Supplementary Material

Ablation of our architecture

As mentioned in Section 3 of the paper we found that using a different aggregation function (different instance of the ConvGRU) for past and future features while keeping the same 2D3D-Resnet18 backbone achieved better results. Here we want to investigate this further and compare the effect of using the same or different aggregation function or backbone for past and future features, *i.e.* sharing weights of the feature extractor for the present representation z_v and and past/future representation z_{pf}. For this ablation we pretrain our models on UCF101 and evaluate the representations via finetuning. The results are shown in Table 1. When using the same aggregation function and the same or different backbone to extract past and future features, we observe a small performance drop compared to the best setting, whereas using both different backbones and different aggregation functions decreases the quality of the representation significantly. We use the best performing setting for all of our experiments.

backbone	agg	top1 Accuracy on UCF101
same	same	62.4
different	same	62.8
same	different	**63.6**
different	different	60.6

Table 1: Ablation of our architecture.

Temporal Negatives

Next, we want to validate the effectiveness of our temporal negatives. We train all methods on UCF101 using spatial transformations that are applied independently to the past, present and future blocks. For evaluation we finetune on UCF101. All methods in Table 2 that do not employ temporal negatives fail, suggesting that the learned representations are bad initializations. Since neither temporal nor spatial negatives are used here, only random video sequences are considered as negatives and the InfoNCE loss in Eq. 1 of the paper gives a lower bound on the mutual information. Adding temporal negatives on the other hand enables our method to learn a representation that is useful for action recognition. These experiments confirm the observation in Section 3.2 that a structured set of hard negatives which are not sampled from the marginal distributions are more effective for representation learning than an accurate approximation of the mutual information.

Model	temporal negatives	spatial augmentations	top1 Accuracy on UCF101
Random Init	-	-	54.4
Ours	✗	-	48.2
Ours	✗	crop	47.3
Ours	✗	crop + flip	47.5
Ours	✗	crop + flip + rot	51.5
Ours	✓	crop + flip	58.2

Table 2: Effectiveness of temporal negatives.