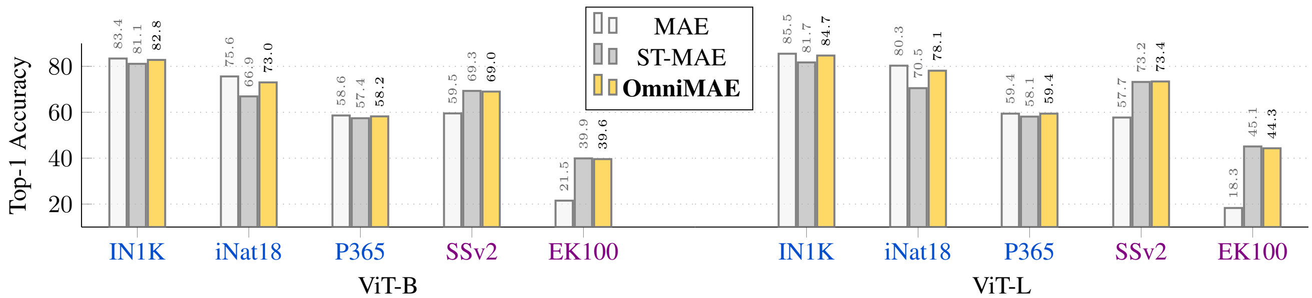


Figure 2. **OmniMAE on image and video downstream tasks.** We finetune the MAE, ST-MAE, and OmniMAE models on image and video benchmarks. We use the ViT architecture with two model sizes: ViT-B and ViT-L. MAE has poor video recognition performance while ST-MAE's performance drops on image datasets. OmniMAE pretraining generalizes to both benchmarks. All models are trained for 800 epochs on the pretraining datasets. The image-only MAE model is inflated [14] to apply MAE to video recognition tasks. The input image is replicated to apply ST-MAE to image recognition benchmarks.

within a mini-batch is an effective technique to improve generalization [8, 12, 43], we replicate a single data point and apply different masks to it each time. Even with replicated samples, the non-masked input to the model is different due to random cropping and masking being different across the replicas. We show in § 4.2 that sample replication leads to gains in runtime without affecting the final transfer accuracy.

3.2. Implementation Details

We note some of the salient implementation details and provide the complete details in Appendix B.

Architecture. We use the ViT [24] architecture for the omnivorous encoder and experiment with its ViT-B, ViT-L, and ViT-H variants. We do not use the `[CLS]` token in the ViT models, yielding a small improvement in runtime without loss in performance. We use a Transformer decoder with 4 layers (8 for ViT-H) of 384, 512, and 512 embedding dimensions for ViT-B, ViT-L, and ViT-H, respectively. The decoder outputs the RGB colors for the pixels in all the input patches. We use sinusoidal positional encoding [78] for the patches in both the encoder and the decoder.

Training details. We train our models with a mini-batch size of 2048. We resize the input images and videos spatially to 224×224 pixels. For videos, we sample a clip of $T = 16$ frames at 6 FPS. We use a patch size of $2 \times 16 \times 16$ for ViT-B and ViT-L, and $2 \times 14 \times 14$ for ViT-H. Images are replicated temporally to meet the patch size.

Masking the input. Compared to prior work, we use extremely high masking for pretraining and only 10% and 5% of the image and video patches are fed to the encoder for pretraining. We uniform randomly mask the patches for images and videos and ablate the masking hyperparameters in § 4.2.

4. Experiments

Pretraining data. We pretrain representations by jointly training on images from the ImageNet (IN1K) [70] dataset and videos from the Something Something-v2 (SSv2) [37] dataset. We choose the SSv2 dataset over web video datasets like Kinetics-400 (K400) [48] for reproducibility, and given SSv2's challenging nature requiring temporal reasoning.

Transfer learning datasets. We consider two image datasets: iNaturalist-2018 (iNat18) [44], a popular fine-grained recognition dataset, and Places-365 (P365) [91], a scene recognition benchmark. For video datasets, we focus on the popular K400 action recognition benchmark, and EPIC-Kitchens-100 (EK100) [22], an egocentric action recognition dataset. We report the top-1 classification accuracy for all transfer datasets. The details about the datasets can be found in Appendix A.

4.1. OmniMAE on Image and Video Recognition

We now evaluate the capabilities of OmniMAE's representations on both image and video recognition tasks.

Baselines. We compare OmniMAE's representations trained jointly on images and videos to MAE, trained solely on images [40]. Additionally, we develop a video-only baseline, SpatioTemporal-MAE (ST-MAE), by training OmniMAE only on videos. For a fair comparison, all models are pretrained for 800 epochs on their respective datasets. The image-only MAE model is inflated [14] for finetuning on video tasks, see Appendix B.2 for details.

Observations. Figure 2 presents the evaluation results for all the models on the image and video benchmarks. The modality specific MAE and ST-MAE models achieve good transfer performance when transferring to datasets which match the pretraining modality. However, both models show a degradation in performance when transferring to a different visual modality. MAE has lower video recognition performance, especially on egocentric videos from EK100 where the smaller ViT-B MAE model is 18.4% worse compared to ST-MAE. Similarly, compared to MAE, ST-MAE is 8.7% worse on the fine-grained classification task of iNat18. With a large model, MAE's performance on the image recognition

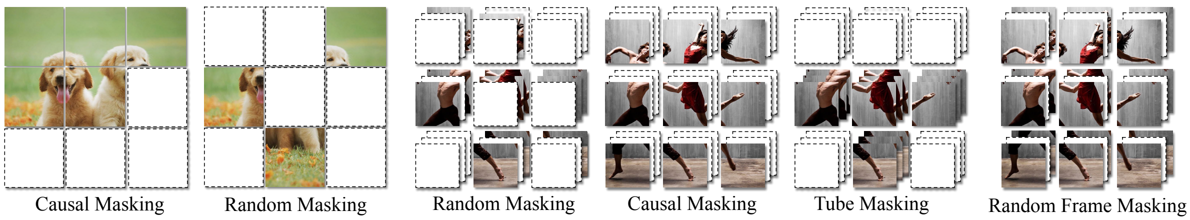
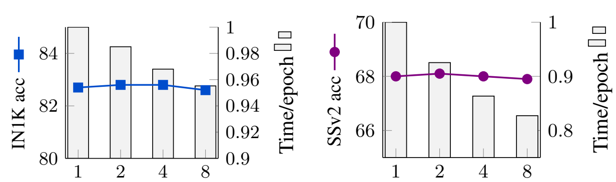


Figure 3. **Different types of masking** for images (left two) and videos. Causal and tube masking use the data's spatio-temporal structure. Random frame masking randomly masks frames in a video. Random masking randomly masks patches and is used by default for OmniMAE.

benchmarks improves but the cross-modality performance degrades further. On EPIC-Kitchens-100 the large MAE model is more than 25% worse than ST-MAE.

When transferred to both image and video recognition tasks, OmniMAE performs favorably to the single modality baselines that use exactly the same architecture. OmniMAE also uses the same finetuning recipe as the single modality baselines. OmniMAE matches the video classification performance of ST-MAE and the image classification performance of MAE for both ViT-B and ViT-L, with OmniMAE's performance improving on both image and video recognition benchmarks with the larger model. These results suggest that our unified model of images and videos serves as a better pretraining across a wider variety of recognition tasks than single modality models.

Qualitative results. Following [40], we re-train OmniMAE without normalizing the pixel targets to obtain easy to visualize RGB reconstructions. We visualize the predictions in Figure 4 using samples that are unseen during training: either `val` sets or different datasets altogether. OmniMAE makes reasonable predictions on the in-distribution ImageNet and SSv2 val sets, as well as the unseen, out-of-distribution K400 and EK100 datasets. As expected, the details in the reconstruction decrease when increasing the masking ratio. However, even with 95% masking, the model can reconstruct coarse details in the input, *e.g.*, in ImageNet, the model reconstructs the coarse structure of the flower, vehicle, dog *etc.*



(a) Image sample replication (b) Video sample replication

Figure 5. **Sample replication.** We study the effect of repeating samples while training our model. In each case, we repeat a sample n times within a mini-batch while fixing the overall mini-batch size and training updates. Replication leads to improved training speeds, especially on video without affecting the final performance.

4.2. Analyzing OmniMAE

We train the ViT-B architecture jointly on the ImageNet and the SSv2 datasets for 800 epochs, where an epoch involves training over all samples from both the datasets. For these analysis experiments, we use the default masking hyperparameters of MAE for images (75%). For ST-MAE, due to redundancy across frames, we use "tube" dropping (*i.e.* dropping all patches at a given spatial location over time) with high masking ratio (90%) All hyperparameters for these experiments are in Appendix B.3. We evaluate all the pretrained models on ImageNet and SSv2 and report the top-1 classification accuracies in Table 1. Please see Appendix D for additional ablations.

Extreme masking. We vary the ratio of masked patches in the input while training the model. We observe that videos benefit from significantly higher amounts of masking than images. Since video frames have a lot of redundant information, the resulting spatio-temporal patches are also redundant and thus higher masking leads to better self-supervised models. Unlike our experiments, prior work found that extremely high masking ratios lead to degradation in performance, for instance MAE [40] saw a significant degradation in performance when masking more than 75% of the patches. However, as we show in Table 1a, OmniMAE models can use extremely high masking of the input while learning good representations.

Reduced compute due to extreme masking. Our models trained with 90% masking on images and 95% masking on videos yield good performance, while being trained with just 19 and 78 unmasked image and video patches respectively, for a patch size of $2 \times 16 \times 16$. As we follow [40] and only pass unmasked patches to the encoder and have a lightweight decoder, extreme masking leads to a dramatically lower computational cost for training the encoder, and consequently the model as a whole. Compared to using all the patches, our masked autoencoder uses $5.9\times$ and $7.8\times$ fewer FLOPS for ViT-B on images and videos respectively, $7.1\times$ and $11.6\times$ for ViT-L, and $7.2\times$ and $11.3\times$ for ViT-H. Compared to MAE, on ViT-B, ViT-L and ViT-H, our higher masking leads to $1.8\times$, $2.0\times$ and $2.0\times$ fewer FLOPS on images. On videos, compared to some concurrent works [28, 76] that use 90%

Ratios		Accuracy	
IN1K	SSv2	IN1K	SSv2
75%	75%	82.6	67.2
75%	90%	82.8	67.9
75%	95%	82.8	68.1
90%	75%	82.8	67.7
90%	90%	82.6	68.5
90%	95%	**82.8**	**68.6**

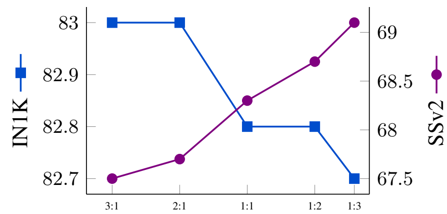
(a) Masking ratios.

Type		Accuracy	
IN1K	SSv2	IN1K	SSv2
Random	Tube	82.8	**67.9**
Causal	Tube	82.1	67.3
Random	Frame	**82.9**	65.8
Random	Causal	82.8	64.4
Random	Random	82.7	67.8

(b) Masking type.

	#params	IN1K	SSv2
Common	7.9M	**82.8**	**67.9**
Separate	33.0M	82.6	67.7
L=2	3.8M	82.7	67.9
L=4	7.9M	**82.8**	67.9
L=8	14.5M	**82.8**	**68.0**
d=384	7.9M	**82.8**	67.9
d=512	13.0M	82.7	**68.1**

(c) Decoder capacity.



(d) Dataset ratio

Table 1. **Ablations.** The default setting for all ablations is highlighted in gray. **(a)** We vary the masking ratios of the input for pretraining the model. This allows us to use an extremely high masking ratio of 90% on images and 95% on videos. **(b)** OmniMAE works well across different masking types. Random masking makes no assumptions about the input patches. Tube masking masks random patches at the same spatial location across all frames. Causal masking for images masks the 'future' patches as determined by a raster left-to-right order. In videos, causal masking masks all the patches from future frames. Frame masking masks all patches for randomly selected frames. **(c)** Different decoder designs (common or separate) and capacities (number of layers L and dimension d of the MLP). A common decoder for both images and videos performs better than a separate decoder. Our final performance is robust to decoder capacity. **(d)** We vary the image/video dataset ratios by replicating the entire datasets by a factor. The number of training updates changes with such replication, and we observe that the model benefits from higher replication of videos.

masking, we obtain $1.3\times$, $1.5\times$ and $1.4\times$ fewer FLOPS for ViT-B/L/H. Given the compute savings and strong performance, we choose 90% and 95% masking for images and videos respectively for our final models.

Type of masking. We study the effect of the type of masking used in training our models. The different types of masking are illustrated in Figure 3. We experiment with Random masking which masks patches in the image or video randomly. For videos, we experiment with two types of masking that exploit the temporal structure. In Tube masking, we mask random patches at the same spatial location across all the frames. In Causal masking, we use a raster order masking, akin to generative image models [16, 67], *i.e.* the patches in the top-left of the image, and earlier frames of the video, are kept while the rest are masked. Finally, in Frame masking, we randomly mask some frames in the video, while keeping all patches from the unmasked frames. As seen in Table 1b, Random and Tube masking perform comparably well on both image and video tasks. We find that in case of Causal masking, the prediction task becomes exceedingly tough due to the uncertainity of the future, and in case of Frame masking, it becomes relatively easy due to the high redundancy of pixel values across frames. Hence in both these cases, the representation learned does not perform as well on video recognition tasks. Given the simplicity of random masking, we use that for both modalities.

Decoder architecture. Since image and video prediction tasks may require specialized parameters, we study two settings: (1) the decoder parameters are shared for image and video pixel prediction; (2) two separate decoders are used for image and video prediction. For the latter we use our default decoder setting for videos (4 layers/384-D), however a larger 8-layer/512-D decoder for images as proposed in MAE [40]. The results in Table 1c show that using a shared decoder for both image and video prediction leads to better transfer learning accuracy. A shared decoder also offers a $4\times$ reduction in the number of parameters compared to using separate decoders while also being simpler to implement.

Decoder capacity. In Table 1c, we vary the decoder capacity and measure its impact on the final transfer learning accuracy. To change the decoder capacity, we vary the number of Transformer layers L used and the dimension d of the MLP used in the Transformer layers. Overall, the final transfer learning accuracy is robust to decoder capacity. A shallow decoder of 4 layers (8 for ViT-H) offers a good trade-off between the decoder size and final accuracies.

Dataset ratios. Since ImageNet and SSv2 have a different number of samples, we study the effect of varying the ratio of the datasets used in training. When increasing the ratio of the datasets, we measure a single epoch as training over the new oversampled set of samples from the datasets. We vary the relative ratio for ImageNet and Something Something-v2 by replicating only one dataset at a time as shown in Table 1d. We observe that increasing the relative dataset ratio has a positive effect on the final transfer accuracy for the oversampled dataset. This is expected as oversampling a dataset proportionally increases its corresponding parameter updates, thus making the representation better tuned for the dataset. We also observe that oversampling Something Something-v2 twice as much as ImageNet leads to an improvement for the video transfer accuracy with no drop in the image transfer performance. Hence, OmniMAE is robust to changes in dataset sizes of individual modalities, and it suggests that longer training on both datasets can further improve the transfer learning performance of our models. For simplicity, by default we do not replicate any dataset for training our models.

Sample replication. We replicate samples for both images and videos, and in each case measure a single epoch as training over the total samples, counting replicated samples, *i.e.*

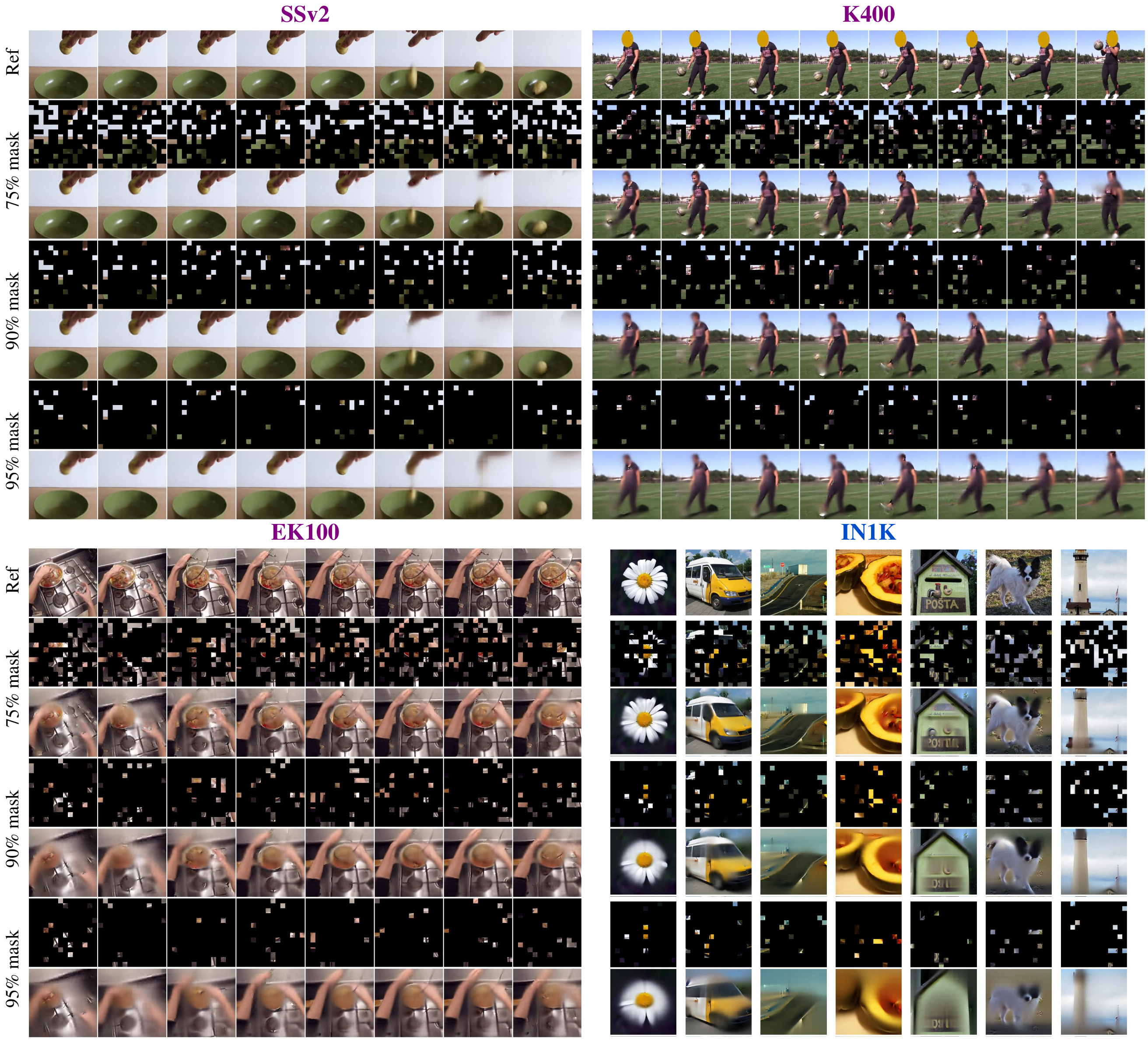


Figure 4. **Reconstruction visualizations** using OmniMAE on different video and image datasets. We show the model predictions for varying masking ratios of the input from 75% to 95% and the ground truth reference (Ref). OmniMAE is trained on ImageNet and SSv2 but the predictions generalize to other datasets like K400 and EK100. Please see the supplement for video visualizations.

replication maintains the same number of training iterations. We train models with different replication factors and show the final transfer accuracy and the normalized training time in Figs. 5a and 5b. Sample replication leads to faster training while maintaining or even improving the final transfer accuracy. Since a large portion of the input sample is masked, replicating the sample multiple number of times still provides enough learning signal for the model and does not lead to a degradation in performance. This becomes even more relevant for OmniMAE's final settings where we use higher masking ratios. For video data, we use an optimized dataloader (see details in Appendix B) with asynchronous I/O and fast video decoding. Even with this optimized dataloader, sample replication leads to 20% faster training. We believe this is an important practical observation as video dataloading is often a bottleneck for training models.

4.3. Comparison to Prior Work

In Table 2, we compare OmniMAE's representation on our image and video classification benchmarks with other

Method	Arch.	Pretrain Data	IN1K	iNat18	P365	K400	SSv2	EK100
DINO	ViT-B	IN1K	82.8	72.6	–	✗	✗	✗
iBOT [92]	ViT-L	IN1K	84.8	–	–	✗	✗	✗
MAE [40]	ViT-B	IN1K	83.6	75.4	57.9	✗	✗	✗
MAE [40]	ViT-L	IN1K	85.9	80.1	59.4	✗	✗	✗
MAE [40]	ViT-H	IN1K	86.9	83.0	59.8	✗	✗	✗
BEiT [6]	ViT-B	IN1K	83.4	72.3	–	✗	✗	✗
BEiT [6]	ViT-L	IN1K	85.2	–	–	✗	✗	✗
VIMPAC [74]	ViT-L/2†	HowTo100M	✗	✗	✗	77.4	68.1	–
MaskedFeat [82]	MViT-v2-L	K400	✗	✗	✗	84.3	–	–
BEVT [80]	Swin-B	IN1K + K400	✗	✗	✗	81.1	71.4	–
OmniMAE	ViT-B	IN1K + K400	82.9	74.2	58.5	80.8	69.0	40.3
OmniMAE	ViT-B	IN1K + SSv2	83.0	74.0	58.4	80.6	69.5	39.3
OmniMAE	ViT-L	IN1K + SSv2	85.2	79.6	59.3	84.0	74.2	45.1
OmniMAE	ViT-H	IN1K + SSv2	86.6	83.2	60.1	85.4	75.5	48.0
Concurrent methods specialized for videos								
VideoMAE-16f [76]	ViT-L	SSv2	✗	✗	✗	–	74.2	–
VideoMAE-16f [76]	ViT-L	K400	✗	✗	✗	84.7	–	–
MAE-Video [28]	ViT-H	K400	✗	✗	✗	85.1	74.1	–

Table 2. **Comparing OmniMAE with prior self-supervised methods** on image and video recognition. Prior work trains a specialized model for a particular visual modality, sometimes using specialized architectures. Our single OmniMAE model is pretrained jointly on images and videos for 1600 epochs (ViT-H for 2400) and performs competitively across all benchmarks while using a simple architecture and pretraining method, even outperforming concurrent work specialized for videos. †ViT-L with half the MLP embedding dimension

state-of-the-art self-supervised methods. We focus on methods which use a Transformer backbone, ViT, or an architecture based on Transformers like Swin. To the best of our knowledge, prior work does not explore joint pre-training except for BEVT [80], which does masked image modeling followed by joint masked image and video modeling. Unlike OmniMAE, BEVT only focuses on video representations and does not evaluate on image recognition.

OmniMAE is pretrained for 1600 epochs (ViT-H for 2400 epochs). Training with IN1K and K400 or IN1K and SSv2 leads to similar results, so we use the latter when scaling model size. OmniMAE performs competitively for *both* image and video recognition when compared to the best models trained separately for either modality. OmniMAE also performs favorably compared to methods like MaskedFeat which are pretrained only for video with specialized architectures on larger video datasets. OmniMAE serves as a competitive initialization for transferring models on all modalities. More notably, OmniMAE's performance improves significantly with larger architectures. Using ViT-H, OmniMAE **performs better or within the margin of error across all benchmarks**. Notably, OmniMAE even outperforms concurrent works which use similar approaches while solely focusing on video representation learning.

5. Conclusion and Future Work

We present OmniMAE, a unified Transformer for images and videos that can be pretrained using masked autoencoding. OmniMAE uses a simple architecture with minimal vision-specific changes and is competitive with specialized architectures and models tailored for images and videos. We believe such generic models and approaches are a critical and under explored area of representation learning. OmniMAE has not been empirically validated on other visual modalities like 3D, or non-visual modalities like audio. The simplicity and generality of OmniMAE can likely enable future multi-modal systems that use shared parameters to model multiple modalities. We only studied masked autoencoder based pre-training strategies for OmniMAE since they are easier to train. However, pixel reconstruction methods do not learn good linearly separable features or similarity metrics [40]. These properties are essential when finetuning with limited labeled data. We believe that exploring other pretraining strategies, particularly in the context of joint modeling, will likely lead to improved models. Similarly, joint finetuning strategies [33] could be used to evaluate such unified models and lead to potentially even superior performance.

Ethical considerations. We study models for visual recognition on images and videos and our technical contributions are neutral from an ethics standpoint. We do not propose a new dataset or make claims about the suitability of our model on production data. Given their self-supervised nature, our learned representations are free from label bias, however they are still susceptible to the bias in the distribution of the visual data We believe all ethical considerations that apply to visual recognition models equally apply to our model.

References

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

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

[3] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020.

[4] Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021.

[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 *ICML*, 2022.

[6] Hangbo Bao, Li Dong, and Furu Wei. BEiT: Bert pre-training of image transformers. In *ICLR*, 2022.

[7] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

[8] Maxim Berman, Hervé Jégou, Andrea Vedaldi, Iasonas Kokkinos, and Matthijs Douze. Multigrain: a unified image embedding for classes and instances. *arXiv preprint arXiv:1902.05509*, 2019.

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

[10] Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 1988.

[11] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

[12] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[13] 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.

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

[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] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020.

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[18] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[19] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021.

[20] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *ICLR*, 2020.

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

[22] 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.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2018.

[24] 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.

[25] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pretraining? *arXiv preprint arXiv:2112.10740*, 2021.

[26] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *ICML*, 2021.

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

[28] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022.

[29] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016.

[30] Patrick Gallinari, Yann LeCun, Sylvie Thiria, and F Fogelman Soulie. Mémoires associatives distribuées: une comparaison (distributed associative memories: a comparison). In *COGNITIVA*. 1987.

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

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

[33] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A Single Model for Many Visual Modalities. In *CVPR*, 2022.

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

[35] Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022.

[36] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, 2019.

[37] 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.

[38] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 2020.

[39] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

[40] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[41] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[42] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. In *NeurIPS*, 1993.

[43] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *CVPR*, 2020.

[44] 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.

[45] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016.

[46] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[47] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[48] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, AMustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[49] Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE*, 1991.

[50] Yann LeCun and Françoise Fogelman-Soulié. Modèles connexionnistes de l'apprentissage. *Intellectica*, 1987.

[51] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C.H. Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021.

[52] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022.

[53] Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv preprint arXiv:2111.12993*, 2021.

[54] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022.

[55] 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.

[56] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021.

[57] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *ICCV*, 2021.

[58] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020.

[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 and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020.

[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] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 1996.

[65] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *NeurIPS*, 2018.

[66] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018.

[67] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, 2018.

[68] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

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

[70] 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.

[71] Ruslan Salakhutdinov and Geoffrey Hinton. Deep Boltzmann machines. In *AISTATS*, 2009.

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

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

[74] 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.

[75] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[76] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

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

[78] 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.

[79] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.

[80] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. BEVT: Bert pretraining of video transformers. In *CVPR*, 2022.

[81] 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.

[82] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022.

[83] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[84] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022.

[85] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. ClusterFit: Improving Generalization of Visual Representations. In *CVPR*, 2020.

[86] 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.

[87] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

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

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

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

[91] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NeurIPS*, 2014.

[92] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *ICLR*, 2022.