Pixel-level Correspondence for Self-Supervised Learning from Video

Yash Sharma [† 1] **Yi Zhu** [2] **Chris Russell** [2] **Thomas Brox** [2 3]

Abstract

While self-supervised learning has enabled effective representation learning in the absence of labels, for vision, video remains a relatively untapped source of supervision. To address this, we propose Pixel-level Correspondence (PICO), a method for dense contrastive learning from video. By tracking points with optical flow, we obtain a correspondence map which can be used to match local features at different points in time. We validate PICO on standard benchmarks, outperforming self-supervised baselines on multiple dense prediction tasks, without compromising performance on image classification.

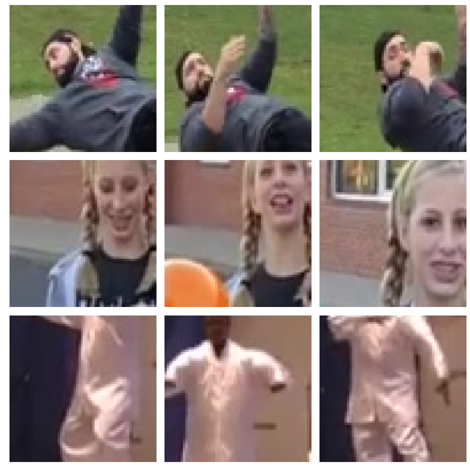

Figure 1. Each row shows patches along point trajectories computed on Kinetics-400, resized for viewing. Clearly, hand-crafted transformations (crops, color distortion) used in practice cannot capture the variation depicted here.

1. Introduction

Deep learning methods have yielded dramatic improvements in a plethora of domains by extracting useful representations from raw data (Bengio et al., 2013; LeCun et al., 2015), albeit assuming the availability of ample supervision. Recent advancements in self-supervised learning (Mikolov et al., 2013; Devlin et al., 2018; Chen et al., 2020a; He et al., 2021) have enabled effective representation learning without curated, labeled datasets (Goyal et al., 2021).

Self-supervised learning obtains supervisory signals from the data itself through the careful construction of prediction tasks which do not rely on manual annotation, yet encourage the model to extract useful features. Specifically, the task of predicting whether a pair, or a set, of examples are views of the "same" image, or "different" images, underlies the recent success of contrastive methods for learning representations of visual data (Wu et al., 2018; Van den Oord et al., 2018; Henaff, 2020; Hjelm et al., 2018; Bachman et al., 2019; He et al., 2020; Chen et al., 2020a).

In contrastive learning, view selection crucially influences the quality of the resulting representations (Tian et al., 2020; Zimmermann et al., 2021; Von Kügelgen et al., 2021). Existing approaches (He et al., 2020; Chen et al., 2020a;b) have constructed views via hand-crafted data augmentations, e.g. cropping sub-regions of the images. Cropping yields views that depict object parts, and thereby induces a learning signal for invariance to occluded objects (Purushwalkam & Gupta, 2020). With that said, augmentations are inherently limited; given a single image, simulating variation in object size, shape, or viewpoint can be difficult. Notably, such variation is ubiquitous in video (see Figure 1). The promise of temporal variation for representation learning has encouraged ample investigation in the context of self-supervision (Misra et al., 2016; Wei et al., 2018; Wang et al., 2019; Vondrick et al., 2018; Isola et al., 2015; Wiskott & Sejnowski, 2002; Klindt et al., 2020; Agrawal et al., 2015; Weis et al., 2021; Lachapelle et al., 2021).

How can we leverage video for learning self-supervised representations of images? While existing work has proposed a multitude of strategies (Wang & Gupta, 2015; Wang et al., 2017; Tschannen et al., 2020; Purushwalkam & Gupta, 2020; Gordon et al., 2020; Romijnders et al., 2021; Xiong et al., 2021; Wu & Wang, 2021; Chen et al., 2021), nearly all exploit instance discrimination methods (Dosovitskiy et al., 2014; Kolesnikov et al., 2019; He et al., 2020; Chen et al., 2020b) designed for global representation learning, or learning encodings at the image-level. However, recent work (Pinheiro et al., 2020; Wang et al., 2021; Xie et al.,

[†] Work done during an internship at Amazon. [1] University of Tübingen [2] Amazon [3] University of Freiburg. Correspondence to: Yash Sharma <yash.sharma@bethgelab.org>.

First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022, Baltimore, Maryland, USA, PMLR 162, 2022.

2021; Xiao et al., 2021; Bai et al., 2022) has demonstrated that dense representation learning, or learning encodings at the region/pixel-level, can improve performance for dense prediction tasks (e.g. segmentation, depth prediction), at the cost of reduced performance for global prediction tasks (e.g. image classification) (Xie et al., 2021; Xiao et al., 2021). Note that this observation echoes the related findings from empirical studies that ImageNet accuracy is not predictive for downstream tasks outside of image/scene classification (Kotar et al., 2021; Atanov et al., 2022).

We thus propose PiCo, a method for dense representation learning from video. Existing work proposed for static images has relied upon aforementioned geometric transformations, e.g. crops, to introduce variation. We demonstrate that temporal variation can also be utilized by tracking points using off-the-shelf optical flow estimators. We find that across a number of downstream tasks, PiCo outperforms existing work restricted to static frames, as well as existing work applied to video assuming static pixel correspondence.

2. Background

As our contribution enables dense representation learning to exploit the natural transformations inherent to video, we will focus on extending a method which learns representations through pixel-level contrastive learning, VADeR (Pinheiro et al., 2020). Thus, we will give a short description of the learning method before proceeding with our contribution, see (Pinheiro et al., 2020) for further details.

Let us represent a pixel u in image $\mathbf{x} \in \mathcal{I} \subset \mathbb{R}^{3\times h\times w}$ by the tuple $(\mathbf{x}, u)$. Let f be an encoder-decoder convolutional network that produces a d-dimensional embedding for every pixel in the image, i.e. $f : (\mathbf{x}, u) \mapsto z \in \mathbb{R}^d$. VADeR's objective is to learn an embedding function that encodes $(\mathbf{x}, u)$ into a representation that is invariant w.r.t. any view $v_1, v_2 \in \mathcal{V}_u$ containing the pixel u. This is achieved through contrastive learning (Gutmann & Hyvärinen, 2010; 2012; Van den Oord et al., 2018), where the objective optimized in practice is to distinguish between views of the same pixel and views of different pixels.

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}_{(v_1,v_2)\sim\mathcal{V}_u}\left[\log \frac{\exp\{\text{sim}(f(v_1,u),f(v_2,u))\}}{\sum_{j=1}^{K}\exp\{\text{sim}(f(v_1,u),f(v'_j,u'_j))\}}\right] \tag{1}$$

where $K-1$ is the number of negative pixels, and the positive pair in the numerator is included in the denominator summation, i.e. $(v'_K, u'_K) = (v_2, u)$. For implementation, the design details for MoCo were followed (He et al., 2020).

For f, the semantic segmentation branch of (Kirillov et al., 2019) was adopted. A feature pyramid network (FPN) (Lin et al., 2017) adds a top-down path to a ResNet-50 (He et al., 2016), generating a pyramid of features (from $1/32$ to $1/4$ resolution). By adding a number of upsampling blocks at each resolution of the pyramid, the pyramid representations are merged into a single dense output representation with dimension 128 and scale $1/4$. The ResNet-50 is initialized with MoCo (He et al., 2020), and pretraining is performed on the ImageNet-1K (IN-1K) (Deng et al., 2009) train split.

3. Method

Here, we detail our procedure for constructing pixel correspondence maps from video for dense contrastive learning.

3.1. Data

For pretraining, we experiment with Kinetics400 (K400) (Kay et al., 2017) and YouTube-8M (YT8M) (Abu-El-Haija et al., 2016). The K400 training set consists of approximately 240,000 videos trimmed to 10 seconds from 400 human action categories. We sample frame sequences at 30 Hz (Kuang et al., 2021). For tractability, we construct a subset of YT8M (YT8M-S) which matches the dataset statistics of K400. Specifically, for 240,000 random videos, we sample 10-second snippets at 30 Hz from shots detected using an off-the-shelf network (Souček & Lokoč, 2020). Further details are provided in the appendix.

3.2. Trajectories

We first compute and store optical flow on K400 and YT8M-S. While in preliminary experiments we found alternatives (Ilg et al., 2017; Sun et al., 2018) to perform comparably, for the presented set of experiments, we use RAFT (Teed & Deng, 2020) trained on a mixed dataset (Contributors, 2021) consisting of FlyingChairs (Dosovitskiy et al., 2015), FlyingThings3D (Mayer et al., 2016), Sintel (Butler et al., 2012), KITTI-2015 (Menze & Geiger, 2015; Geiger et al., 2013), and HD1K (Kondermann et al., 2016). The horizontal and vertical components of the flow were linearly rescaled to a $[0, 255]$ range and compressed using JPEG (after decompression, the flow is rescaled back to its original range) (Simonyan & Zisserman, 2014).

With the precomputed flow, we track points in the video. For each video, we sample an initial set of 1000 points at random locations on random frames. As in (Sundaram et al., 2010), each point is tracked to the next frame using the flow field $\mathbf{w} = (u, v)^T$:

$$(x_{t+1}, y_{t+1})^T = (x_t, y_t)^T + (u_t(x_t, y_t), v_t(x_t, y_t))^T \tag{2}$$

Between pixels, the flow is inferred using bilinear interpolation. Tracking is stopped as soon as a point is occluded, which is detected by checking the consistency of the forward and backward flow. In a non-occlusion case, the backward flow vector should point in the inverse direction of the forward flow vector: $u_t(x_t, y_t) = -\hat{u}_t(x_t + u_t, y_t + v_t)$ and $v_t(x_t, y_t) = -\hat{v}_t(x_t + u_t, y_t + v_t)$, where $\hat{\mathbf{w}}_t = (\hat{u}_t, \hat{v}_t)$

denotes the flow from frame $t+1$ to frame t. We thus use the following threshold:

$$|\mathbf{w}+\hat{\mathbf{w}}|^2 < \gamma(|\mathbf{w}|^2+|\hat{\mathbf{w}}|^2)+\delta \quad (3)$$

3.3. Learning

Existing proposals for visual representation learning with contrastive methods from video typically sample random frames from a given shot for constructing views (Tschannen et al., 2020; Chen et al., 2021; Gordon et al., 2020). Given the endpoints for a set of trajectories in each video, we propose a frame selection strategy for maximizing temporal separation and trajectory density, **anchor sampling**. After sampling an anchor frame, for each trajectory active on said frame, we find the endpoint furthest from said frame. If we are to select N frames for learning, we select the top-N according to endpoint count. With this strategy, as we vary the threshold hyperparameters, the temporal separation between the selected frames varies accordingly.

3.4. Implementation Details

For both implementing the objective and initializing the encoder, we use MoCo-v2 (Chen et al., 2020b;a) instead of MoCo (He et al., 2020). Notably, we found the use of a nonlinear projection head to be critical for performance. As in existing work (Long et al., 2015; Wang et al., 2021; Bai et al., 2022), for dense contrastive learning, we replace the linear layers in the MoCo-v2 global projection head with identical 1×1 convolution layers. Note that we use 2048-dimensional hidden layers for the projector in concert with the 128-dimensional dense output representation; we found that reducing the number of parameters in the dense projection head decreased downstream performance. We also decided to freeze the initialized encoder, thereby maintaining the downstream image/scene classification performance of the image-level encoding.

Finally, note that for the experiments prior to ablation, $\gamma=0$, $\delta=4.0$, and, each iteration, no more than 65536 point pairs (from frame pairs selected from 256 videos) are used.

4. Experiments

We compare PiCo to a set of baselines across datasets & tasks. We provide a visual representation of the comparison in Figure 2. In **Static (Frames)**, a single frame is sampled from each video for view construction, thus, as in (Pinheiro et al., 2020), the variation in corresponding pixels is solely due to the geometric transformations in the MoCo-v2 data augmentation pipeline, i.e. random crops and horizontal flips. In **Static (Video)**, as in PiCo, we sample multiple frames from a given video, but unlike PiCo, the pixel correspondence map is static, i.e. the optical flow field is assumed to consist of zero vectors. By comparing to "Static (Video)", we can isolate the value point tracking is yielding downstream. Additional details regarding evaluation are provided in the appendix.

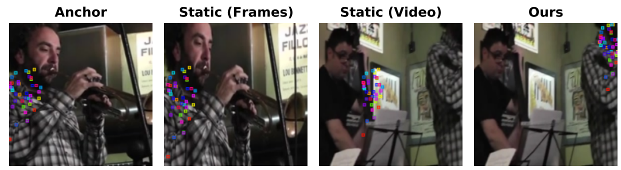

Figure 2. For **Static (Frames)**, variation between local features stems solely from geometric transformations, e.g. random crops. For video, without point tracking, the static pixel correspondence map used by **Static (Video)** becomes imprecise with increased temporal separation. In contrast, we leverage off-the-shelf optical flow estimators to match local features over time.

Table 1. **Linear probe.** COCO semantic segmentation.

Method	Dataset	mIoU	fIoU
MoCo-v2 (IN-1K)		11.2	39.9
Static (Frame)	K400	18.4	49.5
Static (Video)	K400	18.7	49.5
PiCo	K400	**20.9**	**51.6**
Static (Frame)	YT8M-S	17.0	48.3
Static (Video)	YT8M-S	18.3	49.6
PiCo	YT8M-S	**19.9**	**50.5**

4.1. COCO Semantic Segmentation

It is common practice in self-supervised learning to assess the quality of frozen features with a **linear probe** (Goyal et al., 2019; Kolesnikov et al., 2019). Following (Pinheiro et al., 2020), the output of each model is processed by a a 1×1 convolutional layer, $4\times$ upsample, and softmax, where for MoCo-v2, the effective stride is reduced from $1/32$ to $1/4$ by replacing strided convolutions with dilated ones (Chen et al., 2017; Yu et al., 2017). The linear predictor weights are trained using cross entropy.

In Table 1, we observe a tangible improvement in leveraging point trajectories for dense contrastive learning. Interestingly, we find pretraining on K400 largely delivers improved performance relative to YT8M-S. In accordance with previous work (Gordon et al., 2020), we notice that a number of videos in YT8M are unnatural, e.g. "video games" or "cartoons", which clearly yields a domain gap with "everyday scenes containing common objects in their natural context" (Lin et al., 2014).

4.2. Additional Tasks & Benchmarks

Tasks: In Table 2, we evaluate representations on COCO object detection and instance segmentation. For this, we use Mask R-CNN (He et al., 2017) with a frozen FPN backbone (Lin et al., 2017). While PiCo significantly outper-

Table 2. **Mask R-CNN**. K400 pretraining, COCO object detection & instance segmentation with FPN frozen.

Method	Dataset	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Static (Frame)	K400	6.09	15.1	3.73	7.37	14.8	6.64
Static (Video)	K400	7.92	19.7	4.60	9.06	18.7	7.72
PiCo	K400	**10.8**	**24.3**	**7.94**	**11.9**	**23.1**	**10.9**

Table 3. **Additional Benchmarks**: K400 pretraining, linear probing frozen model.

	sem. seg. (mIoU)		depth (RMSE)
Method	VOC	CS	NYU-d v2
Static (Frame)	31.0	34.0	1.000
Static (Video)	34.7	28.7	0.958
PiCo	**35.6**	**35.1**	**0.950**

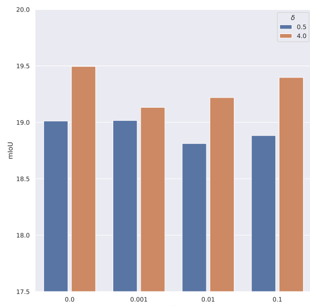

Figure 3. **γ:** COCO linear probing varying the tracking threshold parameters γ and δ. YT-8M-S pretraining, w/o anchor sampling.

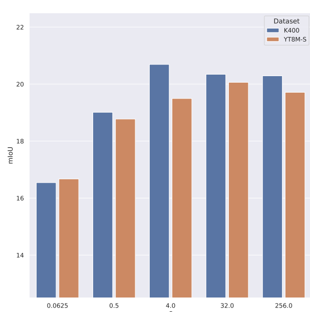

Figure 4. **δ**: COCO linear probing with finer-grained variation in δ. $\gamma = 0$, w/o anchor sampling.

forms the baselines, without being able to adapt the backbone downstream, the absolute scores are low.

Benchmarks: In Table 3, we evaluate on two additional datasets for semantic segmentation (Pascal VOC 2012 (Everingham et al., 2010) and Cityscapes (Cordts et al., 2016)), as well as a dataset for depth prediction (NYU-depth v2 (Silberman et al., 2012)). For depth prediction, we no longer apply the softmax function, and instead minimize the L_1 loss between the linear output and the per-pixel ground-truth depth values (Kotar et al., 2021; Pinheiro et al., 2020). We find that across all tasks, PiCo outperforms the baselines.

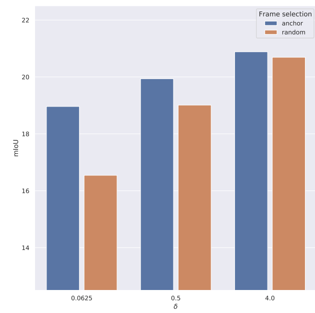

Figure 5. **Anchor Sampling:** COCO linear probing w/ and w/o the anchor sampling strategy. K400 pretraining.

4.3. Ablations

Tracking Threshold: In Figure 3, we evaluate the impact of varying γ and δ. While we do consistently observe improved performance with increased δ (up to a point, see Figure 4), the same cannot be said for γ. Given the computational cost in adjusting the trajectories w.r.t. γ, we were limited in our ablation, and encourage further exploration on the effect of this parameter.

Anchor Sampling: In Figure 5, we isolate the effect of anchor sampling on the downstream performance. We can see that as we increase δ, thereby using longer trajectories for pretraining, the gap between anchor sampling and randomly sampling frames narrows. As δ increases, the likelihood that a point pair will exist between a random pair of frames also increases, while the very same likelihood is invariant to δ when using anchor sampling.

5. Discussion

Limitations While we observe improved performance over our baseline methods, overall performance remains worse than supervised approaches, and there is substantial room for improvement. Specifically, our decoder-only training on video, when compared to the reported scores of encoder-decoder training on IN-1K in a similar experimental setting (Pinheiro et al., 2020), underperforms. In future work, we suggest (i) addressing the domain gap between the video datasets used for pretraining and the image datasets used for benchmarking (Tang et al., 2012; Kalogeiton et al., 2016; Kae & Song, 2020) and (ii) considering alternatives to our strategy of freezing the encoder for maintaining classification performance whilst improving dense prediction.

Conclusion We present PiCo, an approach to dense contrastive learning on video. We show the benefit in constructing pixel correspondence maps over time on a number of tasks and datasets. Our work serves as a first step towards leveraging the temporal variation inherent to video for dense prediction tasks, and in that vein, we encourage further exploration along the aforementioned direction.

Acknowledgements

We thank Peter Gehler, Andrii Zadaianchuk, and Max Horn for compute infrastructure support. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. Y.S. thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support.

References

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

Agrawal, P., Carreira, J., and Malik, J. Learning to see by moving. In *Proceedings of the IEEE international conference on computer vision*, pp. 37–45, 2015.

Atanov, A., Xu, S., Beker, O., Filatov, A., and Zamir, A. Simple control baselines for evaluating transfer learning. *arXiv preprint arXiv:2202.03365*, 2022.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Bai, Y., Chen, X., Kirillov, A., Yuille, A., and Berg, A. C. Point-level region contrast for object detection pre-training. *arXiv preprint arXiv:2202.04639*, 2022.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Butler, D. J., Wulff, J., Stanley, G. B., and Black, M. J. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pp. 611–625. Springer, 2012.

Chen, B., Selvaraju, R. R., Chang, S.-F., Niebles, J. C., and Naik, N. Previts: Contrastive pretraining with video tracking supervision. *arXiv preprint arXiv:2112.00804*, 2021.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Contributors, M. MMFlow: Openmmlab optical flow toolbox and benchmark. https://github.com/open-mmlab/mmflow, 2021.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., and Brox, T. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., and Brox, T. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2758–2766, 2015.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88 (2):303–338, 2010.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Gordon, D., Ehsani, K., Fox, D., and Farhadi, A. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020.

Goyal, P., Mahajan, D., Gupta, A., and Misra, I. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the ieee/cvf International Conference on computer vision*, pp. 6391–6400, 2019.

Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Gutmann, M. U. and Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of machine learning research*, 13(2), 2012.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pp. 4182–4192. PMLR, 2020.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462–2470, 2017.

Isola, P., Zoran, D., Krishnan, D., and Adelson, E. H. Learning visual groups from co-occurrences in space and time. *arXiv preprint arXiv:1511.06811*, 2015.

Kae, A. and Song, Y. Image to video domain adaptation using web supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 567–575, 2020.

Kalogeiton, V., Ferrari, V., and Schmid, C. Analysing domain shift factors between videos and images for object detection. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2327–2334, 2016.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Kirillov, A., Girshick, R., He, K., and Dollár, P. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6399–6408, 2019.

Klindt, D. A., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2020.

Kolesnikov, A., Zhai, X., and Beyer, L. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1920–1929, 2019.

Kondermann, D., Nair, R., Honauer, K., Krispin, K., Andrulis, J., Brock, A., Gussefeld, B., Rahimimoghaddam, M., Hofmann, S., Brenner, C., et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 19–28, 2016.

Kotar, K., Ilharco, G., Schmidt, L., Ehsani, K., and Mottaghi, R. Contrasting contrastive self-supervised representation learning pipelines. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9949–9959, 2021.

Kuang, H., Zhu, Y., Zhang, Z., Li, X., Tighe, J., Schwertfeger, S., Stachniss, C., and Li, M. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3195–3204, 2021.

Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., Le Priol, R., Lacoste, A., and Lacoste-Julien, S. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *First Conference on Causal Learning and Reasoning*, 2021.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., and Brox, T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4040–4048, 2016.

Menze, M. and Geiger, A. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3061–3070, 2015.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Misra, I., Zitnick, C. L., and Hebert, M. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pp. 527–544. Springer, 2016.

Pinheiro, P. O., Almahairi, A., Benmalek, R., Golemo, F., and Courville, A. C. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33:4489–4500, 2020.

Purushwalkam, S. and Gupta, A. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems*, 33:3407–3418, 2020.

Romijnders, R., Mahendran, A., Tschannen, M., Djolonga, J., Ritter, M., Houlsby, N., and Lucic, M. Representation learning from videos in-the-wild: An object-centric approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 177–187, 2021.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Simonyan, K. and Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

Souček, T. and Lokoč, J. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020.

Sun, D., Yang, X., Liu, M.-Y., and Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8934–8943, 2018.

Sundaram, N., Brox, T., and Keutzer, K. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European conference on computer vision*, pp. 438–451. Springer, 2010.

Tang, K., Ramanathan, V., Fei-Fei, L., and Koller, D. Shifting weights: Adapting object detectors from image to video. *Advances in Neural Information Processing Systems*, 25, 2012.

Teed, Z. and Deng, J. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

Tschannen, M., Djolonga, J., Ritter, M., Mahendran, A., Houlsby, N., Gelly, S., and Lucic, M. Self-supervised learning of video-induced visual invariances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13806–13815, 2020.

Van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv e-prints*, pp. arXiv–1807, 2018.

Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. *Advances in Neural Information Processing Systems*, 34, 2021.

Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., and Murphy, K. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 391–408, 2018.

Wang, X. and Gupta, A. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2015.

Wang, X., He, K., and Gupta, A. Transitive invariance for self-supervised visual representation learning. In *Proceedings of the IEEE international conference on computer vision*, pp. 1329–1338, 2017.

Wang, X., Jabri, A., and Efros, A. A. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2566–2576, 2019.

Wang, X., Zhang, R., Shen, C., Kong, T., and Li, L. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021.

Wei, D., Lim, J. J., Zisserman, A., and Freeman, W. T. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8052–8060, 2018.

Weis, M. A., Chitta, K., Sharma, Y., Brendel, W., Bethge, M., Geiger, A., and Ecker, A. S. Benchmarking unsupervised object representations for video sequences. *Journal of Machine Learning Research*, 22(183):1–61, 2021.

Wiskott, L. and Sejnowski, T. J. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.

Wu, H. and Wang, X. Contrastive learning of image representations with cross-video cycle-consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10149–10159, 2021.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Xiao, T., Reed, C. J., Wang, X., Keutzer, K., and Darrell, T. Region similarity representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10539–10548, 2021.

Xie, Z., Lin, Y., Zhang, Z., Cao, Y., Lin, S., and Hu, H. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16684–16693, 2021.

Xiong, Y., Ren, M., Zeng, W., and Urtasun, R. Self-supervised representation learning from flow equivariance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10191–10200, 2021.

Yu, F., Koltun, V., and Funkhouser, T. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 472–480, 2017.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pp. 12979–12990. PMLR, 2021.

A. Additional Details

A.1. Training

YT8M-S Given the size of YT-8M, the authors decided to release frame-level features of the videos instead of the videos themselves (Abu-El-Haija et al., 2016). For our purposes, we extracted YT-8M URLs[1], and downloaded a sampled subset at the scale of K-400. We note that a number of the YT-8M URLs are no longer accessible. We used TransNetV2 (Souček & Lokoč, 2020) off-the-shelf as a high-performing deep learning approach for shot boundary detection.

Tracking The most notable difference with (Sundaram et al., 2010) corresponds to starting point sampling. In (Sundaram et al., 2010), a grid is instantiated on the first frame, and points are re-instantiated as trajectories are stopped. In contrast, we sampled starting points uniformly in space and time, to ensure the same trajectory computation is applicable to variable γ and δ. For storing the trajectories, in particular the consecutive norm differences between forward and backward flow vectors, we used half-precision. Finally, note that the RHS of Equation (3) is dependent on the flow vectors through the γ term, thus tuning γ requires extra computation relative to solely tuning δ.

Training In order to compute the loss, we must map point pairs to feature indices. For this, we simply scale the point indices by $1/4$, given the dense output representation is at $1/4$ resolution.

A.2. Evaluation

For each configuration, we use the default FPN config provided in `Detectron2`[2] as a basis.

A.2.1. COCO Semantic Segmentation

Dataset: Following (Kirillov et al., 2019), semantic annotations are converted from panoptic annotations for the 2017 challenge images, where all "things" are assigned the same semantic label, while each "stuff" category is assigned a unique semantic label.

MoCo-v2: As in (Pinheiro et al., 2020), the dilated resnet architecture is used (Chen et al., 2017; Yu et al., 2017). For each stage where the stride is decreased from 2 to 1, the dilation factor is multiplicatively scaled by 2. With that, the output resolution of the RN-50 is $1/4$, and can thereby be evaluated using the same linear prediction protocol as used for the encoder-decoder architectures.

Configuration: For data augmentation, we perform random absolute crops of size 672×672 after resizing using the default parameters, followed by a random flip.

A.2.2. COCO Instance Segmentation & Object Detection

Configuration: Only discrepancy with the default configuration is freezing the FPN. Thus, in contrast to the semantic segmentation & depth prediction evaluation, where solely a linear predictor is learned, the learned modules here are the proposal generator & ROI heads.

A.2.3. VOC & Cityscapes Semantic Segmentation

VOC Configuration: The minimum size after resizing was decreased to 480, and an absolute crop size of 512×512 was specified. Number of gradient steps was decreased to 40000, with milestone steps decreased to 25000 and 35000. Note that training was performed on the "train_aug" dataset.

Cityscapes Configuration: The minimum size after resizing was decreased to 512, and the maximum size was increased to 2048. Crops of size 512×1024 were performed. Batch size was increased from 16 to 32, base learning rate was decreased from 0.02 to 0.01, and the number of gradient steps was decreased to 65000, with milestone steps decreased to 40000 and 55000.

[1] used following repository.
[2] see here

A.2.4. NYU-Depth V2 Depth Prediction

Dataset: Downloaded from the ViRB framework release[3] (Kotar et al., 2021).

Configuration: Given the variable resolution, both input examples and labels were resized to 224×224 prior to training and testing. For augmentation, only random flips were employed.

[3] see here