Unsupervised Learning from Video with Deep Neural Embeddings

Chengxu Zhuang[1]   Tianwei She[1]   Alex Andonian[2]   Max Sobol Mark[1]   Daniel Yamins[1]
[1]Stanford University   [2] MIT
{chengxuz, shetw, joelmax, yamins}@stanford.edu   andonian@mit.edu

Abstract

Because of the rich dynamical structure of videos and their ubiquity in everyday life, it is a natural idea that video data could serve as a powerful unsupervised learning signal for visual representations. However, instantiating this idea, especially at large scale, has remained a significant artificial intelligence challenge. Here we present the Video Instance Embedding (VIE) framework, which trains deep nonlinear embeddings on video sequence inputs. By learning embedding dimensions that identify and group similar videos together, while pushing inherently different videos apart in the embedding space, VIE captures the strong statistical structure inherent in videos, without the need for external annotation labels. We find that, when trained on a large-scale video dataset, VIE yields powerful representations both for action recognition and single-frame object categorization, showing substantially improving on the state of the art wherever direct comparisons are possible. We show that a two-pathway model with both static and dynamic processing pathways is optimal, provide analyses indicating how the model works, and perform ablation studies showing the importance of key architecture and loss function choices. Our results suggest that deep neural embeddings are a promising approach to unsupervised video learning for a wide variety of task domains.

1. Introduction

A video's temporal sequence often contains information about dynamics and events in the world that is richer than that in its unordered set of frames. For example, as objects and agents move and interact with each other, they give rise to characteristic patterns of visual change that strongly correlate with their visual and physical identities, including object category, geometric shape, texture, mass, deformability, motion tendencies, and many other properties. It is thus an attractive hypothesis that ubiquitiously available natural videos could serve as a powerful signal for unsupervised learning of visual representations for both static and dynamic visual tasks.

However, it has been challenging to embody this hypothesis in a concrete neural network that can consume unlabelled video data to learn useful feature representations, especially in the context of at-scale real-world applications. Perhaps the biggest source of difficulty in making progress on unsupervised video learning, though, is that unsupervised learning has presented a formidable challenge even for the case of single static images. Even for the case of single static images, the gap in representational power between the features learned by unsupervised and supervised neural networks has been very substantial, to the point where the former were unsuitable for use in any at-scale visual task. However, recent advances in learning with deep visual embeddings have begun to produce unsupervised representations that rival the visual task transfer power of representations learned by their supervised counterparts [37, 36, 40, 3]. These methods leverage simple but apparently strong heuristics about data separation and clustering to iteratively bootstrap feature representations that increasingly better capture subtle natural image statistics. As a result, it is now possible to obtain unsupervised deep convolutional neural networks that outperform "early modern" deep networks (such as AlexNet in [13]) on challenging recognition tasks such as ImageNet, even when the latter are trained in a supervised fashion. Moreover, works in supervised video classification, action recognition and video captioning have proposed novel combinations of two- and three-dimensional convolutional structures that are increasingly well-suited to video feature extraction [5].

In this work, we show how the idea of deep unsupervised embeddings can be used to learn features from videos, introducing the Video Instance Embedding (VIE) framework. In VIE, videos are projected into a compact latent space via a deep neural network, whose parameters are then tuned to optimally distribute embedded video instances so that similar videos aggregate while dissimilar videos separate. We find that VIE

learns powerful representations for transfer learning to action recognition in the large-scale Kinetics dataset, as well as for single-frame object classification in the ImageNet dataset. Moreover, where direct comparison to previous methods is possible, we find that VIE substantially improves on the state-of-the-art. We evaluate several possibilities for the unsupervised VIE loss function, finding that those that have been shown to be most effective in the single-frame unsupervised learning [40] are also most effective in the video context. We also explore several neural network embedding and frame sampling architectures, finding that different temporal sampling statistics are better priors for different transfer tasks, and that a two-pathway static-dynamic architecture is optimal. Finally, we present analyses of the learned representations giving some intuition as to how the models work, and a series of ablation studies illustrating the importance of key architectural choices. Codes can be found at `https://github.com/neuroailab/VIE`.

2. Related Work

Unsupervised Learning of Deep Visual Embeddings. In this work, we employ a framework derived from ideas first introduced in the recent literature on unsupervised learning of embeddings for images [37]. In the Instance Recognition (IR) task, a deep nonlinear image embedding is trained to maximize the distances between different images while minimizing distances between augmentations (e.g. crops) of a given image, thus maximizing the network's ability to recognize individual image instances. In the Local Aggregation (LA) task [40], the embedding loss also allows selected groups of images to aggregate, dynamically determining the groupings based on a local clustering measure. Conceptually, the LA approach resembles a blending of IR and the also-recent DeepCluster method [3], and is more powerful than either IR or DeepCluster alone, achieving state-of-the-art results on unsupervised learning with images. Another task Contrastive Multiview Coding (CMC) [28] achieves similar performance to LA within this embedding framework while aggregating different views of a given image. CMC have also been directly applied to videos, where single frame is clustered with its future frame and its corresponding optical flow image. The VIE framework allows the use of any of these embedding objectives, and we test several of them here. Other recently proposed methods such as [12] and [2] have achieved comparable results to LA and CMC through optimizing the mutual information of different views of the images, though they use much deeper and more complex architectures.

Supervised Training of Video Networks. Neural networks have been used for a variety of supervised video tasks, including captioning [19] and 3D shape extraction [23, 1], but the architectures deployed in those works are quite different from those used here. The structures we employ are more directly inspired by work on supervised action recognition. A core architectural choice explored in this literature is how and where to use 2D single-frame vs 3D multi-frame convolution. A purely 2D approach is the Temporal Relational Network (TRN) [39], which processes aggregates of 2D convolutional features using MLP readouts. Methods such as I3D [4] have shown that combinations of both 2D and 3D can be useful, deploying 2D processing on RGB videos and 3D convolution on an optical flow component. A current high-performing architecture for action recognition is the SlowFast network [5], which computes mixed 2D-3D convolutional features from sequences of images at multiple time scales, including a slow branch for low frequency events and a fast branch for higher-frequency events. The dynamic branch of our two-pathway architecture is chosen to mimic the most successful SlowFast network. However, we find it useful to include in our architecture a static pathway that is not equivalent to either of the SlowFast branches.

Unsupervised Learning on Videos. The literature on unsupervised video learning is too extensive to review comprehensively here, so we focus our discussion on several of the most relevant approaches. Temporal autoencoders such as PredNet [22], PredRNN [35], and PredRNN++ [34] are intriguing but have not yet evidenced substantial transfer learning performance at scale. Transfer learning results have been generated from a variety of approaches including the Geometry-Guided CNN [6], motion masks [26], VideoGAN [31], a pairwise-frame siamese triplet network [33], the Shuffle and Learn approach [24], and the Order Prediction Network (OPN) [21]. More recent works, including the Video Rotations Prediction task (3DRotNet) [16], the Video Motion and Appearance task (MoAp) [32], the Space-Time Puzzle task (ST-puzzle) [18], and the Dense Predictive Coding (DPC) task [9], have reported improved performance, with the help of pretraining on large-scale datasets and using spatiotemporal network architectures. All these works only operate on relationships defined within a single video, differentiating them from VIE, which exploits the relationships both within and between different videos through a loss function defined on the *distribution* of video embeddings.

3. Methods

VIE Embedding framework. The general problem of unsupervised learning from videos can be formulated as learning a parameterized function $\phi_{\boldsymbol{\theta}}(\cdot)$ from input videos $\mathbf{V} = \{\mathbf{v}_i | i = 1, 2, .., N\}$, where each $\mathbf{v}_i$ con-
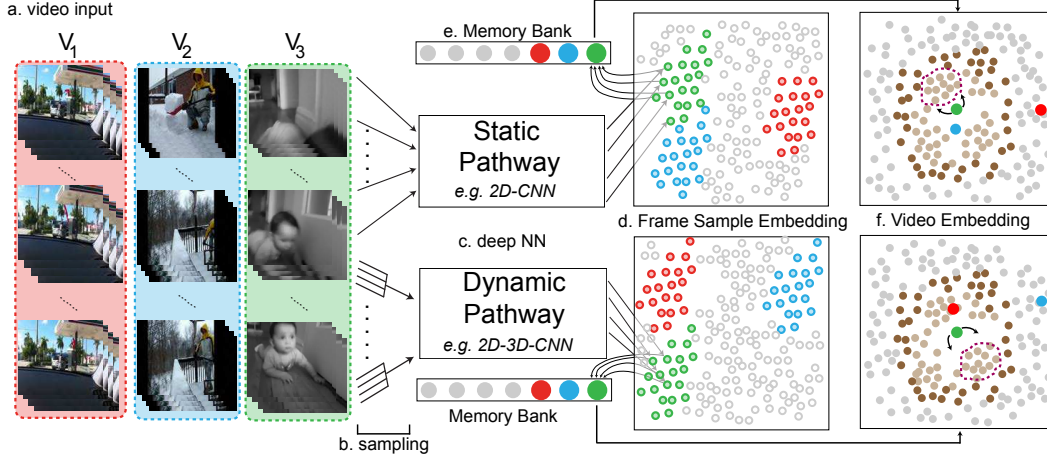
Figure 1: **Schematic of the Video Instance Embedding (VIE) Framework. a.** Frames from individual videos ($\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$) are **b.** sampled into sequences of varying lengths and temporal densities, and input into **c.** deep neural network pathways that are either static (single image) or dynamic (multi-image). **d.** Outputs of frame samples from either pathway are vectors in the D-dimensional unit sphere $S^D \subset \mathbf{R}^{D+1}$. The running mean value of embedding vectors are calculated over online samples for each video, **e.** stored in a memory bank, and **f.** at each time step compared via unsupervised loss functions in the video embedding space. The loss functions require the computation of distribution properties of embedding vectors. For example, the Local Aggregation (LA) loss function involves the identification of Close Neighbors $\mathbf{C}_i$ (light brown points) and Background Neighbors $\mathbf{B}_i$ (dark brown points), which are used to determine how to move target point (green) relative to other points (red/blue).

sists of a sequence of frames $\{f_{i,1}, f_{i,2}, ..., f_{i,m_i}\}$. Our overall approach seeks to embed the videos $\{\mathbf{v}_i\}$ as feature vectors $\mathbf{E} = \{\mathbf{e}_i\}$ in the D-dimension unit sphere $S^D = \{x \in \mathbf{R}^{D+1}$ with $||x||_2^2 = 1\}$. This embedding is realized by a neural network $\phi_{\boldsymbol{\theta}} : \mathbf{v}_i \mapsto S^D$ with weight parameters $\boldsymbol{\theta}$, that receives a sequence of frames $\mathbf{f} = \{f_1, f_2, ..., f_L\}$ and outputs $e = \phi_{\boldsymbol{\theta}}(\mathbf{f}) \in S^D$. Although the number of frames in one video can be arbitrary and potentially large, L must usually be fixed for most deep neural networks. Therefore, the input $\mathbf{f}$ on any single inference pass is restricted to a subset of frames in $\mathbf{v}$ chosen according to a *frame sampling strategy* ρ — that is, a random variable function such that $\mathbf{v}' \subset \mathbf{v}$ for all samples $\mathbf{v}'$ drawn from $\rho(\mathbf{v})$. Given ρ, we then define the associated ***Video Instance Embedding*** (VIE) $\mathbf{e}$ for video $\mathbf{v}$ as the normed (vector-valued) expectation of e under ρ, i.e.

$$\mathbf{e} = \frac{\mathrm{E}_\rho[\phi_{\boldsymbol{\theta}}(\mathbf{f})]}{||\mathrm{E}_\rho[\phi_{\boldsymbol{\theta}}(\mathbf{f})]||_2} \in S^D. \qquad (1)$$

In addition to choosing $\phi_{\boldsymbol{\theta}}$ and ρ, we must choose a loss function $\mathcal{L} : \mathbf{E} \mapsto \mathbf{R}$ such optimizing $\mathcal{L}$ with respect to $\boldsymbol{\theta}$ will cause statistically related videos to be grouped together and unrelated videos to be separated. Note that in theory this function depends on all of embedded vectors, although in practice it is only ever evaluated on a stochastically-chosen batch at any one time, with dataset-wide effects captured via a memory bank.

In the following subsections, we describe natural options for these main components (architecture $\phi_{\boldsymbol{\theta}}$, sampling strategy ρ, and loss function $\mathcal{L}$). As shown in Section 4, such choices can not only influence the quality of learned representations, but also change the focus of the representations along the spectrum between static and dynamic information extraction.

Architecture ϕ and sampling strategy ρ. Recent exploration in supervised action recognition has provided a variety of choices for $\phi_{\boldsymbol{\theta}}$. Although very complex network options are possible [4], since this work is an initial exploration of the interplay between video processing architecture and unsupervised loss functions, we have chosen to concentrate on five simple but distinct model families. They are differentiated mainly by how their frame sampling assumptions represent different types of temporal information about the inputs:

1. *Single-frame 2D-CNNs.* Although deep 2D convolutional neural networks (CNNs) taking one frame as input ignore temporal information in the videos, they can still leverage context information and have achieved nontrivial performance on action recognition datasets [4]. They are also useful baselines for measuring the effect of including temporal information.

2. *3D-CNNs with dense equal sampling.* 3D-CNNs, with spatiotemporal filters, can be applied to dense evenly-sampled frames to capture fine-grained temporal information. This architecture has proven useful in the

R3D networks of [29] and the 3DResNets of [10].

3. *Shared 2D-CNNs with sparse unequal sampling.* The Temporal Relation Network bins videos into half-second clips, chooses L consecutive bins, and randomly samples one frame from each bin. A shared 2D-CNN is then applied to each frame and its outputs are concatenated in order and fed into an MLP that creates the embedding. Unlike the dense sampling approach, this method can capture long-range temporal information through sparse sampling, but as the intervals between frames are uneven, the temporal signal can be noisy.

4. *2D-3D-CNNs with sparse equal sampling.* Addressing the issue of noisy temporal information in the third family, the Slow branch of the SlowFast [5] architecture samples frames equally but sparsely from the input video. These frames are then passed through 2D-CNNs with spatial pooling, applying 3D convolutions downstream once spatial redundancy is reduced.

5. *Multi-pathway architectures.* Combination architectures allow the exploitation of the multiple temporal scales present in natural videos, including the SlowFast approach [5] (combining 2 and 4), and true two-pathway networks with both static and dynamic pathways (combining 1, 2, and 4).

In our experiments (§4), we implement these models with CNN backbones that, while accommodating small unavoidable differences due to the input structure, are otherwise as similar as possible, so that the qualities of learned representations can be fairly compared.

Loss function $\mathcal{L}$. Recent work in unsupervised learning with single images has found useful generic metrics for measuring the quality of deep visual embeddings [37, 40], including the Instance Recognition (IR) and Local Aggegation (LA) loss functions. These methods seek to group similar inputs together in the embedding space, while separating dissimilar inputs. They are based on a simple probabilistic perspective for interpreting compact embedding spaces [37, 40]. Specifically, the probability that an arbitrary feature e is recognized as a sample of $\mathbf{v}_i$ is defined to be:

$$P(i|e, \mathbf{E}) = \frac{\exp(\mathbf{e}_i^T e/\tau)}{\sum_{j=1}^N \exp(\mathbf{e}_j^T e/\tau)} \qquad (2)$$

where temperature $\tau \in (0, 1]$ is a fixed scale hyperparameter. Both $\{\mathbf{e}_i\}$ and e are projected onto the unit sphere S^D. With this definition in mind, we can define the IR and LA loss functions, adapted to the video context via eq. 1.

IR algorithm. The VIE-version of the IR loss is:

$$\mathcal{L}^{\text{IR}}(\mathbf{v}_i, \mathbf{E}) = -\log P(i|e, \mathbf{E}) + \lambda\|\boldsymbol{\theta}\|_2^2 \qquad (3)$$

where λ is a regularization hyperparameter, and where for computational efficiency the denominator

in $P(i|e, \mathbf{E})$ is estimated through randomly choosing a subset of Q out of all N terms (see [37] for further details). Intuitively, optimizing this loss will group embeddings of frame groups sampled from the same video, which then implicitly gathers other similar videos.

LA algorithm. Local Aggregation augments IR by allowing for a more flexible dynamic detection of which datapoints should be grouped together. Define the probability that a feature e is recognized as being in a set of videos $\mathbf{A}$ as $P(\mathbf{A}|e, \mathbf{E}) = \sum_{i\in\mathbf{A}} P(i|e, \mathbf{E})$. For a video $\mathbf{v}_i$ and its embedding $\mathbf{e}_i$, the LA algorithm identifies two sets of neighbors, the *close neighbors* $\mathbf{C}_i$ and *background neighbors* $\mathbf{B}_i$. $\mathbf{C}_i$ is computed via dynamic online k-means clustering, and identifies datapoints expected to be "especially similar" to $\mathbf{v}_i$; $\mathbf{B}_i$ is computed via k-nearest neighbors method, and sets the scale of distance in terms of which closeness judgements are measured. Given these two neighbor sets, the local aggregation loss function measures the negative log-likelihood of a point being recognized as a close neighbor given that is already a background neighbor:

$$\mathcal{L}^{\text{LA}}(x_i, \mathbf{E}) = -\log \frac{P(\mathbf{C}_i \cap \mathbf{B}_i | \mathbf{v}_i, \mathbf{E})}{P(\mathbf{B}_i | \mathbf{v}_i, \mathbf{E})} + \lambda\|\boldsymbol{\theta}\|_2^2. \quad (4)$$

Intuitively, the LA loss function encourages the emergence of soft clusters of datapoints at multiple scales. See [40] for more details on the LA procedure.

Memory Bank. Both the IR and LA loss functions implicitly require access to all the embedded vectors $\mathbf{E}$ to calculate their denominators. However, recomputing $\mathbf{E}$ is intractable for big dataset. This issue is addressed, as in [37, 36, 40], by approximating $\mathbf{E}$ with a memory bank $\bar{\mathbf{E}}$ keeping a running average of the embeddings.

4. Experiments and Results

Experimental settings. To train our models, we use the Kinetics-400 dataset [17], which contains approximately 240K training and 20K validation videos, each around 10 seconds in length and labeled in one of 400 action categories. After downloading, we standardize videos to a framerate of 25fps and reshape all frames so that the shortest edge is 320px. After sampling according to the frame sampling strategy for each model architecture, we then apply the spatial random cropping and resizing method used in [5]. Following [40] we also apply color noise and random horizontal flip, using the same spatial window and color noise parameters for all the frames within one video. At test time, we sample five equally-spaced sequences of frames, resize them so that their shortest side is 256px, and take center 224×224 crop. Softmax logits for the five samples are then averaged to generate the final output prediction.

We use ResNet-18v2 [11] as the convolutional backbone for all our model families, to achieve a balance between model performance and computation efficiency. Implementations for the different model families are denoted VIE-Single ("Family 1"), VIE-3DResNet[1] ("Family 2"), VIE-TRN ("Family 3"), VIE-Slow ("Family 4"), and VIE-SlowFast ("Family 5"). Two-pathway models are created by concatenating single- and multi-frame network outputs, yielding VIE-TwoPathway-S (combining VIE-Single and VIE-Slow) and VIE-TwoPathway-SF (combining VIE-Single and VIE-Slowfast). We follow [40] for general network training hyperparameters including initial learning rate, optimizer setting, learning rate decay schedule, batch size, and weight decay coefficient. Further details of model architecture and training can be found in the supplementary material.

Transfer to action recognition on Kinetics. After training all models on the unlabelled videos from Kinetics, we evaluate the learned representations by assessing transfer learning performance to the Kinetics action recognition task. To compare our method to previous work on this task, we reimplement three strong unsupervised learning algorithms: OPN [21], 3DRotNet [16], and RotNet [7]. As OPN and RotNet require single-frame models, we use ResNet18 as their visual backbones. For 3DRotNet, we use 3DResNet18 with the same input resolution as VIE-3DResNet. Reimplementation details are in the supplementary material. Transfer learning is assessed through the standard transfer procedure of fixing the learned weights and then training linear-softmax readouts from different layers of the fixed model. We implemented this procedure following the choices in [40], but using the Kinetics-specific data augmentation procedure described above. Since some of the models generate outputs with a temporal dimension, directly adding a fully-connected readout layer would lead to more trainable readout parameters as compared to single frame models. To ensure fair comparisons, we thus average the features of such models along the temporal dimension before readout.

Results are shown in Table 1. All VIE variants show significantly better performance than OPN, RotNet, and 3DRotNet. Multi-frame models substantially outperform single-frame versions, an improvement that cannot be explained by the mere presence of additional frames (see the supplementary material Table S1). The two-pathway models achieve the highest performance, with a maximum accuracy of approximately 48.5%. The rank order of unsupervised performances across VIE variants are aligned with that of supervised counterparts, indicating that the unsupervised training

procedure takes advantage of increased architectural power when available. The LA-based VIE-Single model performs better than the IR-based model, consistent with the gap on static object recognition [40]. Finally, though previous work on unsupervised video learning has started to leverage large-scale datasets such as Kinetics [9, 16, 18], the transfer learning performance of the trained models to Kinetics action recognition has never been reported. We therefore hope these results are useful both for understanding the effect of architecture on representation quality and providing a strong unsupervised benchmark for future works.

Transfer to action recognition on UCF101 and HMDB51. To compare VIE to previous methods, we evaluate results on the more commonly-used UCF101 [27] and HMDB51 [20] action recognition benchmarks. We initialize networks by pretrained weights on Kinetics and then finetune them on these datasets, following the procedures used for the most recent work [16, 32, 18]. We notice that details of the data augmentation pipelines used during finetuning can influence final results. Most importantly, we find having color-noise augmentation can improve the finetuing performance. However, the augmentation techniques have not been carefully controlled in previous works. For example, ST-puzzle [18], 3DRotNet [16], and MoAp [32] only use the usual random crop and horizontal flip augmentations, while DPC [9] also uses color-noise. To ensure that our comparisons to these algorithms are fair, we therefore test our models with both augmentation pipelines. Details can be found in the supplement.

Table 2 and 3 show that VIE substantially outperforms other methods. Making these comparisons requires some care, because results reported in previous works are often confounded by the variation of multiple factors at once, making it hard to determine if improvements are really due to a better algorithm, rather than a larger training dataset, a more powerful network architecture, or a difference in input data types. First, holding network architecture and training dataset fixed, VIE-3DResNet surpasses the previous state-of-the-art (ST-Puzzle) by **6.5%** on UCF101 and **11.1%** on HMDB51, approaching the supervised upper-bound. With the better augmentation pipeline, the improvement on UCF101 is **9.7%**. 3DRotNets are trained on Kinetics-600, which contains more than twice as much training data and 50% more categories than that used for ST-Puzzle and VIE models, and are trained with larger inputs (64-frame RGB vs VIE's 16-frame inputs). Nonetheless, VIE still shows improvement of more than **6.2%** when compared with comparable input type. 3DRotNet also reports results for a larger fused model trained with both 64-frame RGB and frame-

---

[1]To be comparable to previous work, our 3DResNet uses a lower input resolution, 112×112. See the supplementary material.

Datasets	Kinetics				ImageNet		
Metric	Super.	Conv3	Conv4	Conv5	Conv3	Conv4	Conv5
Random-Single	–	9.40	8.43	6.84	7.98	7.78	6.23
OPN-Single* [21]	–	16.84	20.82	20.86	13.01	17.63	18.29
RotNet-Single* [7]	–	26.25	30.27	23.33	25.77	27.59	16.13
3DRotNet-3DResNet* [16]	–	28.30	29.33	19.33	23.34	22.05	12.45
VIE-Single (IR)	57.59	23.50	38.72	43.85	22.85	40.49	40.43
VIE-Single	57.59	23.84	38.25	44.41	25.02	40.49	42.33
VIE-TRN	59.43	25.72	39.38	44.91	**27.24**	40.28	37.44
VIE-3DResNet	53.22	**33.01**	41.34	43.40	30.18	35.37	32.62
VIE-Slow	60.84	24.80	40.48	46.36	20.10	37.02	37.45
VIE-Slowfast	**62.36**	28.68	42.07	47.37	22.61	36.84	36.60
VIE-TwoPathway-S	–	26.38	41.80	47.13	23.98	40.52	**44.02**
VIE-TwoPathway-SF	–	29.89	**43.50**	**48.53**	23.23	**40.73**	43.69
Supervised-Single					22.32	37.82	38.26
Supervised-TRN					22.82	41.13	39.15
Supervised-3DResNet			–		28.09	34.40	30.56
Supervised-Slow					21.86	40.77	32.87
Supervised-SlowFast					20.25	37.41	30.75

Table 1: Top-1 transfer learning accuracy (%) on the Kinetics and ImageNet validation sets. "Random" means a randomly initialized ResNet-18 without any training. "Supervised-*" means trained on Kinetics for action recognition. Our supervised performance is not directly comparable to [5] due to different visual backbones used. *: These numbers are generated by us.

difference inputs. VIE-Slowfast nonetheless substantially outperforms this model, using fewer trainable parameters (21M vs 32M) and much less training data (Kinetics-400 vs Kinetics-600). When compared to the very recent DPC algorithm [9] in Table 3, VIE shows improvement of more than **7.3%**, even though DPC uses more frames (40 v.s. 16) as inputs and adds an additional recurrent cell between the action readout layer and its 3DResNet. Again, VIE makes good use of more complex architectures, as can be seen in the SlowFast vs ResNet18 comparison. Moreover, TwoPathway-SF achieves even better performance than SlowFast.

Transfer to static object categorization. To determine the extent to which the VIE procedure learns general visual representations, we also evaluate the learned representations for transfer to image categorization in ImageNet. For models requiring multi-frame inputs, we generate a "static video" by tiling still images across multiple frames. Results are shown in Table 1. As they for the action recognition transfer, the two-pathway models are highest performing for this task as well. Interestingly, however, unlike for the case of action recognition, the multi-frame dynamic models are substantially worse than the single frame models on the ImageNet transfer task, and show a performance drop at the highest convolutional layers. In fact, the transfer performance of the single-frame unsupervised model trained on Kinetics is actually *better* than that of any model *supervised* on Kinetics. Taken together, these re-

Networks	Algorithms	UCF	HMDB
AlexNet[†]	CMC [28]	59.1	26.7
	OPN [21]	56.3	22.1
C3D	MoAp [32]	61.2	33.4
	ST-puzzle [18]	60.6	28.3
3DResNet	Scratch	47.4	21.5
	ST-puzzle	65.8	33.7
	3DRotNet [16]	62.9	33.7
	3DRotNet(64f) [16]	66.0	37.1
	VIE (ours)	**72.3**	**44.8**
	Supervised	*84.4*	*58.7*
SlowFast	Scratch	55.8	21.4
	VIE (ours)	**77.0**	**46.5**
	Supervised	*88.4*	*68.4*
ResNet18	Scratch	46.7	17.3
	VIE (ours)	71.2	38.4
	Supervised	*81.0*	*49.9*
VIE-TwoPathway-SF		**78.2**	**50.5**

Table 2: Top-1 finetuning results on UCF101 and HMDB51 datasets using models pretrained on Kinetics **without** color-noise augmentation. We also provide performance of training from scratch ("Scratch") and from supervisedly trained models on Kinetics ("Supervised"). For 3DRotNet, we compare to its model trained with RGB inputs, where 64f means 64 frames. [†]: AlexNet results are all pretrained on UCF101.

sults strongly motivate the two-pathway architecture, as

Networks	Algorithms	UCF	HMDB
3DResNet	Scratch	60.0	27.0
	DPC [9]	68.2	34.5
	VIE (ours)	**75.5**	**44.6**
	Supervised	*84.8*	*60.2*
SlowFast	Scratch	70.0	37.0
	VIE (ours)	**78.9**	**50.1**
	Supervised	*89.7*	*70.4*
ResNet18	Scratch	57.3	23.9
	VIE (ours)	73.1	41.2
	Supervised	*83.5*	*52.9*
VIE-TwoPathway-SF		**80.4**	**52.5**

Table 3: Top-1 finetuning results on UCF101 and HMDB51 datasets using models pretrained on Kinetics **with** color-noise augmentation.

features that contribute to high performance on action recognition — e.g. processing of dynamical patterns — are not optimal for static-image performance. However, the relatively high performance of the static and two-pathway models shows that VIE can achieve useful generalization, even when train and test datasets are as widely divergent as Kinetics and ImageNet.

5. Analysis

Benefit from long-range temporal structure. A key idea in VIE is to embed entire videos into the latent space, which is intended to leverage contextual information contained in the video. This may even have utility for videos with multiple scenes containing widely-diverging content (common for Kinetics), as the high-dimensional embedding might learn to situate such videos in the latent space so as to retain this structure. As a preliminary test of the validity of this approach, we generated new training datasets by dividing each video into equal-length bins and then use these temporally-clipped datasets to train VIE. Table 4 show that the full-video model outperforms both 2- and 5-bin models, especially on ImageNet transfer learning performance, supporting the choice of embedding entire videos and also indicating that even better performance may be obtained using longer, more contextually complex videos.

Benefit from more data. Although VIE achieves state-of-the-art unsupervised transfer performance on action recognition, and creates representations more useful for static object categorization than supervision on action recognition, its learned representation is (unsurprisingly) worse on ImageNet categorization than its counterpart directly trained on the (much larger) ImageNet training set [40]. To test whether VIE would benefit from more videos, we retrain VIE-Single with subsampled Kinetics (see Table 4). Performance on

ImageNet increases consistently and substantially without obvious saturation, indicating VIE's representation generalizability would benefit substantially if trained on a video dataset of the scale of ImageNet.

Video retrieval. We conduct a video retrieval experiment using distance in the embedding space. Representative examples are shown in Figure 2. Oftentimes, qualitatively similar videos were retrieved, although some failure cases were also observed. Moreover, VIE-Slowfast appears to extract context-free dynamic information, while VIE-Single is more biased by per-frame context, further validating the idea that multi-frame models develop representations focusing on the dynamic features, while single-frame models better extract static information. For example, in the "cleaning shoes" query, the two nearest VIE-Slowfast neighbors share a common dynamic (hand motions) with the query video, while hand and shoe position and the backgrounds all vary. Meanwhile, VIE-Single only captures object semantics (the presence of the hand), lacking information about the movement that hand will make. Retrieval failures likewise exemplify this result: in the bandaging and baking cookies examples, VIE-Slowfast captures high-level motion patterns inaccessible to the static pathway.

6. Conclusion

We have described the VIE method, an approach that combines multi-streamed video processing architectures with unsupervised deep embedding learning, and shown initial evidence that deep embeddings are promising for large-scale unsupervised video learning. In this work, our goal is not so much to illustrate the dominance of one particular novel loss function and architecture, but rather to provide a clear demonstration of the fact that a previously very challenging goal — unsupervised learning on at-scale video tasks — has become substantially more effective that it had previously been, due to a combination of very recent architectural and loss-function ideas. Together with recent results in static image categorization [40, 37, 28, 3], our results suggest that the deep embedding approach is an increasingly viable framework for general unsupervised learning across many visual tasks.

We've also found that performance across different types of visual tasks depends in an understandable fashion on how frame sampling is accomplished by the architecture, and that a two-pathway model with both static and dynamic pathways is relatively better than either alone. This result is, interestingly, consistent with observations from neuroscience, where it has been shown that both ventral stream [38] and dorsal stream [8] brain pathways contribute to visual performance, with the former being more sensitive to static objects and the

Dataset	Kinetics			ImageNet		
Layer	Conv3	Conv4	Conv5	Conv3	Conv4	Conv5
VIE-Single	23.84	38.25	44.41	25.02	40.49	42.33
70%-VIE-Single	26.18	38.87	43.59	23.05	39.63	39.85
30%-VIE-Single	25.54	37.49	40.72	23.33	38.49	36.23
2bin-VIE-Single	24.54	39.16	44.24	25.55	41.43	39.36
5bin-VIE-Single	25.17	38.73	43.33	23.90	40.46	37.83

Table 4: Top-1 accuracy (in %) of transfer learning to Kinetics and ImageNet from VIE-Single models trained using different amount of videos or with videos cut into different number of bins.

Figure 2: Video retrieval results for VIE-Single and VIE-Slowfast models from Kinetics validation set. GT=ground truth label, Pred=model prediction. For each query video, three nearest training neighbors are shown. Red font indicates an error.

latter more sensitive to dynamic stimuli.

However, there are a number of critical limitations in the current method that will need to be overcome in future work. Our choice of Local Aggregation for evaluation in this work should not taken as a statement that it is the only way to achieve useful unsupervised deep neural embeddings. In particular, exploring the use of other very recent unsupervised learning losses, such as CMC [28] and DIM [14], will also be of great interest. We expect many of these methods would be compatible with the general VIE framework, and would likely be usefully complementary to LA.

Another natural direction for improvement of the architecture is to investigate the use of recurrent neural network motifs [15, 25] and attentional mechanisms [30]. It is likely that improved architectures of these or some other type could better take advantage of the rich medium-range temporal structure (e.g. ∼1s-1m in length) of natural video sequences.

Further, our current results are likely impacted by limitations in the Kinetics dataset, especially for harnessing the importance of dynamic processing, since even in the supervised case, single-frame performance is comparatively high. Seeking out and evaluating VIE on additional datasets will be critical — perhaps most importantly, for applications involving large and previously uncurated video data where the potential impact of unsupervised learning is especially high. It will also be critical to test VIE in video task domains other than classification, including object tracking, dynamic 3D shape reconstruction and many others.

References

[1] Amir Akbarzadeh, J-M Frahm, Philippos Mordohai, Brian Clipp, Chris Engels, David Gallup, Paul Merrell, M Phelps, S Sinha, B Talton, et al. Towards urban 3d reconstruction from video. In *Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*, pages 1–8. IEEE, 2006.

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *arXiv preprint arXiv:1812.03982*, 2018.

[6] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5589–5597, 2018.

[7] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[8] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[9] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[10] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[12] Olivier J Hénaff, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

[13] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

[14] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-Supervised Spatiotemporal Feature Learning via Video Rotation Prediction. 2018.

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

[18] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles. 2018.

[19] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715, 2017.

[20] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[21] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017.

[22] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[23] Takashi Matsuyama, Xiaojun Wu, Takeshi Takai, and Shohei Nobuhara. Real-time 3d shape reconstruction, dynamic 3d mesh deformation, and high fidelity visualization for 3d video. *Computer Vision and Image Understanding*, 96(3):393–434, 2004.

[24] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[25] Aran Nayebi, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel L Yamins. Task-driven convolutional recurrent models of the visual system. In *Advances in Neural Information Processing Systems*, pages 5290–5301, 2018.

[26] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2701–2710, 2017.

[27] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

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

[29] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[31] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016.

[32] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised Spatiotemporal Representation Learning for Videos by Predicting Motion and Appearance Statistics. 2019.

[33] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

[34] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and Philip S Yu. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. *arXiv preprint arXiv:1804.06300*, 2018.

[35] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and S Yu Philip. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *Advances in Neural Information Processing Systems*, pages 879–888, 2017.

[36] Zhirong Wu, Alexei A Efros, and Stella X Yu. Improving generalization via scalable neighborhood component analysis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 685–701, 2018.

[37] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[38] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[39] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.

[40] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. *arXiv preprint arXiv:1903.12355*, 2019.