Towards a Hypothesis on Visual Transformation based Self-Supervision: Supplementary Material

Dipan K. Pal
dipanp@andrew.cmu.edu
Sreena Nallamothu
akshaych@andrew.cmu.edu
Marios Savvides
marioss@andrew.cmu.edu

Dept. Electrical and Computer Engg.
Carnegie Mellon University
Pittsburgh, PA, USA

1 Prior Art

Self-supervised learning has recently garnered a lot of attention from the community. For a brief overview of various techniques and methods, we encourage the reader to refer to [13]. Self-supervision has proven to be effective in areas other than vision such as NLP [9, 33], robotics and reinforcement learning [8, 11, 20, 21, 28]. One of the main methods of performing self-supervision to learning useful features is to solve a pretext task. Such a task is chosen that is ideally computationally cheap and more importantly, one that allows the training on a 'good' representation. [10]

Pretext task based methods. Several pretext tasks have been proposed for self-supervision. For instance, solving a patch-based jigsaw puzzle [6, 14, 22], predicting color channels [16, 34, 35], predicting rotations on images [9], learning features through reversing inpainting [26], and learning to count [23]. Utilizing spatial context as a supervision signal was also explored [5]. Studies found that learning robustness to corruptions in input was also an effective pretext task [26, 30]. Geometric transformations were found to be useful to learn representations in the study [7], however it did not predict the instance or the transformation but rather aimed to learn invariance towards them. This is different from a VTSS task which predicts the exact instance of the transformation applied. Another recent task that has shown considerable promise is to match a query representation to other keys in a set belonging to the same image [10]. Other methods utilize clustering even after utilizing a pretext task [2, 24]. There also have been similar pretext tasks proposed on videos such as solving jigsaws on video frames [1, 31]. Augmenting and then predicting rotations on videos was also found to offer a useful self-supervision signal [12]. Contrastive predictive coding [25] and contrastive multiview coding [29] are other successful method which utilized some form of prediction of the data. In the real-world, the laws of physics along with time constantly provide valuable transforming data. These temporal based visual transformations can be yet another source of supervision as explored by [17, 19, 27, 32].

2 Supervision from the Transformation Perspective

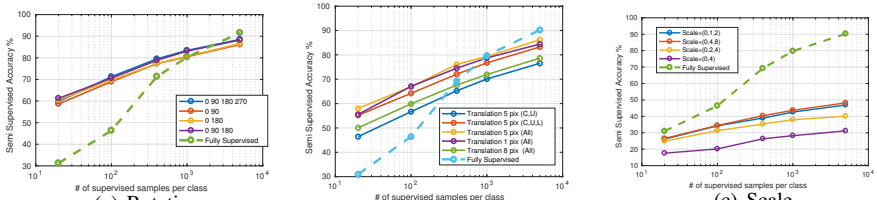
The Transformation Model of Visual Images. We adopt a model of data transformations which accounts for all the variation that is seen in general data. Given an image which when vectorized results in a seed vector x sampled from a seed distribution P_x, it is first acted upon by the transformation $g_k : \mathbb{R}^d \rightarrow \mathbb{R}^d$ *i.e.* $g_k(x;\ \theta_k)$. This transformation $g_k\mathcal{G}$ is parameterized by θ_k and generates a sample from the specific class k. For brevity, we drop the notation for the parameters and express a sample as $g_k(x)$. The transformations g_k are complex and non-linear, and can introduce *features* into a particular sample that to a receiver might appear to be associated to a particular class. For instance, g_k could potentially be features of a particular individual in the case of face recognition, or features of a face at a particular pose for pose estimation.

The Transformation Paradigm of Supervised Classification. In the context of supervised classification, different instantiations of the parameters θ_k along with different seed vectors x give rise to all of the samples that can be observed for class k in training and testing. Training data is assumed to include only a subset of all possible combinations of the parameters. Testing data would be sampled from the remaining space of combinations. Note that we do not account for any relation or overlap between g_k and g_j for classes k and j. For classification, given an input image $g_k(x)$, a classifier F is tasked with predicting the class output k. In other words, the task is to predict which of the k transformations from the set $\mathcal{G}$ was applied.

Self-Supervision from the Visual Transformation Perspective. The transformation framework is general and can be applied to any classification problem, including self-supervision. A general self-supervision task utilizing a particular transformation g would allow all data variation or transformations to be accounted for in the seed distribution P_x independent of g. In fact, the different classes are simply g_k where k is a *particular* instantiation of the transformation g. For instance, self-supervision based on rotation would be modelled as g being the in-plane rotation transformation and k being a particular instantiation of it *e.g.* 90° clockwise. Note that for a self-supervision task under this framework, *all* image data is modelled as seed vectors in the distribution P_x with g_k being a particular instantiation of a transformation, including the identity transformation e.

Overall Observation: Effectiveness of VTSS Rotation. Our results indicate that VTSS Rotation seems to consistently perform better than translation and scale when applied individually. The VTSS hypothesis also begins to offer a probable justification as to why. The degree to which translation and scale are applied as part of the VTSS task to learn these features (which were effective nonetheless), are small. Larger variation we found typically reduced performance given a fixed sized image (see Fig. 1(b) and Fig. 1(c) in the supplementary). These transformations are likely to exist in subtle amounts at similar ranges to those that were applied as part of the VTSS task. This leads to the detrimental effect of transformation conflict (see Fig. 1(b)). On the other hand, VTSS Rotation was found to work well at large ranges (90) in the original study [5]. Nonetheless, rotation seldom occurs naturally in most datasets at such large ranges (including real-world ones). Our VTSS hypothesis therefore predicts that rotation is a particularly well-suited transformation for VTSS for general visual data.

VTSS hypothesis and Invariance based methods. SimCLR [3], PIRL [18] and MoCo [10] has recent techniques which promote invariance towards common transformation but for a large number of unlabelled data. Although the VTSS hypothesis does not hold for these tasks as in, the effect of transformation conflict nonetheless applies. This would be



(a) Rotation (b) Translation (c) Scale

Figure 1: **Results of the Ablation study:** Effect of Transformation Range on VTSS Rotation [2],Translation and Scale on CIFAR 10.

the case when two samples that are negative to wards each other end up very similar under two different transformations. The representation learnt would be sub-optimal in a such case. Nonetheless, given that such the range of transformations are far richer than the pretext task based methods, it is unlikely for those methods to suffer from this effect. Nonetheless, the effect of transformation conflict must be kept in mind. The VTSS hypothesis on the other hand provides valuable insight into what is necessary for a succesful application of a self-supervision method.

3 Appendix: Additional Ablation Studies, Details and Observations

General Hyperparameters. For each transformation and dataset, the evaluation protocol and hyperparameters remained constant. The self-supervised backbone network and the fully-supervised network were both trained on the training set for all training samples (unless specified) for 200 epochs. The learning rate was set at 0.1 and multiplied by 0.02 at 60, 120 and 180 epochs with SGD with a batch-size of 128, momentum of 0.9 and weight decay of 5×10^{-4}. However, for every transformation to added in for a particular VTSS task, *e.g.* rotation, the entire batch was transformed by that augmentation and added to the batch. However, for the VTSS hypothesis confirmation studies only the transformations were added into this original 128 sized batch as an ablation study. For this, if there are K different instantiations of the transformations to be added in, then the 128 sized batch was divided by K and each shard was transformed by one instantiation. Once the self-supervised network was trained, the weights till the second conv block were frozen and a classifier on top was added. The tables in the main paper (Tables 1 and 2) were not performed with any data augmentation. However,for the ablation studies in this supplementary, we performed standard data augmentation of random crops with a 2 pixel padding with randomized horizontal flips. Interestingly, we find that VTSS translation is effective even with such augmentation.

Further Details in Architecture. All networks consisted of units or blocks called a conv block. Each conv block consisted of 3 conv layers with 192 channels, each followed by batch normalization and ReLU. The fully-supervised and self-supervised backbone networks consisted of 4 conv blocks (unless specified otherwise) with average pooling of kernel size 3, stride 2 and padding 1 after each block. The semi-supervised classifier was trained on conv block 2 features after training the self-supervision model. The semi-supervised classifier added consisted of a single conv block with 192 channels, global average pooling followed by a single linear layer.

Ablation A: Effect of Transformation Range. We trained a backbone feature extractor

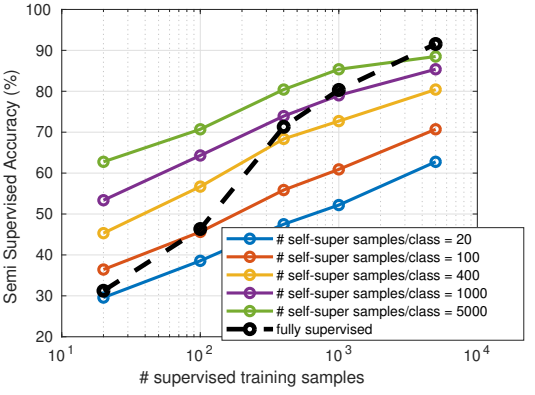


Figure 2: **Results of Ablation B.** Effect of number of self-supervised and supervised samples.

network (RotNet) with the VTSS rotation task for different sets of rotations[1] on CIFAR 10. There was no independent rotations added (as in our VTSS hypothesis confirmation study) other than the rotations added by the VTSS task itself. We also perform this experiment similarly for the VTSS translation and scale tasks. In this case, we steadily increase the number of directions the translation is added in and correspondingly increase the number of classes for prediction. Lastly, the pixel range of the translation was also varied. For VTSS Scale, the number of scales and the range was varied.

Results. The results of this experiment are presented in Fig. 1(a), Fig. 1(b) and Fig. 1(c). We find that though predicting more rotation angles help, the improvement is marginal. Further, predicting even a single instantiation of the transformation resulted in the learning of useful representations. These results are consistent with those reported in the original study [9]. In this case of translation however, we find a significant increase initially after which there are diminishing returns. It is interesting to note that VTSS translation learnt useful features even with a single pixel shift (on each side) (while using traditional data augmentation). In fact, the 5 pixel shift performs only marginally better. However, a 8 pixel shift (therefore a total crop of just 32-16=16) deems excessive for a 32 sized dataset and drastically decreases performance. For VTSS Scale, it was difficult to find an overall trend. Nonetheless, we find that representations learned are in general poor. This we hypothesis is largely due to the existence of scale variation already in CIFAR. However, it is interesting to note that a scale variation of even 1 and 2 pixels can be useful for representation learning.

Ablation B: Effect of number of self-supervised and supervised samples. Typically, self supervision tasks are trained on as much data as possible. This is primarily due to the availability of inexpensive self-labels. However, we explore the case where there is a imbalance of data between the self-supervision task and semi-supervision tasks. We increase the number of samples available per class through $\{20, 100, 400, 1000, 5000\}$ samples for both the VTSS Rotation and the downstream semi-supervision tasks.

Ablation B: Results. Fig. 2 showcases the results of this experiment. Interestingly, we find that the performance increases linearly at almost identical rates for both the self-supervision and the downstream semi-supervision tasks. For instance, the performance of a

[1] We provide additional ablation studies in the supplementary

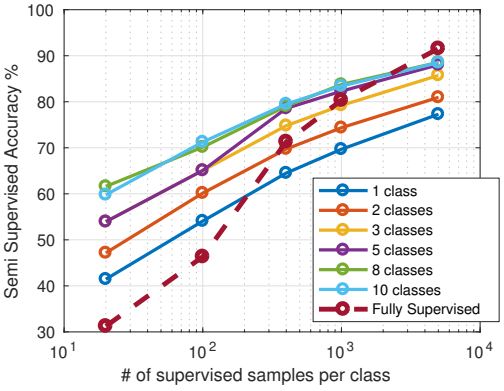


Figure 3: **Results of Ablation C.** Effect of number of classes used for self-supervision.

1000 samples/class for VTSS and just 20 samples/class for semi-supervised learning is very similar to 20 samples/class for VTSS and 1000 samples/class for semi-supervised learning. We find similar trends for other settings. This highlights the benefits of VTSS tasks. For a particular amount of data with an inexpensive self-labelling scheme, VTSS provides a level of performance to the downstream classifier similar to that of a semi-supervised model which was trained with the same amount of *labeled* data. Nonetheless, such linear parallels between VTSS and downstream semi-supervision are encouraging.

Ablation C: Effect of number of classes used for self-supervision. VTSS tasks are typically applied to a wide array of data. However, what is the level of returns that a VTSS task provides given a steady increase in both diversity and amount of data? For this experiment, we train the same 4 conv block layered network with VTSS Rotation given $\{1,2,3,5,8,10\}$ classes. We are interested in the trends with which the performance increases w.r.t the downstream semi-supervised accuracy.

Ablation C: Results. The results of this experiment are presented in Fi.g 3. We find that the downstream semi-supervision performance increases as expected. However, the returns are diminishing. Indeed, though when there are 5000 samples/class available for training the semi-supervised network, the performance saturates with just 5 classes=5000 samples in total for the VTSS task. Similar trends are observed for the cases where there are lower samples/class available for semi-supervised learning. This suggests that though VTSS tasks are powerful, they seem to be hitting a barrier to the diversity of features that the model can learn. Attention must be paid to other aspects of the learning problem such as the size of the network, architecture etc. [15] in order to further allow improvements leveraging more self-supervised data.

Additional Observation: Effectiveness of VTSS tasks in general. Taking a step back from a detailed inter-transformation analysis, we observe the performance of self-supervision followed by semi-supervised tasks. Recall that the original VTSS backbone network consisted of 4 conv blocks of three conv layers each. Then, only the first two conv blocks were used as a fixed feature extractor for the semi-supervised classifier on top that was trained. This classifier consisted of a single conv block that is identical to the other blocks. Therefore, in Table 1 and Table 2 in the main paper, the fully supervised networks or FS with 3 blocks had similar complexity to the overall self and then semi-supervised models, and provided a fair comparison from the perspective of model complexity. Yet, in the case of SVHN and FMNIST, we find that VTSS Rotation performs better than FS 3 blocks for the full crop

setting (including scale for FMNIST). For the center crop experiments, VTSS R+T performs better than FS 3 blocks for SVHN. This showcases the overall effectiveness of VTSS tasks in general compared to fully-supervised networks of similar complexity.

References

[1] Unaiza Ahsan, Rishi Madhok, and Irfan Essa. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 179–189. IEEE, 2019.

[2] 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.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[6] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[7] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014.

[8] Frederik Ebert, Sudeep Dasari, Alex X Lee, Sergey Levine, and Chelsea Finn. Robustness via retrying: Closed-loop robotic manipulation with self-supervised learning. *arXiv preprint arXiv:1810.03043*, 2018.

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

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2019.

[11] Eric Jang, Coline Devin, Vincent Vanhoucke, and Sergey Levine. Grasp2vec: Learning object representations from self-supervised grasping. *arXiv preprint arXiv:1811.06964*, 2018.

[12] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11387*, 2018.

[13] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *arXiv preprint arXiv:1902.06162*, 2019.

[14] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Learning image representations by completing damaged jigsaw puzzles. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 793–802. IEEE, 2018.

[15] Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. *arXiv preprint arXiv:1901.09005*, 2019.

[16] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision*, pages 577–593. Springer, 2016.

[17] 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.

[18] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. *arXiv preprint arXiv:1912.01991*, 2019.

[19] 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.

[20] Adithyavairavan Murali, Lerrel Pinto, Dhiraj Gandhi, and Abhinav Gupta. Cassl: Curriculum accelerated self-supervised learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6453–6460. IEEE, 2018.

[21] Ashvin Nair, Dian Chen, Pulkit Agrawal, Phillip Isola, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Combining self-supervised learning and imitation for vision-based rope manipulation. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2146–2153. IEEE, 2017.

[22] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[23] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5898–5906, 2017.

[24] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9359–9367, 2018.

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[26] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[27] 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.

[28] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. *arXiv preprint arXiv:1906.04161*, 2019.

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

[30] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008.

[31] Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian, and Alan L Yuille. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1910–1919, 2019.

[32] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018.

[33] Jiawei Wu, Xin Wang, and William Yang Wang. Self-supervised dialogue learning. *CoRR*, abs/1907.00448, 2019. URL http://arxiv.org/abs/1907.00448.

[34] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[35] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.