Time Does Tell: Self-Supervised Time-Tuning of Dense Image Representations

Mohammadreza Salehi, Efstratios Gavves, Cees G. M. Snoek, Yuki M. Asano
QUVA Lab, University of Amsterdam
{s.salehidehnavi, e.gavves, c.g.m.snoek, y.m.asano}@uva.nl

Abstract

Spatially dense self-supervised learning is a rapidly growing problem domain with promising applications for unsupervised segmentation and pretraining for dense downstream tasks. Despite the abundance of temporal data in the form of videos, this information-rich source has been largely overlooked. Our paper aims to address this gap by proposing a novel approach that incorporates temporal consistency in dense self-supervised learning. While methods designed solely for images face difficulties in achieving even the same performance on videos, our method improves not only the representation quality for videos – but also images. Our approach, which we call time-tuning, starts from image-pretrained models and fine-tunes them with a novel self-supervised temporal-alignment clustering loss on unlabeled videos. This effectively facilitates the transfer of high-level information from videos to image representations. Time-tuning improves the state-of-the-art by 8-10% for unsupervised semantic segmentation on videos and matches it for images. We believe this method paves the way for further self-supervised scaling by leveraging the abundant availability of videos. The implementation can be found here: https://github.com/SMSD75/Timetuning

1. Introduction

Dense self-supervised learning, whereby meaningful deep features for each pixel or patch of input are learned in an unsupervised manner, has recently received increasing attention [74, 24, 57, 51]. By learning spatially consistent features for different views of an input, strong gains in unsupervised semantic segmentation have been achieved using unlabeled images. However, so far, an even more information-rich source for unsupervised training has been largely overlooked: videos. With their additional time dimension and being the most rapidly growing form of digital content, they are well-suited to scaling dense self-supervised learning even further.

Some efforts have already been made to learn from the video domain by using different frames from a video as augmentations [18, 63, 49] or by involving temporal correspondence [71, 58]; however, they mostly did it in a supervised way [29, 73, 19, 40, 65, 38, 31, 28, 41, 37, 54, 44], which is not scalable specifically for dense tasks where the number of targets can increase significantly as the number of pixels grows. To this end, self-supervised learning approaches offer a solution by reducing the need for supervision. However, these methods typically rely on the notion of “views”, which involves learning similar features for corresponding locations over time. This usually leads to a chicken-and-egg problem, where the correspondences are required for learning dense features – which in turn enable good correspondences [30].

In images, the challenge is trivially solved by considering the correspondence function based on the augmentation function. For instance, in the case of color augmentations, this correspondence is simply given by the identity. However, shifts in time cannot be viewed as mere augmenta-
tions. As we demonstrate through a new evaluation protocol, using image models on frames alone is not nearly as effective. A similar finding has also been reported, albeit for non-dense works [20, 34], which assumed the passage of time as an image augmentation. These works have generally reported reduced performances, even when compared to simple image-based pretraining methods. Similarly, video-level tasks [17, 18, 55] assume sufficiently similar semantics between different frames. This is also not true for dense tasks, as static features can only be assumed where nothing is moving – which is rare due to possible object, camera, and background motion between the frames.

To address this challenge, we propose to model the additional time-dimension explicitly to identify which pixels should retain similar embeddings and which should not. We propose two separate modules to tackle the correspondence and the dense learning, respectively. For the former, we introduce the Feature-Forwarder (FF) module, which breaks the mentioned chicken-and-egg loop by leveraging the good tracking performance of pretrained image models, and allows an approximate second “view” that can then be treated as a target for the further dense self-supervised loss. On top of this, we introduce a spatio-temporally dense clustering module, which learns unsupervised clusters across samples, locations and time. Using these two components and starting from image-pretrained features, our proposed method allows time-tuning (TIME T) the dense representation in a self-supervised manner, see Figure 1.

Finally, we demonstrate that TIME T paves the way for further scaling of self-supervised learning by leveraging the abundant availability of video datasets and transferring their knowledge to the image domain. This results in consistently achieving state-of-the-art performances not only for the task of unsupervised semantic segmentation of videos, but also for unsupervised image semantic segmentation, a feat previously out of reach for methods trained on videos.

Overall, this paper makes the following contributions:

- We show that image-based unsupervised dense segmentation models applied to videos exhibit degraded performance and lack temporal consistency in their segmentation maps.

- Building on this observation, we propose a novel dense self-supervised learning method that utilizes temporal consistency as a learning signal.

- We demonstrate that our method enables the scaling of self-supervised learning by leveraging abundant video datasets and effectively transferring knowledge to the image domain. Our approach consistently achieves state-of-the-art performance for both images and videos, opening up new opportunities in the field.

2. Related Works

Dense self-supervised learning. These methods build upon image-level self-supervised representation learning by incorporating existing losses to enhance the spatial features, demonstrating a commendable advancement. DenseCL [61] works on spatial features by constructing dense correspondences across views using the contrastive objective given in MoCo [23], while PixPro [62] utilizes the augmentation wrapper to get the spatial correspondence of the pixel intersection between two views. Similarly, MaskContrast [57], Leopart [74], DetCon [24] and Odin [25] also ensure spatial feature similarities via contrastive learning and spatial correspondences. Self-Patch [67] treats the spatial neighbors of the patch as positive examples for learning more semantically meaningful relations among patches. Inspired by SelfPatch, AD-CLR [69] proposes patch-level contrasting via query crop and cross-attention mechanism. Pursuing the same objective through none end-to-end approaches. Both [68] and [59] propose an unsupervised salient object segmentation pipeline that extracts noisy object masks from the inputs and fine-tunes a specifically designed object segmentation head with several self-training steps to make an unsupervised semantic segmentation model. Unlike the existing image-based approaches, we propose a method that improves the dense prediction performance of a pretrained encoder by explicitly modeling the temporal dimension and learning from diverse natural dynamics and variations found in videos.

Video to image knowledge transfer. Learning from videos instead of images has recently received attention since they contain far more information than still images and hold the potential for learning rich representations of the visual world. To this end, several non-dense works have been released [46, 4, 27]. VITO [46] shows that naively applying image domain self-supervised learning methods on videos can lead to a performance drop coming from the distribution shift, which by applying data processing techniques is relieved but not fully solved. Following the same way, [27] trains masked autoencoders [22] with contrastive learning [10] on video datasets and shows a decent performance on both image and video tasks. While video to image knowledge transfer is the end goal of such methods and our paper; our dense task is different and is as of yet unaddressed.

Unsupervised video object segmentation. Unsupervised video object segmentation methods do not require any manual annotations. However, they are only designed to tackle a foreground/background segmentation task, which refers to the segmentation of the most prominent, general objects in video sequences [66, 43, 50, 1, 39]. For instance, [66]
Figure 2: Time-Tuning overview. We conduct self-supervised video semantic segmentation by tuning image-pretrained models with temporal information from videos. (a): Our general pipeline of adapting an image-pretrained ViT transformer on video data using our dense clustering loss. The encoder is specified by $\Psi$ and kept frozen except for the last two layers. Also, $z_t$ shows the output of Sinkhorn-Knopp algorithm, which is forwarded to time-step $T$ to be compared with the features that are obtained using the last time-step. (b): Detailed view into our Feature-Forwarder module that is used for aligning cluster-maps from past frames to logit-features of future ones.

uses an AutoEncoder-based architecture based on slot attention so that pixels that show the same motion flows based on optical-flow are grouped together. In this way, dominant moving objects can be detected and separated from the background. In our experiments, we benchmark two representative instances of this line of work, i.e., [66, 1], to provide better insights into the differences of unsupervised object detection methods and our proposed benchmark for video semantic segmentation.

To conclude, we propose the first self-supervised video semantic segmentation method, which tunes image-trained models such as [9, 72] based on the temporal information that exists in unlabeled videos to enhance their effectiveness in dense prediction tasks.

3. Method

Our method works by densely clustering features in a manner that is consistent with time. At a high level, it works by transforming features from a past time $t'=t$ to the current frame $t'=T$ and forcing these to be consistent with the currently observed ones at $t'=T$. The overall learning signal comes from the dense temporal clustering of forwarded features and allows “timetuning” image-pretrained models to learn from videos. Figure 2 shows the proposed method’s architecture. In the following, we describe each component in detail.

3.1. Feature Forwarding

The goal of this component, the feature-forwarder (FF), is to propagate the dense features in time as accurately as possible, given some RGB input frames. More formally, let $f$ be the feature forwarding function, $I_t$ be a frame at time $t$ for a given video, and $z_t$ the features corresponding to $I_t$, then the task is to propagate $z_t$ to a future time $T$,

$$f([I_t, \ldots, I_T], z_t) \rightarrow \hat{z}_T.$$  (1)

Note that $\hat{z}_T$ and $z_T$ are not necessarily the same for $T > t$; since $\hat{z}_T$ represents an approximation of $z_T$ using the previous features. Obviously, the further $T$ is from $t$, the more variance is seen and therefore a better learning signal might be provided; however, for such larger intervals, the idea of feature forwarding encounters two main challenges: error accumulation and object non-permanence.

Challenge 1: Error accumulation. While computing how features have changed simply based on the start and end frames is the most straight-forward (i.e., $I_t$ and $I_T$), it does not fully leverage the knowledge contained in the intermediate frames. Instead, it is common [30, 60, 36] to forward features for every time-step $\delta t$, i.e.,

$$f([I_t, \ldots, I_T], z_t) = \bigcup_{t'=t}^{t'=T} f([I_{t'}-\delta t, I_{t'}], z_t).$$  (2)

where $\bigcup$ is a composition operator such as element-wise multiplication. However, this results in the accumulation of errors over time.

Challenge 2: Object non-permanence. One difficulty that arises when using videos as training data is the fact that objects sometimes simply disappear from the screen. Compared to images, which carry a heavy photographer’s bias, videos are more prone to natural variation due to camera and object motion resulting in temporary object non-permanence in a particular video clip. A method that simply assumes an object (or a feature) has to be present in any given frame simply because it was present in the previous
We address these challenges using a novel, yet simple component, which we call the Stabilizing Feature-Forwarder. The first choice in designing this module is the function and data used for composing information across time. For instance, a simple approach would be to use pixel-wise optical flow and aggregate the flow at the feature level to predict how the features change over time. However, as we also show in our experiments, optical flow is prone to accumulating error with time and also cannot easily recover from occlusions [6]. Instead, we utilize the fact that we wish to pretrain the visual encoder and occlusions [6]. We propose a self-supervised clustering task across views that include information about the dynamics with time, its relation to mask-propagation methods. While previous methods such as DINO [9] and STC [30] have utilized a similar technique for propagating ground-truth masks for evaluating, for example on the DAVIS dataset, there are three key differences to our forwarding method.

First, instead of propagating binary maps of foreground-vs-background, our feature-forwarder has as inputs soft, noisy and multi-label self-supervised segmentation maps, which require the forwarder to tolerate overlapping of different object probabilities throughout training. Second, previous methods have used this approach mainly for inference; we, however, use this module as a trainable component and show how our loss improves this forwarder with training time. Finally, we do not follow a typical re-normalizing step across feature dimensions (typically done after Eq. (5)) as this harms the scale of logits that are being propagated and leads to heavily diluted target distributions.

3.2. Self-supervised dense clustering

While the Feature-Forwarder produces target features that include information about the dynamics with time, its computation utilizes $\Phi(I_T)$ and could lead to trivial solutions, i.e., $f(\{I_t,...,I_T\}, z_t) = \Phi(I_T)$. To counteract this, we propose a self-supervised clustering task across views in time. For this, we utilize the basic online clustering algorithm based optimal-transport [12], utilized in works such as SeLa [2], SwAV [8] and DINOv2 [45]. In particular, let $\Psi$ be a visual encoder with a clustering head $g$ that yields a $K$ dimensional output, then the clustering loss is given by:

$$\mathcal{L}(x_i) = -\tilde{y}_i \log (g(\Psi(x_i))),$$

which is a standard cross-entropy loss with regards to self-supervised pseudo-label $\tilde{y}_i$. These labels, in turn, are generated by solving an entropy-regularised optimal transport problem on the batch $B$ [2]:

$$\min_{\tilde{y}} (\tilde{y}_i - \log g(\Psi(x_i))) + \frac{1}{\lambda} \text{KL}(\tilde{y}_i r e^{\mathcal{T}}),$$

where $r = \frac{1}{K} \mathbb{1}$, $c = \frac{1}{|B|} \mathbb{1}$. (8)

Here $\lambda$ controls the entropy regularisation and $r, c$ are the marginals for the prototypes and the batch, respectively. Note that solving this problem can be done extremely quickly on the GPU and yields soft pseudo-labels $\tilde{Y}$, such that $\text{argmax}(\tilde{Y}) = \tilde{y}_i$.

These might yield sufficiently similar semantics [17] this is not true for dense tasks, as static features can only be assumed for videos where nothing is moving, which is rare.

3.3. Overall TIME T loss

We combine the previous two modules to arrive at our full method (as shown in Fig. 2): First, self-supervised

Concretely, we first compute the L2-normalised dense features of a pretrained visual encoder $\Phi$ to compute spatial similarities across time (the $'$ in $t'$ left out for clarity):

$$F_{ij}^t = \langle \Phi(I_i), \Phi(I_T)_j \rangle / \tau,$$

which yields matrices $F^{t,...,t-\delta t} \in \mathbb{R}^{N \times N}$ with values between [0,1] indicating semantic similarities between the $N$ spatial features that are sharpened by a temperature $\tau$. Note that because we compare each frame with the final frame, the effect of object non-permanence can be minimized: even if the object is only present in every other frame, there is enough signal to forward its features.

Next to propagate from time $t$ to $T$, these similarities are stacked and normalised along time, as follows:

$$\tilde{F}_{ij}^t = \exp(N(F_{ij}^t)) / \sum_{i',t'} \exp(N(F_{ij'}^t)),$$

where $t \leq t' \leq T$ and $N$ is a neighborhood thresholding method which forces short-term spatial-smoothness, i.e., $N(F_{ij}^t)=0$ if $i$ is not within a local window around $j$ with the size $k$. Finally, the propagated feature are computed $\tilde{z}_T$ as:

$$\tilde{z}_T(j) = \sum_{t',i} \tilde{F}_{ij'}^t \tilde{z}_{t'}(i), \quad j \in \{1,...,N\}.$$ (5)

This means that for arriving at the target feature $\tilde{z}_T$, we not only use the previous frame as the source, but instead use the past $(T-t)/\delta t$ frames and aggregate these.

Relation to mask-propagation methods. While previous methods such as DINO [9] and STC [30] have utilized a similar technique for propagating ground-truth masks for evaluating, for example on the DAVIS dataset, there are three key differences to our forwarding method. First, instead of propagating binary maps of foreground-vs-background, our feature-forwarder has as inputs soft, noisy and multi-label self-supervised segmentation maps, which require the forwarder to tolerate overlapping of different object probabilities throughout training. Second, previous methods have used this approach mainly for inference; we, however, use this module as a trainable component and show how our loss improves this forwarder with training time. Finally, we do not follow a typical re-normalizing step across feature dimensions (typically done after Eq. (5)) as this harms the scale of logits that are being propagated and leads to heavily diluted target distributions.
Sinkhorn-Knopp clustering is conducted on early features \( g(\Psi(I_t)) \), yielding soft pseudo-labels \( SK(g(\Psi(I_t))) = \tilde{Y}_t \). These are then forwarded in time using our Feature-Forwarder to arrive at dense targets \( FF(\tilde{Y}_t) \), which are used in the final loss. Compactly:

\[
L_{\text{TIME}}(I_T) = -\sum_{i,j} FF(\tilde{Y}_t) \log(g(\Psi(I_T))). \tag{9}
\]

4. Experiments

4.1. Setup

Datasets. We train our method and baselines on YTVOS [64], one of the largest video segmentation datasets available, and evaluate on DA VIS17 [48] and YTVOS. For YTVOS the ground truth masks are only available for the first frames of the test and validation sets, and therefore, a fixed random 20% of the training set is used for testing. For transfer learning experiments, we use the validation set of Pascal VOC 2012 [16]. As the dataset has been commonly used as a main reference for recent works in dense self-supervised image segmentation [74, 57, 61]. For completeness, we also report the performance on egocentric datasets that have less object-centric bias and are prevalent in real-world scenarios. For egocentric experiments, we train on EPIC-KITCHENS-100 [13] and evaluate on VISOR [14]. Further details are provided in Appendix A.

Models and baselines. Currently, there is a lack of available unsupervised semantic video semantic segmentation methods. Nevertheless, we have included a comprehensive comparison of our method with state-of-the-art techniques in both image and video domains for unsupervised image semantic segmentation and unsupervised video object segmentation. To evaluate the image-based models, we utilized either official reported numbers or provided pretrained models of STEGO [21] and Leopart [74]. We have taken every measure to ensure fairness in our comparison. To do so, all the used pretrained models have the same pretraining dataset (ImageNet-1k), and the number of their backbone parameters is roughly similar. For those models that use extra datasets, for instance, Leopart, we select the pretrained backbones that closely match the specification of YTVOS training dataset. Additionally, we trained image-based models on the same video datasets, where we converted video data into their corresponding image data and reported their performance wherever possible. To ensure comprehensive analysis, we have included the recent concurrent work Flowdino [71], which utilizes optical flow to refine DINO features on unlabelled videos, in our comparison benchmarks as well. This is done only in cases where there is a shared experimental setup.

Evaluation procedure. Despite unsupervised object segmentation being a well-established evaluation in the image domain [57, 74], evaluating unsupervised video multi-label object segmentation is challenging due to the absence of an established evaluation protocol for video object semantic segmentation without supervision. In this regard, we propose a set of evaluation protocols for unsupervised video multi-label object segmentation, which exploits existing video object segmentation datasets for evaluation purposes (see details in Appendix B). In our experiments, we discard any projection head used during training and evaluate ViT’s spatial tokens directly, similar to [57, 32, 61], using four methods: classification with a linear head, classification with an FCN head, overclustering, and clustering with as many clusters as the number of ground truth objects. To fine-tune different heads on top of the frozen spatial tokens, we follow [57]. For unsupervised semantic segmentation, we apply \( K \)-Means on all spatial tokens, where \( K \) is chosen to be higher or equal to the ground-truth number of classes, following the common protocol in image clustering [32, 56]. For grouping clusters to ground-truth classes, we match them either by pixel-wise precision or Hungarian matching on merged cluster maps [35]. For video datasets, we allow the matching to be per-frame, per-video, and per-dataset. The evaluation metric is permutation-invariant, following [32], and the results are averaged over five different seeds. Clustering evaluations are preferred as they require less supervision and fewer hyperparameters than training a linear classifier and operate directly on the learned embedding spaces. We report our results in mean Intersection over Union (mIoU) unless otherwise specified.

Model training. We train a ViT-Small with patch size 16 and initialized from ImageNet-pretrained DINO weights [9] using the proposed self-supervised loss function. For the ablations and the main experiments, our models are trained for 12 and 30 epochs respectively. Further training details are provided in Appendix A. Code will be made available.

4.2. Ablations

We first examine the essential parameters of our method by training TimeT on YTVOS and assessing its ability to perform semantic segmentation on Pascal VOC. In addition to presenting the results through clustering and overclustering, we also demonstrate linear classification outcomes.

Number of prototypes. We first ablate the influence of the number of prototypes on downstream segmentation performance. Results in Table 1a indicate a sharp increase in performance with a rise in the number of prototypes, but once a sufficiently large number is reached, performance stabilizes. We observe peak performance with a moderate number of 200 prototypes. The stability of our method over
forwarding the network’s output logits sharpened feature maps and does not work when simply
portantly, our Feature-Forwarder relies on forwarding SK-features, these gains are small (+4.6%)
sumption can already lead to some gains. However, com-
many, training videos are not very dynamic, and a static as-
formance compared to the first row. This shows that
feature map from the source frame is simply forwarded to
"Identity" shows the performance that is obtained when the
numbers here can be traced back to using an un-entropy-
clustering step. The reason for obtaining considerably lower
puts (\[\Phi(x)\]) does not work as well as SK-
Choice of propagation feature. Finally, in Table 2, we vary the type of feature that we forward to a future time. In
the first row, we report the performance when not using any feature forwarding – thus solely training on single-frame in-
puts (i.e., the same as row 1 in Table 1c). In the second row, “Identity” shows the performance that is obtained when the
feature map from the source frame is simply forwarded to the future without any changes, which shows an increase in the performance compared to the first row. This shows that
framing videos are not very dynamic, and a static as-
umption can already lead to some gains. However, com-
pared to forwarding the Sinkhorn-Knopp (SK) clustered features, these gains are small (+4.6% vs. + 17%). Importantly, our Feature-Forwarder relies on forwarding SK-
(contact) followed by a clustering step. The reason for obtaining considerably lower
numbers here can be traced back to using an un-entropy-
regularized clustering algorithm. In this case, the logits tend to cause a few prototypes to dominate the cluster centers, exacerbated after the propagation, resulting in highly noisy and uninformative propagated logits. This shows that care-
ful design of the Feature-Forwarder is indeed required.

Applying \(\text{TIME} T\) to different models. As shown in Table 3, our method is generalisable to different backbones and self-supervised learning initializations, enabling it to be
an effective method to transfer the knowledge of the videos to images in an unsupervised way, which reduces the cost of labeling for different downstream tasks.

### Table 1: Ablations of the key parameters of our method.

The model is trained for 12 epochs on Pascal VOC, and results for unsupervised segmentation with clustering (\(K=21\)), overclustering (\(K=500\)), and linear pixel-wise classification (LC) are shown. The stability of our method over a range of prototypes (50-300), inter-frame time intervals (\(\delta T \in \{0.5s-1.0s\}\)), and the number of training frames (4-8) at a fixed clip duration (\(T\)) shows the robustness of the method.

<table>
<thead>
<tr>
<th>(P)</th>
<th>(#) frames (\delta T)</th>
<th>LC</th>
<th>K=21</th>
<th>K=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 0s 50.7 5.6 26.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>4 0.2s 56.5 7.4 36.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>4 0.5s 59.7 9.2 42.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>4 1.0s 57.1 8.3 38.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(P)</th>
<th>(#) frames (\delta T)</th>
<th>LC</th>
<th>K=21</th>
<th>K=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.7 5.6 26.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identity</td>
<td>55.3 7.4 36.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\Psi(x))</td>
<td>53.1 6.6 35.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\Phi(x))</td>
<td>59.7 9.2 42.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: Propagating different features in FF on Pascal VOC.

We find that our Sinkhorn-Knopp (SK) based module outperforms static training (‘None’ or ‘Identity’). Moreover, propagating logits \(\Psi(x)\) does not work as well as SK-
regularised features.

<table>
<thead>
<tr>
<th>FF type</th>
<th>LC</th>
<th>K=21</th>
<th>K=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>50.7</td>
<td>5.6</td>
<td>26.2</td>
</tr>
<tr>
<td>Identity</td>
<td>55.3</td>
<td>7.4</td>
<td>36.1</td>
</tr>
<tr>
<td>(\Psi(x))</td>
<td>53.1</td>
<td>6.6</td>
<td>35.5</td>
</tr>
<tr>
<td>(\Phi(x))</td>
<td>59.7</td>
<td>9.2</td>
<td>42.8</td>
</tr>
</tbody>
</table>

### Table 3: Applying \(\text{TIME} T\) to different pretrainings on Pascal VOC. \(\text{TIME} T\) can boost (\(\uparrow\)) the performance of dif-
erent backbones with different initialization by a consider-
able margin, showing the generality of our approach.

<table>
<thead>
<tr>
<th>Pretrain</th>
<th>Backbone</th>
<th>At Init</th>
<th>+(\text{TIME} T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSN [3]</td>
<td>ViT-S/16</td>
<td>26.0</td>
<td>55.4</td>
</tr>
<tr>
<td>iBOT [72]</td>
<td>ViT-S/16</td>
<td>32.1</td>
<td>62.1</td>
</tr>
<tr>
<td>DINO [9]</td>
<td>ViT-S/8</td>
<td>22.5</td>
<td>55.8</td>
</tr>
<tr>
<td>DINO [9]</td>
<td>ViT-B/16</td>
<td>28.9</td>
<td>59.1</td>
</tr>
</tbody>
</table>
Figure 3: **TIME T unsupervised segmentations.** As we regularize DINO’s backbone to be consistent across time on YTVOS, it obtains strong performance on both image and video segmentation datasets, yielding high class consistencies (indicated by the segmentation colors) and tight borders. We provide more qualitative results in Appendix D.

Table 4: **Unsupervised video semantic segmentation results** in mIoU for clustering (K=GT) and overclustering. TIME T not only gets better numbers on YTVOS, but also achieves considerably better results on DA VIS. This shows the better quality of the learned features at transferability to other video datasets. The clip-length is set to 16 and 4 for DA VIS and YTVOS, respectively. For clustering, the Hungarian algorithm [35] matches the unsupervised segmentation clusters (K) with the ground truth (GT) per frame (F), clip (C) or across the whole dataset (D). For overclustering, we use K=10 for the frame-wise (F) and clip-wise (C) evaluations, and for dataset-wise (D) evaluation, K=200 for DA VIS and K=500 for YTVOS. * denotes finetuning pretrained models on the same video frames as input as our method. The matching protocol is greedy many-to-one, see Appendix B for details.

### 4.3. Large-scale experiments

**Unsupervised video semantic segmentation.** In this section, we train our TIME T method on the YTVOS dataset and evaluate it for unsupervised video object segmentation on both DA VIS and YTVOS. The results are shown in Table 4. It should be noted that the number of prototypes we analyzed in Table 1a is not the number of clusters used to report accuracies for Table 4, as the actual number is usually unknown in the case of unsupervised learning. For the clustering experiment, we set K to the number of ground truth objects, which makes evaluation metrics easier to interpret and is commonly done in image clustering and segmentation. This evaluation is repeated but for larger numbers of clusters, in the “overclustering” scenario. Note that in dense self-supervised learning, overclustering can be particularly important because the learned representations are used as a feature extractor for downstream tasks such as semantic segmentation or object detection. By using more fine-grained representations, the network may be able to extract more discriminative features such as object parts [74], which can lead to better performances.

From Table 4, we observe a clear trend: our method, trained on YTVOS not only achieves superior performances on YTVOS, but also beats existing image-trained models on DA VIS by a large margin. In particular, the state-of-the-art self-supervised clustering method, Leopart, has 4% lower performance on per dataset DA VIS evaluation for K=GT and 1.6% lower performance on YTVOS. The gap even becomes larger when Leopart is trained on the same video dataset with 8% to 16% lower numbers in per dataset and per clip numbers across different datasets. This means that when grouping objects of the same class over time and
the whole dataset, the image-based self-supervised methods have trouble generalising to videos, where objects are not centered in the frame, can appear in varying poses, and are more difficult in general. A strong contender is Leopart, which matches the performance of DINO for per frame and per clip clustering, however, improves considerably with the per dataset clustering. We attribute this to the fact that their dense learning objective improves mainly the generalisation capabilities of the learned representation, however falling short of generalising to temporal variations. We make similar observations for overclustering, improving the DINO baseline which we utilize as initialisation by 12% to 14% across different datasets in the per dataset metric.

We conclude that the proposed method outperforms all other methods in learning robust and discriminative representations for dense video understanding, and image-based self-supervised learning may not have sufficient generalisation capacity. This shows that if used right, time is a particularly strong inductive bias for self-supervised learning. Figure 3 shows segmentations returned by the proposed model trained on YTVOS and tested on YTVOS, DA VIS, and VISOR respectively. The proposed method groups objects accurately, and importantly, the frame segmentations are considerably consistent over time (see more examples in Appendix D). This highlights the importance and relevance of temporal fine-tuning, not just for higher accuracy, but crucially for consistency and robustness.

**Transfer from video training to images.** Despite the common belief that features learned from videos perform worse when transferred to images [34, 20], the results in Table 5 demonstrate our method achieves high performance that match state-of-the-art methods directly trained on images, specifically for the K=500, FCN, and LC metrics. We also compare our performance gains in videos against models designed for unsupervised image segmentation. The results show that models highly biased towards image datasets cannot provide the same performance when trained on videos, lagging behind our approach by 7% to 30%. Our results demonstrate that achieving high transfer performance on challenging tasks through video-based self-supervised learning is not only feasible but can also maintain high performance across modalities. These findings suggest that our method can drive further advances in self-supervised learning and inspire new directions for research in this field. Figure 3 shows the qualitative results on Pascal VOC. For more visualisations we refer to Appendix D.

**Salient object segmentation.** In Table 6, we compare foreground masks obtained with various DINO ViT based methods. We use the cluster-based foreground extraction protocol from [74] (details provided in Appendix C). First, we find that our method outperforms the original DINO at-

<table>
<thead>
<tr>
<th>Pascal VOC</th>
<th>DAVIS</th>
<th>YTVOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DINO [9]</td>
<td>52.1</td>
<td>34.5</td>
</tr>
<tr>
<td>Leopart [74]</td>
<td>59.6</td>
<td>37.3</td>
</tr>
<tr>
<td>STEGO [21]</td>
<td>49.1</td>
<td>30.4</td>
</tr>
<tr>
<td><strong>TMT (ours)</strong></td>
<td><strong>63.9</strong></td>
<td><strong>44.5</strong></td>
</tr>
</tbody>
</table>

Table 5: **Transfer from video training to images.** Numbers taken from [74, 52, 70]. \*: finetuning pretrained models on the same video frames as input as our method. †: uses a 40% larger superset of our train data. Details on the different datasets and methods are provided in the Appendix.

Table 6: **Salient object segmentation.** We report performance using the Jaccard score [9] and use the official pretrained models for evaluation.

**Generalisation to egocentric datasets.** We train our method on EPIC-Kitchens-100 [13] and evaluate on VISOR [14]. We use the official code to convert VISOR to a DA VIS-like structure, in which we can report our numbers for per frame and per clip evaluation. Note that, after conversion, the object IDs do not maintain global consistency in the whole dataset; therefore, we cannot report the per dataset number. As Table 7 shows, TMT outperforms image-based competitors on egocentric datasets as well.
Table 7: Generalisation to egocentric datasets. Unsupervised semantic segmentation results in mIoU for clustering (K=GT) on VISOR [14] after training on EPIC-Kitchens-100 [13]. The clip-length is set to 4. Our method gets better results on egocentric datasets as well.

<table>
<thead>
<tr>
<th>Model</th>
<th>Encoder</th>
<th>Params</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>ViT-S16</td>
<td>21M</td>
<td>35.1</td>
</tr>
<tr>
<td>MoCo-v3*</td>
<td>ViT-S16</td>
<td>21M</td>
<td>19.5</td>
</tr>
<tr>
<td>DINO* [9]</td>
<td>ViT-S16</td>
<td>21M</td>
<td>47.9</td>
</tr>
<tr>
<td>CrOC* [52]</td>
<td>ViT-S16</td>
<td>21M</td>
<td>50.0</td>
</tr>
<tr>
<td>Leopart* [74]</td>
<td>ViT-S16</td>
<td>21M</td>
<td>63.6</td>
</tr>
<tr>
<td>DINO [9]</td>
<td>ViT-B16</td>
<td>86M</td>
<td>55.9</td>
</tr>
<tr>
<td>MAE [22]</td>
<td>ViT-B16</td>
<td>86M</td>
<td>6.6</td>
</tr>
<tr>
<td>Hummingbird [5]</td>
<td>ViT-B16</td>
<td>86M</td>
<td>70.5</td>
</tr>
</tbody>
</table>

Table 8: Visual In-Context Learning evaluation. Numbers are taken from [5]. *: numbers are produced by this paper. This approach is equivalent to a non-parametric nearest-neighbor based evaluation on Pascal VOC. The details of the evaluation benchmark can be found in the provided implementation codes.

5. Conclusion

This paper has aimed to learn dense representations that are learned from videos; yet can be generalised to the image domain as well. As video content is growing rapidly and contains more information compared to images, learning generalisable knowledge from them facilitates further scaling of self-supervised learning methods. To this effect, we have proposed a self-supervised clustering loss to encourage temporally consistent features between different frames of a video. To efficiently find corresponding views between clip frames, we have proposed to recycle pretrained transformer features to leverage their natural tracking ability at each clip. Our empirical results indicate that this method achieves significant gains on the challenging task of video object segmentation across three different evaluation protocols and three datasets. Moreover, by transferring the learned model to the image domain, we have demonstrated the generalisability of the learned features by surpassing or matching the state-of-the-art for unsupervised image segmentation.

6. Acknowledgement

This work is financially supported by Qualcomm Technologies Inc., the University of Amsterdam and the allowance Top consortia for Knowledge and Innovation (TKIs) from the Netherlands Ministry of Economic Affairs and Climate Policy.

References


[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.


[28] Ping Hu, Fabian Cabu, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks...


[53] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE/CVF conference on
computer vision and pattern recognition, pages 8934–8943, 2018. 15


