Similarity contrastive estimation for image and video soft contrastive self-supervised learning

Julien Denize, Jaonary Rabarisoa, Astrid Orcesi, Romain Hérault

Machine Vision and Applications (2023) 34:111
https://doi.org/10.1007/s00138-023-01444-9

ORIGINAL PAPER

Similarity contrastive estimation for image and video soft contrastive self-supervised learning

Julien Denize[1,2] · Jaonary Rabarisoa[1] · Astrid Orcesi[1] · Romain Hérault[2]

Received: 30 March 2023 / Revised: 4 July 2023 / Accepted: 3 August 2023 / Published online: 26 September 2023

Abstract
Contrastive representation learning has proven to be an effective self-supervised learning method for images and videos. Most successful approaches are based on Noise Contrastive Estimation (NCE) and use different views of an instance as positives that should be contrasted with other instances, called negatives, that are considered as noise. However, several instances in a dataset are drawn from the same distribution and share underlying semantic information. A good data representation should contain relations between the instances, or semantic similarity and dissimilarity, that contrastive learning harms by considering all negatives as noise. To circumvent this issue, we propose a novel formulation of contrastive learning using semantic similarity between instances called Similarity Contrastive Estimation (SCE). Our training objective is a soft contrastive one that brings the positives closer and estimates a continuous distribution to push or pull negative instances based on their learned similarities. We validate empirically our approach on both image and video representation learning. We show that SCE performs competitively with the state of the art on the ImageNet linear evaluation protocol for fewer pretraining epochs and that it generalizes to several downstream image tasks. We also show that SCE reaches state-of-the-art results for pretraining video representation and that the learned representation can generalize to video downstream tasks. Source code is available here: https://github.com/juliendenize/eztorch.

Keywords Deep learning · Self-supervised learning · Contrastive · Representation

1 Introduction

Self-Supervised learning (SSL) is an unsupervised learning procedure in which the data provide its own supervision to learn a practical representation of the data. A pretext task is designed to make this supervision. The pretrained model is then fine-tuned on downstream tasks, and several works have shown that a self-supervised pretrained network can outperform its supervised counterpart for image [1–3] and video [4, 5]. It has been successfully applied to various image and video applications such as image classification, action classification, object detection and action localization.

Contrastive learning is a state-of-the-art self-supervised paradigm based on Noise Contrastive Estimation (NCE) [6] whose most successful applications rely on instance discrimination [7–10]. Pairs of views from same images or videos are generated by carefully designed data augmentations [4, 8, 11]. Elements from the same pairs are called *positives*, and their representations are pulled together to learn view invariant features. Other instances called *negatives* are considered as noise, and their representations are pushed away from positives. Frameworks based on contrastive learning paradigm require a procedure to sample positives and negatives to learn a good data representation. Videos add the time dimension that offers more possibilities than images to generate positives such as sampling different clips as positives [4, 12], using different temporal context [13–15].

A large number of negatives are essential [16], and various strategies have been proposed to enhance the number of negatives [7, 8, 17, 18]. Sampling hard negatives [18–22]

✉ Julien Denize
julien.denize@cea.fr

Jaonary Rabarisoa
jaonary.rabarisoa@cea.fr

Astrid Orcesi
astrid.orcesi@cea.fr

Romain Hérault
romain.herault@insa-rouen.fr

[1] Université Paris-Saclay, CEA, List, F-91120 Palaiseau, France

[2] LITIS, INSA Rouen, Normandie Université, 76801 Saint Etienne du Rouvray, France

improves the representations but can be harmful if they are semantically false negatives which causes the "class collision problem" [23–25].

Other approaches that learn from positive views without negatives have been proposed by predicting pseudo-classes of different views [1, 3, 26], minimizing the feature distance of positives [2, 4, 27] or matching the similarity distribution between views and other instances [28]. These methods free the mentioned problem of sampling hard negatives.

Based on the weaknesses of contrastive learning using negatives, we introduce a self-supervised soft contrastive learning approach called Similarity Contrastive Estimation (SCE) that contrasts positive pairs with other instances and leverages the push of negatives using the inter-instance similarities. Our method computes relations defined as a sharpened similarity distribution between augmented views of a batch. Each view from the batch is paired with a differently augmented query. Our objective function will maintain for each query the relations and contrast its positive with other images or videos. A memory buffer is maintained to produce a meaningful distribution. Experiments on several datasets show that our approach outperforms our contrastive and relational baselines MoCov2 [29] and ReSSL [28] on images. We also demonstrate using relations for video representation learning is better than contrastive learning.

Our contributions can be summarized as follows:

- We propose a self-supervised soft contrastive learning approach called Similarity Contrastive Estimation (SCE) that contrasts pairs of augmented instances with other instances and maintains relations among instances for either image or video representation learning.
- We demonstrate that SCE outperforms on several benchmarks its baselines MoCov2 [29] and ReSSL [28] on images on the same architecture.
- We show that our proposed SCE is competitive with the state of the art on the ImageNet linear evaluation protocol and generalizes to several image downstream tasks.
- We show that our proposed SCE reaches state-of-the-art results for video representation learning by pretraining on the Kinetics400 dataset as we beat or match previous top-1 accuracy for finetuning on HMDB51 and UCF101 for ResNet3D-18 and ResNet3D-50. We also demonstrate it generalizes to several video downstream tasks.

2 Related work

2.1 Image self-supervised learning

Early self-supervised learning In early works, different *pretext tasks* to perform Self-Supervised Learning have been proposed to learn a good data representation. They consist in transforming the input data or part of it to perform supervision such as: instance discrimination [30], patch localization [31], colorization [32], jigsaw puzzle [33], counting [34], angle rotation prediction [35].

Contrastive learning Contrastive learning is a learning paradigm [1, 2, 7, 8, 11, 16, 17, 21, 22, 36–39] that outperformed previously mentioned *pretext tasks*. Most successful methods rely on instance discrimination with a *positive* pair of views from the same image contrasted with all other instances called *negatives*. Retrieving lots of negatives is necessary for contrastive learning [16], and various strategies have been proposed. MoCo(v2) [7, 29] uses a small batch size and keeps a high number of negatives by maintaining a memory buffer of representations via a momentum encoder. Alternatively, SimCLR [8, 40] and MoCov3 [41] use a large batch size without a memory buffer, and without a momentum encoder for SimCLR.

Sampler for contrastive learning All negatives are not equal [23], and hard negatives, negatives that are difficult to distinguish with positives, are the most important to sample to improve contrastive learning. However, they are potentially harmful to the training because of the "class collision" problem [23–25]. Several samplers have been proposed to alleviate this problem such as debiasing negatives sampling [25] further improved by selecting hard negatives [19], or using the nearest neighbor as positive for NNCLR [22]. Truncated-triplet [39] optimizes a triplet loss using the k-th similar element as negative that showed significant improvement. It is also possible to generate views by adversarial learning as AdCo [21] showed. Some other works [42, 43] proposed a denoised contrastive loss that reduces or reverses the gradient for medium and highly similar negatives. They use hard margins between different categories of negatives. Instead, we propose a soft contrastive loss that seeks to estimate relations between instances and consider all negatives equally.

Contrastive learning without negatives Various siamese frameworks perform contrastive learning without the use of negatives to avoid the class collision problem. BYOL [2] trains an online encoder to predict the output of a momentum updated target encoder. SwAV [1] enforces consistency between online cluster assignments from learned prototypes. DINO [3] proposes a self-distillation paradigm to match distribution on pseudo class from an online encoder to a momentum target encoder. Barlow-Twins [44] aligns the cross-correlation matrix between two paired outputs to the identity matrix that VICReg [45] stabilizes by adding an intra-batch decorrelation loss function.

Regularized contrastive learning Several works regularize contrastive learning by optimizing a contrastive objective along with an objective that considers the similarities among instances. CO2 [24] adds a consistency regularization term that matches the distribution of similarity for a query and

its positive. PCL [46] and WCL [47] combines unsupervised clustering with contrastive learning to tighten representations of similar instances.

Relational learning and knowledge distillation Contrastive learning implicitly learns the relations, also called semantic similarity, between instances based on the meaning or semantics they convey by optimizing alignment and matching a prior distribution [48, 49]. ReSSL [28] introduces an explicit relational learning objective by maintaining consistency of pairwise similarities between strong and weak augmented views. The pairs of views are not directly aligned which harms the discriminative performance. Other approaches relied on self-supervised knowledge distillation [50–52] for which a student model seeks to predict the distribution of similarities among instances computed by a larger pretrained teacher. As such, in opposition with contrastive and relational learning and therefore our approach, knowledge distillation is not an end-to-end approach and requires a former pretraining.

Masked modeling Masked modeling [53, 54] has shown impressive results in Natural Language Processing tasks using the transformer architecture [55]. More recently, it has been successfully applied to the vision domain thanks to advances on vision transformers [56, 57] which use attentions on tokens made by projecting patches of images in a token space. Specifically designed pretext tasks relying on mask modeling for images have been proposed [58–60]. The general idea of mask modeling is masking a part of the input and predicting the masked parts either at token level or at pixel level. It has shown competitive performance on transformer architectures with contrastive learning.

In our work, we optimize a contrastive learning objective using negatives that alleviate class collision by pulling related instances. We do not use a regularization term but directly optimize a soft contrastive learning objective that leverages the contrastive and relational aspects. As we performed a study using convolutional networks, we did not perform a comparative study with Mask Modeling approaches which rely on transformers that require supplementary computational resources.

2.2 Video self-supervised learning

Video Self-Supervised Learning follows the advances of Image Self-Supervised Learning and often picked ideas from the image modality with adjustment and improvement to make it relevant for videos and make best use of it.

Pretext tasks As for images, in early works several *pretext tasks* have been proposed on videos. Some were directly picked from images such as rotation [61], solving Jigsaw puzzles [62], but others have been designed specifically for videos. These specific pretext-tasks include predicting motion and appearance [63], the shuffling of frame [64, 65] or clip [66, 67] order, predicting the speed of the video [68, 69]. These methods have been replaced over time by more performing approaches that are less limited by a specific pretext task to learn a good representation. Recently, TransRank [5] introduced a new paradigm to perform temporal and spatial pretext tasks prediction on a clip relatively to other transformations to the same clip and showed promising results.

Contrastive learning Video Contrastive Learning [4, 9, 10, 12–15, 70–72] has been widely studied in the recent years as it gained interest after its better performance than standard pretext tasks in images. Several works studied how to form positive views from different clips [4, 10, 12, 13] to directly apply contrastive methods from images. CVRL [12] extended SimCLR to videos and propose a temporal sampler for creating temporally overlapped but not identical positive views which can avoid spatial redundancy. Also, [4] extended SimCLR, MoCo, SwaV and BYOL to videos and studied the effect of using random sampled clips from a video to form views. They pushed further the study to sample several positives to generalize the Multi-crop procedure introduced for images by [1]. Some works focused on combining contrastive learning and predicting a pretext task [73–77, 82]. To help better represent the time dimension, several approaches were designed to use different temporal context width [13–15] for the different views.

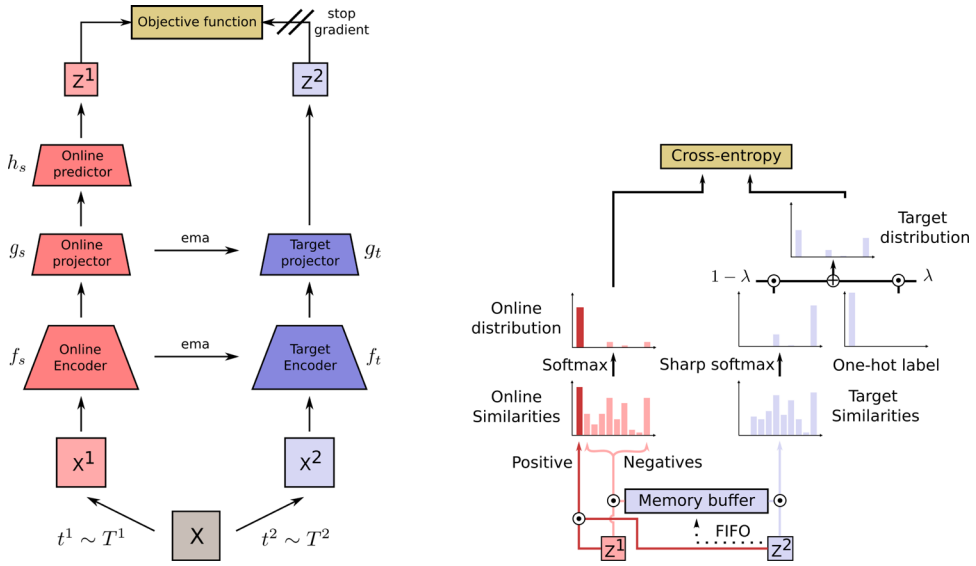
Multi-modal learning To improve self-supervised representation learning, several approaches made use of several modalities to better capture the spatio-temporal information provided by a video. It can be from text [78, 79], audio [14, 73, 80], and optical flow [10, 14, 26, 70, 73, 81, 82].

Masked modeling Transformers have been extended from images to videos for learning spatio-temporal representations [83, 84]. Approaches on videos for Masked Modeling [85–87] essentially converted pretext tasks from images to videos by considering spatio-temporal masking of tokens instead of simply spatial tokens.

In our work, we propose a soft contrastive learning objective using only RGB frames that directly generalizes our approach from image with changes related to data processing and architectures. To the best of our knowledge, we are the first to introduce the concept of soft contrastive learning using relations for video self-supervised representation learning. As for images, we did not perform a thorough comparative study with Mask Modeling as these methods rely on transformers and we worked with convolutional networks.

3 Methodology

In this section, we will introduce our baselines: MoCov2 [29] for the contrastive aspect and ReSSL [28] for the relational aspect. We will then present our self-supervised soft contrastive learning approach called Similarity Contrastive



(a) Siamese pipeline (b) SCE objective function

Fig. 1 SCE follows a siamese pipeline illustrated in **a**. A batch $\mathbf{x}$ of images is augmented with two different data augmentation distributions T^1 and T^2 to form $\mathbf{x^1} = t^1(\mathbf{x})$ and $\mathbf{x^2} = t^2(\mathbf{x})$ with $t^1 \sim T^1$ and $t^2 \sim T^2$. The representation $\mathbf{z^1}$ is computed through an online encoder f_s, projector g_s and optionally a predictor h_s such as $\mathbf{z^1} = h_s(g_s(f_s(\mathbf{x^1})))$. A parallel target branch updated by an exponential moving average of the online branch, or ema, computes $\mathbf{z^2} = g_t(f_t(\mathbf{x^2}))$ with f_t and g_t the target encoder and projector. In the objective function of SCE illustrated in **b**, $\mathbf{z^2}$ is used to compute the inter-instance target distribution by applying a sharp softmax to the cosine similarities between $\mathbf{z^2}$ and a memory buffer of representations from the momentum branch. This distribution is mixed via a $1-\lambda$ factor with a one-hot label factor λ to form the target distribution. Similarities between $\mathbf{z^1}$ and the memory buffer plus its positive in $\mathbf{z^2}$ are also computed. The online distribution is computed via softmax applied to the online similarities. The objective function is the cross entropy between the target and the online distributions

Estimation (SCE). All these methods share the same architecture illustrated in Fig. 1a. We provide the pseudo-code of our algorithm in Appendix B.

3.1 Contrastive and relational learning

Consider $\mathbf{x} = \{\mathbf{x_k}\}_{k\in\{1,\dots,N\}}$ a batch of N images. Siamese momentum methods based on Contrastive and Relational learning, such as MoCo [7] and ReSSL [28], respectively, produce two views of $\mathbf{x}$, $\mathbf{x^1} = t^1(\mathbf{x})$ and $\mathbf{x^2} = t^2(\mathbf{x})$, from two data augmentation distributions T^1 and T^2 with $t^1 \sim T^1$ and $t^2 \sim T^2$. For ReSSL, T^2 is a weak data augmentation distribution compared to T^1 to maintain relations. $\mathbf{x^1}$ passes through an online network f_s followed by a projector g_s to compute $\mathbf{z^1} = g_s(f_s(\mathbf{x^1}))$. A parallel target branch containing a projector g_t and an encoder f_t computes $\mathbf{z^2} = g_t(f_t(\mathbf{x^2}))$. $\mathbf{z^1}$ and $\mathbf{z^2}$ are both l_2-normalized.

The online branch parameters θ_s are updated by gradient (∇) descent to minimize a loss function $\mathcal{L}$. The target branch parameters θ_t are updated at each iteration by exponential moving average of the online branch parameters with the *momentum value* m, also called *keep rate*, to control the update such as:

$$\theta_s \leftarrow optimizer(\theta_s, \nabla_{\theta_s}\mathcal{L}), \tag{1}$$

$$\theta_t \leftarrow m\theta_t + (1-m)\theta_s. \tag{2}$$

MoCo uses the InfoNCE loss, a similarity-based function scaled by the temperature τ that maximizes agreement between the positive pair and push negatives away:

$$L_{InfoNCE} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(\mathbf{z_i^1}\cdot\mathbf{z_i^2}/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{z_i^1}\cdot\mathbf{z_j^2}/\tau)}\right). \tag{3}$$

ReSSL computes a target similarity distribution $\mathbf{s^2}$ that represents the relations between weak augmented instances, and the distribution of similarity $\mathbf{s^1}$ between the strongly augmented instances with the weak augmented ones. Temperature parameters are applied to each distribution: τ for $\mathbf{s^1}$ and τ_m for $\mathbf{s^2}$ with $\tau > \tau_m$ to eliminate noisy relations. Indeed, as the temperature decreases, it exponentially increases softmax values for highly similar instances and decreases exponentially values for low similar instances which makes them negligible in the target distribution. The loss function is the cross-entropy between $\mathbf{s}^2$ and $\mathbf{s}^1$:

$$s_{ik}^1 = \frac{\mathbb{1}_{i\neq k}\cdot\exp(\mathbf{z_i^1}\cdot\mathbf{z_k^2}/\tau)}{\sum_{j=1}^{N}\mathbb{1}_{i\neq j}\cdot\exp(\mathbf{z_i^1}\cdot\mathbf{z_j^2}/\tau)}, \tag{4}$$

$$s_{ik}^2 = \frac{\mathbb{1}_{i\neq k} \cdot \exp(\mathbf{z_i^2} \cdot \mathbf{z_k^2}/\tau_m)}{\sum_{j=1}^{N} \mathbb{1}_{i\neq j} \cdot \exp(\mathbf{z_i^2} \cdot \mathbf{z_j^2}/\tau_m)}, \tag{5}$$

$$L_{ReSSL} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{k=1 \\ k\neq i}}^{N} s_{ik}^2 \log\left(s_{ik}^1\right). \tag{6}$$

A memory buffer of size $M >> N$ filled by $\mathbf{z^2}$ is maintained for both methods.

3.2 Similarity contrastive estimation

Contrastive Learning methods damage relations among instances which Relational Learning correctly build. However, Relational Learning lacks the discriminating features that contrastive methods can learn. If we take the example of a dataset composed of cats and dogs, we want our model to be able to understand that two different cats share the same appearance, but we also want our model to learn to distinguish details specific to each cat. Based on these requirements, we propose our approach called Similarity Contrastive Estimation (SCE).

We argue that there exists a true distribution of similarity $\mathbf{w_i^*}$ between a query $\mathbf{q_i}$ and the instances in a batch of N images $\mathbf{x} = \{\mathbf{x_k}\}_{k\in\{1,\dots,N\}}$, with $\mathbf{x_i}$ a positive view of $\mathbf{q_i}$. If we had access to $\mathbf{w_i^*}$, our training framework would estimate the similarity distribution $\mathbf{p_i}$ between $\mathbf{q_i}$ and all instances in $\mathbf{x}$, and minimize the cross-entropy between $\mathbf{w_i^*}$ and $\mathbf{p_i}$ which is a soft contrastive learning objective:

$$L_{SCE^*} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} w_{ik}^* \log\left(p_{ik}\right). \tag{7}$$

L_{SCE^*} is a soft contrastive approach that generalizes InfoNCE and ReSSL objectives. InfoNCE is a hard contrastive loss that estimates $\mathbf{w_i^*}$ with a one-hot label and ReSSL estimates $\mathbf{w_i^*}$ without the contrastive component.

We propose an estimation of $\mathbf{w_i^*}$ based on contrastive and relational learning. We consider $\mathbf{x^1} = t^1(\mathbf{x})$ and $\mathbf{x^2} = t^2(\mathbf{x})$ generated from $\mathbf{x}$ using two data augmentations $t^1 \sim T^1$ and $t^2 \sim T^2$. Both augmentation distributions should be different to estimate different relations for each view as shown in Sect. 4.1.1. We compute $\mathbf{z^1} = h_s(g_s(f_s(\mathbf{x^1})))$ from the online encoder f_s, projector g_s and optionally a predictor h_s [2, 41]). We also compute $\mathbf{z^2} = g_t(f_t(\mathbf{x^2}))$ from the target encoder f_t and projector g_t. $\mathbf{z^1}$ and $\mathbf{z^2}$ are both l_2-normalized.

The similarity distribution $\mathbf{s_i^2}$ that defines relations between the query and other instances is computed via Eq. (5). The temperature τ_m sharpens the distribution to only keep relevant relations. A weighted positive one-hot label is added to $\mathbf{s_i^2}$ to build the target similarity distribution $\mathbf{w_i^2}$:

$$w_{ik}^2 = \lambda \cdot \mathbb{1}_{i=k} + (1-\lambda) \cdot s_{ik}^2. \tag{8}$$

The online similarity distribution $\mathbf{p_i^1}$ between $\mathbf{z_i^1}$ and $\mathbf{z^2}$, including the target positive representation in opposition with ReSSL, is computed and scaled by the temperature τ with $\tau > \tau_m$ to build a sharper target distribution:

$$p_{ik}^1 = \frac{\exp(\mathbf{z_i^1} \cdot \mathbf{z_k^2}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}. \tag{9}$$

The objective function illustrated in Fig. 1b is the cross-entropy between each $\mathbf{w^2}$ and $\mathbf{p^1}$:

$$L_{SCE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{N} w_{ik}^2 \log\left(p_{ik}^1\right). \tag{10}$$

The loss can be symmetrized by passing $\mathbf{x^1}$ and $\mathbf{x^2}$ through the momentum and online encoders and averaging the two losses computed.

A memory buffer of size $M >> N$ filled by $\mathbf{z^2}$ is maintained to better approximate the similarity distributions.

The following proposition explicitly shows that SCE optimizes a contrastive learning objective while maintaining inter-instance relations:

Proposition 1 *L_{SCE} defined in Eq. (10) can be written as:*

$$L_{SCE} = \lambda \cdot L_{InfoNCE} + \mu \cdot L_{ReSSL} + \eta \cdot L_{Ceil}, \tag{11}$$

with $\mu = \eta = 1 - \lambda$ and

$$L_{Ceil} = -\frac{1}{N} \sum_{i=1}^{N} \log\left(\frac{\sum_{j=1}^{N} \mathbb{1}_{i\neq j} \cdot \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{z_i^1} \cdot \mathbf{z_j^2}/\tau)}\right).$$

The proof separates the positive term and negatives. It can be found in Appendix C. L_{Ceil} leverages how similar the positives should be with hard negatives. Because our approach is a soft contrastive learning objective, we optimize the formulation in Eq. (10) and have the constraint $\mu = \eta = 1 - \lambda$. It frees our implementation from having three losses to optimize with two hyperparameters μ and η to tune. Still, we performed a small study of the objective defined in Eq. (11) without this constraint to check if L_{Ceil} improves results in Sect. 4.1.1.

4 Empirical study

In this section, we will empirically prove the relevance of our proposed Similarity Contrastive Estimation (SCE) self-supervised learning approach to learn a good data representation for both images and videos representation learning.

4.1 Image study

In this section, we first make an ablative study of our approach SCE to find the best hyperparameters on images. Secondly, we compare SCE with its baselines MoCov2 [29] and ReSSL [28] for the same architecture. Finally, we evaluate SCE on the ImageNet Linear evaluation protocol and assess its generalization capacity on various tasks.

4.1.1 Ablation study

To make the ablation study, we conducted experiments on ImageNet100 that has a close distribution to ImageNet, studied in Sect. 4.1.3, with the advantage to require less resources to train. We keep implementation details close to ReSSL [28] and MoCov2 [29] to ensure fair comparison.

Dataset ImageNet [88] is a large dataset with 1k classes, almost 1.3M images in the training set and 50K images in the validation set. ImageNet100 is a selection of 100 classes from ImageNet whose classes have been selected randomly. We took the selected classes from [37] referenced in Appendix A.

Implementation details for pretraining We use the ResNet-50 [89] encoder and pretrain for 200 epochs. We apply by default *strong* and *weak* data augmentations defined in Table 1. We do not use a predictor, and we do not symmetry the loss by default. Specific hyper-parameter details can be found in Appendix D.1.

Evaluation protocol To evaluate our pretrained encoders, we train a linear classifier following Chen et al. [29] and Zheng et al. [28] that is detailed in Appendix D.1. **Leveraging contrastive and relational learning** SCE defined in Eq. (8) leverages contrastive and relational learning via the λ coefficient. We studied the effect of varying the λ coefficient on ImageNet100. Temperature parameters are set to $\tau = 0.1$ and $\tau_m = 0.05$. We report the results in Table 2. Performance increases with λ from 0 to 0.5 after which it starts decreasing. The best λ is inside $[0.4, 0.5]$, confirming that balancing the contrastive and relational aspects provides better representation. In next experiments, we keep $\lambda = 0.5$.

We performed a small study of the optimization of Eq. (11) by removing L_{ceil} ($\eta = 0$) to validate the relevance of our approach for $\tau = 0.1$ and $\tau_m \in \{0.05, 0.07\}$. The results are reported in Table 3. Adding the term L_{ceil} consistently improves performance, empirically proving that our approach is better than simply adding $L_{InfoNCE}$ and L_{ReSSL}. This performance boost varies with temperature parameters, and our best setting improves by $+0.9$ percentage points (p.p.) in comparison with adding the two losses.

Asymmetric data augmentations to build the similarity distributions Contrastive learning approaches use strong data augmentations [8] to learn view invariant features and prevent the model to collapse. However, these strong data augmentations shift the distribution of similarities among instances that SCE uses to approximate w_i^* in Eq. (8). We need to carefully tune the data augmentations to estimate a relevant target similarity distribution. We listed different distributions of data augmentations in Table 1. The *weak* and *strong* augmentations are the same as described by ReSSL

Table 1 Different distributions of data augmentations applied to SCE

Parameter	Weak	Strong	Strong-α	Strong-β	Strong-γ
Random crop probability	1	1	1	1	1
Flip probability	0.5	0.5	0.5	0.5	0.5
Color jittering probability	0	0.8	0.8	0.8	0.8
Brightness adjustment max intensity	–	0.4	0.4	0.4	0.4
Contrast adjustment max intensity	–	0.4	0.4	0.4	0.4
Saturation adjustment max intensity	–	0.4	0.2	0.2	0.2
Hue adjustment max intensity	–	0.1	0.1	0.1	0.1
Color dropping probability	0	0.2	0.2	0.2	0.2
Gaussian blurring probability	0	0.5	1	0.1	0.5
Solarization probability	0	0	0	0.2	0.2

The *weak* distribution is the same as ReSSL [28], and *strong* is the standard contrastive data augmentation [8]. The *strong-*α and *strong-*β are two distributions introduced by BYOL [2]. Finally, *strong-*γ is a mix between *strong-*α and *strong-*β

Table 2 Effect of varying λ on the Top-1 accuracy on ImageNet100

λ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Top-1	81.5	81.8	82.5	82.8	**82.9**	**82.9**	82.2	81.6	81.8	81.8	81.1

The optimal λ is in $[0.4, 0.5]$, confirming that learning to discriminate and maintaining relations is best. Results style: **best**, second best

Table 3 Effect of loss coefficients in Eq. (11) on the Top-1 accuracy on ImageNet100

Method	Loss coefficients			Top-1	
	λ	μ	η	$\tau_m = 0.05$	$\tau_m = 0.07$
InfoNCE	1	0	0	81.1	81.1
	0.5	0.5	0	82.8	82.5
SCE	0.5	0.5	0.5	**82.9**	**83.4**
ReSSL	0	1	0	80.8	78.4
	0	1	1	81.5	79.6

L_{Ceil} consistently improves performance that varies given the temperature parameters. Results style: **best**, second best

Table 4 Effect of using different distributions of data augmentations for the two views and of the loss symmetrization on the Top-1 accuracy on ImageNet100

Online Aug	Teacher Aug	Sym	Top-1
Strong	*Weak*	No	82.9
Strong-γ	*Weak*	No	**83.0**
Weak	*Strong*	No	73.4
Strong	*Strong*	No	80.5
Strong-α	*Strong-*β	No	80.7
Strong	*Weak*	Yes	83.7
Strong	*Strong*	Yes	83.0
Strong-α	*Strong-*β	yes	**84.2**

Using a *weak* view for the teacher without symmetry is necessary to obtain good relations. With loss symmetry, asymmetric data augmentations improve the results, with the best obtained using strong-α and *strong-*β. Results style: **best**, second best

[28]. *strong-*α and *strong-*β have been proposed by BYOL [2]. *strong-*γ combines *strong-*α and *strong-*β.

We performed a study in Table 4 on which data augmentations are needed to build a proper target distribution for the non-symmetric and symmetric settings. We report the Top-1 accuracy on Imagenet100 when varying the data augmentations applied on the online and target branches of our pipeline. For the non-symmetric setting, SCE requires the target distribution to be built from a *weak* augmentation distribution that maintains consistency across instances.

Once the loss is symmetrized, asymmetry with strong data augmentations has better performance. Indeed, using *strong-*α and *strong-*β augmentations is better than using *weak* and *strong* augmentations, and same *strong* augmentations have lower performance. We argue symmetrized SCE requires asymmetric data augmentations to produce different relations for each view to make the model learn more information. The effect of using stronger augmentations is balanced by averaging the results on both views. Symmetrizing the loss boosts the performance as for [2, 27].

Sharpening the similarity distributions The temperature parameters sharpen the distributions of similarity

Table 5 Effect of varying the temperature parameters τ_m and τ on the Top-1 accuracy on ImageNet100

$\tau = 0.1$		$\tau = 0.2$	
τ_m	Top-1	τ_m	Top-1
0.03	82.3	0.03	**81.3**
0.04	82.5	0.04	81.2
0.05	82.9	0.05	81.2
0.06	82.5	0.06	81.2
0.07	**83.4**	0.07	81.1
0.08	82.7	0.08	80.9
0.09	82.5	0.09	81.2
0.10	82.1	0.10	81.2

τ_m is lower than τ to produce a sharper target distribution without noisy relations. SCE does not collapse when $\tau_m \to \tau$. Results style: **best**, second best

exponentially. SCE uses the temperatures τ_m and τ for the target and online similarity distributions with $\tau_m < \tau$ to guide the online encoder with a sharper target distribution. We made a temperature search on ImageNet100 by varying τ in $\{0.1, 0.2\}$ and τ_m in $\{0.03, ..., 0.10\}$. The results are in Table 5. We found the best values $\tau_m = 0.07$ and $\tau = 0.1$ proving SCE needs a sharper target distribution. In Appendix E, this parameter search is done for other datasets used in comparison with our baselines. Unlike ReSSL [28], SCE does not collapse when $\tau_m \to \tau$ thanks to the contrastive aspect. Hence, it is less sensitive to the temperature choice.

4.1.2 Comparison with our baselines

We compared on 6 datasets how SCE performs against its baselines. We keep similar implementation details to ReSSL [28] and MoCov2 [29] for fair comparison.

Small datasets Cifar10 and Cifar100 [90] have 50K training images, 10K test images, 32×32 resolution and 10–100 classes, respectively.

Medium datasets STL10 [91] has a 96×96 resolution, 10 classes, 100K unlabeled data, 5k labeled training images and 8K test images. Tiny-Imagenet [92] is a subset of ImageNet with 64×64 resolution, 200 classes, 100k training images and 10K validation images.

Implementation details Architecture implementation details can be found in Appendix D.1. For MoCov2, we use $\tau = 0.2$ and for ReSSL their best τ and τ_m reported [28]. For SCE, we use the best temperature parameters from Sect. 4.1.1 for ImageNet and ImageNet100 and from Appendix E for the other datasets. The same architecture for all methods is used except for MoCov2 on ImageNet that kept the ImageNet100 projector to improve results.

Results are reported in Table 6. Our baselines reproduction is validated as results are better than those reported by

Table 6 Comparison of SCE with its baselines MoCov2 [29] and ReSSL [28] on the Top-1 Accuracy on various datasets

Method	ImageNet	ImageNet100	Cifar10	Cifar100	STL10	Tiny-IN
MoCov2 [29]	67.5	–	–	–	–	–
MoCov2 [*]	68.8	80.5	87.6	61.0	86.5	45.9
ReSSL [28]	69.9	–	90.2	63.8	88.3	46.6
ReSSL*	70.2	81.6	90.2	64.0	89.1	49.5
SCE (Ours)	**70.5**	**83.4**	**90.3**	**65.5**	**89.9**	**51.9**

SCE outperforms on all benchmarks its baselines. Results style: **best**, second best
*Denotes our reproduction

the authors. SCE outperforms its baselines on all datasets, proving that our method is more efficient to learn discriminating features on the pretrained dataset. We observe that our approach outperforms more significantly ReSSL on smaller datasets than ImageNet, suggesting that it is more important to learn to discriminate among instances for these datasets. SCE has promising applications to domains with few data such as in medical applications.

4.1.3 ImageNet linear evaluation

We compare SCE on the widely used ImageNet linear evaluation protocol with the state of the art. We scaled our method using a larger batch size and a predictor to match state-of-the-art results [2, 41].

Implementation details We use the ResNet-50 [89] encoder, apply *strong-α* and *strong-β* augmentations defined in Table 1. We follow the same training hyperparameters used by [41] and detailed in Appendix D.2. The loss is symmetrized and we keep the best hyperparameters from Sect. 4.1.1: $\lambda = 0.5$, $\tau = 0.1$ and $\tau_m = 0.07$.

Multi-crop setting We follow [21] setting and sample 6 different views detailed in Appendix D.2.

Evaluation protocol We follow the protocol defined by Chen et al. [41] and detailed in Appendix D.2.

We evaluated SCE at epochs 100, 200, 300 and 1000 on the Top-1 accuracy on ImageNet to study the efficiency of our approach and compare it with the state of the art in Table 7. At 100 epochs, SCE reaches **72.1**% up to **74.1**% at 1000 epochs. Hence, SCE has a fast convergence and few epochs of training already provides a good representation. SCE is the Top-1 method at 100 epochs and Top-2 for 200 and 300 epochs proving the good quality of its representation for few epochs of pretraining.

At 1000 epochs, SCE is below several state-of-the-art results. We argue that SCE suffers from maintaining a λ coefficient to 0.5 and that relational or contrastive aspects do not have the same impact at the beginning and at the end of pretraining. A potential improvement would be using a scheduler on λ that varies over time.

We added multi-crop to SCE for 200 epochs of pretraining. It enhances the results, but it is costly in terms of time and memory. It improves the results from 72.7% to our best result **75.4**% (**+2.7p.p.**). Therefore, SCE learns from having local views and they should maintain relations to learn better representations. We compared SCE with state-of-the-art methods using multi-crop in Table 8. SCE is competitive with top state-of-the-art methods that trained for 800+ epochs by having slightly lower accuracy than the best method using multi-crop (−0.3p.p) and without multi-crop (−0.5p.p). SCE is more efficient than other methods, as it reaches state-of-the-art results for fewer pretraining epochs.

Table 7 State-of-the-art results on the Top-1 Accuracy on ImageNet under the linear evaluation protocol at different pretraining epochs: 100, 200, 300, 800+

Method	100	200	300	800–1000
SimCLR [8]	66.5	68.3	–	70.4
MoCov2 [27]	67.4	69.9	–	72.2
SwaV [1]	66.5	69.1	–	71.8
BYOL [2]	66.5	70.6	72.5	74.3
Barlow-Twins [44]	–	–	71.4	73.2
AdCo [21]	–	68.6	–	72.8
ReSSL [28]	–	71.4	–	–
WCL [47]	68.1	70.3	–	72.2
VICReg [45]	–	–	–	73.2
UniGrad [93]	70.3	–	–	–
MoCov3 [41]	68.9	–	72.8	74.6
NNCLR [22]	69.4	70.7	–	75.4
Triplet [39]	–	**73.8**	–	**75.9**
SCE (ours)	**72.1**	72.7	**73.3**	74.1

SCE is Top-1 at 100 epochs and Top-2 for 200 and 300 epochs. For 800+ epochs, SCE has lower performance than several state-of the-art methods. Results style: **best**, second best

4.1.4 Transfer learning

We study the generalization of our proposed SCE on several tasks: linear transfer learning (Table 9), low-shot (Table 10), and object detection and instance segmentation (Table 11). We use our multi-crop checkpoint pretrained for 200 epochs on ImageNet.

Table 8 State-of-the-art results on the Top-1 Accuracy on ImageNet under the linear evaluation protocol with multi-crop

Method	Epochs	Top-1
200 epochs		
SwaV [1]	200	72.7
AdCo [21]	200	73.2
WCL [47]	200	73.3
Triplet [39]	200	74.1
ReSSL [28]	200	74.7
SCE (ours)	200	**75.4**
800+ epochs		
WCL [47]	800	74.7
SwaV [1]	800	75.3
DINO [3]	800	75.3
UniGrad [93]	800	75.5
NNCLR [22]	1000	75.6
AdCo [21]	800	**75.7**

SCE is competitive with the best state-of-the-art methods by pretraining for only 200 epochs instead of 800+. Results style: **best**, second best

Table 10 Transfer learning on low-shot image classification on Pascal VOC2007

Method	$K = 16$	$K = 32$	$K = 64$	full
MoCov2 [29]	76.1	79.2	81.5	84.6
PCLv2 [46]	78.3	80.7	82.7	85.4
ReSSL [28]	79.2	82.0	83.8	86.3
SwAV [1]	78.4	81.9	84.4	87.5
WCL [47]	**80.2**	83.0	85.0	87.8
SCE (ours)	79.5	**83.1**	**85.5**	**88.2**

All methods have been pretrained for 200 epochs. SCE is Top-1 when using 32–64-all images per class and Top-2 for 16 images. Results style: **best**, second best

Low-shot evaluation Low-shot transferability of our backbone is evaluated on Pascal VOC2007. We followed the protocol proposed by Zheng et al. [28]. We select 16, 32, 64 or all images per class to train the classifier. Our results are compared with other state-of-the-art methods pretrained for 200 epochs in Table 10. SCE is Top-1 for 32, 64 and all images per class and Top-2 for 16 images per class, proving the generalization of our approach to few-shot learning.

Linear classifier for many-shot recognition datasets We follow the same protocol as Grill et al. [2] and Ericsson et al. [96] to study many-shot recognition in transfer learning on the datasets FGVC Aircraft [97], Caltech-101 [98], Standford Cars [99], CIFAR-10 [90], CIFAR-100 [90], DTD [100], Oxford 102 Flowers [101], Food-101 [102], Oxford-IIIT Pets [103], SUN397 [104] and Pascal VOC2007 [105]. These datasets cover a large variety of number of training images (2–75 k) and number of classes (10–397). We report the Top-1 classification accuracy except for Aircraft, Caltech-101, Pets and Flowers for which we report the mean per-class accuracy and the 11-point MAP for VOC2007.

We report the performance of SCE in comparison with state-of-the-art methods in Table 9. SCE outperforms on 7 datasets all approaches. In average, SCE is above all state-of-the-art methods as well as the supervised baseline, meaning SCE is able to generalize to a wide range of datasets.

Object detection and instance segmentation We performed object detection and instance segmentation on the COCO dataset [94]. We used the pretrained network to initialize a Mask R-CNN [95] up to the C4 layer. We follow the protocol of [39] and report the Average Precision for detection AP^{Box} and instance segmentation AP^{Mask}.

We report our results in Table 11 and observe that SCE is the second best method after Truncated-Triplet [39] on both metrics, by being slightly below their reported results and above the supervised setting. Therefore, our proposed SCE is able to generalize to object detection and instance segmentation task beyond what the supervised pretraining can (**+1.6p.p.** of AP^{Box} and **+1.3p.p.** of AP^{Mask}).

4.2 Video study

In this section, we first make an ablation study of our approach SCE to find the best hyperparameters on videos. Then, we compare SCE to the state of the art after pretraining on Kinetics400 and assess generalization on various tasks.

Table 9 Linear classifier trained on popular many-shot recognition datasets in comparison with SimCLR [8], supervised training, BYOL [2] and NNCLR [22]

Method	Food	CIFAR10	CIFAR100	SUN	Cars	Air	VOC	DTD	Pets	Caltech	Flow	Avg
SimCLR	72.8	90.5	74.4	60.6	49.3	49.8	81.4	75.7	84.6	89.3	92.6	74.6
Supervised	72.3	93.6	78.3	61.9	66.7	61.0	82.8	74.9	91.5	**94.5**	94.7	79.3
BYOL	75.3	91.3	78.4	62.2	**67.8**	60.6	82.5	75.5	90.4	94.2	**96.1**	79.5
NNCLR	76.7	93.7	79.0	62.5	67.1	**64.1**	83.0	75.5	**91.8**	91.3	95.1	80.0
SCE (ours)	**77.7**	**94.8**	**80.4**	**65.3**	65.7	59.6	**84.0**	**77.1**	90.9	92.7	**96.1**	**80.4**

SCE is Top-1 on 7 datasets and in average. Results style: **best**, second best

4.2.1 Ablation study

Pretraining dataset To make the ablation study, we perform pretraining experiments on Mini-Kinetics200 [106], later called Kinetics200 for simplicity. It is a subset of Kinetics400 [107] meaning they have a close distribution with less resources required on Kinetics200 to train. Kinetics400 is composed of 216k videos for training and 18k for validation for 400 action classes. However, it has been created from Youtube and some videos have been deleted. We use the dataset hosted[1] by the CVD foundation.

Evaluation datasets To study the quality of our pretrained representation, we perform linear evaluation classification on the Kinetics200 dataset. Also, we finetune on the first split of the UCF101 [108] and HMDB51 [109] datasets. UCF101 is an action classification dataset that contains 13k3 different videos for 101 classes and has 3 different training and validation splits. HMDB51 is also an action classification dataset that contains 6k7 different videos from 51 classes with 3 different splits.

Pretraining implementation details We used the ResNet3D-18 network [110] following the Slow path of Feichtenhofer et al. [111]. We kept hyperparameters close to the ones used for ImageNet in Sect. 4.1.3. More details can be found in Appendix D.3. We pretrain for 200 epochs with a batch size of 512. The loss is symmetrized. To form two different views from a video, we follow Feichtenhofer et al. [4] and randomly sample two clips from the video that lasts 2.56 seconds and keep only 8 frames.

Linear evaluation and finetuning evaluation protocols We follow Feichtenhofer et al. [4] and details can be found in Appendix D.3. For finetuning on UCF101 and HMDB51, we only use the first split in ablation study.

Baseline and supervised learning We define an SCE baseline which uses the hyperparameters $\lambda = 0.5$, $\tau = 0.1$, $\tau_m = 0.07$. We provide performance of our SCE baseline as well as supervised training in Table 12. We observe that our baseline has lower results than supervised learning with -8.1p.p for Kinetics200, -1.2p.p for UCF101 and -3.1p.p for HMDB51 which shows that our representation has a large margin for improvement.

Leveraging contrastive and relational learning As for the image study, we varied λ from Eq. (8) in the set $\{0, 0.125, ..., 0.875, 1\}$ to observe the effect of leveraging the relational and contrastive aspects and report results in Table 13. Using relations during pretraining improves the results rather than only optimizing a contrastive learning objective. The performance on Kinetics200, UCF101 and HMDB51 consistently increases by decreasing λ from 1 to 0.25. The best λ obtained is 0.125. Moreover, $\lambda = 0$ performs better than $\lambda = 1$. These results suggest that for video pretraining with standard image contrastive learning augmentations, relational learning performs better than contrastive learning and leveraging both further improve the quality of the representation.

Target temperature variation We studied the effect of varying the target temperature with values in the set $\tau_m \in \{0.03, 0.04, ..., 0.08\}$ while maintaining the online temperature $\tau = 0.1$. We report results in Table 14. We observe that

Table 11 Object detection and Instance Segmentation on COCO [94] training a Mask R-CNN [95]

Method	AP^{Box}	AP^{Mask}
Random	35.6	31.4
Supervised	40.0	34.7
Rel-Loc [31]	40.0	35.0
Rot-Pred [35]	40.0	34.9
NPID [17]	39.4	34.5
MoCo [7]	40.9	35.5
MoCov2 [29]	40.9	35.5
SimCLR [8]	39.6	34.6
BYOL [2]	40.3	35.1
SCE (ours)	<u>41.6</u>	<u>36.0</u>
Triplet [39]	**41.7**	**36.2**

SCE is Top-2 on both tasks, slightly below Truncated-Triplet [39] and better than supervised training. Results style: **best**, <u>second best</u>

Table 12 Comparison of our baseline and supervised training on the Kinetics200, UCF101 and HMDB51 Top-1 accuracy

Method	K200	UCF101	HMDB51
SCE baseline	63.9	86.3	57.0
Supervised	**72.0**	**87.5**	**60.1**

Supervised training is consistently better

Table 13 Effect of varying λ on the Kinetics200, UCF101 and HMDB51 Top-1 accuracy

λ	K200	UCF101	HMDB51
0.000	64.2	86.2	<u>57.5</u>
0.125	**64.8**	**86.9**	**58.2**
0.250	64.3	<u>86.7</u>	**58.2**
0.375	<u>64.7</u>	86.3	56.8
0.500	63.9	86.3	57.0
0.625	63.4	86.2	55.7
0.750	63.1	85.8	56.2
0.875	62.1	85.7	55.3
1.000	61.9	85.0	55.4

The best λ is 0.125 meaning contrastive and relational leverage increases performance. Results style: **best**, <u>second best</u>

[1] Link to the Kinetics400 dataset hosted by the CVD foundation: https://github.com/cvdfoundation/kinetics-dataset.

Table 14 Effect of varying τ_m on the Top-1 accuracy on Kinetics200, UCF101 and HMDB51 while maintaining $\tau = 0.1$

τ_m	K200	UCF101	HMDB51
0.03	63.4	86.1	56.9
0.04	63.8	**86.6**	56.6
0.05	**64.3**	86.4	**57.1**
0.06	64.1	86.2	56.4
0.07	63.9	86.3	57.0
0.08	63.8	85.9	55.8

The best τ_m is 0.05 meaning that a sharper target distribution is required. Results style: **best**, second best

Table 15 Effect of strength for color jittering for *strong-*α and *strong-*β augmentations on the Kinetics200, UCF101 and HMDB51 Top-1 accuracy

strength	K200	UCF101	HMDB51
0.50	63.9	86.3	57.0
0.75	64.6	86.8	57.8
1.00	**64.8**	**87.0**	**58.1**

Strong color jittering improves performance. Results style: **best**, second best

the best temperature is $\tau_m = 0.05$, indicating that a sharper target distribution is required for video pretraining. We also observe that varying τ_m has a lower impact on performance than varying λ.

Spatial and temporal augmentations We tested varying and adding some data augmentations that generates the pairs of views. As we are dealing with videos, these augmentations can be either spatial or temporal. We define the *jitter* augmentation that jitters by a factor the duration of a clip, *reverse* that randomly reverses the order of frames and *diff* that randomly applies RGB difference on the frames. RGB difference consists in converting the frames to grayscale and subtracting them over time to approximate the magnitude of optical flow. In this work, we consider RGB difference as a data augmentation that is randomly applied during pretraining. In the literature, it is often used as a modality to provide better representation quality than RGB frames [5, 61, 70]. Here, we only apply it during pretraining as a random augmentation. Evaluation only sees RGB frames.

We tested to increase the color jittering strength in Table 15. Using a strength of 1.0 improved our performance on all the benchmarks, suggesting that video pretraining requires harder spatial augmentations than images.

We tested our defined temporal augmentations with *jitter* of factor 0.2, meaning sampling clips between 0.80×2.56 and 1.20×2.56 seconds, randomly applying $reverse$ with 0.2 probability and randomly applying *diff* with 0.2 or 0.5 probability. We report results in Table 16. Varying the clip duration had no noticeable impact on our benchmarks, but reversing the order of frames decreased the performance on UCF101 and HMDB51. This can be explained by the fact that this augmentation can prevent the model to correctly represent the arrow of time. Finally, applying *diff* with 0.2 probability considerably improved our performance over our baseline with **+1.5p.p.** on Kinetics200, **+2.0p.p.** on UCF101 and **+4.4p.p.** on HMDB51. It outperforms supervised learning for generalization with **+0.8p.p.** on UCF101 and **+1.3p.p.** on HMDB51. Applying more often *diff* decreases performance. These results show that SCE benefits from using views that are more biased towards motion than appearance. We believe that it is particularly efficient to model relations based on motion.

Table 16 Effect of using the temporal augmentations by applying clip duration jittering *jitter*, randomly reversing the order of frames *reverse* or randomly using RGB difference *diff* on the Kinetics200, UCF101 and HMDB51 Top-1 accuracy

Jitter	Reverse	Diff	K200	UCF101	HMDB51
0.0	0.0	0.0	63.9	86.3	57.0
0.2	0.0	0.0	64.2	86.4	56.9
0.0	0.2	0.0	64.0	85.7	55.4
0.0	0.0	0.2	65.4	**88.3**	**61.4**
0.0	0.0	0.5	64.1	87.7	60.8
Supervised			**72.0**	87.5	60.1

The *diff* augmentation consistently improves results on the three benchmarks and outperforms supervised pretraining. The other augmentations unchange or decrease performance in average. Results style: **best**, second best

Bringing all together We studied varying one hyperparameter from our baseline and how it affects performance. In this final study, we combined our baseline with the different best hyperparameters found which are $\lambda = 0.125$, $\tau_m = 0.05$, color strength $= 1.0$ and applying *diff* with 0.2 probability. We report results in Table 17 and found out that using harder augmentations increased the optimal λ value as using $\lambda = 0.5$ performs better than $\lambda = 0.125$. This indicates that relational learning by itself cannot learn a better representation through positive views that share less mutual information. The contrastive aspect of our approach is proven efficient for such harder positives. We take as best configuration $\lambda = 0.5$, $\tau_m = 0.05$, *diff* applied with probability 0.2 and color strength $= 1.0$ as it provides best or second best results for all our benchmarks. It improves our baseline by **+2.1p.p.** on Kinetics200 and UCF101, and **+5.0p.p.** on HMDB51. It outperforms our supervised baseline by **+0.9p.p.** on UCF101 and **+1.9p.p.** on HMDB51.

Table 17 Effect of combining best hyper-parameters found in the ablation study which are $\lambda = 0.125$, $\tau_m = 0.05$, *color strength*$= 1.0$ and adding randomly time difference on the Kinetics200, UCF101 and HMDB51 Top-1 accuracy

λ	τ_m	Diff	Strength	K200	UCF101	HMDB51
0.125	0.05	0.2	1.0	65.0	87.4	<u>61.1</u>
0.125	0.07	0.2	1.0	64.7	88.2	60.6
0.500	0.05	0.2	1.0	<u>66.0</u>	<u>88.4</u>	**62.0**
0.500	0.07	0.2	1.0	65.4	**88.6**	61.0
SCE Baseline				63.9	86.3	57.0
Supervised				**72.0**	87.5	60.1

Using time difference and stronger color jittering increases the optimal λ value which indicates contrastive learning is efficient to deal with harder views and helps relational learning. The best value $\tau_m = 0.05$ performs favorably for Kinetics200 and HMDB51. Results style: **best**, <u>second best</u>

4.2.2 Comparison with the state of the art

Pretraining dataset To compare SCE with the state of the art, we perform pretraining on Kinetics400 [107] introduced in Sect. 4.2.1.

Evaluation datasets UCF101 [108] and HMDB51 [109] have been introduced in Sect. 4.2.1.

AVA (v2.2) [112] is a dataset used for spatiotemporal localization of humans actions composed of 211k training videos and 57k validation videos for 60 different classes. Bounding box annotations are used as targets, and we report the mean Average Precision (mAP) for evaluation.

Something-Something V2 (SSv2) [113] is a dataset composed of human-object interactions for 174 different classes. It contains 169k training and 25k validation videos.

Pretraining implementation details We use the ResNet3D-18 and ResNet3D-50 network [110] and more specifically the slow path of Feichtenhofer et al. [111]. We kept the best hyperparameters from Sect. 4.2.1 which are $\lambda = 0.5$, $\tau_m = 0.05$, RGB difference with probability of 0.2, and color strength $= 1.0$ on top of the $strong - \alpha$ and $strong - \beta$ augmentations. From the randomly sampled clips, we specify if we keep 8 or 16 frames.

Action recognition We compare SCE on the linear evaluation protocol on Kinetics400 and finetuning on UCF101 and HMDB51. We kept the same implementation details as in Sect. 4.2.1. We compare our results with the state of the art in Table 18 on various architectures. To propose a fair comparison, we indicate for each approach the pretraining dataset, the number of frames and resolution used during pre-training as well as during evaluation. For the unknown parameters, we leave the cell empty. We compared with some approaches that used the other visual modalities Optical Flow and RGB difference and the different convolutional backbones S3D [116] and R(2+1)D-18 [117].

On ResNet3D-18 even when comparing with methods using several modalities, by using 8×224^2 frames we obtain state-of-the-art results on the three benchmarks with **59.8**% accuracy on Kinetics400, **90.9**% on UCF101, **65.7**% on HMDB51. Using 16×112^2 frames, which is commonly used with this network, improved by +0.9p.p on HMDB51 and decreased by −3.2p.p on kinetics400 and −1.8 on UCF101 and keep state-of-the-art results on all benchmarks, except on UCF101 with −0.5p.p compared with Duan et al. [5] using RGB and RGB difference modalities.

On ResNet3D-50, we obtain state-of-the-art results using 16×224^2 frames on HMDB51 with **74.7**% accuracy even when comparing with methods using several modalities. On UCF101, with **95.3**% SCE is on par with the state of the art, −0.2p.p. than Feichtenhofer et al. [4], but on Kinetics400 −1.9p.p for **69.6**%. We have the same computational budget as they use 4 views for pretraining. Using 8 frames decreased performance by −2.0p.p., −1.2p.p. and −4.2p.p on Kinetics400,UCF101 and HMDB51. It maintains results that outperform on the three benchmarks ρMoCo and ρBYOL with 2 views. It suggests that SCE is more efficient with fewer resources than these methods. By comparing our best with approaches on the S3D backbone that better fit smaller datasets, SCE has slightly lower performance than the state of the art: −1.0p.p. on UCF101 and −0.3p.p. on HMDB51.

Video retrieval We performed video retrieval on our pretrained backbones on the first split of UCF101 and HMDB51. To perform this task, we extract from the training and testing splits the features using the 30-crops procedure as for action recognition, detailed in Appendix D.3. We query for each video in the testing split the N nearest neighbors ($N \in \{1, 5, 10\}$) in the training split using cosine similarities. We report the recall R@N for the different N in Table 19.

We compare our results with the state of the art on ResNet3D-18. Our proposed SCE with 16×112^2 frames is Top-1 on UCF101 with **74.5**%, **85.6**% and **90.5**% for R@1, R@5 and R@10. Using 8×224^2 frames slightly decreases results that are still state of the art. On HMDB51, SCE with 8×224^2 frames outperforms the state of the art with **40.1**%, **63.3**% and **75.4**% for R@1, R@5 and R@10. Using 16×112^2 frames decreased results that are competitive with the previous state-of-the-art approach [114] for −2.3p.p., +1.5p.p. and −1.4p.p. on R@1, R@5 and R@10.

We provide results using the larger architecture ResNet3d-50 which increases our performance on both benchmarks and outperforms the state of the art on all metrics to reach **83.9**%, **92.2**% and **94.9**% for R@1, R@5 and R@10 on UCF101 as well as **45.9**%, **69.9**% and **80.5**% for R@1, R@5 and R@10 on HMDB51. Our soft contrastive learning approach makes our representation learn features that cluster similar instances even for generalization.

Table 18 Performance of SCE for the linear evaluation protocol on Kinetics400 and finetuning on the three splits of UCF101 and HMDB51 (color figure online)

Method	T_p	Res_p	T_e	Res_e	Modality	Pretrain	K400	UCF101	HMDB51
Backbone: S3D / S3D-G									
SpeedNet [68]	64	-	16	224^2	R	K400	-	81.1	48.8
CoCLR [81]	32	128^2	32	128^2	R	K400	-	87.9	54.6
CoCLR [81]	32	128^2	32	128^2	R+F	K400	-	90.6	62.9
TEC [77]	32	128^2	32	128^2	R	K400	-	86.9	63.5
ρBYOL ($\rho = 4$) [4]	32	224^2	32	256^2	R	K400	-	**96.3**	**75.0**
Backbone: R(2+1)D-18									
VideoMoCo [13]	32	112^2	-	-	R	K400	-	78.7	49.2
RSPNet [75]	16	112^2	16	224^2	R	K400	-	81.1	44.6
TransRank [5]	16	112^2	-	-	R	K200	-	87.8	60.1
TransRank [5]	16	112^2	-	-	R+RD	K200	-	90.7	64.2
TEC [77]	16	112^2	16	112^2	R	K400	-	88.2	62.2
ρBYOL ($\rho = 4$) [4]	32	224^2	32	256^2	R	K400	-	**94.4**	**72.2**
Backbone: ResNet3D-18									
ST-Puzzle [62]	16	-	16	112^2	R	K400	-	65.8	33.7
3D-RotNet [61]	16	112^2	-	-	R	K400	-	66.0	37.1
3D-RotNet [61]	16	112^2	-	-	R+D	K400	-	76.7	47.0
VTHCL [9]	8	224^2	8	224^2	R	K400	-	80.6	48.6
TransRank [5]	16	112^2	-	-	R	K200	-	85.7	58.1
TransRank [5]	16	112^2	-	-	R+RD	UCF101	-	88.5	63.0
TransRank [5]	16	112^2	-	-	R+RD	K200	-	89.6	63.5
TEC [77]	16	128^2	16	128^2	R	K400	-	87.1	63.6
ProViCo [114]	16	112^2	-	-	R	K400	-	87.2	59.4
ρMoCo ($\rho = 2$) [4]	8	224^2	8	256^2	R	K400	56.2	87.1	-
SCE (Ours)	8	224^2	8	256^2	R	K200	–	88.4	62.0
SCE (Ours)	16	112^2	16	128^2	R	K400	56.6	89.1	**66.6**
SCE (Ours)	8	224^2	8	256^2	R	K400	**59.8**	**90.9**	65.7
Backbone: ResNet3D-50									
VTHCL [9]	8	224^2	8	224^2	R	K400	–	82.1	49.2
CATE [72]	8	224^2	32	256^2	R	K400	–	88.4	61.9
CVRL [12]	16	224^2	32	256^2	R	K400	66.1	92.2	66.7
CVRL [12]	16	224^2	32	256^2	R	K600	70.4	93.4	68.0
$CORP_f$ [82]	16	224^2	32	256^2	R+F	K400	66.6	93.5	68.0
ConST-CL [115]	16	224^2	32	256^2	R	K400	66.6	94.8	71.9
BraVe [14]	16	224^2	32	224^2	R	K400	-	93.7	72.0
BraVe [14]	16	224^2	32	224^2	R+F	K400	-	94.7	72.7
BraVe [14]	16	224^2	32	224^2	R	K600	-	94.1	74.0
BraVe [14]	16	224^2	32	224^2	R+F	K600	-	95.1	74.3
ρMoCo ($\rho = 2$) [4]	8	224^2	8	256^2	R	K400	65.8	91.0	-
ρMoCo ($\rho = 2$) [4]	16	224^2	16	256^2	R	K400	67.6	93.3	-
ρBYOL ($\rho = 2$) [4]	8	224^2	8	256^2	R	K400	65.8	92.7	-
ρBYOL ($\rho = 4$) [4]	8	224^2	8	256^2	R	K400	70.0	94.2	72.1
ρBYOL ($\rho = 4$) [4]	8	224^2	16	256^2	R	K400	**71.5**	**95.5**	73.6
SCE (Ours)	8	224^2	8	256^2	R	K400	67.6	94.1	70.5
SCE (Ours)	16	224^2	16	256^2	R	K400	69.6	95.3	**74.7**

Res_p, Res_e means the resolution for pretraining and evaluation. T_p, T_e means the number of frames used for pretraining and evaluation. For *Modality*, "R" means RGB, "F" means Optical Flow, "RD" means RGB difference. Best viewed in color, gray rows highlight multi-modal trainings and green rows our results. SCE obtains state-of-the-art results on ResNet3D-18 and on the finetuning protocol for ResNet3D-50. Results style: **best**,second best

Table 19 Performance of SCE for video retrieval on the first split of UCF101 and HMDB51 (color figure online)

Method	Res_p	T_p	Res_e	T_e	Pretrain	UCF101 R@1	UCF101 R@5	UCF101 R@10	HMDB51 R@1	HMDB51 R@5	HMDB51 R@10
Backbone: ResNet3D-18											
MemDPC [10]	40	224^2	40	224^2	UCF101	20.2	40.4	52.4	7.7	25.7	40.6
RSPNet [75]	16	112^2	16	224^2	K400	41.1	59.4	68.4	-	-	-
MFO [71]	16	112^2	16	112^2	K400	41.5	60.6	71.2	20.7	40.8	55.2
TransRank [5]	16	112^2	-	-	UCF101	46.5	63.7	-	19.4	45.4	59.1
ViCC [26]	16	128^2	16	128^2	UCF101	50.3	70.9	78.7	22.7	46.2	60.9
TransRank [5]	16	112^2	-	-	K200	54.0	71.8	-	25.5	52.3	65.8
TCLR [15]	16	112^2	-	-	UCF101	56.2	72.2	79.0	22.8	45.4	57.8
TEC [77]	16	128^2	16	128^2	UCF101	63.6	79.0	84.8	32.2	60.3	71.6
ProViCo [114]	16	112^2	-	-	UCF101	63.8	75.1	84.8	35.9	55.2	74.3
ProViCo [114]	16	112^2	-	-	K400	67.6	81.4	90.1	**40.1**	60.6	75.2
SCE (Ours)	16	112^2	16	128^2	K400	**74.5**	**85.9**	**90.5**	37.8	62.1	73.8
SCE (Ours)	8	224^2	8	256^2	K400	74.4	85.6	90.0	**40.1**	**63.3**	**75.4**
Backbone: ResNet3D-50											
CATE [72]	8	224^2	32	256^2	K400	54.9	68.3	75.1	33.0	56.8	69.4
SCE (Ours)	8	224^2	8	256^2	K400	81.5	89.7	92.8	43.0	67.0	79.0
SCE (Ours)	16	224^2	16	256^2	K400	**83.9**	**92.2**	**94.9**	**45.9**	**69.9**	**80.5**

Res_p, **Res_e** means the resolution for pretraining and evaluation. **T_p**, **T_e** means the number of frames used for pretraining and evaluation. We report the recall R@1, R@5, R@10. We obtain state-of-the-art results for ResNet3D-18 on both benchmarks and further improve our results using the larger network ResNet3D-50. Results style: **best**, second best

Table 20 Performance of SCE in comparison with Feichtenhofer et al. [4] for linear evaluation on Kinetics400 and finetuning on the first split of UCF101, AVA and SSv2 (color figure online)

Method	views	T	*Linear protocol* K400	*Finetuning accuracy* UCF101	AVA (mAP)	SSv2
Supervised	1	8	**74.7**	94.8	22.2	52.8
ρSimCLR ($\rho = 3$)	3	8	62.0 (−12.7)	87.9 (−6.9)	17.6 (−4.6)	52.0 (−0.8)
ρSwAV ($\rho = 3$)	3	8	62.7 (−12.0)	89.4 (−5.4)	18.2 (−4.0)	51.7 (−1.1)
ρBYOL ($\rho = 3$)	3	8	68.3 (−6.4)	93.8 (−1.0)	**23.4** (+1.2)	55.8 (+3.0)
ρMoCo ($\rho = 3$)	3	8	67.3 (−7.4)	92.8 (−2.0)	20.3 (−1.9)	54.4 (+1.8)
SCE (Ours)	2	8	67.6 (−7.1)	94.1 (−0.7)	20.3 (−1.9)	53.9 (+1.1)
SCE (Ours)	2	16	69.6 (−5.1)	**95.5** (+0.7)	21.6 (−0.6)	**57.2** (+4.4)

SCE is on par with ρMoCo for fewer views. Increasing the number of frames outperforms ρBYOL on Kinetics400, UCF101 and SSv2. Results style: **best**, second best

Generalization to downstream tasks. We follow the protocol introduced by Feichtenhofer et al. [4] to compare the generalization of our ResNet3d-50 backbone on Kinetics400, UCF101, AVA and SSv2 with ρSimCLR, ρSwAV, ρBYOL, ρMoCo and supervised learning in Table 20. To ensure a fair comparison, we provide the number of views used by each method and the number of frames per view for pretraining and evaluation.

For 2 views and 8 frames, SCE is on par with ρMoCo with 3 views on Kinetics400, AVA and SSv2 but is worst than ρBYOL especially on AVA. For UCF101, results are better than ρMoCo and on par with ρBYOL. These results indicate that our approach proves more effective than contrastive learning as it reaches similar results than ρMoCo using one less view. Using 16 frames, SCE outperforms all approaches, including supervised training, on UCF101 and

SSv2 but performs worse on AVA than ρByol and supervised training. This study shows that SCE can generalize to various video downstream tasks which is a criteria of a good learned representation.

5 Conclusion

In this paper, we introduced a self-supervised soft contrastive learning approach called Similarity Contrastive Estimation (SCE). It contrasts pairs of asymmetrical augmented views with other instances while maintaining relations among instances. SCE leverages contrastive learning and relational learning and improves the performance over optimizing only one aspect. We showed that it is competitive with the state of the art on the linear evaluation protocol on ImageNet, on video representation learning and to generalize to several image and video downstream tasks. We proposed a simple but effective initial estimation of the true distribution of similarity among instances. An interesting perspective would be to propose a finer estimation of this distribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00138-023-01444-9.

Acknowledgements This publication was made possible by the use of the Factory-AI supercomputer, financially supported by the Ile-de-France Regional Council, and the HPC resources of IDRIS under the allocation 2022-AD011013575 made by GENCI.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

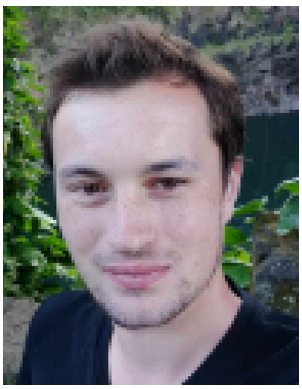
1. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (2020)
2. Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent—a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (2020)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision, pp. 6706–6716 (2021). https://doi.org/10.1109/ICCV48922.2021.00951
4. Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R.B., He, K.: A large-scale study on unsupervised spatiotemporal representation learning. In: Conference on Computer Vision and Pattern Recognition, pp. 3299–3309 (2021). https://doi.org/10.1109/CVPR46437.2021.00331
5. Duan, H., Zhao, N., Chen, K., Lin, D.: Transrank: self-supervised video representation learning via ranking-based transformation recognition. In: Conference on Computer Vision and Pattern Recognition, pp. 2990–3000 (2022). https://doi.org/10.1109/CVPR52688.2022.00301
6. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: 13th International Conference on Artificial Intelligence and Statistics, pp. 297–304 (2010)
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: Conference on Computer Vision and Pattern Recognition, pp. 9726–9735 (2020). https://doi.org/10.1109/CVPR42600.2020.00975
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning, pp. 1597–1607 (2020)
9. Yang, C., Xu, Y., Dai, B., Zhou, B.: Video representation learning with visual tempo consistency. arXiv:2006.15489 (2020)
10. Han, T., Xie, W., Zisserman, A.: Memory-augmented dense predictive coding for video representation learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) 16th European Conference on Computer Vision, pp. 312–329 (2020). https://doi.org/10.1007/978-3-030-58580-8_19
11. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (2020)
12. Qian, R., Meng, T., Gong, B., Yang, M., Wang, H., Belongie, S.J., Cui, Y.: Spatiotemporal contrastive video representation learning. In: Conference on Computer Vision and Pattern Recognition, pp. 6964–6974 (2021). https://doi.org/10.1109/CVPR46437.2021.00689
13. Pan, T., Song, Y., Yang, T., Jiang, W., Liu, W.: Videomoco: Contrastive video representation learning with temporally adversarial examples. In: Conference on Computer Vision and Pattern Recognition, pp. 11205–11214 (2021). https://doi.org/10.1109/CVPR46437.2021.01105
14. Recasens, A., Luc, P., Alayrac, J., Wang, L., Strub, F., Tallec, C., Malinowski, M., Patraaucean, V., Altché, F., Valko, M., Grill, J., Oord, A., Zisserman, A.: Broaden your views for self-supervised video learning. In: International Conference on Computer Vision, pp. 1235–1245 (2021). https://doi.org/10.1109/ICCV48922.2021.00129
15. Dave, I.R., Gupta, R., Rizve, M.N., Shah, M.: TCLR: temporal contrastive learning for video representation. Comput. Vis. Image Underst. (2022). https://doi.org/10.1016/j.cviu.2022.103406
16. Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv:1807.03748 (2018)
17. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Conference on

Computer Vision and Pattern Recognition, pp. 3733–3742 (2018). https://doi.org/10.1109/CVPR.2018.00393
18. Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D.: Hard negative mixing for contrastive learning. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (2020)
19. Robinson, J.D., Chuang, C., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. In: 9th International Conference on Learning Representations (2021)
20. Wu, M., Mosse, M., Zhuang, C., Yamins, D., Goodman, N.D.: Conditional negative sampling for contrastive learning of visual representations. In: 9th International Conference on Learning Representations (2021)
21. Hu, Q., Wang, X., Hu, W., Qi, G.: Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In: Conference on Computer Vision and Pattern Recognition, pp. 1074–1083 (2021)
22. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: With a little help from my friends: nearest-neighbor contrastive learning of visual representations. In: 2021 International Conference on Computer Vision, pp. 9568–9577 (2021). https://doi.org/10.1109/ICCV48922.2021.00945
23. Cai, T.T., Frankle, J., Schwab, D.J., Morcos, A.S.: Are all negatives created equal in contrastive instance discrimination? arXiv:2010.06682 (2020)
24. Wei, C., Wang, H., Shen, W., Yuille, A.L.: CO2: consistent contrast for unsupervised visual representation learning. In: 9th International Conference on Learning Representations (2021)
25. Chuang, C., Robinson, J., Lin, Y., Torralba, A., Jegelka, S.: Debiased contrastive learning. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (2020)
26. Toering, M., Gatopoulos, I., Stol, M., Hu, V.T.: Self-supervised video representation learning with cross-stream prototypical contrasting. In: Winter Conference on Applications of Computer Vision, pp. 846–856 (2022). https://doi.org/10.1109/WACV51458.2022.00092
27. Chen, X., He, K.: Exploring simple SIAMESE representation learning. In: Conference on Computer Vision and Pattern Recognition, pp. 15750–15758 (2021). https://doi.org/10.1109/CVPR46437.2021.01549
28. Zheng, M., You, S., Wang, F., Qian, C., Zhang, C., Wang, X., Xu, C.: RESSL: relational self-supervised learning with weak augmentation. In: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems, pp. 2543–2555 (2021)
29. Chen, X., Fan, H., Girshick, R.B., He, K.: Improved baselines with momentum contrastive learning. arXiv:2003.04297 (2020)
30. Dosovitskiy, A., Fischer, P., Springenberg, J.T., Riedmiller, M.A., Brox, T.: Discriminative unsupervised feature learning with exemplar convolutional neural networks. Trans. Pattern Anal. Mach. Intell. **38**(9), 1734–1747 (2016). https://doi.org/10.1109/TPAMI.2015.2496141
31. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: International Conference on Computer Vision, pp. 1422–1430 (2015). https://doi.org/10.1109/ICCV.2015.167
32. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: 14th European Conference on Computer Vision, pp. 649–666 (2016). https://doi.org/10.1007/978-3-319-46487-9_40
33. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: 14th European Conference on Computer Vision, pp. 69–84 (2016). https://doi.org/10.1007/978-3-319-46466-4_5
34. Noroozi, M., Pirsiavash, H., Favaro, P.: Representation learning by learning to count. In: International Conference on Computer Vision, pp. 5899–5907 (2017). https://doi.org/10.1109/ICCV.2017.628
35. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: 6th International Conference on Learning Representations (2018)
36. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In: 7th International Conference on Learning Representations (2019)
37. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: 16th European Conference on Computer Vision, pp. 776–794 (2020). https://doi.org/10.1007/978-3-030-58621-8_45
38. Misra, I., Maaten, L.: Self-supervised learning of pretext-invariant representations. In: Conference on Computer Vision and Pattern Recognition, pp. 6706–6716 (2020). https://doi.org/10.1109/CVPR42600.2020.00674
39. Wang, G., Wang, K., Wang, G., Torr, P.H.S., Lin, L.: Solving inefficiency of self-supervised representation learning. In: International Conference on Computer Vision, pp. 9485–9495 (2021). https://doi.org/10.1109/ICCV48922.2021.00937
40. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (2020)
41. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: International Conference on Computer Vision, pp. 9620–9629 (2021). https://doi.org/10.1109/ICCV48922.2021.00950
42. Yang, M., Li, Y., Huang, Z., Liu, Z., Hu, P., Peng, X.: Partially view-aligned representation learning with noise-robust contrastive loss. In: Conference on Computer Vision and Pattern Recognition, pp. 1134–1143 (2021). https://doi.org/10.1109/CVPR46437.2021.00119
43. Yang, M., Li, Y., Hu, P., Bai, J., Lv, J., Peng, X.: Robust multi-view clustering with incomplete information. Trans. Pattern Anal. Mach. Intell. **45**(1), 1055–1069 (2023). https://doi.org/10.1109/TPAMI.2022.3155499
44. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: 38th International Conference on Machine Learning, pp. 12310–12320 (2021)
45. Bardes, A., Ponce, J., LeCun, Y.: VICReg: Variance-invariance-covariance regularization for self-supervised learning. In: International Conference on Learning Representations (2022)
46. Li, J., Zhou, P., Xiong, C., Hoi, S.C.H.: Prototypical contrastive learning of unsupervised representations. In: 9th International Conference on Learning Representations (2021)
47. Zheng, M., Wang, F., You, S., Qian, C., Zhang, C., Wang, X., Xu, C.: Weakly supervised contrastive learning. In: International Conference on Computer Vision, pp. 10022–10031 (2021). https://doi.org/10.1109/ICCV48922.2021.00989
48. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: Proceedings of the 37th International Conference on Machine Learning, pp. 9929–9939 (2020)
49. Chen, T., Li, L.: Intriguing properties of contrastive losses. arXiv:2011.02803 (2020)
50. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Conference on Computer Vision and Pattern Recognition, pp. 3967–3976 (2019). https://doi.org/10.1109/CVPR.2019.00409
51. Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y., Liu, Z.: SEED: self-supervised distillation for visual representation. In: International Conference on Learning Representations (2021)

52. Koohpayegani, S.A., Tejankar, A., Pirsiavash, H.: Compress: Self-supervised learning by compressing representations. In: Advances in Neural Information Processing Systems (2020)
53. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, vol. 1 (Long and Short Papers), pp. 4171–4186 (2019). https://doi.org/10.18653/v1/n19-1423
54. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, pp. 5998–6008 (2017)
56. 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 16×16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
57. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: hierarchical vision transformer using shifted windows. In: International Conference on Computer Vision, pp. 9992–10002 (2021). https://doi.org/10.1109/ICCV48922.2021.00986
58. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: Image BERT pre-training with online tokenizer. In: International Conference on Learning Representations (2022)
59. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: BERT pre-training of image transformers. In: International Conference on Learning Representations (2022)
60. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: Conference on Computer Vision and Pattern Recognition, pp. 15979–15988 (2022). https://doi.org/10.1109/CVPR52688.2022.01553
61. Jing, L., Tian, Y.: Self-supervised spatiotemporal feature learning via video rotation prediction. arXiv:1811.11387 (2018)
62. Kim, D., Cho, D., Kweon, I.S.: Self-supervised video representation learning with space-time cubic puzzles. In: 31st Innovative Applications of Artificial Intelligence Conference, pp. 8545–8552 (2019). https://doi.org/10.1609/aaai.v33i01.33018545
63. Wang, J., Jiao, J., Bao, L., He, S., Liu, Y., Liu, W.: Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In: Conference on Computer Vision and Pattern Recognition, pp. 4006–4015 (2019). https://doi.org/10.1109/CVPR.2019.00413
64. Lee, H., Huang, J., Singh, M., Yang, M.: Unsupervised representation learning by sorting sequences. In: International Conference on Computer Vision, pp. 667–676 (2017). https://doi.org/10.1109/ICCV.2017.79
65. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: Unsupervised learning using temporal order verification. In: 14th European Conference on Computer Vision, pp. 527–544 (2016). https://doi.org/10.1007/978-3-319-46448-0_32
66. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y.: Self-supervised spatiotemporal learning via video clip order prediction. In: Conference on Computer Vision and Pattern Recognition, pp. 10334–10343 (2019). https://doi.org/10.1109/CVPR.2019.01058
67. Jenni, S., Meishvili, G., Favaro, P.: Video representation learning by recognizing temporal transformations. In: 16th European Conference on Computer Vision, pp. 425–442 (2020). https://doi.org/10.1007/978-3-030-58604-1_26
68. Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W.T., Rubinstein, M., Irani, M., Dekel, T.: Speednet: learning the speediness in videos. In: Conference on Computer Vision and Pattern Recognition, pp. 9919–9928 (2020). https://doi.org/10.1109/CVPR42600.2020.00994
69. Yao, Y., Liu, C., Luo, D., Zhou, Y., Ye, Q.: Video playback rate perception for self-supervised spatio-temporal representation learning. In: Conference on Computer Vision and Pattern Recognition, pp. 6547–6556 (2020). https://doi.org/10.1109/CVPR42600.2020.00658
70. Lorre, G., Rabarisoa, J., Orcesi, A., Ainouz, S., Canu, S.: Temporal contrastive pretraining for video action recognition. In: Winter Conference on Applications of Computer Vision, pp. 651–659 (2020). https://doi.org/10.1109/WACV45572.2020.9093278
71. Qian, R., Li, Y., Liu, H., See, J., Ding, S., Liu, X., Li, D., Lin, W.: Enhancing self-supervised video representation learning via multi-level feature optimization. In: International Conference on Computer Vision, pp. 7970–7981 (2021). https://doi.org/10.1109/ICCV48922.2021.00789
72. Sun, C., Nagrani, A., Tian, Y., Schmid, C.: Composable augmentation encoding for video representation learning. In: International Conference on Computer Vision, pp. 8814–8824 (2021). https://doi.org/10.1109/ICCV48922.2021.00871
73. Piergiovanni, A.J., Angelova, A., Ryoo, M.S.: Evolving losses for unsupervised video representation learning. In: Conference on Computer Vision and Pattern Recognition, pp. 130–139 (2020). https://doi.org/10.1109/CVPR42600.2020.00021
74. Wang, J., Jiao, J., Liu, Y.: Self-supervised video representation learning by pace prediction. In: 16th European Conference on Computer Vision, pp. 504–521 (2020). https://doi.org/10.1007/978-3-030-58520-4_30
75. Chen, P., Huang, D., He, D., Long, X., Zeng, R., Wen, S., Tan, M., Gan, C.: Rspnet: relative speed perception for unsupervised video representation learning. In: 33rd Conference on Innovative Applications of Artificial Intelligence, pp. 1045–1053 (2021)
76. Huang, D., Wu, W., Hu, W., Liu, X., He, D., Wu, Z., Wu, X., Tan, M., Ding, E.: Ascnet: self-supervised video representation learning with appearance-speed consistency. In: International Conference on Computer Vision, pp. 8076–8085 (2021). https://doi.org/10.1109/ICCV48922.2021.00799
77. Jenni, S., Jin, H.: Time-equivariant contrastive video representation learning. In: International Conference on Computer Vision (2021). https://doi.org/10.1109/ICCV48922.2021.00982
78. Sun, C., Baradel, F., Murphy, K., Schmid, C.: Learning video representations using contrastive bidirectional transformer. arXiv:1906.05743 (2019)
79. Miech, A., Alayrac, J., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Conference on Computer Vision and Pattern Recognition, pp. 9876–9886 (2020). https://doi.org/10.1109/CVPR42600.2020.00990
80. Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (2020)
81. Han, T., Xie, W., Zisserman, A.: Self-supervised co-training for video representation learning. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (2020)
82. Hu, K., Shao, J., Liu, Y., Raj, B., Savvides, M., Shen, Z.: Contrast and order representations for video self-supervised learning. In: International Conference on Computer Vision, pp. 7919–7929 (2021). https://doi.org/10.1109/ICCV48922.2021.00784
83. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., Schmid, C.: Vivit: a video vision transformer. In: International Conference

on Computer Vision, pp. 6816–6826 (2021). https://doi.org/10.1109/ICCV48922.2021.00676
84. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Conference on Computer Vision and Pattern Recognition, pp. 3192–3201 (2022). https://doi.org/10.1109/CVPR52688.2022.00320
85. Feichtenhofer, C., Fan, H., Li, Y., He, K.: Masked autoencoders as spatiotemporal learners. In: Advances in Neural Information Processing Systems (2022)
86. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Advances in Neural Information Processing Systems (2022)
87. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Jiang, Y., Zhou, L., Yuan, L.: BEVT: BERT pretraining of video transformers. In: Conference on Computer Vision and Pattern Recognition, pp. 14713–14723 (2022). https://doi.org/10.1109/CVPR52688.2022.01432
88. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: Computer Society Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
89. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
90. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Cs.Toronto.Edu, pp. 1–58 (2009)
91. Coates, A., Ng, A.Y., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, pp. 215–223 (2011)
92. Abai, Z., Rajmalwar, N.: Densenet models for tiny imagenet classification. arXiv:1904.10429 (2019)
93. Tao, C., Wang, H., Zhu, X., Dong, J., Song, S., Huang, G., Dai, J.: Exploring the equivalence of siamese self-supervised learning via A unified gradient framework. In: Conference on Computer Vision and Pattern Recognition (2022). https://doi.org/10.1109/CVPR52688.2022.01403
94. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: 13th European Conference on Computer Vision, pp. 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48
95. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask r-CNN. In: International Conference on Computer Vision, pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.322
96. Ericsson, L., Gouk, H., Hospedales, T.M.: How well do self-supervised models transfer? In: Conference on Computer Vision and Pattern Recognition, pp. 5414–5423 (2021). https://doi.org/10.1109/CVPR46437.2021.00537
97. Maji, S., Rahtu, E., Kannala, J., Blaschko, M.B., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv:1306.5151 (2013)
98. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. Comput. Vis. Image Underst. (2007) https://doi.org/10.1016/j.cviu.2005.09.012
99. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: International Conference on Computer Vision Workshops, pp. 554–561 (2013)
100. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Conference on Computer Vision and Pattern Recognition, pp. 3606–3613 (2014). https://doi.org/10.1109/CVPR.2014.461
101. Nilsback, M., Zisserman, A.: Automated flower classification over a large number of classes. In: 6th Indian Conference on Computer Vision, Graphics and Image Processing, pp. 722–729 (2008). https://doi.org/10.1109/ICVGIP.2008.47
102. Bossard, L., Guillaumin, M., Gool, L.V.: Food-101—mining discriminative components with random forests. In: 13th European Conference on Computer Vision, pp. 446–461 (2014). https://doi.org/10.1007/978-3-319-10599-4_29
103. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: Conference on Computer Vision and Pattern Recognition, pp. 3498–3505 (2012). https://doi.org/10.1109/CVPR.2012.6248092
104. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: large-scale scene recognition from abbey to zoo. In: Conference on Computer Vision and Pattern Recognition, pp. 3485–3492 (2010). https://doi.org/10.1109/CVPR.2010.5539970
105. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. (2010) https://doi.org/10.1007/s11263-009-0275-4
106. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification. In: 15th European Conference on Computer Vision, pp. 318–335 (2018). https://doi.org/10.1007/978-3-030-01267-0_19
107. 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. arXiv:1705.06950 (2017)
108. Soomro, K., Zamir, A.R., Shah, M.: UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402 (2012)
109. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T.A., Serre, T.: HMDB: a large video database for human motion recognition. In: International Conference on Computer Vision, pp. 2556–2563 (2011). https://doi.org/10.1109/ICCV.2011.6126543
110. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and imagenet? In: Conference on Computer Vision and Pattern Recognition, pp. 6546–6555 (2018). https://doi.org/10.1109/CVPR.2018.00685
111. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: International Conference on Computer Vision, pp. 6201–6210 (2019). https://doi.org/10.1109/ICCV.2019.00630
112. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: Ava: a video dataset of spatio-temporally localized atomic visual actions. In: Conference on Computer Vision and Pattern Recognition, pp. 6047–6056 (2018). https://doi.org/10.1109/CVPR.2018.00633
113. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fründ, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense. In: International Conference on Computer Vision, pp. 5843–5851 (2017). https://doi.org/10.1109/ICCV.2017.622
114. Park, J., Lee, J., Kim, I., Sohn, K.: Probabilistic representations for video contrastive learning. In: Conference on Computer Vision and Pattern Recognition, pp. 14691–14701 (2022). https://doi.org/10.1109/CVPR52688.2022.01430
115. Yuan, L., Qian, R., Cui, Y., Gong, B., Schroff, F., Yang, M., Adam, H., Liu, T.: Contextualized spatio-temporal contrastive learning with self-supervision. In: Conference on Computer Vision and

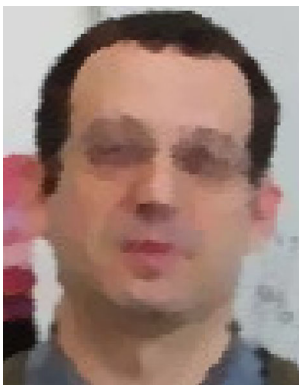
Pattern Recognition, pp. 13957–13966 (2022). https://doi.org/10.1109/CVPR52688.2022.01359
116. Zhang, D., Dai, X., Wang, X., Wang, Y.: S3D: single shot multi-span detector via fully 3d convolutional networks. In: British Machine Vision Conference, p. 293 (2018)
117. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018). https://doi.org/10.1109/CVPR.2018.00675



Julien Denize is a Ph.D. student in Machine Learning for Computer Vision at the CEA LIST and Normandie Université. He obtained his Master's Degree in Engineering in Computer Science in 2020 from the Engineering School Télécom SudParis with a major in Artificial Intelligence. Before starting his Ph.D., he was a research intern at CEA LIST on the topic of people re-identification and cross-domain adaptation via generative models. His research has led to several successful applications for image and video analysis, which will lead him to defend his Ph.D. thesis in late 2023. His main research interest is self-supervised learning for deep representation learning in computer vision.

Jaonary Rabarisoa is a research scientist at the CEA LIST, a research institute in France. He received his Master's degrees in Numerical Analysis from the University of Pierre et Marie Curie (France) in 2004 and in Machine Learning and Vision from the ENS Cachan (France) in 2005. He worked as a research engineer at the CERTIS laboratory of the École Nationale des Ponts et Chaussées (France) from 2005 to 2008. He then joined the CEA LIST in 2009, where he is currently a senior research scientist. His research interests include machine learning for computer vision applications and decision making. He has co-authored several papers in peer-reviewed journals and conferences, such as the Conference on Computer Vision and Pattern Recognition, the Winter Conference on Applications of Computer Vision, and the Conference on Lifelong Learning Agents. He also manages several national and European funded research projects. Jaonary is passionate about using machine learning to solve real-world problems. He is currently working on projects related to image and video understanding, autonomous driving, and robotics. He is also interested in developing new machine learning algorithms that are more robust and efficient.

Astrid Orcesi is a research-engineer in computer vision at CEA LIST. After a graduate of Grenoble INP-Phelma engineering school, she joined the vision team in 2016 on the ITEA3 Emospaces project involving video detection of a person's activities in an apartment. In 2018, she joined the joint laboratory with Thales, where she works on analyzing interactions in images until 2021. In 2022, she works on the detection of abnormal events. In 2020, she was appointed leader of the ANR TeamSports project, which aims to analyze the group dynamics of team sports teams. These various projects have enabled her to supervise 2 PhD students, 2 engineers and 6 interns. Her 7 years' experience led her to defend a VAE PhD in 2023 on learning methods applied to vision-based analysis of human behavior.



Romain Hérault is an Associate Professor (Maître de conférences) who has dedicated his work to the eld of machine learning. Since September 2008, he has been associated with INSA de Rouen's computer department ITI and the LITIS lab. In a new phase of his academic journey, Romain is preparing to transition to the GREYC lab in Caen, starting from September 2023. Romain's academic achievements include a Ph.D. in Information and System Technology from Université de Technologie de Compiègne in 2007. His research primarily revolves around machine learning, with a focus on deep neural networks, kernel methods, and dictionary learning. He has effectively applied machine learning techniques to various domains such as signal processing, pattern recognition, data mining, medical imaging, and human movement analysis. Romain's work underscores his commitment to advancing the understanding and applications of machine learning.