Efficient Image Pre-Training with Siamese Cropped Masked Autoencoders

Alexandre Eymaël*[1], Renaud Vandeghen*[1], Anthony Cioppa[1,2], Silvio Giancola[2], Bernard Ghanem[2], and Marc Van Droogenbroeck[1]

[1] University of Liège, Belgium
[2] KAUST, Saudi Arabia
r.vandeghen@uliege.be

Abstract. Self-supervised pre-training of image encoders is omnipresent in the literature, particularly following the introduction of Masked autoencoders (MAE). Current efforts attempt to learn object-centric representations from motion in videos. In particular, SiamMAE recently introduced a Siamese network, training a shared-weight encoder from two frames of a video with a high asymmetric masking ratio (95%). In this work, we propose CropMAE, an alternative approach to the Siamese pre-training introduced by SiamMAE. Our method specifically differs by exclusively considering pairs of cropped images sourced from the same image but cropped differently, deviating from the conventional pairs of frames extracted from a video. CropMAE therefore alleviates the need for video datasets, while maintaining competitive performances and drastically reducing pre-training and learning time. Furthermore, we demonstrate that CropMAE learns similar object-centric representations without explicit motion, showing that current self-supervised learning methods do not learn such representations from explicit object motion, but rather thanks to the implicit image transformations that occur between the two views. Finally, CropMAE achieves the highest masking ratio to date (98.5%), enabling the reconstruction of images using only two visible patches. Our code is available at `https://github.com/alexandre-eymael/CropMAE`.

Keywords: Self-supervised learning, Masked autoencoders, Siamese networks, Video segmentation, Label propagation.

1 Introduction

Self-supervised learning (SSL) has become increasingly popular in the last few years thanks to its capacity to learn meaningful and robust representation without the need for labels, sometimes even leading to performances on downstream tasks surpassing its supervised counterpart. This is especially interesting in domains in which data labelling is costly, such as image segmentation or object detection, or when the exact task to solve is not known beforehand [1]. Among popular self-supervised paradigms, visual contrastive learning [7,20,24] and masked image modeling (MIM) [23,30,48] have received significant interest due to their

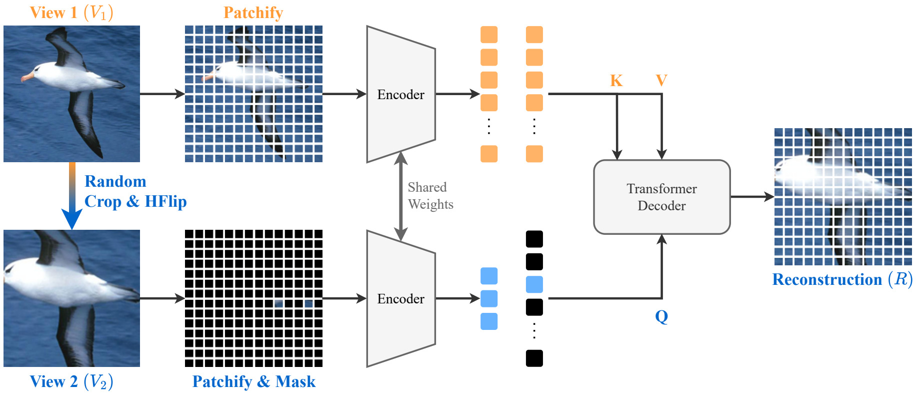


Fig. 1: CropMAE self-supervised pre-training. Given an input image (V_1), a second image is generated by performing a random crop and, optionally, a horizontal flip on the original image (V_2). We then patchify [13] both views and mask [23, 30] an extremely high portion of the second image (above 98.5%). Both views are encoded by a Siamese [5] ViT encoder, with added positional embedding [13]. A transformer [19] decoder reconstructs the masked image R using self-attention layers on the tokens of the masked image and cross-attention layers between the tokens of the masked and unmasked images.

impressive performance. While highly effective, MIM methods often require a large amount of data and/or extensive training time to achieve satisfactory performance [16,21,41]. This necessity largely stems from their objective to develop a conceptual understanding of the data distribution they are trained on, enabling them to reconstruct images at the pixel level. This challenge is particularly pronounced with Vision Transformers (ViTs) [13] as encoders, as they perform suboptimally with limited data due to the lack of visual inductive biases that they exhibit [13]. A major weakness of contrastive learning techniques is that they rely on carefully chosen transformations to achieve good performances [7,20,47].

Recently, Siamese Masked autoencoders (SiamMAE) [21] achieved state-of-the-art performance in numerous propagation tasks [27,35,52] by learning object-centric representations from videos. This method leverages a Siamese encoder [5] to process pairs of frames that are asymmetrically masked. Despite its impressive performance, SiamMAE faces two main limitations. Firstly, it is designed to only process video frames, not standalone images. Yet, image datasets are typically orders of magnitude larger than video datasets, and less computationally expensive to decode, making image-based pre-training more effective and scalable than video-based pre-training. Secondly, while SiamMAE reduces the need for the intense data augmentation found in contrastive learning methods, it still requires learning a conceptual understanding of the visual world, similar to most MIM techniques, thus requiring extensive training (2,000 epochs) on large datasets such as K400 [29] to reach state-of-the-art performances.

In this work, we propose a novel self-supervised learning method, called *CropMAE*, that reframes the siamese-based paradigm introduced in SiamMAE in order to alleviate the need for video dataset, while keeping competitive performances on downstream tasks. Specifically, we use random views of the same image to simulate viewpoint changes, object transformations, motion, and occlusions. Our method can therefore leverage both image and video datasets, and train at a significantly faster pace than SiamMAE. Moreover, we demonstrate that CropMAE learns meaningful object-centric representations for downstream video tasks without explicit motion. Finally, unlike most MIM techniques, the pretext task of CropMAE is directly tractable based on the visible frame without the need to learn conceptual information about the world, which we believe is the reason for its faster training. An overview of our method is presented in Figure 1.

Contributions. We summarize our contributions as follows. **(i)** We introduce a novel pre-training method, CropMAE, based on sole images, which alleviates the need for video decoding and significantly accelerates training. The novel pretext task we introduce learns faster while quickly reaching good performances. **(ii)** We empirically demonstrate the feasibility of learning meaningful representations for downstream video tasks from still images or data distributions traditionally not associated with videos. Notably, this approach yields better results than training directly on video frames. **(iii)** We show, for the first time, that employing an extremely high masking ratio (98.5%, *i.e.*, using only two visible patches for a ViT/16), surpassing those explored in existing studies, can be optimal and generate a meaningful and challenging self-supervised task.

2 Related Work

Visual representation learning. Visual self-supervised learning focuses on learning rich and generalizable representations of images or videos. This is typically achieved through pretext tasks [7, 32, 33, 39], enabling the learned representations to be applicable to a broad set of downstream tasks [11,14,52], either by fine-tuning the learned models for specific tasks, or by freezing the weights and training a linear classifier or an MLP on top of it. Key downstream tasks in the visual domain include image classification [3, 6, 7, 9, 18, 20, 23, 34, 48, 51], video classification [15,16,18,34,41,44], object detection [9,20,23], and video segmentation [4,6,9,21,28]. Our method, CropMAE, is a new visual self-supervised representation learning method for propagation tasks [27,35,52].

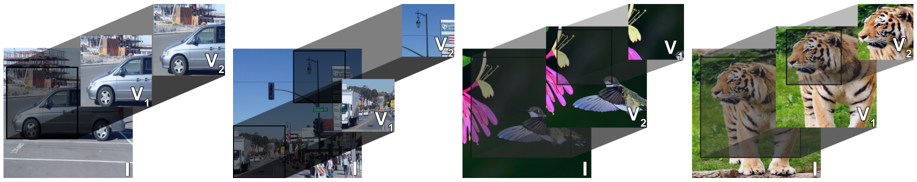
Contrastive Self-Supervised Learning. Contrastive self-supervised learning [22] has been recognized as an effective method for feature extraction, applicable both to images [9,20] and videos [10,38]. This approach encourages the encoder to learn robust representations of the input data by minimizing the distance between representations of different augmented versions of the same image. Initially, it was common to enforce distinct images to have different representations in order to avoid representation collapse [7, 12, 46]. However, subsequent discoveries [20, 24] have shown that robust learning can be achieved even without imposing this constraint. Contrastive self-supervised learning has also been widely used for correspondence learning [26, 45], as it inherently learns to build representations that are invariant and robust to perturbations. Contrary to contrastive learning, CropMAE does not rely as extensively on data augmentations and is not subject to representation collapse issues.

Masked Image Modeling. Drawing inspiration from the field of natural language processing [30], masked image modeling (MIM) techniques have emerged as highly effective learners in the vision domain [3, 23, 49]. This approach involves dividing images into small patches [13], with a high proportion of them being masked, and subsequently reconstructing them using a denoising autoencoder [43]. Notably, after the training phase, the decoder is discarded, leaving the encoder to serve as a feature extractor. MIM has been applied with success across a broad range of fields, and has had numerous extensions and improvements [2, 8, 15–18, 21, 28, 34, 36, 41, 44].

Siamese Masked Autoencoders. Building upon the work of masked autoencoders [23], Siamese Masked Autoencoders (SiamMAE) [21] have emerged as a new state-of-the-art in video propagation tasks such as video object segmentation [35], pose keypoint propagation [27], and semantic part propagation [52]. SiameseMAE uses a Siamese encoder [5] to process either pairs [21] or groups [28] of frames, randomly selected from a video. A key feature of SiameseMAE is its asymmetric masking technique: the initial frame undergoes no masking, thereby serving as a complete reference, while a substantial portion (up to 95%) of the second frame is masked. This setup encourages the network to accurately reconstruct the masked subsequent frames using the fully visible initial frame as a reference. The efficacy of SiameseMAE is believed to stem from its ability to effectively model object motion from videos and visual correspondence, learning the "propagation" and boundaries of objects from their observed positions in the past to their future locations, based on the few visible patches [21]. In this work, we show that explicit motion derived from videos is not mandatory for Siamese masked autoencoders to learn object-centric representations. Particularly, we demonstrate that the ability to recognize object boundaries and acquire propagation skills can be effectively learned from still images.

3 Method

We propose a novel self-supervised method, namely *CropMAE*, capable of learning valuable representations both from images and video frames. First, we create two augmented views (V_1 and V_2) of an input image (I) by randomly cropping, resizing and horizontally flipping the original image (Sec. 3.1). Second, we patchify [13] both views V_1 and V_2 (Sec. 3.2) and mask [23, 30] an extremely high portion of the second view (V_2) (Sec. 3.3). Both views are encoded in a Siamese [5] ViT encoder, with an additional positional embedding [13]. Third, a



(a) Same. **(b)** Random. **(c)** Local-to-Global. **(d)** Global-to-Local.

Fig. 2: Illustration of our four cropping strategies. For a given input image I, we generate an unmasked view V_1 and a masked view V_2 following one of four different cropping strategies: (a) Same Views, where $V_1 = V_2$; (b) Random Views, where V_1 and V_2 are two independent random crops; (c) Local-to-Global, where V_1 is a random crop within V_2, and (d) Global-to-Local, where V_2 is a random crop within V_1.

transformer [19] decoder reconstructs a target image R (Sec. 3.4). The Siamese network and the decoder are trained by minimizing the L2 norm between the target V_2 and the reconstructed image R. After such pre-training, the decoder is discarded, and we use the encoder as a feature extractor on downstream tasks. With this setup, we demonstrate that meaningful data augmentations, particularly random crops, can generate rich and useful object-centric representations for propagation tasks *without* explicit motion. Figure 1 illustrates the main components of our method.

3.1 Cropping

Random crops have been widely used in visual self-supervised learning, especially in contrastive learning, where they are essential to reach excellent performances and develop robust representations [7,9,20]. Specifically, we examine four strategies inspired by the contrastive learning literature [7].

- **Same Views.** This setup corresponds to a direct adaptation of SiamMAE to images, in which the input image I is cropped once and serves both as V_1 and V_2. An illustration is given in Figure 2a.
- **Random Views.** For a given input image I, two independent random cropped views are generated for V_1 and V_2. This setup poses a challenge, particularly when the views are adjacent, *i.e.*, that there is minimal to no overlap between the two crops as illustrated in Figure 2b.
- **Local-to-Global Views.** In this setup, the masked view V_2 is a random crop of the original image I, and the unmasked view V_1 is another random crop of the masked view V_2. An illustration is provided in Figure 2c.
- **Global-to-Local Views.** Inversely, the unmasked view V_1 is a random crop of the original image I, and the masked view V_2 is another random crop of the unmasked view V_1. An illustration is provided in Figure 2d.

Note that our experiments indicate that the Global-to-Local view strategy leads to the best performance.

3.2 Patching

The two views V_1 and V_2 are patched following the original ViT [13]. Specifically, each view is converted into $N \times N$ patches that are fed into the encoder. Similar to SiamMAE, we augment the linear projections of these patches with positional embeddings [42], and append a `[CLS]` token.

3.3 Masking

Since both views are highly spatially redundant, a high masking ratio (above 75%) is usually necessary to create a challenging pretext task and to achieve optimized performances with masked autoencoders [23]. This is even more important in videos where both the spatial and temporal dimensions are highly redundant, requiring even higher masking ratios (90%) [16,41,44]. SiamMAE [21] employs a highly asymmetrical masking strategy, where the first frame is left completely visible while the second one is masked at 95%, which corresponds to 9 visible patches out of the 196 available when using a ViT/16 [13]. Using such a high masking ratio encourages the model to propagate the visible patches from the first frame to the second one and to learn temporal correspondences through motion [21]. However, employing a high masking ratio can make some examples ambiguous or may require additional knowledge beyond merely "propagating" patches from the unmasked view. For instance, if an object is only partially visible in the first view, while it is completely present (but masked) in the second one, the task becomes intractable if the model relies solely on the first view to reconstruct it. This prompts the model to learn a conceptual representation of the objects it encounters [23], enabling it to "hallucinate" what it partially sees when propagating past patches is either impossible or insufficiently informative.

Unlike previously introduced MAE methods, CropMAE does not need to learn any conceptual information about objects. Indeed, since our pretext task reconstructs a local view from a global one, there is no ambiguity as the local view is always present within the global view. Provided that the model **(i)** successfully identifies the location of the local view within the global view based on the visible patches and **(ii)** accurately determines the transformations required to reconstruct the local view from the global view, the task is directly tractable based on the inputs that the model receives without any prior conceptual knowledge. This naturally makes the pretext task significantly easier than in other MAE approaches such as MAE [23], VideoMAE [41], or SiamMAE [21], where rich conceptual representations should be used to solve the task. For that reason, we employ an even higher masking ratio. More specifically, our method performs best with only a few visible patches, typically 1 or 2 out of 196, which corresponds to a masking ratio between 98% and 99%. Note that increasing the masking ratio from 95% to 98.5% decreases the number of visible patches by a factor of 4.5, reducing them from 9 to just two visible patches.

3.4 Encoder and Decoder Architectures

Following [21], we use a Siamese ViT [13] encoder to process our two views and a vanilla Transformer [42] composed of cross-attention and self-attention layers as our decoder. Specifically, our decoder alternates between self-attention, where tokens of the masked image attend to each other, and cross-attention layers, where the tokens of the masked image attend to tokens of the visible image. We train the Siamese architecture by minimizing the L2 loss between the normalized [23] pixel values of the view V_2 and the reconstruction R.

4 Experiments

4.1 Experimental setup

Implementation details. Following previous methods [6, 21, 41], we use the ViT-S/16 as encoder architecture [13] for most of our experiments and fair comparisons with respect to other methods in the field. For the decoder, we employ a 4-layer Transformer [42] with a dimension $d_{\text{model}} = 256$, where each block comprises a cross-attention layer, a feed-forward layer (of dimension $d_{\text{ff}} = 2048$), and a self-attention layer. GELU activation functions [25] are utilized alongside a dropout rate of 10% [40]. We use the AdamW [31] optimizer and a base learning rate of $1.5e^{-4}$. The exhaustive list of hyper-parameters that we use can be found in the Appendix.

Baselines. We compare our method with several state-of-the-art methods including MAE-ST [16], MAE [23], VideoMAE [41], and SiamMAE [21]. To the best of our knowledge, no official open-source code is available for SiamMAE, so we reimplemented it to compare the evolution of our performance during training, using the exact same hyperparameters described in the SiamMAE paper (refer to the supplementary material). Our results are consistent with the ones reported in their paper [21]. However, we train for 400 epochs instead of 2000 to save computational resources. Results for longer training can be found in the Appendix.

Datasets. We pre-train our models on Kinetics-400 [29] (K400), on ImageNet [37] (IN), or on a subset of ImageNet (IN Subset). IN Subset contains $239,787$ randomly selected images, which corresponds to the number of videos in K400, for fair comparison between methods trained on K400 and ImageNet. During pre-training, we randomly sample an image (or a frame on K400), which is then processed following our methodology described in Section 3.

Downstream tasks. We evaluate our method on three propagation downstream tasks: video object segmentation (DAVIS-2017 [35]), human pose propagation (JHMDB [27]) and semantic part propagation (VIP [52]). These propagation tasks are framed as a semi-supervised problem, where the first annotated frame is provided, and the model is expected to propagate the segmentation mask to subsequent frames.

Table 1: Comparison with prior work. We evaluate our method on three downstream tasks: video object segmentation (DAVIS-2017 [35]), human pose propagation (JHMDB [27]) and semantic part propagation (VIP [52]). Specifically, we compare our method with other methods trained on 400 epochs, on K400 [29] or on our ImageNet [11] Subset (IN Sub) for fair comparison. † refers to results reported in [21]. ‡ refers to our implementation.

				DAVIS			VIP	JHMDB	
Method	Backbone	Dataset	Epochs	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$	mIoU	PCK@0.1	PCK@0.2
MAE-ST [16] †	ViT-L/16	K400	800	54.6	55.5	53.6	33.2	44.4	72.5
MAE [23] †	VIT-B/16	IN	1600	53.5	52.1	55.0	28.1	44.6	73.4
VideoMAE [41] †	ViT-S/16	K400	800	39.3	39.7	38.9	23.3	41.0	67.9
SiamMAE [21] †	ViT-S/16	K400	2000	**62.0**	**60.3**	**63.7**	**37.3**	**47.0**	**76.1**
SiamMAE [21] ‡	ViT-S/16	K400	400	57.9	56.0	60.0	33.2	**46.1**	**74.0**
CropMAE (ours)	ViT-S/16	K400	400	58.6	55.8	61.4	**33.7**	42.9	71.1
CropMAE (ours)	ViT-S/16	IN Sub	400	**60.4**	**57.6**	**63.3**	33.3	43.6	72.0
CropMAE (ours)	ViT-B/16	IN Sub	400	60.9	57.9	63.8	32.8	44.3	72.3

4.2 Results

We compare our method to previous works and present quantitative results in Table 1. We then provide some qualitative results of the reconstructed image and the downstream tasks respectively in Figures 3 and 4. The first part of Table 1 displays results as reported in their original papers, under optimal training conditions in terms of both training duration and data volume. In the second part, we report the results achieved by our reproduced implementation of SiamMAE and CropMAE under our constrained training: either on K400 or on our ImageNet Subset, for a fixed duration of 400 epochs, and for both ViT-S/16 and ViT-B/16.

When trained for 2,000 epochs on K400, SiamMAE achieves state-of-the-art performances on the three downstream tasks, and outperforms previous MAE methods such as MAE-ST [16], MAE [23] and VideoMAE [41]. However, considering a fixed budget of 400 epochs, CropMAE achieves significantly better results than SiamMAE on DAVIS-2017 [35], both when trained on K400 and on our ImageNet Subset (+0.7% and +2.5% respectively). We believe that by explicitly transforming images through cropping, our pre-training method more quickly understands features useful for segmentation, such as object boundaries. On VIP [52], CropMAE still performs better than SiamMAE, although by a smaller margin (+0.1 when trained on ImageNet, and +0.5 when trained on K400). On JHMDB [27], CropMAE only outperforms VideoMAE. We explain these inferior performances by noting that SiamMAE uses two different frames, resulting in complex human pose modifications, which likely helps the network understand human motion and perform better on JHMDB. Conversely, our random crops do not mimic these transformations. Yet, they help the network learn object boundaries more explicitly, making it more suited for segmentation tasks such as DAVIS.

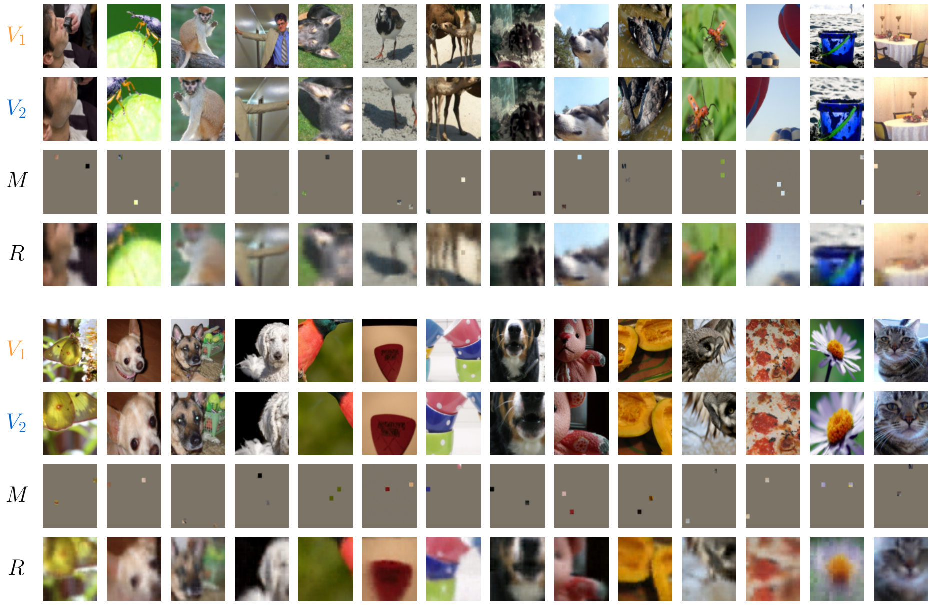


Fig. 3: Reconstructions of CropMAE. We train CropMAE with a ViT-S/16 without normalizing pixel values and a masking ratio of 98.5%. We visualize the reconstructions of some images from ImageNet. The images are displayed in the following order from top to bottom: Input Image (V_1), Random Resized Crop (V_2), Masked Image (M), and Reconstruction (R).

4.3 Attention Maps

In SiamMAE, Gupta *et al.* [21] argue that their model learns the concept of object boundaries through object motion in videos. To support this claim, they present attention maps extracted at some layers of their model, demonstrating that attention predominantly focuses on object boundaries. In a similar way, we train a ViT-S/8 with CropMAE on our ImageNet Subset and visualize the self-attention maps of the [CLS] token from a specific head of the last encoder layer. We show the results in Figure 5. Our findings indicate that our model learns to identify object boundaries as well as SiamMAE without explicit motion (*i.e.*, without relying on video frames). This implies that learning object boundaries is not solely attributable to the motion observed in videos; instead, it can also stem from the transformations and deformations operated on a single image. Hence, this phenomenon is present in both SiamMAE, where it happens naturally between two frames, and in CropMAE, where motion is artificially induced through random cropping. The main difference remains that CropMAE is trained on images instead of videos.

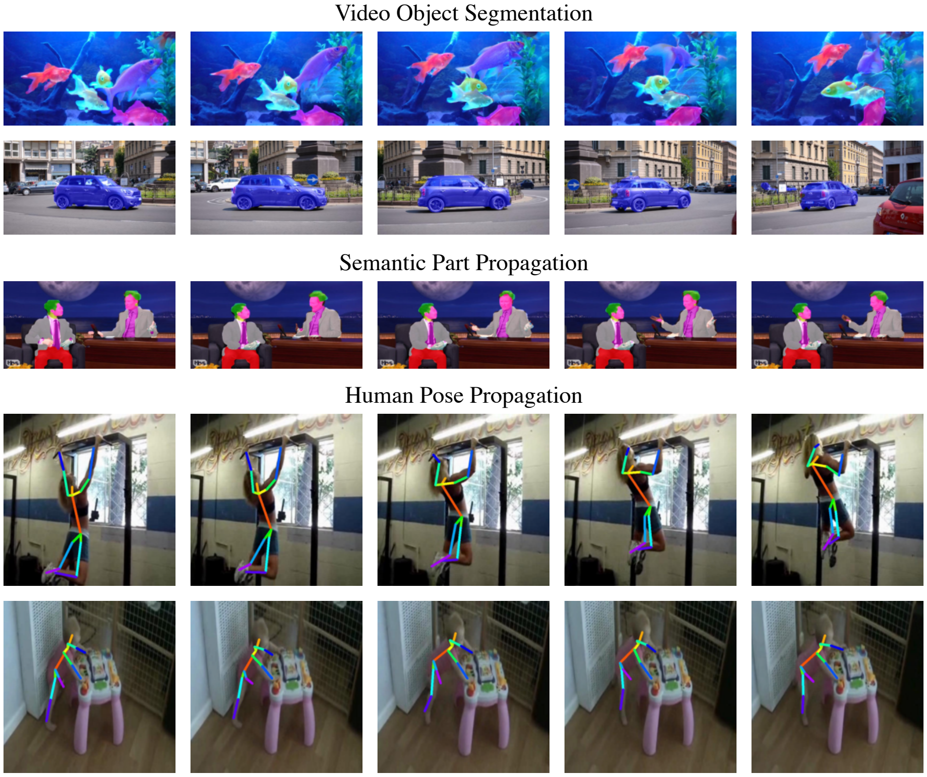


Fig. 4: Qualitative results. We train CropMAE with a ViT-S/16 and qualitatively validate our results on three propagation downstream tasks: video object segmentation (DAVIS-2017 [35]), semantic part propagation [52], and human pose propagation (JHMDB [27]).

4.4 Learning Speed

We evaluate the evolution of the performances of CropMAE and SiamMAE. In particular, we compare SiamMAE trained on K400, CropMAE trained on K400, and CropMAE trained on ImageNet Subset, all for 400 epochs. The performance on the DAVIS-2017 object propagation task [35] is reported every 50 epochs in Figure 6. Remarkably, our approach demonstrates superior performance when trained on the ImageNet Subset compared to training using K400 video frames. This improvement can be attributed to two main factors: **(i)** the greater diversity of the ImageNet dataset, containing a broader spectrum of objects, and **(ii)** its focus on currated object-centric images, which likely results in more relevant crops and reconstruction tasks. In contrast, random cropping in K400 frequently yields images without any objects, diminishing the effectiveness of the learning process.

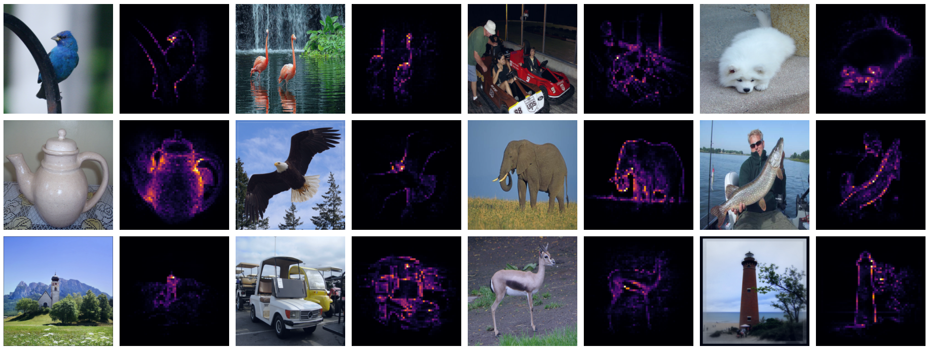


Fig. 5: Self-attention maps from CropMAE with a ViT-S/8 trained on our ImageNet subset. We visualize the self-attention of the `[CLS]` token from a selected head in the last encoder layer of a ViT-S/8, which was trained on our ImageNet subset without using any supervision to learn this specific token. These self-attention maps reveal that our model can learn object boundaries without the need for prior motion information during pre-training.

Our approach demonstrates significantly faster learning than SiamMAE. In particular, our method achieves a $\mathcal{J}\&\mathcal{F}_m$ value of 58.0 after only 150 epochs on our ImageNet Subset and 250 epochs on K400. In contrast, SiamMAE reaches the same performance level after 350 epochs. We attribute this trend to our pretext task, which does not require any conceptual knowledge to be completely tractable and uses object transformations much more explicitly than SiamMAE, leading to faster propagation comprehension. In contrast, SiamMAE must learn the concept of motion and understand object transformations more implicitly between two frames through more complex perturbations such as occlusions and viewpoint changes.

4.5 Training time

We compare the training times of CropMAE and SiamMAE. On the one hand, CropMAE uses an extremely high masking ratio, and only needs a single frame of a video clip to train, or even a standalone image. On the other hand, SiamMAE uses a lower masking ratio and needs two different frames to work. Both these factors significantly impact the training time, as seeking distant frames may require decoding a larger portion of the video, and the number of operations performed by the attention layers increases quadratically with the number of visible patches [23]. We measure the total time taken by both approaches to train and report our results in Table 2. As it can be seen, CropMAE trains almost 30% faster than SiamMAE on K400 for a fixed computational budget, thanks to its use of fewer patches and frames. When pre-training on images (*i.e.*, on the IN Subset), which are significantly faster to decode, CropMAE achieves a tremendous speed-up of 2380% on our hardware while also reaching better performances.

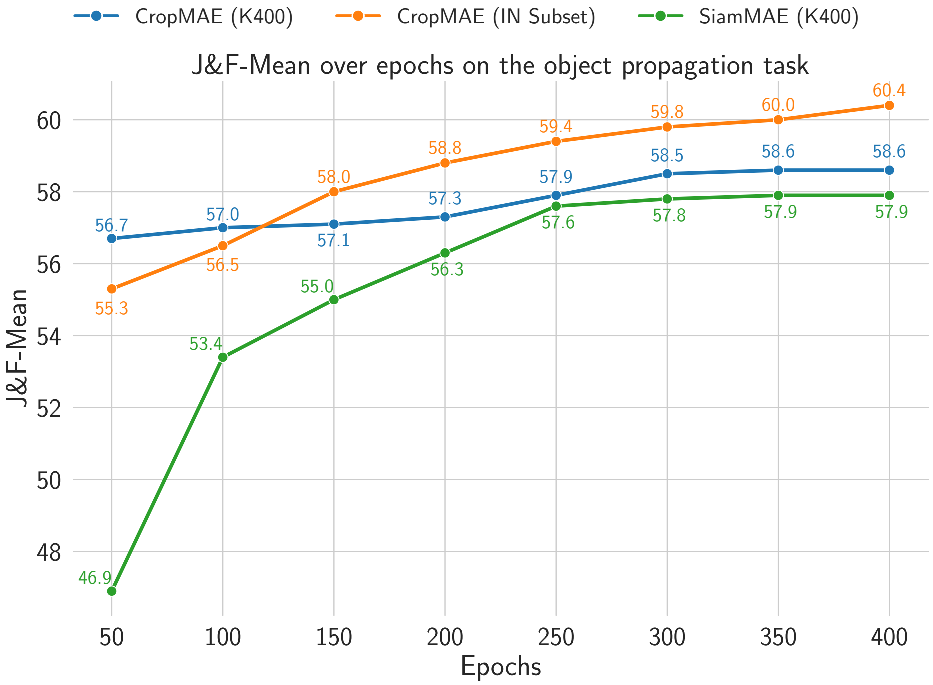


Fig. 6: Performances of CropMAE and SiamMAE on DAVIS during pre-training. For a fixed number of 400 epochs, CropMAE trains faster and consistently yields better results than SiamMAE [21], when trained on K400 frames or ImageNet Subset images.

Table 2: Speedup of CropMAE compared to SiamMAE. We train both methods for 400 epochs on K400, and on ImageNet Sub for CropMAE, and report the speedups observed on the whole training process.

Method	Dataset	Number of images	Mask Ratio	GFLOPS	Speedup
SiamMAE	K400	2	95%	5.8	×1.0
CropMAE	K400	1	98.5%	5.6	×1.29
CropMAE	IN Subset	1	98.5%	5.6	×23.8

4.6 Ablation Studies

We perform several ablation studies on the different components of CropMAE and report the results in Table 3. Unless stated otherwise, we use the default parameters presented in the Appendix. Specifically, we train CropMAE on our ImageNet subset for 400 epochs and report the results obtained on the DAVIS-2017 [35] object propagation task.

Table 3: Ablation Study. We analyze the different components of our method to understand their impact on the downstream performance. We use a ViT-S/16 [13] with the default configuration, as presented in Section 4.1, and report the results obtained on the DAVIS-2017 [35] validation set.

Crop Strategy	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Same Views	36.6	35.8	37.5
Random Views	60.0	57.2	62.8
Local-to-Global	55.9	53.8	58.0
Global-to-Local	**60.4**	**57.6**	**63.3**

(a) Crop Strategy. A simple extension of SiamMAE to images does not work. Reconstructing the local view from the global view works best for CropMAE.

Mask Ratio	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
0.75 (49)	45.3	44.3	46.3
0.90 (19)	47.1	46.1	48.0
0.95 (9)	51.2	49.9	52.4
0.985 (2)	**60.4**	**57.6**	**63.3**
0.99 (1)	58.6	55.9	61.5

(b) Mask Ratio and number of visible patches. Our method works best when an extremely large portion of the patches is masked.

Decoder Depth	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
2	59.1	56.7	61.6
4	**60.4**	**57.6**	**63.3**
8	57.0	54.5	59.4

(c) Decoder Depth. Our method works best with a small depth.

Decoder Embed Dim	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
128	58.5	56.0	61.0
256	**60.4**	**57.6**	**63.3**
384	59.0	56.3	61.7

(d) Decoder Embedding Dimension. Our method works best with a small decoder embedding dimension.

Augmentation	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Color Jitter	56.2	53.1	59.2
Gaussian Blur	59.6	56.7	62.4
None	60.3	57.4	63.2
Horizontal flip	**60.4**	**57.6**	**63.3**

(e) Data Augmentations. Our method works best with horizontal flips randomly applied on both random crops.

Cropping Strategy. We study the effect on performance of different cropping strategies in Table 4a. We can see that reconstructing the same views (Figure 2a) yields very poor performances (36.6), suggesting that the model failed to learn any propagation capabilities. Reconstructing the Local-to-Global view (Figure 2c) results in significantly improved performance (55.9). The Random Views (Figure 2b) and Global-to-Local (Figure 2d) approaches achieve the highest scores (60.0 and 60.4, respectively). Interestingly, these setups are the only ones enabling a completely tractable task without any prior knowledge, meaning the reconstruction can solely rely on the unmasked image. In fact, tractability is *sometimes* guaranteed in the random setting, while it is *always* true for the Global-to-Local approach, which likely explains its superior performance.

Masking Ratio. We examine the importance of the masking ratio in Table 4b. Our method exhibits suboptimal performance at a 75% masking ratio, despite this being the preferred choice for the traditional image MAE framework [23]. Similarly, it underperforms at the 90% ratio used in video frameworks [16, 41, 44]. We can see an improvement with a masking ratio of 95%, as adopted in SiamMAE [21], but the optimal results are reached with a visibility reduced to merely a few patches, *i.e.*, two (60.4) or one (58.6), equivalent of masking ratios of 98.5% and 99%, respectively. We attribute this trend to the fact that our pretext task is simpler than those used in other frameworks as it does not require any conceptual knowledge and can be fully achieved with the help of the visible image, thus requiring an extremely high masking ratio to be challenging.

Decoder Architecture. Next, we study different decoder architectures, specifically their depth and embedding dimension. We report our results in Tables 4c and 4d. Similarly to other MAE works [23,50], we found that the optimal decoder (256-d, 4 blocks) is smaller than the encoder (384-d, 12 blocks).

Data Augmentations. We evaluate our method with additional data augmentations commonly used in contrastive learning [7, 20] and present our results in Table 4e. Similar to SiamMAE [21], we observe that using color jitter significantly reduces performance. The use of Gaussian blur also leads to a decline in performance but to a lesser extent. When we do not apply the random horizontal flip, we observe a minimal drop in performance.

5 Conclusion

In this work, we introduce CropMAE, a self-supervised method for quickly learning rich features for video propagation tasks by reconstructing a crop of an image that has been masked at an extremely high proportion (over 98.5%). We empirically demonstrate that our method can learn useful features for video downstream tasks without requiring explicit video motion. These features can be learned from still images, resulting in even richer information. Thanks to our tractable pretext task, our method trains faster than existing methods and is applicable to both video frames and still images. Finally, we show on-par performances with state-of-the-art methods for three video propagation downstream tasks.

Limitations and future work. Despite being designed to work with small quantities of data and facilitate fast training, we believe the scalability of our method warrants further investigation. This includes both model scalability (*i.e.*, patch size and ViT size) and data scalability (*i.e.*, the amount of data available and the differences between images and video frames). More effort should be directed towards understanding the unique contributions of video frames instead of still images, especially concerning scalability, and determining their necessity to develop rich and robust representations.

Acknowledgements

A. Cioppa is funded by the F.R.S.-FNRS. The research reported in this publication was supported by funding from KAUST Center of Excellence on GenAI, under award number 5940, and the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence. The present research benefited from computational resources made available on Lucia, the Tier-1 supercomputer of the Walloon Region, infrastructure funded by the Walloon Region under the grant agreement n°1910247. We acknowledge EuroCC Belgium for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium.

References

1. Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., et al.: A cookbook of self-supervised learning. CoRR **abs/2304.12210** (2023). `https://doi.org/10.48550/arXiv.2304.12210`
2. Bandara, W.G.C., Patel, N., Gholami, A., Nikkhah, M., Agrawal, M., Patel, V.M.: AdaMAE: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 14507–14517. Inst. Electr. Electron. Eng. (IEEE), Vancouver, Can. (Jun 2023). `https://doi.org/10.1109/cvpr52729.2023.01394`
3. Bao, H., Dong, L., Piao, S., Wei, F.: BEiT: BERT pre-training of image transformers. In: Int. Conf. Learn. Represent. (ICLR) (May 2022), `https://openreview.net/forum?id=p-BhZSz59o4`
4. Bao, Z., Tokmakov, P., Jabri, A., Wang, Y.X., Gaidon, A., Hebert, M.: Discovering objects that can move. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 11779–11788. Inst. Electr. Electron. Eng. (IEEE), New Orleans, LA, USA (june 2022). `https://doi.org/10.1109/cvpr52688.2022.01149`
5. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. In: Cowan, J., Tesauro, G., Alspector, J. (eds.) Advances in Neural Information Processing Systems. vol. 6. Morgan-Kaufmann (1993), `https://proceedings.neurips.cc/paper_files/paper/1993/file/288cc0ff022877bd3df94bc9360b9c5d-Paper.pdf`
6. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 9630–9640. Inst. Electr. Electron. Eng. (IEEE), Montreal, QC, Canada (october 2021). `https://doi.org/10.1109/iccv48922.2021.00951`
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Int. Conf. Mach. Learn. (ICML). Proc. Mach. Learn. Res., vol. 119, pp. 1597–1607 (Jul 2020)
8. Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., Wang, J.: Context autoencoder for self-supervised representation learning. Int. J. Comput. Vis. **132**(1), 208–223 (Aug 2023). `https://doi.org/10.1007/s11263-023-01852-4`

9. Chen, X., He, K.: Exploring simple siamese representation learning. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 15745–15753. Inst. Electr. Electron. Eng. (IEEE), Nashville, TN, USA (Jun 2021). https://doi.org/10.1109/cvpr46437.2021.01549
10. Dave, I., Gupta, R., Rizve, M.N., Shah, M.: TCLR: Temporal contrastive learning for video representation. Comput. Vis. Image Underst. **219**, 1–9 (Jun 2022). https://doi.org/10.1016/j.cviu.2022.103406
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 248–255. Inst. Electr. Electron. Eng. (IEEE), Miami, FL, USA (Jun 2009). https://doi.org/10.1109/CVPR.2009.5206848, https://doi.org/10.1109/CVPR.2009.5206848
12. Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 2070–2079. Inst. Electr. Electron. Eng. (IEEE), Venice, Italy (Oct 2017). https://doi.org/10.1109/iccv.2017.226
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (ICLR). Austria (May 2021)
14. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (Jun 2010). https://doi.org/10.1007/s11263-009-0275-4
15. Fan, D., Wang, J., Liao, S., Zhu, Y., Bhat, V., Santos-Villalobos, H., M. V., R., Li, X.: Motion-guided masking for spatiotemporal representation learning. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 5596–5606. Inst. Electr. Electron. Eng. (IEEE), Paris, Fr. (october 2023). https://doi.org/10.1109/iccv51070.2023.00517
16. Feichtenhofer, C., fan, h., Li, Y., He, K.: Masked autoencoders as spatiotemporal learners. In: Adv. Neural Inf. Process. Syst. (NeurIPS). vol. 35, pp. 35946–35958. Curran Assoc. Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/e97d1081481a4017df96b51be31001d3-Paper-Conference.pdf
17. Feng, Z., Zhang, S.: Evolved part masking for self-supervised learning. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 10386–10395. Inst. Electr. Electron. Eng. (IEEE), Vancouver, Can. (Jun 2023). https://doi.org/10.1109/cvpr52729.2023.01001
18. Girdhar, R., El-Nouby, A., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: OmniMAE: Single model masked pretraining on images and videos. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 10406–10417. Inst. Electr. Electron. Eng. (IEEE), Vancouver, Can. (Jun 2023). https://doi.org/10.1109/cvpr52729.2023.01003
19. Girdhar, R., Ramanan, D.: Attentional pooling for action recognition. In: Adv. Neural Inf. Process. Syst. (NeurIPS). vol. 30, pp. 1–12. Curran Assoc. Inc., Long Beach, CA, USA (Nov 2017)
20. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent – a new approach to self-supervised learning. In: Adv. Neural Inf. Process. Syst. (NeurIPS). vol. 33, pp. 21271–21284. Curran Assoc. Inc. (2020)
21. Gupta, A., Wu, J., Deng, J., Fei-Fei, L.: Siamese masked autoencoders. In: Adv. Neural Inf. Process. Syst. (NeurIPS). vol. 37. Curran Assoc. Inc., New Orleans, LA, USA (2023), https://openreview.net/forum?id=yC3q7vInux

22. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR). vol. 2, pp. 1735–1742. Inst. Electr. Electron. Eng. (IEEE), New York, NY, USA (Jun 2019). https://doi.org/10.1109/cvpr.2006.100
23. He, K., Chen, X., Xie, S., Li, Y., Dollar, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 15979–15988. Inst. Electr. Electron. Eng. (IEEE), New Orleans, LA, USA (Jun 2022). https://doi.org/10.1109/cvpr52688.2022.01553
24. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 9726–9735. Inst. Electr. Electron. Eng. (IEEE), Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/cvpr42600.2020.00975
25. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). CoRR **abs/1606.08415** (2016). https://doi.org/10.48550/arXiv.1606.08415
26. Jabri, A., Owens, A., Efros, A.A.: Space-time correspondence as a contrastive random walk. In: Adv. Neural Inf. Process. Syst. (NeurIPS). vol. 34. Curran Assoc. Inc. (2020)
27. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 3192–3199. Inst. Electr. Electron. Eng. (IEEE), Sydney, NSW, Aust. (Dec 2013). https://doi.org/10.1109/iccv.2013.396
28. Jiang, Z., Wang, B., Xiang, T., Niu, Z., Tang, H., Li, G., Li, L.: Concatenated masked autoencoders as spatial-temporal learner. CoRR **abs/2311.00961** (2023). https://doi.org/10.48550/arXiv.2311.00961
29. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. CoRR **abs/1705.06950** (2017). https://doi.org/10.48550/arXiv.1705.06950
30. Kenton, L., Devlin, J., Chang, M.W., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. vol. 1, pp. 4171–4186. Minneapolis, Minnesota (Jun 2019)
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Int. Conf. Learn. Represent. (ICLR). New Orleans, LA, USA (May 2019)
32. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: Unsupervised learning using temporal order verification. In: Eur. Conf. Comput. Vis. (ECCV). Lect. Notes Comput. Sci., vol. 9905, pp. 527–544. Springer Int. Publ. (2016). https://doi.org/10.1007/978-3-319-46448-0_32
33. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Eur. Conf. Comput. Vis. (ECCV). Lect. Notes Comput. Sci., vol. 9910, pp. 69–84. Springer Int. Publ. (2016). https://doi.org/10.1007/978-3-319-46466-4_5
34. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. (2024), https://openreview.net/forum?id=a68SUt6zFt
35. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 DAVIS challenge on video object segmentation. CoRR **abs/1704.00675** (2017). https://doi.org/10.48550/arXiv.1704.00675

36. Qing, Z., Zhang, S., Huang, Z., Wang, X., Wang, Y., Lv, Y., Gao, C., Sang, N.: MAR: Masked autoencoders for efficient action recognition. IEEE Trans. Multimedia **26**, 218–233 (2024). `https://doi.org/10.1109/tmm.2023.3263288`
37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (Apr 2015). `https://doi.org/10.1007/s11263-015-0816-y`
38. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., Brain, G.: Time-contrastive networks: Self-supervised learning from video. In: IEEE Int. Conf. Robot. Autom. (ICRA). pp. 1134–1141. Inst. Electr. Electron. Eng. (IEEE), Brisbane, QLD, Australia (May 2018). `https://doi.org/10.1109/icra.2018.8462891`
39. Spyros, G., Praveer, S., Nikos, K.: Unsupervised representation learning by predicting image rotations. In: Int. Conf. Learn. Represent. (ICLR). Vancouver, Can. (May 2018), `https://openreview.net/forum?id=S1v4N2l0-`
40. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (Jan 2014)
41. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Adv. Neural Inf. Process. Syst. (NeurIPS). vol. 35, pp. 10078–10093. Curran Assoc. Inc. (2022)
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017). `https://doi.org/10.48550/arXiv.1706.03762`
43. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning - ICML '08. p. 1096–1103. ACM Press, Helsinki, Finland (Jul 2008). `https://doi.org/10.1145/1390156.1390294`
44. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: VideoMAE V2: Scaling video masked autoencoders with dual masking. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 14549–14560. Inst. Electr. Electron. Eng. (IEEE), Vancouver, Can. (Jun 2023). `https://doi.org/10.1109/cvpr52729.2023.01398`
45. Wang, X., Jabri, A., Efros, A.A.: Learning correspondence from the cycle-consistency of time. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 2561–2571. Inst. Electr. Electron. Eng. (IEEE), Long Beach, CA, USA (Jun 2019). `https://doi.org/10.1109/cvpr.2019.00267`
46. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 3733–3742. Inst. Electr. Electron. Eng. (IEEE), Salt Lake City, UT, USA (Jun 2018). `https://doi.org/10.1109/cvpr.2018.00393`
47. Xiao, T., Wang, X., Efros, A.A., Darrell, T.: What should not be contrastive in contrastive learning. In: Int. Conf. Learn. Represent. (ICLR). Vienna, Austria (May 2021)
48. Xie, R., Wang, C., Zeng, W., Wang, Y.: An empirical study of the collapsing problem in semi-supervised 2D human pose estimation. In: IEEE/CVF Int. Conf. Comput. Vis. (ICCV). pp. 11220–11229. Inst. Electr. Electron. Eng. (IEEE), Montreal, QC, Canada (Oct 2021). `https://doi.org/10.1109/iccv48922.2021.01105`
49. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: SimMIM: a simple framework for masked image modeling. In: IEEE/CVF Conf. Comput. Vis.

Pattern Recognit. (CVPR). pp. 9643–9653. Inst. Electr. Electron. Eng. (IEEE), New Orleans, LA, USA (Jun 2022). https://doi.org/10.1109/cvpr52688.2022.00943

50. Yao, R., Lin, G., Xia, S., Zhao, J., Zhou, Y.: Video object segmentation and tracking: A survey. ACM Transactions on Intelligent Systems and Technology **11**(4), 36:1–47 (May 2020). https://doi.org/10.1145/3391743
51. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: iBOT: Image bert pre-training with online tokenizer. In: Int. Conf. Learn. Represent. (ICLR). Vienna, Austria (May 2022), https://openreview.net/forum?id=ydopy-e6Dg
52. Zhou, Q., Liang, X., Gong, K., Lin, L.: Adaptive temporal encoding network for video instance-level human parsing. In: Proceedings of the 26th ACM international conference on Multimedia. p. 1527–1535. ACM (october 2018). https://doi.org/10.1145/3240508.3240660