Towards a Hypothesis on Visual Transformation based Self-Supervision

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

Abstract

We propose the first qualitative hypothesis characterizing the behavior of visual transformation based self-supervision, called the VTSS hypothesis. Given a dataset upon which a self-supervised task is performed while predicting instantiations of a transformation, the hypothesis states that if the predicted instantiations of the transformations are already present in the dataset, then the representation learned will be less useful. The hypothesis was derived by observing a key constraint in the application of self-supervision using a particular transformation. This constraint, which we term the transformation conflict for this paper, forces a network to learn degenerative features thereby reducing the usefulness of the representation. The VTSS hypothesis helps us identify transformations that have the potential to be effective as a self-supervision task. Further, it helps to generally predict whether a particular transformation based self-supervision technique would be effective or not for a particular dataset. We provide extensive evaluations on CIFAR 10, CIFAR 100, SVHN and FMNIST confirming the hypothesis and the trends it predicts. We also propose novel cost-effective self-supervision techniques based on translation and scale, which when combined with rotation outperform all transformations applied individually. Overall, the aim of this paper is to shed light on the phenomenon of visual transformation based self-supervision.

1 Introduction

The Mystery of Self-Supervision. Self-supervision loosely refers to the class of representation learning techniques, where it is cost effective to produce effective supervision for models using some function of the data itself. Indeed in many cases, the data becomes its own ground-truth. While a lot of efforts are being directed towards developing more effective techniques [3, 8, 12, 13, 15], there has not been enough attention on the problem of understanding these techniques or at least a sub-set of them at a more fundamental level. Indeed while there have been many efforts which introduced self-supervision in different forms [5, 14, 16], there have been only a few efforts which shed more light into related phenomenon [6]. In one such work, the authors focus on the trends (and the lack of) that different architecture choices have on the performance of the learnt representations [9]. One emerging technique that has proven

to learn useful representations while being deceptively elementary is the study of RotNet [2, 5]. RotNet takes an input image and applies specific rotations to it. The network is then tasked with predicting the correct rotation applied. In doing so, it was shown to learn useful representations which could be used in many diverse downstream applications. It was argued that however, that in learning to predict rotations, the network is forced to attend to the shapes, sizes and relative positions of visual concepts in the image. This explanation though intuitive, does not help us to better understand the behavior of the technique. For instance, it does not help us answer the question: *Under what conditions would a particular method succeed in learning useful representations?*

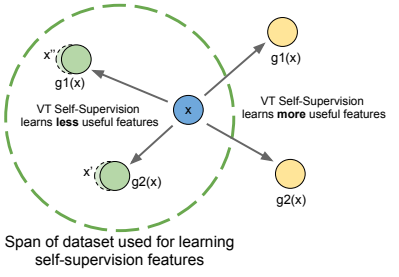
Towards an Initial Hypothesis of Self-Supervision. This study hopes to provide initial answers or directions to such questions. This hypothesis is the first attempt to characterize and move towards a theory of the behavior of self-supervision techniques utilizing visual transformations. The VTSS hypothesis also offers practical applications. For instance, the hypothesis can predict or suggest reasonable transformations to use as a prediction task in VTSS given a particular dataset. Indeed, we introduce two novel VTSS tasks based on translation and scale respectively which we study in our experiments. The hypothesis can also help predict the relative trends in performance between VTSS tasks based on one or more transformations on a particular dataset based on some of the dataset properties. We confirm the VTSS hypothesis and a few trends that it predicts in our experiments. It is worthwhile to note that more recent self-supervised representation learning methods such as SimCLR [2], PIRL [10] and MoCo [7] do not predict instantiations of any transformation or task and at first glance seem to be immune to the effects of transformation conflict. However, we argue this is not the case theoretically although it in practice they are expected to be more robust [1]. Nonetheless, our deep focus in this study is primarily on visual transformation prediction based self-supervision methods.

Our Contributions. 1) We provide an initial step towards a theory of behavior of visual transformation based self-supervision (VTSS) techniques in the form of a qualitative hypothesis. The hypothesis describes a condition when a self-supervision technique based on a particular visual transformation would succeed. 2) We use the hypothesis to propose two novel self-supervision tasks based on translation and scale and argue why they might be effective in particular cases. 3) We provide extensive empirical evaluation on CIFAR 10, CIFAR 100, SVHN and FMNIST providing empirical evidence towards confirming the hypothesis for all transformations studied, *i.e.* translation, rotation and scale. We further propose to combine transformations within the same self-supervision tasks leading to performance gains over individual transformations. 4) Finally, we provide an array of ablation studies and observations on VTSS using rotation, translation and scale. For instance, we find that improvements in semi-supervised classification performance provided by unlabelled data used for self-supervision is very similar to that provided by having the same amount of labelled data used for semi-supervised training.

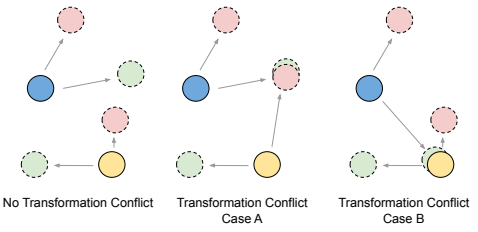
2 The VTSS Hypothesis

For our purpose, we define *usefulness* of a representation as the semi-supervised classification test accuracy C, of a downstream classifier using that representation on a classification task of higher abstraction. A classification task can be loosely termed to be at a higher

[1] More discussion and related work in the supplementary.



(a) The VTSS Hypothesis



(b) Trasnformation Conflict

Figure 1: **a) The Visual Transformation Self-Supervision (VTSS) Hypothesis:** Given a dataset upon which a self-supervised task is performed while predicting instantiations of a transformation, the hypothesis states that if the predicted instantiations of the transformations are already present in the dataset, then the representation learned will be less useful. **b) The Effect of Transformation Conflict:** This effect can prevent a self-supervision method from learning a useful representation. There are at least two potential ways of such an effect manifesting in the dataset. Case A: where two different transformations of two different samples are identical or very similar. Case B: where a transformed version of a sample is identical or very similar to another untransformed sample.

abstraction level if it uses a more complicated transformation set $\mathcal{G}$, than the one used for the self-supervision task.

The VTSS Hypothesis: Let $\mathcal{G} = \{g_k\} \cup \{e\} \;\; \forall k = 1...|\mathcal{G}|$ be a set of transformations acting on a vector space $\mathbb{R}^d$ with e being the identity transformation and further with $g(x') \in \mathbb{R}^d \;\; \forall x' \in \mathbb{R}^d \;\; \forall g \in \mathcal{G}$. Let $\mathcal{X}$ be the set of *all* seed vectors $x \in \mathbb{R}^d$. Finally, we simulate a dataset with a set of *pre-existing* transformations $\mathcal{H}$, by letting $\mathcal{H} = \{h \mid h(x) \in \mathcal{X} \;\; \forall x \in \mathcal{X}\}$. Now, for a usefulness measure of C of a representation $F(x)$ that is trained using a transformation based self-supervision task which predicts instantiations of $\mathcal{G}$, the VTSS hypothesis predicts

$$C \propto (|\mathcal{G} \cap \mathcal{H}|)^{-1} \tag{1}$$

In other words, consider a dataset of images and a visual transformation based self-supervision (VTSS) task that predicts instantiations of $\mathcal{G}$. Then if the dataset already contains a lot of variations in its samples due to any of the transformations in $\mathcal{G}$, then the VTSS hypothesis predicts that the features learnt on that dataset using the VTSS task corresponding to $\mathcal{G}$ will *not* produce useful features or the usefulness will be *diminished*. In the hypothesis statement, the transformation set $\mathcal{H}$ is the set of all possible transformations and variations that exist in the dataset $\mathcal{X}$. Thus, if $\mathcal{G}$ and $\mathcal{H}$ have a lot of transformations in common, C decreases. In other words, C is inversely proportional to the number of transformations common between $\mathcal{G}$ and $\mathcal{H}$. It is important to note however that every *instantiation* of a transformation is considered different. Therefore, a rotation by 45° clockwise is a different transformation than a rotation by 90° clockwise. Each instantiation can be used as a prediction target while constructing the corresponding VTSS task.

Approaching RotNet from a new perspective. The hypothesis deters the use of transformations for VTSS tasks which are already present in the data. This might discourage us from utilizing in-plane rotations as a VTSS task since small yet appreciable amounts of in-plane rotation exist in most real-world datasets. However, we must recall that each *instantiation* of the transformation is considered different. Hence if we consider a rotation angles large

enough such as $\{0°, 90°, 180°, 270°\}$, they are unlikely to exist in the dataset. Thus, the VTSS hypothesis predicts that in-plane rotations would be an effective VTSS task provided the range of rotation is large enough and this indeed has been the observation [5].

Identifying Effective Transformations for VTSS. Following this train of thought, it is natural to ask what other transformations can be used for VTSS? Translation and scale are two relatively simple transformations that are easier to apply (especially translation). It would however be a fair observation to make that both transformations are in fact the most common transformations of variation in real-world visual data, which according to our hypothesis would result in an ineffective learning task. However, owing to manual and automated labelling efforts, there are many datasets in which the visual concept of interest is fairly centered in the image across all samples. This creates an opportunity for the direct use of translation as a computationally inexpensive transformation to apply to self-supervision, while being consistent with the VTSS hypothesis. Scale variation as well can be controlled and accounted for. Nonetheless, in many datasets even when the object is localized, there is relatively more scale variation than translation jitter. As part of this study, we propose the use of both translation and scale as VTSS tasks for use whenever the conditions are favorable.

Predicting Trends in Relative Performance of VTSS on Datasets. Currently, a barrage of self-supervision tasks are applied to a particular dataset as part of a trial and error process towards obtaining a desirable level of performance. It seems that there exists no heuristic to predict even at minimum a trend of effectiveness. The VTSS hypothesis provides the an initial heuristic to predict trends in relative performance on any given dataset, given some properties of the dataset. There are at times the possibility of estimating how much variation due to a particular transformation might exist in a given dataset. This might be possible due to control or knowledge of the data collection process, coarse estimation through techniques like PCA or more sophisticated techniques such as disentanglement through an information bottleneck [1]. In such cases, the VTSS hypothesis can assist in rejecting particular VTSS tasks and prioritize others. For instance, a rotation based VTSS technique is predicted to not be beneficial on a dataset such as Rotated MNIST. Indeed, in our experiments, we observe cases when rotation based self-supervision fails completely.

Understanding the VTSS hypothesis. We discussed a few ways the VTSS hypothesis could be useful. We now provide a qualitative explanation for the same. Consider two samples x and x' belonging to a dataset $\mathcal{X}$. Let there be a network F which will be trained for a VTSS task utilizing the transformation set $\mathcal{G} = \{g_i \mid i = \{1...k\}\} \cup \{e\}$ where e is the identity transformation and $|\mathcal{G}| = k+1$. Therefore, $F(x)$ learns to predict one out of $k+1$ outputs. Specifically for any x, given $g_i(x)$ as input (where g_i could be identity), $F(x)$ would need to predict the correct *instantiation* of $\mathcal{G}$ including the identity. Now, the VTSS hypothesis predicts that as long as $\nexists g' \in \mathcal{G}$ s.t. $g'(x) = x'$ or $g'(x') = x$, the VTSS task will learn useful features. In other words, as long there exists no transformation instantiation in $\mathcal{G}$ such that x and x' can be related to one another through it, a useful feature will be learned. To see why, we assume $\exists g_k \in \mathcal{G}$ s.t. $g_k(x) = x'$. Under this assumption, the output of $F(g_k(x))$ should be k *i.e.* $F(g_k(x)) = k$. However, we also have $F(x') = F(e(x')) = e$ *i.e.* predicting the identity class since $e \in \mathcal{G}$. Notice that a conflict arises with these two equations, 1) $F(x') = F(e(x')) = e$ and also $F(g_k(x)) = F(x') = F(e(x')) = k$. Therefore, for the same input x', the network is expected to output two separate classes. We term this phenomenon as a **transformation conflict** for this paper (see Fig. 1(b)), and we observe it in our experiments. This condition over the course of many iterations will learn noisy filters. This is because in practice, there exists small differences between $g_k(x)$ and x'. The network will be forced to amplify such differences while trying to minimize the loss, leading to noise being learned as features.

The Transformation Conflict: In Fig. 1(b), consider a VTSS task of predicting between three instantiations (including the identity) of a transformation g on a two samples (blue and yellow dots) from a dataset. When g is applied to the samples, it results in the corresponding transformed samples (light red and green dotted dots). Each color signifies the specific label or instantiation of a transformation that the network is tasked with predicting (in the figure there are two colored labels, red and green, with the identity transformation being the sample itself). For instance, RotNet [5] predicts between 4 angles including $0°$. The self-supervised network will take in as input each transformed or original sample (all dots), and predict the corresponding label (transformation instantiation). **Left:** In this case, each dataset sample is transformed into points that are distinct and away from other transformed samples or data points. Hence, there is no transformation conflict. The VTSS hypothesis in this case predicts that the features learnt will be useful. **Center: Case A.** Here, one of the samples (yellow) transforms into a point (corresponding light red) close to a transformed version (near by light green) of a separate data sample (blue). This presents a way of incurring transformation conflict within the dataset. For the similar inputs of the closely overlapping dotted red and green dots, the network is expected to predict/output both red and green labels. This causes the network to learn degenerate features, as it is trying to maximize discrimination between the two close-by samples. **Right: Case B.** Here, one of the samples (blue) transforms into a point (corresponding dotted light green) close to another original data point (yellow). This presents a second way of incurring transformation conflict within the dataset. For the similar inputs of the closely overlapping yellow and dotted green dots, the network is expected to predict/output both green and identity labels. In both cases A and B, the VTSS hypothesis predicts less useful features learnt by the self-supervision task.

3 Experimental Validation

Our goal through an extensive experimental validation is threefold. 1) To confirm (or find evidence otherwise to) the VTSS hypothesis for VTSS tasks based on rotation and translation and. 2) To explore the efficacy of solving VTSS tasks with individual transformations and additive combinations. 3) To perform ablation studies on VTSS task based on rotation, translation and scale to help gain insights into effects on semi supervision performance [2]. For these experiments, we utilize the CIFAR 10, CIFAR 100, FMNIST and SVHN datasets[3]. Our effort is not to maximize any individual performance metric or achieve state-of-the-art performance on any task, but rather to discover overall trends in behavior leading to deeper insights into the phenomenon of self-supervision through visual transformations.

General Experimental Protocol. For each transformation and dataset, the overall experimental setup remained unchanged. We follow the training protocol introduced in the RotNet study [5] where a 4 convolution block backbone network is first trained with a VTSS task based on some transformation (rotation, translation and/or scale). This network is tasked with predicting specific transformation instances following the self-supervision protocol. After training, the network weights are frozen and the feature representations from the second convolution block is utilized for training a downstream convolution classifier which is tasked with the usual supervised classification task *i.e.* predicting class labels for a particular dataset. Our choice of exploring the performance trends of the second block is informed by the original RotNet study where the second conv block exhibited maximum performance for CIFAR

[2] We provide these ablation study results and discussions in the supplementary due to space constraints

[3] We provide additional experimental details in the supplementary.

10 [5]. However, since our focus is on discovering overall trends rather than maximizing individual performance numbers, this choice is inconsequential for our study. Thus, the overall learning setting is semi-supervised learning since part of the pipeline utilized the frozen self-supervised weights. The final semi-supervised test accuracies reported on the test data of each dataset utilized this semi-supervised pipeline.

Architecture: The network architecture for all experiments consists of four convolutional blocks, followed by global average pooling and a fully connected layer. Each convolutional block is a stack of three convolution layers, each of which is followed by Batch normalization and a ReLU. Finally, there exists a pooling layer between two blocks[4].

EXP 1: Confirming the VTSS Hypothesis

Goal: Recall that the VTSS hypothesis predicts that a VTSS task would learn useful features using a particular transformation $\mathcal{G}$ only when the predicted instantiations of $\mathcal{G}$ do not already exist in data. In this experiment, we test this hypothesis for the VTSS tasks based on rotation [5] and translation. The overall approach for this experiment is to break the assumption of the VTSS hypothesis *i.e.* the assumption that instantiations from $\mathcal{G}$ are *not* present in the original data or $\mathcal{H}$. We do this by introducing increasing elements from $\mathcal{G}$ in the original data itself, *independent* of the fact that the VTSS task would additionally apply and predict instantiations of $\mathcal{G}$ to learn a useful representation. This artificially increases $|\mathcal{G} \cap \mathcal{H}|$. Checking if C (semi-supervised performance of the learned representations) varies inversely allows one to confirm whether the VTSS hypothesis holds.

Experimental Setup: We explore two VTSS tasks based on rotation and translation respectively. The *prediction* range of these transformations are as follows[5]:

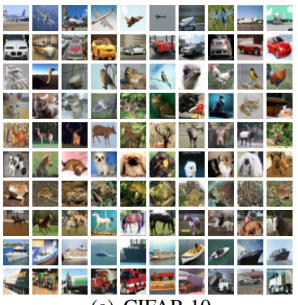
1) **VTSS Rotation: [5]** Image rotations by $\{0°, 90°, 180°, 270°\}$ leading to 4-way classification. The input image was rotated by one of the four angles. The VTSS task was to predict the correct rotation angle applied. This is essentially the same VTSS task employed by RotNet.

2) **VTSS Translation:** Image translations by 5 pixels with the directions {up, down, left, right, no translation (center)} leading to a 5-way classification task. From the original image, a center crop with a 5 pixel margin was cropped, which was now considered to be the 'no translation' input (center crop). Translations by 5 pixels were applied to this center patch in one of the directions between up, down, left and right. The VTSS task was to predict which direction the image was translated. The 5 pixel margin allows for a 5 pixel translation with no artifacts. This task based on translations is novel and is part of our contribution.

For each transformation $\mathcal{G}$, more instantiations of $\mathcal{G}$ were sequentially added in in the *original data* independent of the corresponding VTSS task. Therefore for each image, there are in fact two separate stages where a transformation is added a) the proposed ablation study itself and b) the VTSS task independent of the ablation study. We now explain the protocol in detail for rotation which has 4 runs (experiments). **Run 1) Baseline. 0°** The original data contains no rotations added in. This is used for the standard RotNet VTSS task of predicting a 4-way task after rotating the image by one of $\{0°, 90°, 180°, 270°\}$ rotations. This model is evaluated for semi-supervision accuracy and is set as the baseline. **Run 2) 0°, 90°**. Next, the same procedure is followed however, the *original data* that is sent to the VTSS task, *already* contains all images at 0° *and* 90° rotations. It is crucial to note however, that the VTSS task of rotating one of $\{0°, 90°, 180°, 270°\}$ rotations and then predicting the rotation remains unchanged. The VTSS task then transforms and predicts based on the original 0°

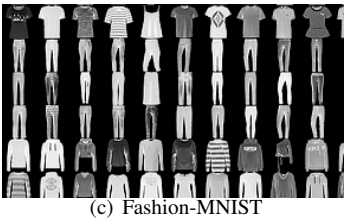
[4] More details are provided in the supplementary.

[5] We provide more details in the supplementary.



(a) CIFAR 10

(b) SVHN



(c) Fashion-MNIST

Figure 2: Representative samples from the CIFAR 10 and SVHN dataset. Samples from the Fashion-MNIST dataset. Note that compared to CIFAR 10 and SVHN, FMNIST contains considerably less scale variation. Thus, the VTSS hypothesis predicts that VTSS Scale would perform better, than that of CIFAR 10 and SVHN, which is indeed the case from Table. 3. This is yet another confirmation of the VTSS hypothesis.

Rot	N	C10	SVHN
Base (0°)	4	$88.50_{(91.99)}$	$90.08_{(86.29)}$
0	4	$88.55_{(46.93)}$	$89.88_{(43.46)}$
0,90,180	4	$87.87_{(33.87)}$	$89.28_{(33.49)}$
All	4	$10_{(25)}$	$7.75_{(25)}$
Trans	N	C10	SVHN
Base (C)	5	$85.13_{(60.55)}$	$91.16_{(76.98)}$
CU	5	$83.10_{(34.86)}$	$91.02_{(43.17)}$
CUR	5	$82.48_{(33.16)}$	$90.60_{(39.59)}$
CUDL	5	$81.48_{(33.24)}$	$90.62_{(36.47)}$
All	5	$79.29_{(32.44)}$	$90.16_{(37.19)}$
FS 3 blocks	10	91.35	80.83
FS 4 blocks	10	91.66	90.57

VTSS	C10	SVHN	FMNIST	C100
R [□]	89.15	91.29	*91.94*	63.62
T	86.20	91.16	88.98	57.10
S	43.89	28.42	83.48	17.72
R+T	**89.58**	*91.56*	**92.18**	*64.79*
S+T	71.53	87.56	88.02	45.09
R+S	89.00	89.16	91.81	63.78
R+T+S	*89.39*	**91.72**	91.63	**64.87**
FS 3 B	89.96	91.43	77.37	64.93
FS 4 B	90.26	92.50	92.21	65.95
S (F)	31.16	7.76	87.87	15.01
R (F)	88.50	90.08	93.70	61.04
FS 3 (F)	91.35	80.83	92.65	52.60
FS 4 (F)	91.66	90.57	94.62	67.95

Table 1: **(Left table) VTSS Hypothesis Confirmation.** The column of transformation instantiations on the left were added into the *original data* independent of the additional augmentation by the VTSS task. For each dataset, the number denotes the semi-supervision accuracy following the protocol described while having a N-way transformation prediction that is fixed for Rotation and Translation. FS indicate Fully Supervised and F denotes (full). The smaller number in the bracket denotes the N-way (number of transformations) test accuracy during the self-supervision task. **(Right table) Visual Transformation based Self-Supervision (VTSS) through a combination of transformations.** The column on the left denoted the transformation that the self-supervised backbone was trained with. In the full crop (full) setting, the entire image was utilized for training and testing. In the base setting, the center crop of the image with a margin of 5 pixels on all sides was used. This was done for a better comparison with VTSS Translation which required a 5 pixel margin to allow room for translations as the VTSS task. **bold** and *italics* indicate the best and second-best performances respectively. VTSS tasks using a combination of transformations performed the best for all four datasets.

images and the images at 90° identically. **Run 3) 0°, 90°, 180°** Now, the same procedure (VTSS task followed by semi-supervision evaluation) is followed by having all images rotated

at each of $\{0°, 90°, 180°\}$ **Run 4) 0°,90°,180°,270°** Finally, yet another run uses all four rotations added in to all images of the original data. This protocol is followed similarly for translation (predicting 5-way between no translation, up, down, left and right) where the particular transformations that were measured by the VTSS task were added into the original data sequentially. Further details are provided in the supplementary. The performance metric considered for each transformation $\mathcal{G}$ is the semi-supervised accuracy (obtained using the protocol explained in the general experimental settings) on the CIFAR 10 and SVHN test sets.

Results and Discussion. Table. 3 showcases the results of these experiments. The left column indicates which transformation instantiations (each for rotation and translation) were added into the original data as part of this ablation study. The semi-supervised accuracy indicates the performance of the learned features towards the downstream classification task (protocol introduced in [5]). The number in the bracket is the self-supervised accuracy which indicates the test accuracy on the VTSS task itself. Higher accuracy indicates the model is able to distinguish between the between transformations added in. We make a few observations.

Observation 1: We find that VTSS Rotation performs well when there are no rotations already present in the data (both for CIFAR 10 and SVHN). However, the method completely breaks down for both datasets when *all* rotations are present. This indicates that the model has learnt noisy features.

Observation 2: Notice that the self-supervision classification accuracy for all transformations steadily decreases as more rotations were added into the original data. This is in alignment with Eq. 1. Indeed, as more ablation transformations are added in the original data, it becomes difficult for the VTSS task to learn useful features due to the transformation conflict effect. Observation 1 and 2 together confirm the VTSS hypothesis.

Pre-existing Transformations in SVHN and CIFAR: For the next observation we take a look at the SVHN [11] and CIFAR 10 datasets illustrated with a few samples in Fig. 2(b) and Fig. 2(a). For SVHN, we find that there exist considerable scale variation and blur within each digit class. This blur also acts as scale variation as it simulates the process when a small low resolution object is scaled up leading to blur. However, note that since the dataset was created by extending each digit bounding box in the appropriate directions leading to a square, each digit of interest is almost exactly centered. Thus, there pre-exists very little translation in the dataset. Coupled with the fact that digits have lesser variation than general objects, the visual concepts of interest are more centered. CIFAR 10 on the other hand has more complicated objects also with some scale variation already present in the vanilla dataset. The complex nature of the visual classes results in relatively more translation jitter of visual concepts of interest than SVHN. Lastly, both datasets have some rotation variation however not as extreme as 90° or beyond.

Observation 3: We observe that VTSS Translation performs closer to the fully supervised performance for SVHN compared to CIFAR 10. Keeping in mind that SVHN has relatively less translation than CIFAR 10, this is consistent with and supports the VTSS hypothesis.

EXP 2: Exploring VTSS Tasks with Multiple Transformations Simultaneously

Goal: Typically, self supervision using visual transformations has been applied with a single transformation type, for instance exclusively rotations for RotNet [5]. Given that in this study, we have demonstrated the existence of a VTSS technique for translation and scale as well, it is natural to ask the question: *how does the performance differ when using multiple transformations in conjunction?*. We explore answers and also observe phenomenon that the VTSS hypothesis predicts.

Experimental Setup: For the datasets CIFAR 10, 100, FMNIST and SVHN, we train the standard backbone network with 4 convolution blocks with VTSS tasks of Rotation

$\{0°, 90°, 180°, 270°\}$, Translation {up, down, left, right, no translation (center)} with a shift of 5 pixels and Scale {0 pix, 2 pix zoom, 4 pix zoom}. However, VTSS Translation needs a small margin (to translate without artifacts). We apply this crop margin (of 5 pixels) to all data for all tasks. Therefore, the default images for all tasks is the 5 pixel margin center crop of the original image. For VTSS Scale, this crop was designated to be the 0 pix zoom. A 2 pix zoom (or 4 pix zoom) would perform yet another crop with a 2 pix (or 4 pix) margin on each side before resizing the image back to the center crop size. We combine two or more transformations in an additive fashion. For instance, if VTSS Rotation predicts 4 classes and VTSS Translation predicts 5 classes, the task VTSS Rotation + Translation will predict 4+5-1 = 8 classes overall (where we combine the identity class of all transformations into a single class). We perform experiments with all 4 combinations between the three transformations. Additionally, we run each transformation individually to serve as a baseline under the center crop setting. All individual and combination transformations were run in the center crop setting to be consistent in data size for the VTSS Translation task for the combination experiments. However, in practice, when VTSS Scale and Rotation [5] would be applied independently, the entire image would be used and not just the center crop. Thus we provide additional results with just the individuals transformations of VTSS Rotation and Scale on the full sized crop of side 32 (without any center cropping). The corresponding fully-supervised results with the full crop were also provided.

Pre-existing Transformations in Fashion-MNIST: A few sample images from the Fashion-MNIST dataset are shown in Fig. 2(c). One notices immediately that the dataset contains little to no translation jitter, no rotation and importantly, very little scale variation as compared to CIFAR and SVHN (see corresponding figure in main paper). This implies that the VTSS hypothesis would predict that VTSS Scale would be effective. The FMNIST dataset hence is a good dataset to prove effectiveness of the VTSS Scale task. The dataset contains 60,000 training images and 10,000 testing images. each image is sized 28, which for our experiments was rescaled to 32. This does not affect overall trends in our experiments since all images were resized equally.

Results: Individual Transformations. The results of these experiments are presented in Table. 3. We find that VTSS Rotation overall performs consistently high. However, given that SVHN has lesser translation (see discussion on pre-existing transformations in datasets), VTSS Translation performs better on SVHN than CIFAR 10 and 100. This is indeed consistent with the VTSS hypothesis. Note that scale performs worse on both the CIFAR datasets and SVHN. Recalling the prior discussion regarding the presence of scale variation in both CIFARs and SVHN, this result is consistent with the VTSS hypothesis. In fact, due to the presence of more blur which acts as scale variation, VTSS Scale works worse on SVHN than CIFAR 10, which has no common blurry artifacts. This observation as well is consistent with the VTSS hypothesis. Interestingly however, that given the observation that FMNIST has considerably less scale variation than CIFAR and SVHN, the VTSS hypothesis predicts that VTSS Scale would perform better on FMNIST than on CIFAR and SVHN. Indeed, this is what we observe. Both the full crop and the center crop VTSS Scale performance on FMNIST are significantly higher than that of CIFAR and SVHN. This provides further evidence towards the confirmation of the VTSS hypothesis.

Results: Combinations of Transformations. We find that a combination of VTSS R + T works better than isolated VTSS Rotation for all four datasets. This also true for VTSS R+T+S for all datasets except FMNIST. This is the first evidence that utilizing multiple transformations simultaneously as a single VTSS task can provide performance gains over any individual transformation. Given that it is computationally inexpensive to train under such

a setting, this result is encouraging. Notice also that even though VTSS Scale performs worse on SVHN than CIFAR, the combination VTSS S+T performs better on CIFAR. Nonetheless, due to the inherent presence of scale in CIFAR and SVHN, the VTSS hypothesis predicts that VTSS tasks involving scale would suffer in performance. However, scale achieves more success on FMNIST due to the absense of inherent scale (a hypothesis prediction). This is something we do observe in Table. 3. From these experiments, we conclude that there is benefit in combining VTSS tasks for different transformations, however it must be done so while being aware of what transformations or factors of variation already exist in the data. Indeed, VTSS tasks using some sort of combination of transformations consistently outperformed all individual transformations for all four datasets.

References

[1] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

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

[3] 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.

[4] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10364–10374, 2019.

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

[6] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. *arXiv preprint arXiv:1905.01235*, 2019.

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

[8] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *arXiv preprint arXiv:1906.12340*, 2019.

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

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

[11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[12] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1134–1141. IEEE, 2018.

[13] Trieu H Trinh, Minh-Thang Luong, and Quoc V Le. Selfie: Self-supervised pretraining for image embedding. *arXiv preprint arXiv:1906.02940*, 2019.

[14] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. The visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

[15] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2555, 2019.

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