INCORPORATING SCALABILITY IN UNSUPERVISED SPATIO-TEMPORAL FEATURE LEARNING

Sujoy Paul, Sourya Roy and Amit K. Roy-Chowdhury

Dept. of Electrical and Computer Engineering, University of California, Riverside, CA 92521

ABSTRACT

Deep neural networks are efficient learning machines which leverage upon a large amount of manually labeled data for learning discriminative features. However, acquiring substantial amount of supervised data, especially for videos can be a tedious job across various computer vision tasks. This necessitates learning of visual features from videos in an unsupervised setting. In this paper, we propose a computationally simple, yet effective, framework to learn spatio-temporal feature embedding from unlabeled videos. We train a Convolutional 3D Siamese network using positive and negative pairs mined from videos under certain probabilistic assumptions. Experimental results on three datasets demonstrate that our proposed framework is able to learn weights which can be used for same as well as cross dataset and tasks.

***Index Terms*—** unsupervised, feature learning, spatio-temporal, scalable

1. INTRODUCTION

Large labeled datasets and computational power can be attributed as the main reason behind recent successes of Deep Neural Networks in various computer vision tasks [14, 23, 11, 7, 28, 9]. Unsupervised learning [2] of visual features from huge amount of unlabeled videos available today, using deep networks can be a potential solution to the data hungriness of supervised algorithms. Autoencoders, Restricted Boltzmann Machines (RBM) and the likes [3, 8, 26] trained in a greedy layer-wise fashion have been one of the popular methods for learning visual features from images in an unsupervised manner. However, such approaches fail to discover higher level structures from the data, necessary in recognition tasks.

Recent works in unsupervised feature learning from image data take a slightly different path. Most of the approaches belonging to this category first define a semantically challenging task such that its training instances can be directly extracted from the unlabeled data. For example, Pathak et al. [20] used motion information to learn visual representation of objects via a segmentation approach. Patch level puzzle solving in images has also been explored to learn visual features in [4, 19]. Objects tracked over time played a key role in providing the supervisory signal for visual feature learning in Wang and Gupta's work [29]. Ego-motion guided unsupervised learning has been explored in [1, 10]. Li et al. [16] proposed an image similarity based method using low level features to learn visual features of objects in an unsupervised setting. Unlike images, unsupervised spatio-temporal representation learning from videos has not been thoroughly studied in literature. Reconstruction and prediction has been used to learn features from videos using Long Short Term Memory (LSTM) networks in [25]. Misra et al. [17] exploited temporal ordering of frames in a video to learn features. It may be noted that in contrast to the previously mentioned approaches which are mainly applicable for learning only spatial features, the task of learning spatio-temporal features from videos in an unsupervised setting is significantly more challenging. The primary reason is that it is very difficult to define a tractable task for videos in the first place compared to images.

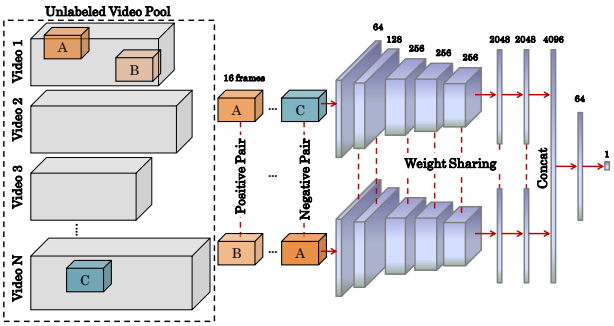

Fig. 1: This figure presents our proposed framework for unsupervised feature learning. A,B,C are three example spatio-temporal volumes. As the sub-volumes within a certain space-time boundary share similar semantic concepts, but possess different appearance and motion content, we select them as positive pairs. s-t volumes belonging to different videos constitute a negative pair (details in Section 2).

Visual continuity is prevalent in natural videos where semantic correlation between spatio-temporally associated volumes exist [5]. In this work, we build our hypothesis around this notion of spatio-temporal (s-t) correlatedness and argue that the likelihood of sharing higher semantic information between s-t related volumes is more compared to volumes from other videos. We structure our proposed framework based on this hypothesis to learn discriminative appearance and motion features in an entirely unsupervised manner using the 3D convolutional networks [27]. Our key contributions are below -

• We introduce a novel unsupervised feature learning framework by exploiting spatio-temporal relationships between intra and inter video segments.
• We propose an efficient strategy to mine positive and negative semantic pairs whose scalable nature makes our framework usable to ubiquitously present large unlabeled datasets.
• Finally, our proposed method can be integrated as a pre-training module with several supervised learning tasks.

2. METHODOLOGY

Our goal is to learn discriminative feature embedding of spatio-temporal volumes from unlabeled videos. In order to accomplish the task, we train a Siamese network based on the 3D CNN, which involves mining of positive and negative samples that can act as a supervisory data to learn semantically discriminative features. We device a simple pair mining strategy which is easy to implement and scalable to large datasets.

Siamese Network Training. Let us consider that we have a set of N labeled triplets $\{(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, y_i)\}_{i=1}^N$ where $\boldsymbol{x}_i^1, \boldsymbol{x}_i^2 \in \mathbb{R}^m$ and $y_i \in \{+1, -1\}$. In a Siamese network [6], generally the same function $\mathcal{T}$ is used to project $\boldsymbol{x}_i^1, \boldsymbol{x}_i^2$ to obtain a lower dimensional representation $\boldsymbol{f}_i^j = \mathcal{T}(\boldsymbol{x}_i^j; \boldsymbol{W}_1), j \in 1, 2$. They are converted to the desired output using another function $\mathcal{G}$. These outputs are compared with the ground-truth annotations to compute the loss $\mathcal{L}$, which needs to be minimized in order to learn the weights of the network for the particular task in hand. The optimal weights can be represented as,

$$\boldsymbol{W}_1^*, \boldsymbol{W}_2^* = \underset{\boldsymbol{W}_1, \boldsymbol{W}_2}{\arg\min} \sum_{i=1}^{N} \mathcal{L}(\mathcal{G}(\boldsymbol{f}_i^1, \boldsymbol{f}_i^2; \boldsymbol{W}_2), y_i) \quad (1)$$

This training strategy of Siamese networks enforces the transformation $\mathcal{T}$ to semantically group the training instances in the feature space, analogous to the perceptual grouping ability of human cognitive system. Although the process of obtaining binary labels demands lesser human effort compared to obtaining individual class labels, acquiring pair-wise labels still requires a lot of manual labeling effort. In our approach, we mine the positive and negative training pairs in an unsupervised manner as explained below.

Unsupervised Siamese Pair Mining. Consider a set of N unlabeled videos $\{\boldsymbol{V}_i\}_{i=1}^N$, such that $\boldsymbol{V}_i\{x, y, t, x_0, y_0, t_0\}$ denote the spatio-temporal volume of shape (x_0, y_0, t_0) at position (x, y, t) for the i^{th} video.

Positive Pair Mining. In the context of training a Siamese network, positive pair can be defined as a tuple consisting of two semantically similar instances. Generally, in natural videos, distinct s-t volumes within a certain boundary contain different appearance and motion features. To elaborate, arrangements of objects across different spatial segments may vary. Similarly, separate temporal portions capture different motion structure. Despite containing different s-t patterns, such volumes represents similar semantic entities. Fig. 2 demonstrates this idea.

Although, two separate s-t volumes within a certain boundary from the same video can be used as positive pairs, the network may not be able to generalize well for videos with

(a) (b)

Fig. 2: This figure illustrates the motivational concept behind our positive pair mining strategy in 2D. (a) and (b) are RGB and optical flow frame of a video. The bounding boxes denote a sample positive pair. As we can observe, the appearance and motion content of the two pairs differ, but they belong to the same semantic category of *Soccer Juggling*.

variations. To deal with this issue, we expand our pool of positive pairs by applying the following transformations- **1.** Color transform $\mathcal{F}_1$ in the HSV domain involving three parameters (discussed in Section 3), **2.** A non-parametric transformation $\mathcal{F}_2$ defined as XA, where A is an anti-diagonal matrix containing only 1s and X is an image. This operation flips the image in the horizontal direction. It may be noted that the same transformation is carried out in all the images of the spatio-temporal volume. Finally, the poitive pairs we use may be defined as

$$\{\boldsymbol{V}_i\{x^1, y^1, t^1, x_0, y_0, t_0\}, \mathcal{F}(\boldsymbol{V}_i\{x^2, y^2, t^2, x_0, y_0, t_0\})\}$$

where $\mathcal{F}$ is chosen probabilistically from the set of transformations $\{\mathcal{I}, \mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_1 \circ \mathcal{F}_2\}$ and $\mathcal{I}$ is the identity transformation. The probabilities associated with selection of each transformation can be found in Algorithm 1.

Negative Pair Mining. Our negative pair mining strategy is based on the following idea. Consider that the unlabeled video pool came from m distinct distributions representing semantic concepts and n_k, $k \in \{1, \ldots, m\}$ be the number of instances belonging to the k^{th} distribution. The maximum probability that a pair of s-t volumes extracted from two videos randomly from the entire dataset, belong to the same distribution is

$$p \leq \left(\frac{\max_k n_k}{\sum_{k=1}^m n_k}\right)^2 m$$

Assuming that $\max_k n_k << \sum_{k=1}^m n_k$, which may be the case in natural unlabeled video pool, $p_{max} \to 0$. With this assumption, the negative pairs may be defined as,

$$\{\boldsymbol{V}_i\{x^1, y^1, t^1, x_0, y_0, t_0\}, \mathcal{F}(\boldsymbol{V}_{j \neq i}\{x^2, y^2, t^2, x_0, y_0, t_0\})\}$$

Learning Spatio-Temporal Features. Learning spatio-temporal features for videos is important for several recognition tasks in computer vision. Convolutional 3D (C3D) [27] network have been successful in learning both motion and appearance features. We use the smaller C3D network defined in their paper as a transformation $\mathcal{T}$ in Eqn. 1. The output of this transformation is feature vectors $\boldsymbol{f}^1, \boldsymbol{f}^2 \in \mathbb{R}^{2048}$ for the two input s-t volumes of the Siamese network. We concatenate the two feature vectors to obtain a single feature vector $\in \mathbb{R}^{4096}$. In order to learn the transformation $\mathcal{G} : \mathbb{R}^{4096} \to \mathbb{R}$ in Eqn. 1, we use two fully connected (fc) layers such that output of first

fc layer $\in \mathbb{R}^{64}$. Finally, we minimize the hinge loss [21] for binary classification, which may be presented as,

$$\mathcal{L} = \sum_{i=1}^{B} C_i \max(0, 1 - t_i y_i) + \lambda ||\boldsymbol{W}||_F \qquad (2)$$

where t_i is the scalar output of the siamese network for the i^{th} pair in the batch and $\boldsymbol{W}$ represents all the weights of the network. The second part of the equation is the regularization term and we set $\lambda = 0.0005$ in our experiments. B is the mini-batch size of Stochastic Gradient Descent. C_i is $\frac{1}{N_p}$ and $\frac{1}{N_n}$ respectively for number of positive and negative samples in the mini-batch. t is the final scalar output of the Siamese network. We use dropout of 0.5 in the fc layers, except the final output layer.

Algorithm 1 Online Pair Mine Algorithm

Input: 1. Unlabeled video dataset $\{\boldsymbol{V}_i\}_{i=1}^{N}$,
2. Positve Pair Ratio (p)
Output: Training data batch $\{\boldsymbol{S}_i^1, \boldsymbol{S}_i^2, y_i\}_{i=1}^{B}$
1. $N_p = \text{round}(pB)$, $N_n = B - N_p$, $i \leftarrow 1$
while $i \leq N_p$ **do**
  $m \sim \mathcal{U}[1, N]$
  $\boldsymbol{S}_i^1,\ \boldsymbol{S}_i^2 \leftarrow \text{STVolume}(\boldsymbol{V}_m),\ \text{STVolume}(\boldsymbol{V}_m)$
  if $\boldsymbol{S}_i^1 \cap \boldsymbol{S}_i^2 \neq \Phi$ **then**
    Goto Step **5**
  end if
  $y_i \leftarrow +1$, $i \leftarrow i + 1$
end while
while $i \leq N_n$ **do**
  $m \sim \mathcal{U}[1, N]$, $n \sim \mathcal{U}\{[1, N] - m\}$
  $\boldsymbol{S}_i^1,\ \boldsymbol{S}_i^2 \leftarrow \text{STVolume}(\boldsymbol{V}_m),\ \text{STVolume}(\boldsymbol{V}_n)$
  if $r_1 \sim \mathcal{N}(0, 1) > 0$ **then**
    $\boldsymbol{S}_i^2 \leftarrow \mathcal{F}_1(\boldsymbol{S}_i^2)$
  end if
  if $r_2 \sim \mathcal{N}(0, 1) > 0$ **then**
    $\boldsymbol{S}_i^2 \leftarrow \mathcal{F}_2(\boldsymbol{S}_i^2)$
  end if
  $y_i \leftarrow -1$, $i \leftarrow i + 1$
end while

Note: STVolume($\boldsymbol{V}$) is a function which randomly extracts a spatio-temporal volume from the video $\boldsymbol{V}$.

3. EXPERIMENTS AND RESULTS

In this section, we present results and analysis of the proposed unsupervised feature learning algorithm. We mainly focus on the activity recognition and video similarity classification.

Unsupervised Learning. We use the UCF101 [24] dataset for training our C3D Siamese network in an unsupervised manner. We follow Algorithm 1 to mine the pairs required for training our C3D Siamese network. To optimize the loss function, we use Adam Optimizer [12] in a Stochastic Gradient Descent setting with mini-batch size of 10 and 30-70% split in positive and negative samples respectively.

Color Transformations. The color transformation $\mathcal{F}_1$ mentioned in Section 2 are as follows. **1.** Adding a random number $\in (-0.1, 0.1)$ to all the pixels of an image, **2.** Taking element-wise exponent of the Saturation and Value component with a random number $\in (0.5, 2)$, **3.** Scaling the Saturation and Value components by a random number $\in (0.7, 2)$. All random numbers are generated from uniform distributions.

Feature Representation Visualization. We visualize the $pool5$ feature response of our unsupervised C3D network to identify the regions it detect as semantically relevant. We visualize only single frame in Fig. 3. We obtain 256 activation maps from the pool5 layer. Then, we average over all the receptive fields corresponding to the units with maximum response from each activation map. Finally, we segment the averaged receptive fields to obtain the bounding boxes. Results depict that our network is able to identify areas of the image which involve human-object interaction and motion.

Nearest Neighbor. In order to understand the semantic concepts learned by our network trained in an unsupervised fashion, we retrieve the nearest neighbor of query videos. For each video of the UCF101 dataset, we extract 10 random s-t volumes and pass it through the learned network to obtain 10 feature vectors $\in \mathbb{R}^{2048}$ from the $fc7$ layer followed by mean pooling to obtain a single feature vector. Then, given a query video, we find its nearest neighbor in the feature space. Sample results of the nearest neighbor video retrieval task are presented in Fig. 4. This shows that although the positive training pairs belong to different spatio-temporal volumes of the same video of UCF101, the network is able to generalize to similar semantic concepts belonging to different videos.

Finetuning. A deep neural network is generally trained on a large number of labeled instances and then finetuned on a smaller task-specific dataset [22]. In scenarios where constrained budget limit the acquisition of labeled samples, we may have to train the network using only a small task-specific dataset, which may not yield good performance. In this section, we demonstrate that the proposed method can be used to first learn a feature embedding in an unsupervised way from a large unlabeled dataset, and then the network weights can be finetuned on a smaller task-specific dataset to obtain better performance compared to a randomly initialized network. We also show that unsupervised feature learning using our framework on a dataset followed by finetuning with labels on the same dataset can perform better compared to supervised training from randomly initialized weights. In this section, we use the HMDB51 dataset [15] for activity recognition which is a smaller dataset than UCF101.

Results. We use the learned weights from our unsupervised Siamese network, trained on UCF101, to finetune using the HMDB51 dataset. We also train the C3D network from randomly initialized weights using the HMDB51 dataset. We compare our results against four baseline methods for unsupervised feature learning, which are - Temporal Coherence (TCH) [18], Invariant Mapping (IM) [6], Object Patch (OP) [29] and Shuffle & Learn (S&L) [17]. The results are presented in Table 1. Better performance of the proposed method over the randomly initialized network suggests that the pro-

Fig. 3: This figure presents the top response region of pool5 of our network learned from unlabeled data.

Table 1: Activity recognition accuracy on UCF101 and HMDB51. The column with C3D indicate results when the network is trained using randomly initialized weights. 'Ours' for HMDB51 mean UCF101 Unsupervised + HMDB51 Supervised, and that for UCF101 mean UCF101 Unsupervised + UCF101 Supervised.

Algorithms →	Chance	C3D [27]	STIP [24]	TCH [18]	IM [6]	OP [29]	S&L [17]	Ours
HMDB51	1.9	19.2	20.0	15.9	16.3	15.6	18.1	**20.6**
UCF101	01.0	44.1	43.9	45.4	45.7	40.7	**50.2**	48.5

Fig. 4: In this figure we present the nearest neighbors of a query video using the features learned by our unsupervised network.

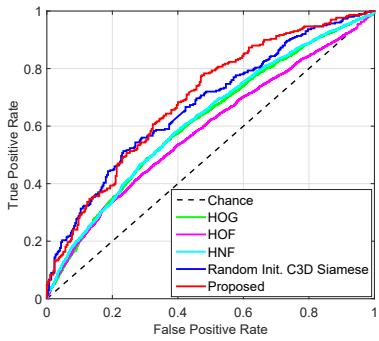

Fig. 5: ROC curve on ASLAN dataset

Table 2: Pair mining computational time for a training sample

Algorithms	Optical Flow	Ours
Time per pair (in second)	8.0×10^{-2}	1.8×10^{-5}

posed unsupervised learning framework can serve as a module to structure the feature space for superior supervised learning performance. We also finetune the network weights learned in an unsupervised manner using UCF101 class labels. We compare with other unsupervised baselines mentioned previously. The results are presented in Table 1. As can be observed that, the proposed method performs better than randomly initialized C3D by a **margin of 4.4%**, which clearly indicates that our method can be used to enhance the performance of video related supervised tasks. Our method also performs better than other baselines except S&L. However, S&L and other works in literature involve optical flow in their pair mining strategy which adds to the computational time. On the other hand, our randomized pair mining strategy is computationally efficient and scalable to large datasets. Table 2 presents this comparison.

Table 3: Action Similarity results on ASLAN

Algo.	C3D [27]	HOG [13]	HOF [13]	HNF [13]	Ours
Acc.	57.3	56.6	56.8	58.9	**63.5**
AUC	67.2	61.6	58.5	62.1	**69.3**

Action Similarity Classification Our proposed unsupervised framework learns similarities in videos which can be used to solve the problem of similarity labeling, given a pair of videos. In this section, we explore, whether our network, learned in an unsupervised manner, can help to achieve better performance compared to randomly initialized network on this task. We use the Action Similarity Labeling (ASLAN) dataset [13] for this experiment.

Results. We present the results of our network on ASLAN dataset along with the result obtained after training the Siamese C3D network from randomly initialized weights in Table 3. We also compare the results with other baselines from existing literature [13]. The ROC curve is presented in Fig. 5. It is evident from the experimental results that the proposed method performs better than randomly initialized Siamese C3D network which suffers from the scarcity of training data.

4. CONCLUSION

In this work we present a novel approach to learn spatio-temporal feature learning from unlabeled videos. Experimental results suggest that the embeddings learned by our framework are transferable to new datasets and can be finetuned to achieve superior performance than training with random initialization using the new dataset. Furthermore, the performance of supervised learning on a certain dataset can be improved by first using our unsupervised learning scheme on the same dataset.

Acknowledgment. This work was partially supported by NSF grants IIS-21185 and CNS-33218.

5. REFERENCES

[1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *ICCV*, pages 37–45, 2015.

[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *PAMI*, 35(8):1798–1828, 2013.

[3] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, et al. Greedy layer-wise training of deep networks. *NIPS*, 19:153, 2007.

[4] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015.

[5] D. W. Dong and J. Atick. Spatiotemporal coupling and scaling of natural images and human visual sensitivities. *NIPS*, pages 859–865, 1997.

[6] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pages 1735–1742. IEEE, 2006.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[8] F. J. Huang, Y.-L. Boureau, Y. LeCun, et al. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *CVPR*, pages 1–8. IEEE, 2007.

[9] S. Huang, X. Li, Z. Zhang, F. Wu, S. Gao, R. Ji, and J. Han. Body structure aware deep crowd counting. *IEEE Transactions on Image Processing*, 27(3):1049–1059, 2018.

[10] D. Jayaraman and K. Grauman. Learning image representations tied to ego-motion. In *ICCV*, pages 1413–1421, 2015.

[11] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[12] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] O. Kliper-Gross, T. Hassner, and L. Wolf. The action similarity labeling challenge. *PAMI*, 34(3):615–621, 2012.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE, 2011.

[16] D. Li, W.-C. Hung, J.-B. Huang, S. Wang, N. Ahuja, and M.-H. Yang. Unsupervised visual representation learning by graph-based consistent constraints. In *ECCV*, pages 678–694. Springer, 2016.

[17] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, pages 527–544. Springer, 2016.

[18] H. Mobahi, R. Collobert, and J. Weston. Deep learning from temporal coherence in video. In *ICML*, pages 737–744. ACM, 2009.

[19] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016.

[20] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning features by watching objects move. 2017.

[21] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

[22] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPRW*, pages 806–813, 2014.

[23] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[24] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[25] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. In *ICML*, pages 843–852, 2015.

[26] E. W. Tramel, M. Gabrié, A. Manoel, F. Caltagirone, and F. Krzakala. A deterministic and generalized framework for unsupervised learning with restricted boltzmann machines. *arXiv preprint arXiv:1702.03260*, 2017.

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[28] M. Tu and X. Zhang. Speech enhancement based on deep neural networks with skip connections. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5565–5569. IEEE, 2017.

[29] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, pages 2794–2802, 2015.