

---

# GPS: Guided Positive Sampling To Move Away From Data-Augmentation In SSL

---

Anonymous Authors<sup>1</sup>

## Abstract

This paper introduces *Guided Positive Sampling Self-Supervised Learning* (GPS-SSL), a method aimed at incorporating prior knowledge into Self-Supervised Learning (SSL) positive sample selection. Unlike current SSL methods relying solely on Data-Augmentations (DA) to generate positive samples, GPS-SSL creates a metric space aligning distances with semantic relationships and enabling informed positive sample selection through nearest neighbor sampling. A direct byproduct of GPS-SSL –and its core motivation– is the reduced importance of devising optimal DA recipes to learn performant representations. Since the proposed method solely alters the positive pair sampling, it can be coupled off-the-shelf with many SSL methods. Evaluation against baseline SSL methods on diverse datasets demonstrates the effectiveness of GPS-SSL, especially in scenarios with minimal DA; thus offering potential for further research on advancing SSL beyond careful DA design.

## 1. Introduction

Self-supervised learning (SSL) has recently shown to be one of the most effective learning paradigms across many data domains (Radford et al., 2021; Girdhar et al., 2023; Assran et al., 2023; Chen et al., 2020; Grill et al., 2020; Bardes et al., 2021; Balestriero et al., 2023). SSL belongs to the broad category of annotation-free representation learning approaches, which have enabled machine learning models to use abundant and easy-to-collect unlabeled data, facilitating the training of ever-growing deep neural network architectures.

Despite the SSL promise, current approaches require handcrafted a priori knowledge to learn useful representations. This a priori knowledge is often injected through the positive sample – *i.e.*, semantically related samples – generation strategies employed by SSL methods (Chen

et al., 2020). In fact, SSL representations are learned so that such positive samples get as similar as possible in embedding space, all while preventing a collapse of the representation to simply predicting a constant for all inputs. The different strategies to achieve that goal lead to different flavors of SSL methods (Chen et al., 2020; Grill et al., 2020; Bardes et al., 2021; Zbontar et al., 2021; Chen & He, 2021). In computer vision, positive sample generation mostly involves sampling an image from the dataset, and applying multiple handcrafted and heavily tuned data augmentations (DAs) to it, such as rotations and random crops, which preserve the main content of the image.

The impact of designing DAs which are effective for the dataset at hand is enormous –as measured by its effect on performance (Garrido et al., 2023; Dangovski et al., 2021; Xiao et al., 2020; Tamkin et al., 2020; Kirichenko et al., 2023)–, to the point of producing a near random representation, in the worst case scenario (Balestriero et al., 2023). As such, tremendous time and resources have been devoted to designing optimal DA recipes, most notably for ubiquitous datasets such as ImageNet (Deng et al., 2009). From a practitioner’s standpoint, positive sample generation could thus be considered solved if one were to deploy SSL methods *only* on such popular datasets. Unfortunately – and as we will thoroughly demonstrate throughout this paper –, common DA recipes used in those settings fail to transfer to other datasets. We hypothesize that as the dataset domains get semantically further from ImageNet, on which the current set of optimal DAs are designed, the effectiveness of DAs reduces. For example, since ImageNet consists of object-centric natural images focusing on  $\sim 1000$  different object categories, we observe and report a reduction of performance on datasets consisting of more specialized images, such as hotel room images (Stylianou et al., 2019; Kamath et al., 2021), images of different types of airplanes (Maji et al., 2013), or medical images (Yang et al., 2023). Since searching for the optimal DAs is computationally intense (Tamkin et al., 2020), there remains an important bottleneck when it comes to deploying SSL to new or under-studied domains. This becomes particularly noticeable when applying SSL methods on data gathered for real-world applications.

In this paper, we introduce a strategy to obtain positive samples, which generalizes the well established NNCLR SSL

<sup>1</sup>Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

method (Dwibedi et al., 2021). While NNCLR proposes to obtain positive samples by leveraging known DAs and nearest neighbors in the embedding space of the network being trained, we propose to perform nearest neighbour search in the embedding space of a pre-defined mapping of each image to its possible positive samples. The mapping may be generated by a clone of the network being trained – therefore recovering NNCLR – but perhaps most interestingly may also be generated by any available pre-trained network or be hand-crafted. This flexibility allows to (i) enable simple injection of prior knowledge into positive sampling –without relying on tuning the DA– and most importantly (ii) makes the underlying SSL method much more robust to under-tuned DAs parameters. By construction, the proposed method – coined GPS-SSL for Guided Positive Sampling Self-Supervised Learning–, can be coupled off-the-shelf with any SSL method used to learn representations, e.g., BarlowTwins (Zbontar et al., 2021), SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), and VICReg (Bardes et al., 2021). We validate the proposed GPS-SSL approach on a benchmark suite of under-studied datasets, namely FGVCAircraft, PathMNIST, TissueMNIST, and show remarkable improvements over baseline SSL methods. We further evaluate our model on a real-world dataset, Revised-Hotel-ID (R-HID) (Feizi et al., 2022) and show clear improvements of our method compared the baseline SSL methods. Finally, we validate the approach on commonly used image datasets, i.e., Cifar10 and TinyImageNet, with known effective DAs recipes, and show that GPS remains competitive. Through comprehensive ablations, we show that GPS-SSL takes a step towards shifting the focus of designing *well-crafted DAs* to having a better *prior knowledge* embedding space in which choosing the nearest neighbour becomes an attractive positive sampling strategy.

The contributions of this paper can be summarized as:

- We propose a positive sampling strategy, GPS-SSL, that enables SSL models to use prior knowledge about the target-dataset to help with the learning process and reduce the reliance on carefully hand-crafted data augmentation recipes. The prior knowledge is a mapping between images and a few of their closest nearest neighbors that could be computed with a pre-trained network or even be hand-crafted.
- We evaluate GPS-SSL by coupling it with different SSL methods on a benchmark suite of understudied datasets. We show that GPS-augmented approaches significantly outperform the baseline methods when using minimal augmentations, highlighting the potential of GPS to learn representations from under-studied or real-world data. Moreover, when compared to SSL baselines leveraging strong augmentations or on well-studied datasets, the GPS-augmented approaches remain competitive.

- We further evaluate our model on datasets with under-studied applications of hotel retrieval which is of great importance to fight human trafficking. Similar to benchmark datasets, we see on this less studied dataset that GPS-SSL outperforms the baseline SSL methods by a significant margin.

We provide the code for GPS-SSL to reproduce our results on the (anonymized) GitHub: <https://anonymous.4open.science/r/gps-ssl-1E68>, for the research community.

## 2. Related Work

Self Supervised Learning (SSL) is a particular form of unsupervised learning methods in which a given Deep Neural Network (DNN) learns meaningful representations of their inputs without labels.

The variants of SSL are numerous. At the broader scale, SSL defines a pretext task on the input data and train themselves by solving the defined task. In SSL for computer vision, the pretext tasks generally involve creating different views of images and encoding both so that their embeddings are close to each other. However, that criteria alone would not be sufficient to learning meaningful representations as a degenerate solution is for the DNN to simply collapse all samples to a single embedding vector. As such, one needs to introduce an “anti-collapse” term. Different types of solutions have been proposed for this issue, splitting SSL methods into multiple groups, three of which are: 1) Contrastive(Chen et al., 2020; Dwibedi et al., 2021; Kalantidis et al., 2020): this group of SSL methods prevent collapsing by considering all other images in a mini-batch as negative samples for the positive image pair and generally use the InfoNCE (Oord et al., 2018) loss function to push the negative embeddings away from the positive embeddings. 2) Distillation(Grill et al., 2020; He et al., 2020; Chen & He, 2021): these methods often have an asymmetric pair of encoders, one for each positive view, where one encoder (teacher) is the exponential moving average of the other encoder (student) and the loss only back-propagates through the student encoder. In general, this group prevents collapsing by creating asymmetry in the encoders and defines the pre-text task that the student encoder must predict the teach encoder’s output embedding. 3) Feature Decorrelation(Bardes et al., 2021; Zbontar et al., 2021): These methods focus on the statistics of the embedding features generated by the encoders and defines a loss function to encourage the embeddings to have certain statistical features. By doing so, they explicitly force the generated embeddings not to collapse. For example, (Bardes et al., 2021) encourages the features in the embeddings to have high variance, while being invariant to the augmentations and also having a low covariance among different features in the embeddings. Besides these groups,

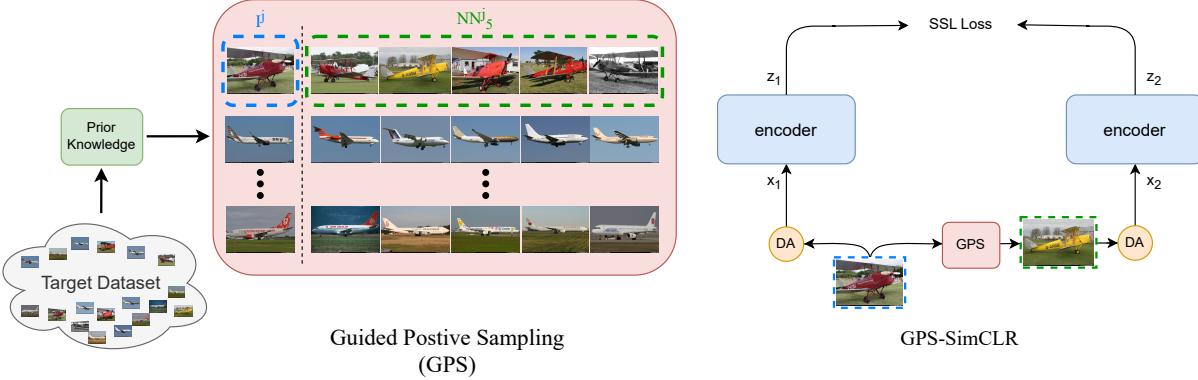


Figure 1. Our strategy, GPS-SSL, for positive sampling based on prior knowledge DA-based methods.

there are multiple other techniques for preventing collapsing, such as clustering methods (Caron et al., 2020; Xie et al., 2016), gradient analysis methods (Tao et al., 2022).

Although the techniques used for preventing collapse may differ among these groups of methods, they generally require the data augmentations to be chosen and tuned carefully in order to achieve high predictive performance (Chen et al., 2020). Although choosing the optimal data augmentations and hyper-parameters may be considered a solved problem for popular datasets such as Cifar10 (Krizhevsky et al., 2009) or ImageNet (Deng et al., 2009), the SSL dependency on DA remains their main limitation to be applied to large real-world datasets that are not akin natural images. Due to the importance of DA upon the DNN’s representation quality, a few studies have attempted mitigation strategies. For example, (Cabannes et al., 2023b) ties the impact of DA with the implicit prior of the DNN’s architecture, suggesting that informed architecture may reduce the need for well designed DA although no practical answer was provided. (Cabannes et al., 2023a) proposed to remove the need for DA at the cost of requiring an oracle to sample the positive samples from the original training set. Although not practical, this study brings a path to train SSL without DA. Also (Van Gansbeke et al., 2020) proposes a two-stage learning process where first a clustering method with a pretext task is applied to the dataset and soft labels are acquired for performing an unsupervised learning on top of it. Additionally, a key limitation with DA lies in the need to be implemented and fast to produce. In fact, the strong DA strategies required by SSL are one of the main computational time bottleneck of current training pipelines (Bordes et al., 2023). Lastly, the over-reliance on DA may have serious fairness implications since, albeit in a supervised setting, DA was shown to impact the DNN’s learned representation in favor of specific classes in the dataset (Balestrieri et al., 2022; Kirichenko et al., 2023).

All in all, SSL would greatly benefit from a principled strategy to embed a priori knowledge into generating positive pairs that does not rely on DA. We propose a first

step towards such Guided Positive Sampling (GPS) below.

### 3. Guided Positive Sampling for SSL

We propose a novel strategy, *Guided Positive Sampling Self-Supervised Learning* (GPS-SSL), that takes advantage of prior knowledge for positive sampling to make up for the sub-optimality of generating positive pairs solely from DA in SSL.

#### 3.1. Nearest Neighbor Positive Sampling in Any Desired Embedded Space

As theoretically shown in several studies (HaoChen et al., 2021; Balestrieri & LeCun, 2022; Kiani et al., 2022), the principal factors that impact the quality of the learned representation resides in how the positive pairs are defined. In fact, we recall that in all generality, SSL losses that are minimized can mostly be expressed as

$$\mathcal{L}_{SSL} = \sum_{(\mathbf{x}, \mathbf{x}') \in \text{PositivePairs}} \text{Distance}(f_\theta(\mathbf{x}), f_\theta(\mathbf{x}')) - \text{Diversity}(\{f_\theta(\mathbf{x}), \mathbf{x} \in \mathbb{X}\}), \quad (1)$$

for the current training or mini-batch  $\mathbb{X}$ , with a distance measure such as the  $\ell_2$  norm or the cosine similarity, and a diversity measure such that the rank of the embeddings or proxies of their entropy. All in all, defining the right set of PositivePairs is what determines the ability of the final representation to solve downstream tasks. The common solution is to repeatedly apply DA onto a single datum to generate such positive pairs:

$$\text{PositivePairs} \triangleq \{(DA(\mathbf{x}), DA(\mathbf{x})), \forall \mathbf{x} \in \mathbb{X}\}, \quad (2)$$

where the DA operator includes the random realisation of the DA such as the amount of rotation or zoom being applied onto its input image. However, this strategy often reaches its limits since such DAs need to be easily implemented for the specific data being used, and needs to be known a priori. When considering an image dataset, the challenge of designing DA for less common datasets, e.g., FGVCaircraft,

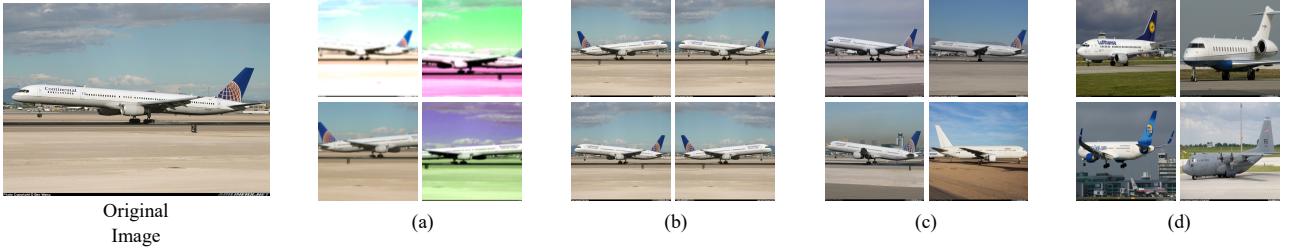


Figure 2. An example (a) *StrongAug* and (b) *RHflipAug* applied to an image from the FGVCaircraft dataset. Furthermore, (c) and (d) depict examples of the 4 nearest neighbors calculated by CLIP and VAE embeddings, respectively.

led practitioners to instead train the model on a dataset such as ImageNet, where strong DAs have already been discovered, and then transfer the model to other datasets. This however has its limits when considering images from completely different domains –e.g. medical images.

We propose GPS-SSL, an alternative strategy to sample positive pairs, which can be used off-the-shelf with any baseline SSL method –e.g., SimCLR, VICReg. GPS-SSL defines positive pairs through nearest neighbour sampling in an a priori known embedding space denoted as  $g_\gamma$ .

First, we define the collection of samples that are less than  $\tau > 0$  away from a query sample  $\mathbf{x} \in \mathbb{X}$  in the chosen embedding space as

$$\mathcal{B}(\mathbf{x}) \triangleq \{\mathbf{x}' \in \mathbb{X} : \|g_\gamma(\mathbf{x}) - g_\gamma(\mathbf{x}')\|_2^2 < \tau\}. \quad (3)$$

From eq. (3), GPS-SSL obtains positive pairs by selecting the furthest point from  $\mathbf{x}$  in  $\mathcal{B}(\mathbf{x})$  as in

$$\begin{aligned} \text{PositivePairs}_{\text{GPS}} &\triangleq \{(\text{DA}(\mathbf{x}), \text{DA}(\mathbf{x}')), \\ &\forall (\mathbf{x}, \mathbf{x}') \in \mathbb{X}^2 : \mathbf{x}' = \arg \max_{\mathbf{u} \in \mathcal{B}(\mathbf{x})} \|g_\gamma(\mathbf{u}) - g_\gamma(\mathbf{x})\|_2^2\}, \end{aligned} \quad (4)$$

In short, we replace the set of positive pairs generated from applying a given DA to a same input, by applying a given DA onto two different inputs found so that one is the nearest neighbor of the other in some embedding space provided by  $g_\gamma$ . From this, we obtain a first direct result below making GPS-SSL recover a powerful existing method known as NNCLR (Dwibedi et al., 2021).

**Proposition 1.** *For any employed DA, GPS-SSL which replaces eq. (2) by eq. (4) in any SSL loss (eq. (1)) recovers (i) input space nearest neighbor positive sampling when  $g_\gamma$  is the identity and  $\tau \gg 0$ , (ii) standard SSL when  $g_\gamma$  is the identity but  $\tau \rightarrow 0$ , and (iii) NNCLR when  $g_\gamma = f_\theta$  and  $\tau \rightarrow 0$ .*

The above result provides a first strong argument demonstrating how GPS-SSL does not reduce the capacity of SSL; in fact, it introduces a novel axis of freedom—namely the

design of  $(g_\gamma, \tau)$ —to extend current SSL beyond what is amenable solely by tweaking the baseline SSL network  $f_\theta$ , or the used DA. The core motivation of the presented method is that the ability to design  $g_\gamma$  reduces the laborious task of designing effective DA recipes. In fact, if we consider the case where the original DA is part of the original dataset

$$\forall \mathbf{x} \in \mathbb{X}, \exists \rho : \text{DA}(\mathbf{x}; \rho) \in \mathbb{X}, \quad (5)$$

i.e., for any sample in the training set  $\mathbb{X}$ , there exists at least one DA configuration ( $\rho$ ) that produces another training set sample; GPS-SSL can recover standard SSL albeit without employing any DA.

**Theorem 1.** *Performing standard SSL (employing eq. (2) into eq. (1)) with a given DA and a training set for which eq. (5) holds, is equivalent to performing GPS-SSL (employing eq. (2) into eq. (1)) without any DA and by setting  $g_\gamma$  to be invariant to that DA, i.e.  $g_\gamma(\text{DA}(\mathbf{x})) = g_\gamma(\mathbf{x})$ .*

By construction from eq. (5) and assuming that one has the ability to design such an invariant  $g_\gamma$ , it is clear that the nearest neighbour within the training set for any  $\mathbf{x} \in \mathbb{X}$  will be the corresponding samples  $\text{DA}(\mathbf{x})$  therefore proving theorem 1. That result is quite impractical but nevertheless provides a great motivation to GPS-SSL. Note that  $g_\gamma$  not only has the ability to mitigate the DA design, but can also be used jointly with DA, hence allowing one to embed as much a priori knowledge as possible through both  $g_\gamma$  and said DA simultaneously.

**The design of  $g_\gamma$ .** The proposed strategy (eq. (4)) is based on finding the nearest neighbors of different candidate inputs in a given embedding space, i.e., a prescribed mapping  $g_\gamma$ . Throughout our study, we will focus on the most direct solution of employing a previously pre-trained mapping. The pre-training may or may not have occurred on the same dataset being considered for SSL. Naturally, the alignment between both datasets affects the quality and reliability of the embeddings. If one does not have access to such pre-trained models, an alternative solution is to first learn abstracted representation of the data, e.g., using an MAE (He et al., 2022) or VAE (Kingma & Welling, 2013), and

then use the said representations for  $g_\gamma$ . In this setting, the motivation lies in the final SSL representation being superior to the encoder ( $g_\gamma$ ) alone for solving downstream tasks.

We provide some examples of the resulting positive pairs with our strategy in Figure 1. In this figure, we use a pre-trained model to calculate the set of  $k$  nearest neighbors for each image  $x$  in the target dataset. Then for each image  $x$ , the model randomly chooses the positive image from the nearest neighbors in embedding space (recall eq. (4)). Finally, both the original image and the produced positive sample are augmented using the chosen DA and passed as a positive pair of images through the encoders. Note that as per proposition 1, GPS-SSL may choose the image itself as its own positive sample, but the probability of it happening reduces as  $\tau$  increases. As we will demonstrate in the later sections, the proposed positive sampling strategy often outperforms the baseline DA-based positive pair sampling strategy on multiple datasets.

**Relation to NNCLR.** The commonality of NNCLR and GPS-SSL has been brought forward in proposition 1. In short, they both choose the nearest neighbor of input images as the positive sample. However, the embedding space in which the nearest neighbor is chosen is different. In NNCLR, the model being trained creates the embedding space which is thus updated at every training step, i.e.,  $g_\gamma = f_\theta$ . However, GPS-SSL generalizes that in the sense that the nearest neighbors can stem from any prescribed mapping, without the constraint that it is trained as part of the SSL training, or even that it takes the form of a DNN. The fact that NNCLR only considers the model being trained to obtain its positive samples also makes it heavily dependent on complex and strong augmentations to produce non degenerate results. Yet, our ability to prescribe other mappings for the nearest neighbor search makes GPS-SSL much less tied to the employed DA. We summarize and contrast with alternative SSL methods in Figure 3.

## 4. Empirical Validation on Benchmarked Datasets

In our experiments, we train the baseline SSL methods and the proposed GPS-SSL with two general sets of augmentations: *StrongAug*, which are augmentations that have been tuned on either Cifar10 (Krizhevsky et al., 2009) or on ImageNet in the case of TinyImageNet (Le & Yang, 2015), and the under-studied datasets (i.e., FGVCAircraft (Maji et al., 2013), PathMNIST (Yang et al., 2023), TissueMNIST (Yang et al., 2023), and R-HID). *RHFlipAug*, representing the scenario where we do not know the correct augmentations and use minimal ones. The set of *StrongAug* consists of random-resized cropping, random-horizontal flipping, color jittering, gray-scaling, Gaussian blurring, and solarizing, while *RHFlipAug* only uses random-horizontal flipping.

Table 1. Classification accuracy of a ResNet18 in different ablation settings, comparing GPS-SimCLR when different pre-trained backbones (GPS-BB) are used to obtain embeddings for nearest-neighbor calculation, i.e., prior knowledge.

GPS-BB	FGVCAircraft	
	RHFlipAug	StrongAug
ViT-B <sub>MAE</sub>	10.53	29.55
ViT-L <sub>MAE</sub>	14.70	35.28
RN50 <sub>SUP</sub>	18.15	41.47
RN50 <sub>VAE</sub>	11.04	32.06
RN50 <sub>CLIP</sub>	<b>19.38</b>	<b>50.08</b>

Table 2. Classification accuracy of a ResNet18 in different ablation settings. Best performance in *StrongAug* setting of SimCLR and GPS-SimCLR given different learning rates (LR).

LR	FGVCAircraft	
	SimCLR	GPS-SimCLR
0.003	21.39	35.7
0.01	30.18	43.68
0.03	39.27	49.57
0.1	39.81	50.08
0.3	39.87	48.10

In order to thoroughly validate GPS-SSL as an all-purpose strategy for SSL, we consider SimCLR, BYOL, NNCLR, and VICReg as baseline SSL models, and for each of them, we consider the standard SSL positive pair generation (eq. (2)) and the proposed one (eq. (4)). We opted for a *randomly-initialized* ResNets backbone (He et al., 2016) as encoder. We also bring forward the fact that most SSL methods are generally trained on a large dataset for which strong DAs are known and well-tuned, such as ImageNet, and the learned representation is then transferred to solve tasks on smaller and domain-specific datasets. In many cases, training those SSL models directly on those less known datasets lead to catastrophic failures, as the optimal DAs have not yet been discovered. Lastly, we consider five different embeddings for  $g_\gamma$ : one obtained using supervised learning on ImageNet, one CLIP vision-language model (Radford et al., 2021) trained on LAION-400M (Schuhmann et al., 2021), one with VAEs (Kingma & Welling, 2013) trained on Object365, and two with MAEs (He et al., 2022) trained on ImageNet as well. We also show that our method is more robust to hyper-parameter changes (see Tables 1 and 2). Since the embedding models are trained on ImageNet or LAION-400M, the results reported throughout this study remain practical since the labels of the target datasets, on which SSL models are trained and evaluated, are never observed for the training of neither  $g_\gamma$  nor  $f_\theta$ .

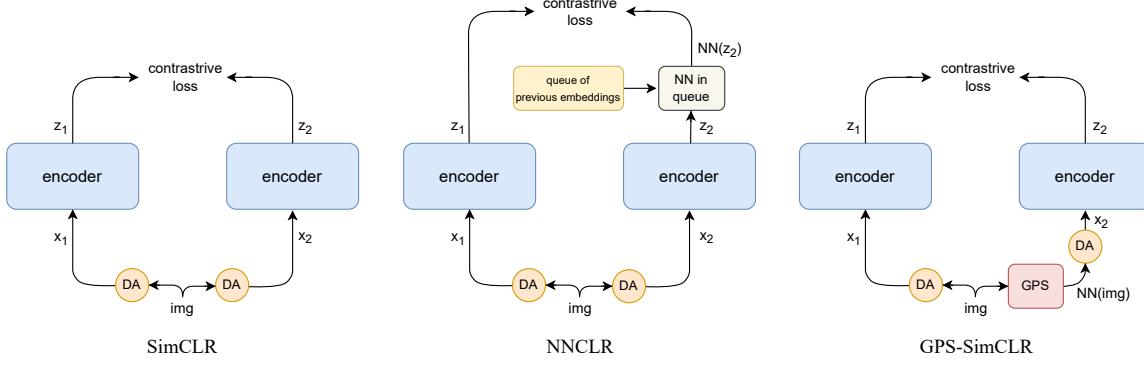


Figure 3. Architectures of SimCLR, NNCLR, and GPS-SimCLR. This figure demonstrates where the data augmentation (DA) happens in each method and also how the nearest neighbor (NN) search is different between NNCLR and GPS-SimCLR.

Table 3. Classification accuracy of baseline SSL methods with and without GPS-SSL on four datasets on ResNet50 using pretrained RN50<sub>CLIP</sub> embeddings for positive sampling. We consider both *StrongAug* (Strong Augmentation) and *RHFlipAug* (Weak Augmentation) settings. The set of DA used for *StrongAug* are random-resized-crop, random-horizontal-flip, color-jitter, gray-scale, gaussian-blur, and solarization. For the *RHFlipAug* setting, the only DA used is random horizontal flip. We mark the **first**, **second**, and **third** best performing models accordingly.

Aug.	Method	Datasets			
		Cifar10 (10 classes)	FGVCAircraft (100 classes)	PathMNIST (9 classes)	TissueMNIST (8 classes)
<i>RHFlipAug</i>	SimCLR	47.01	5.61	63.42	50.35
	BYOL	41.79	6.63	67.08	48.00
	NNCLR	28.46	6.33	56.70	37.98
	Barlow Twins	41.73	5.34	53.27	43.57
	VICReg	37.51	6.18	46.46	39.79
	GPS-SimCLR (ours)	<b>85.08</b>	<b>18.18</b>	<b>87.79</b>	<b>53.14</b>
	GPS-BYOL (ours)	84.07	13.50	87.67	53.05
	GPS-Barlow (ours)	<b>84.45</b>	<b>17.34</b>	<b>88.77</b>	<b>56.63</b>
<i>StrongAug</i>	GPS-VICReg (ours)	<b>85.58</b>	<b>18.81</b>	<b>88.91</b>	<b>56.44</b>
	SimCLR	90.24	<b>47.11</b>	<b>93.64</b>	<b>58.53</b>
	BYOL	90.50	34.23	<b>93.29</b>	56.63
	NNCLR	90.03	34.80	92.87	52.57
	Barlow Twins	88.34	18.12	92.03	<b>61.69</b>
	VICReg	<b>91.21</b>	38.74	<b>93.22</b>	<b>60.18</b>
	GPS-SimCLR (ours)	<b>91.17</b>	<b>55.60</b>	92.30	55.59
	GPS-BYOL (ours)	<b>91.15</b>	44.28	92.40	55.03
	GPS-Barlow (ours)	88.52	15.47	91.98	57.04
	GPS-VICReg (ours)	89.71	<b>47.29</b>	92.55	55.79

**Strong Augmentation Experiments.** The DAs in the *StrongAug* configuration consist of strong augmentations that usually distort the size, resolution, and color characteristics of the original image. First, we note that in this setting, GPS-SSL generally does not harm the performance of the baseline SSL model on common datasets, i.e. Cifar10 (Table 3). In fact, GPS-SSL performs comparable to the best-performing baseline SSL model on Cifar10, i.e., VICReg. We believe that the main reason lies in the fact that the employed DA has been specifically designed for those datasets (and ImageNet). However, we observe that GPS-SSL out-

performs (on FGVCAircraft and TissueMNIST) or is comparable to (on PathMNIST) the baseline SSL methods for the under-studied and real-word datasets (Table 3). The reason for this is that, the optimal set and configuration of DA for one dataset is not necessarily the optimal set and configuration for another, and while SSL solely relies on DA for its positive samples, GPS-SSL is able to alleviate that dependency through  $g_\gamma$  and uses positive samples that can be more useful than default DAs, as seen in Figure 2. Note that our method's runtime is similar to the baseline SSL method on the dataset it is learning and does not hinder

330 the training process (Figure 4).

331  
 332 **Weak-Augmentation Experiments.** We perform all experiments under the *RHFlipAug* setting as well, showing  
 333 GPS-SSL also produces high quality representations in  
 334 this setting, validating theorem 1. As shown in Table 3,  
 335 GPS-SSL significantly outperforms all baseline SSL meth-  
 336 ods across both well-studied and under-studied datasets.  
 337 These results show that our GPS-SSL strategy, though  
 338 conceptually simple, coupled with the *RHFlipAug* setting,  
 339 approximates strong augmentations used in the *StrongAug*  
 340 configuration. This creates a significant advantage for GPS-SSL  
 341 to be applied to real-world datasets where strong aug-  
 342 mentations are not readily available, but where the invariances  
 343 learned by  $g_\gamma$  generalizes to them.

344  
 345 **Ablation Study** In this section we explore multiple ablation  
 346 experiments in order to show GPS-SSL improves SSL and  
 347 is indeed a future direction for improving SSL methods.  
 348 First, we compare SSL and GPS training on Cifar10 and  
 349 FGVCAircraft starting from a randomly initialized back-  
 350 bone (realistic setting), supervised ImageNet pre-trained  
 351 weights, or CLIP pre-trained weights to explore whether the  
 352 improvement of GPS-SSL are due to better positive sam-  
 353 pling or simply because of using a strong prior knowledge.  
 354 We show in Table 6 that GPS-SSL performs better than the  
 355 baseline SSL methods, even when they both have access to  
 356 the pre-trained network weights. This proves that the im-  
 357 provement in performance of GPS-SSL compared to base-  
 358 line SSL methods is indeed due to better positive sampling.

359  
 360 Next, in Table 1, we compare GPS-SimCLR with five differ-  
 361 ent embeddings for  $g_\gamma$ . We observe that as the pre-trained  
 362 embeddings become higher quality, the performance of our  
 363 method increases in both the *RHFlipAug* and *StrongAug*  
 364 setting. However, note that even given the weakest embed-  
 365 dings, i.e., the ViT-B<sub>MAE</sub> embeddings, GPS-SimCLR still  
 366 outperforms the baseline SimCLR in the *RHFlipAug* setting,  
 367 highlighting that the nearest neighbors add value to the  
 368 learning process when DA recipes are not readily available.

369  
 370 We further explore if the improvement of GPS-SSL holds  
 371 when methods are trained longer. To that end, we train  
 372 a ResNet18 for 1000 epochs with SimCLR and VICReg  
 373 with *StrongAug*, along with their GPS versions, on Cifar10  
 374 and FGVCAircraft and compare the results with the  
 375 performance from 400 epochs. As seen in Table 4, the  
 376 improvement of GPS-SSL compared to the baseline SSL  
 377 method holds on FGVCAircraft dataset and remains compa-  
 378 rable on Cifar10, showcasing the robustness of GPS-SSL.

379 Moreover, we compare GPS-SSL and baseline SSL meth-  
 380 ods on a larger scale dataset. We train and evaluate Sim-  
 381 CLR and VICReg and their GPS versions with different  $g_\gamma$   
 382 on TinyImageNet for 200 epochs with a ResNet50 under  
 383 *StrongAug* and *RHFlipAug* settings. As seen in Table 5, un-

Table 4. Test accuracy comparison of GPS-SSL after 1K epochs versus 400 training epochs. We show the improvements of GPS-SimCLR are still significant on FGVCAircraft and comparable on Cifar10.

Method	Cifar10		FGVCAircraft	
	400 eps	1000 eps	400 eps	1000 eps
SimCLR	88.26	<b>91.25</b>	39.87	45.55
GPS-SimCLR	<b>89.57</b>	91.10	<b>50.08</b>	<b>51.64</b>
VICReg	89.34	<b>90.61</b>	33.21	41.19
GPS-VICReg	<b>89.68</b>	89.84	<b>45.48</b>	<b>49.29</b>

Table 5. Test accuracy comparison of GPS-SSL and baseline SSL after 200 training epochs on TinyImageNet (TinyIN) with ResNet50. The GPS backbone (GPS-BB) and dataset used for pretraining it (GPS-DS) is also specified.

Method	GPS-BB	GPS-BB	TinyIN	
			<i>RHFlipAug</i>	<i>StrongAug</i>
SimCLR	—	—	3.17	42.25
VICReg	—	—	2.81	44.02
GPS-SimCLR	ViT-L <sub>MAE</sub>	TinyIN	28.32	41.69
GPS-VICReg	ViT-L <sub>MAE</sub>	TinyIN	27.67	42.71
GPS-SimCLR	RN50 <sub>CLIP</sub>	LAIION-400M	<b>40.73</b>	48.09
GPS-VICReg	RN50 <sub>CLIP</sub>	LAIION-400M	40.26	<b>48.47</b>

der the *RHFlipAug* setting, all GPS versions significantly outperform the baseline SSL methods. We can further see that using *StrongAug*, GPS-SSL with CLIP and MAE embeddings outperforms and is comparable to the baseline SSL methods, respectively.

Finally, we aim to measure the sensitivity of the performance of a baseline SSL method to a hyper-parameter, i.e., learning rate, with and without GPS-SSL. In this experiment, we report the best performance of SimCLR and GPS-SimCLR given different learning rates in the *StrongAug* setting. We observe that GPS-SSL when applied to a baseline SSL method is as robust, if not more robust, to hyper-parameter changes. The results are reported in Table 2. We perform further ablations, e.g., comparing GPS-SSL with the linear probing performance, in Appendix A.4.

## 5. Case Study on the Hotels Image Dataset

In this section, we study how GPS-SSL compares to baseline SSL methods on an under-studied real-world dataset. We opt the R-HID (Feizi et al., 2022) dataset for our evaluation which gathers hotel images for the purpose of countering human-trafficking. R-HID provides a single train set alongside 4 evaluation sets, each with a different level of difficulty, ranging from known (seen) hotel chains and branches to unknown chains and branches.

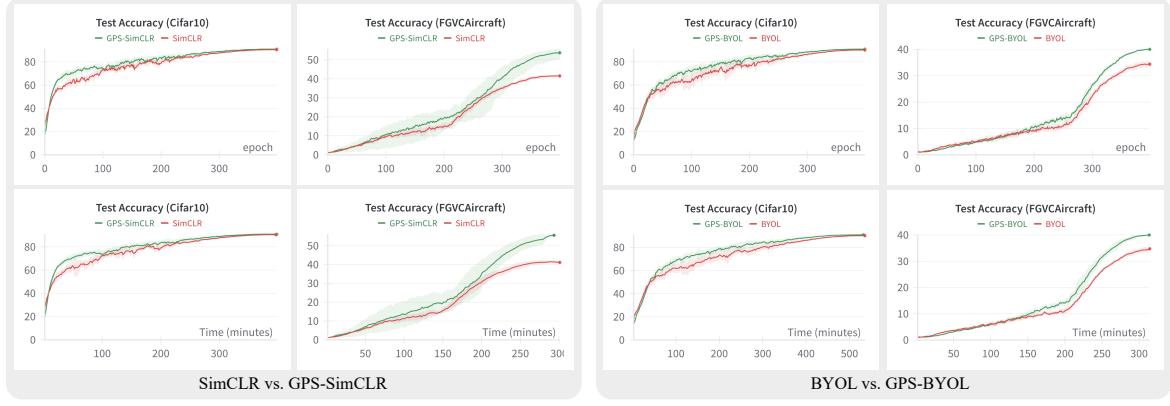


Figure 4. Comparing the runtime of BYOL vs. GPS-BYOL and SimCLR vs. GPS-SimCLR on FGVC Aircraft and Cifar10. In general, we see while the runtime of GPS-SSL remains the same as the original baseline SSL method, it improves the performance.

Table 6. Comparing SimCLR on ResNet50 with and without GPS-SimCLR from different model initializations with minimal (*Weak*) and strong (*Strong*) augmentations. *RAND*, *SUP*, and *CLIP* represent random weights, ImageNet supervised weights, and CLIP pre-trained weights.

Method	Weight Init.	Cifar10		FGVCAircraft	
		Weak	Strong	Weak	Strong
SimCLR	<i>RAND</i>	46.69	87.39	5.67	27.36
		<b>85.2</b>	<b>90.48</b>	<b>17.91</b>	<b>43.56</b>
GPS-SimCLR	<i>SUP</i>	43.99	94.02	17.91	59.92
		<b>91.3</b>	<b>95.53</b>	<b>39.45</b>	<b>66.88</b>
SimCLR	<i>CLIP</i>	45.57	90.26	6.21	41.04
		<b>89.44</b>	<b>91.23</b>	<b>24.15</b>	<b>49.63</b>

Table 7. R@1 on different splits on R-HID Dataset for SSL methods. The splits are  $\mathcal{D}_{SS}$ : {branch: seen, chain: seen},  $\mathcal{D}_{SU}$ : {branch: unseen, chain: seen},  $\mathcal{D}_{UU}$ : {branch: unseen, chain: unseen}, and  $\mathcal{D}_{??}$ : {branch: unknown, chain: unknown}. We mark the best-performing score in **bold**.

Method	$\mathcal{D}_{SS}$	$\mathcal{D}_{SU}$	$\mathcal{D}_{UU}$	$\mathcal{D}_{??}$
SimCLR	3.28	16.76	20.30	16.00
BYOL	3.69	19.27	23.02	18.47
Barlow Twins	3.04	15.54	18.96	15.06
VICReg	3.41	17.52	20.45	16.53
GPS-SimCLR	<b>4.84</b>	<b>23.67</b>	<b>26.30</b>	<b>22.28</b>
GPS-BYOL	3.89	19.64	23.18	19.38
GPS-Barlow	<b>4.49</b>	<b>21.98</b>	<b>25.23</b>	<b>20.82</b>
GPS-VICReg	<b>5.33</b>	<b>25.71</b>	<b>28.29</b>	<b>23.78</b>

We evaluate the baseline SSL models with and without GPS-SSL with the R-HID dataset and report the Recall@1 (R@1) for the different splits introduced. Based on the findings from 3, we adapt the *StrongAug* setting along with the prior knowledge generated by a CLIP-pretrained ResNet50.

As seen in Table 7, SSL baselines always get an improvement when used with GPS-SSL. The reason the baseline

SSL methods underperform compared to their GPS-SSL version is that the positive samples generated only using DA lack enough diversity since the images from R-HID dataset have various features and DA recipes limit the information the network learns; however, paired with GPS-SSL, we see a clear boost in performance across all different splits due to the information extracted from the nearest neighbors.

## 6. Conclusions

In this paper, we proposed GPS-SSL, a strategy to obtain positive samples for Self-Supervised Learning. In particular, GPS-SSL moves away from the usual DA-based positive sampling by instead producing positive samples from the nearest neighbors of the data as measured in a given embedding space. That is, GPS-SSL introduces a new axis of research to advance SSL methods that is complementary to the design of DA recipes and losses. Through this strategy, we were able to train SSL methods on relatively under-explored datasets such as medical images—without having to search and tune for the right DA. Those results open new avenues to employ SSL on datasets for which effective DA recipes are not available. In fact, we observe that while GPS-SSL meets or surpasses SSL performances across our experiments, the performance gap is more significant when the optimal DAs are not known, e.g., in PathMNIST and TissueMNIST. Besides practical applications, GPS-SSL finally provides a novel strategy to embed prior knowledge into SSL.

**Limitations.** The main limitation of our method is akin to the one of SSL, it requires the knowledge of the embedding space in which positive samples are produced using the nearest neighbors. This limitation is on par with standard SSL’s reliance on DA, but its formulation is somewhat dual (recall theorem 1) in that one may know how to design such an embedding without knowing the appropriate DA for the dataset, and vice-versa. Alternative techniques like training separate and simple deep networks to provide such embeddings prior to the SSL learning could be considered for future research.

## 440 7. Impact Statement

441 This paper presents work whose goal is to advance the field  
 442 of Machine Learning. There are many potential societal  
 443 consequences of our work, none which we feel must be  
 444 specifically highlighted here.

## 445 446 References

- 447 Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent,  
 448 P., Rabat, M., LeCun, Y., and Ballas, N. Self-supervised  
 449 learning from images with a joint-embedding predictive  
 450 architecture. *arXiv preprint arXiv:2301.08243*, 2023.
- 451 Balestiero, R. and LeCun, Y. Contrastive and non-  
 452 contrastive self-supervised learning recover global and  
 453 local spectral embedding methods. *Advances in Neural  
 454 Information Processing Systems*, 35:26671–26685, 2022.
- 455 Balestiero, R., Bottou, L., and LeCun, Y. The effects  
 456 of regularization and data augmentation are class  
 457 dependent. In Koyejo, S., Mohamed, S., Agarwal,  
 458 A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances  
 459 in Neural Information Processing Systems*, volume 35, pp.  
 460 37878–37891. Curran Associates, Inc., 2022. URL  
 461 [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/f73c04538a5e1cad40ba5586b4b517d3-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/f73c04538a5e1cad40ba5586b4b517d3-Paper.pdf).
- 462 Balestiero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar,  
 463 S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian,  
 464 Y., et al. A cookbook of self-supervised learning. *arXiv  
 465 preprint arXiv:2304.12210*, 2023.
- 466 Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-  
 467 invariance-covariance regularization for self-supervised  
 468 learning. *arXiv preprint arXiv:2105.04906*, 2021.
- 469 Bordes, F., Balestiero, R., and Vincent, P. Towards democ-  
 470 ratizing joint-embedding self-supervised learning, 2023.
- 471 Cabannes, V., Bottou, L., Lecun, Y., and Balestiero, R.  
 472 Active self-supervised learning: A few low-cost relation-  
 473 ships are all you need, 2023a.
- 474 Cabannes, V., Kiani, B., Balestiero, R., Lecun, Y., and  
 475 Bietti, A. The SSL interplay: Augmentations, induc-  
 476 tive bias, and generalization. In Krause, A., Brunskill,  
 477 E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J.  
 478 (eds.), *Proceedings of the 40th International Conference  
 479 on Machine Learning*, volume 202 of *Proceedings of  
 480 Machine Learning Research*, pp. 3252–3298. PMLR,  
 481 23–29 Jul 2023b. URL <https://proceedings.mlr.press/v202/cabannes23a.html>.
- 482 Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P.,  
 483 and Joulin, A. Unsupervised learning of visual features  
 484 by contrasting cluster assignments. *Advances in neural  
 485 information processing systems*, 33:9912–9924, 2020.
- 486 Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A  
 487 simple framework for contrastive learning of visual rep-  
 488 resentations. In *International conference on machine  
 489 learning*, pp. 1597–1607. PMLR, 2020.
- 490 Chen, X. and He, K. Exploring simple siamese represen-  
 491 tation learning. In *Proceedings of the IEEE/CVF con-  
 492 ference on computer vision and pattern recognition*, pp.  
 493 15750–15758, 2021.
- 494 Chen, X., Xie, S., and He, K. An empirical study of training  
 495 self-supervised vision transformers. In *Proceedings of the  
 496 IEEE/CVF international conference on computer vision*,  
 497 pp. 9640–9649, 2021.
- 498 Dangovski, R., Jing, L., Loh, C., Han, S., Srivastava, A.,  
 499 Cheung, B., Agrawal, P., and Soljačić, M. Equivariant  
 500 contrastive learning. *arXiv preprint arXiv:2111.00899*,  
 501 2021.
- 502 Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei,  
 503 L. Imagenet: A large-scale hierarchical image database.  
 504 In *2009 IEEE conference on computer vision and pattern  
 505 recognition*, pp. 248–255. Ieee, 2009.
- 506 Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and  
 507 Zisserman, A. With a little help from my friends: Nearest-  
 508 neighbor contrastive learning of visual representations. In  
 509 *Proceedings of the IEEE/CVF International Conference  
 510 on Computer Vision*, pp. 9588–9597, 2021.
- 511 Feizi, A., Casanova, A., Romero-Soriano, A., and Rabbany,  
 512 R. Revisiting hotels-50k and hotel-id. *arXiv preprint  
 513 arXiv:2207.10200*, 2022.
- 514 Garrido, Q., Najman, L., and Lecun, Y. Self-supervised  
 515 learning of split invariant equivariant representations.  
 516 *arXiv preprint arXiv:2302.10283*, 2023.
- 517 Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V.,  
 518 Joulin, A., and Misra, I. Imagebind: One embedding  
 519 space to bind them all. In *Proceedings of the IEEE/CVF  
 520 Conference on Computer Vision and Pattern Recognition*,  
 521 pp. 15180–15190, 2023.
- 522 Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P.,  
 523 Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z.,  
 524 Gheshlaghi Azar, M., et al. Bootstrap your own latent-a  
 525 new approach to self-supervised learning. *Advances in  
 526 neural information processing systems*, 33:21271–21284,  
 527 2020.
- 528 HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable  
 529 guarantees for self-supervised deep learning with spec-  
 530 tral contrastive loss. *Advances in Neural Information  
 531 Processing Systems*, 34:5000–5011, 2021.

- 495 He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning  
 496 for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,  
 497 pp. 770–778, 2016.  
 498
- 500 He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation  
 501 learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738,  
 502 2020.  
 503
- 505 He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick,  
 506 R. Masked autoencoders are scalable vision learners. In  
 507 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.  
 508
- 510 Kalantidis, Y., Sarıyıldız, M. B., Pion, N., Weinzaepfel,  
 511 P., and Larlus, D. Hard negative mixing for contrastive  
 512 learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020.  
 513
- 515 Kamath, R., Rolwes, G., Black, S., and Stylianou, A. The  
 516 2021 hotel-id to combat human trafficking competition  
 517 dataset. *arXiv preprint arXiv:2106.05746*, 2021.  
 518
- 519 Kiani, B. T., Balestrieri, R., Chen, Y., Lloyd, S., and LeCun,  
 520 Y. Joint embedding self-supervised learning in the kernel  
 521 regime. *arXiv preprint arXiv:2209.14884*, 2022.  
 522
- 523 Kingma, D. P. and Welling, M. Auto-encoding variational  
 524 bayes. *arXiv preprint arXiv:1312.6114*, 2013.  
 525
- 526 Kirichenko, P., Ibrahim, M., Balestrieri, R., Bouchacourt,  
 527 D., Vedantam, R., Firooz, H., and Wilson, A. G. Understanding  
 528 the detrimental class-level effects of data augmentation.  
 529 *arXiv preprint arXiv:2401.01764*, 2023.  
 530
- 531 Krizhevsky, A., Hinton, G., et al. Learning multiple layers  
 532 of features from tiny images. 2009.  
 533
- 534 Le, Y. and Yang, X. S. Tiny imagenet visual recogni-  
 535 tion challenge. 2015. URL <https://api.semanticscholar.org/CorpusID:16664790>.  
 536
- 537 Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi,  
 538 A. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.  
 539
- 540 Oord, A. v. d., Li, Y., and Vinyals, O. Representation learn-  
 541 ing with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.  
 542
- 543 Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G.,  
 544 Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J.,  
 545 et al. Learning transferable visual models from natural  
 546 language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.  
 547
- 548 Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk,  
 549 R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and  
 550 Komatsuzaki, A. Laion-400m: Open dataset of clip-  
 551 filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.  
 552
- 553 Stylianou, A., Xuan, H., Shende, M., Brandt, J., Souvenir,  
 554 R., and Pless, R. Hotels-50k: A global hotel recognition  
 555 dataset. *arXiv preprint arXiv:1901.11397*, 2019.  
 556
- 557 Tamkin, A., Wu, M., and Goodman, N. Viewmaker net-  
 558 works: Learning views for unsupervised representation  
 559 learning. *arXiv preprint arXiv:2010.07432*, 2020.  
 560
- 561 Tao, C., Wang, H., Zhu, X., Dong, J., Song, S., Huang, G.,  
 562 and Dai, J. Exploring the equivalence of siamese self-  
 563 supervised learning via a unified gradient framework. In  
 564 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14431–14440, 2022.  
 565
- 566 Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proes-  
 567 mans, M., and Van Gool, L. Scan: Learning to classify  
 568 images without labels. In *European conference on com-  
 569 puter vision*, pp. 268–285. Springer, 2020.  
 570
- 571 Xiao, T., Wang, X., Efros, A. A., and Darrell, T. What  
 572 should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020.  
 573
- 574 Xie, J., Girshick, R., and Farhadi, A. Unsupervised deep  
 575 embedding for clustering analysis. In *International con-  
 576 ference on machine learning*, pp. 478–487. PMLR, 2016.  
 577
- 578 Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister,  
 579 H., and Ni, B. Medmnist v2-a large-scale lightweight  
 580 benchmark for 2d and 3d biomedical image classification.  
 581 *Scientific Data*, 10(1):41, 2023.  
 582
- 583 Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Bar-  
 584 low twins: Self-supervised learning via redundancy reduc-  
 585 tion. In *International Conference on Machine Learning*,  
 586 pp. 12310–12320. PMLR, 2021.  
 587

550 **A. Appendix**551 **A.1. R-HID Splitting method**

553 R-HID (Feizi et al., 2022) is created carefully to make sure no data leakage occurs. They mention how the total data is  
 554 divided into the train and the multiple test splits. More specifically, first a set of chains (along with *all* their branches) are  
 555 reserved for the  $\mathcal{D}_{UU}$  to make sure the chains (super-classes) and branches (classes) are not seen during training. Next, out  
 556 of the remaining chains, a set of the branches are chosen to add *all* of their images to the  $\mathcal{D}_{SU}$  test split (since the training  
 557 set will have other images from other branches from the same chain, but not the same branch images). Finally, out of the  
 558 remaining branches, the images in each are split between  $\mathcal{D}_{SS}$  and train, creating the final test split that has a subset of the  
 559 branches seen during training. With this procedure, they make sure of the table of overlapping below. More details regarding  
 560 the splits is provided in the original paper.  
 561

562 **A.2. Computational and Memory Cost of GPS-SSL**  
 563

564 The computational cost of GPS-SSL and standard SSL are not significantly different; the main distinction lies in the  
 565 positive sampling and acquiring the positive samples. In our approach, we investigate acquiring the positive samples from a  
 566 pre-trained network, which entails an additional cost of computing the embeddings of the target dataset and computing the  
 567  $k$ -nearest neighbors of each data point. However, it's important to note that the computation of the  $k$ -nearest neighbors  
 568 needs to be performed only once during the training procedure and can be cached for any SSL method on the same dataset  
 569 and GPS-BB, rendering the additional computation negligible.

570 Regarding additional memory usage, GPS-SSL requires storing the indexes of the  $k$ -nearest neighbors of each data point in  
 571 the training set. Thus, the order of magnitude of the memory usage is  $O(N)$  ( $k \ll N$ ) for a dataset of size  $N$ . However,  
 572 it's noteworthy that this array does not need to be loaded onto the GPU during training.  
 573

574 **A.3. Hyper-Parameter Search**  
 575

576 In all main experiments (Table 3), we train for both 400 and 1000 epochs with a batch size of 256 using one RTX  
 577 8000 GPU for all methods. To ensure we are choosing the correct hyper-parameters for a fair comparison, we search  
 578 over a vast range of hyper-parameter combinations ( $lr \in \{1e^{-3}, 3e^{-3}, 3e^{-2}, 1e^{-2}, 3e^{-1}, 1e^{-1}, 1\}$ ,  $classifier\_lr \in$   
 579  $\{3e^{-2}, 1e^{-2}, 3e^{-1}, 1e^{-1}, 1, 3\}$ ,  $weight\_decay \in \{1e^{-4}, 1e^{-3}\}$ ) and for GPS-SSL with all SSL baselines we also search  
 580 over  $k \in \{1, 4, 9, 49\}$ ). For experiments using *RHFlipAug* and *StrongAug*, we use nearest neighbors calculated based on  
 581 embeddings created from a ResNet50 that have been CLIP pre-trained as the prior knowledge. Finally, for each method, we  
 582 report the best classification accuracy for Cifar10, FGVCAircraft, PathMNIST, and TissueMNIST, and Recall@1 (R@1)  
 583 for R-HID in Tables 3, 7, and 12. To calculate both metrics, we first train the encoder on the target dataset using the SSL  
 584 method, with or without GPS-SSL. Then, for classification accuracy, we train a linear classifier on top of it, and for R@1,  
 585 we encode all the images from the test set and calculate the percentage of images which their first nearest neighbor is from  
 586 the same class.  
 587

588 **A.4. Ablation Study**  
 589590 **A.4.1. DIFFERENT BACKBONE**

591 First, we provide the same experiments as in Table 3, but trained with a ResNet18 instead of a ResNet50 and provide the  
 592 results in Table 8. We see the same results for ResNet50 (discussed for Table 3) also hold when ran on a smaller architecture,  
 593 i.e., ResNet18. This shows the improvements of GPS-SSL over baseline SSL methods is more reliable and robust.  
 594

595 **A.4.2. ADDITIONAL DATASETS**  
 596

597 To further evaluate our method, we train and evaluate GPS-SSL on a large-scale dataset, i.e., ImageNet100, a fine-grained  
 598 image classification dataset, i.e., Food101, and finally an additional image classification dataset, i.e., Cifar100. For  
 599 ImageNet100, we train a ResNet50 for 1000 epochs and for the other two additional datasets, we train a ResNet18 for 400  
 600 epochs. As seen in Tables ?? GPS-SSL helps improve the performance of the original SSL methods in both *RHFlipAug* and  
 601 *StrongAug* settings.  
 602

605  
 606 *Table 8.* Classification accuracy of baseline SSL methods with and without GPS-SSL on four datasets on *ResNet18* using pretrained  
 607 *RN50<sub>CLIP</sub>* embeddings for positive sampling. We consider both *StrongAug* (Strong Augmentation) and *RHflipAug* (Weak Augmen-  
 608 tation) settings. The set of DA used for *StrongAug* are random-resized-crop, random-horizontal-flip, color-jitter,  
 609 gray-scale, gaussian-blur, and solarization. For the *RHflipAug* setting, the only DA used is random horizontal  
 610 flip. We mark the **first**, **second**, and **third** best performing models accordingly.  
 611  
 612  
 613  
 614  
 615

616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654	613 614 615 <b>Aug.</b>  <i>RHflipAug</i>	<b>Method</b>	612 <b>Datasets</b>				
			Cifar10 (10 classes)	FGVCAircraft (100 classes)	PathMNIST (9 classes)	TissueMNIST (8 classes)	
605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654	605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654	605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654	SimCLR	47.62	7.70	62.99	52.30
		BYOL	49.72	8.99	77.77	51.00	
		NNCLR	71.74	8.10	56.92	42.59	
		Barlow Twins	42.00	7.53	64.82	49.43	
		VICReg	36.04	4.95	56.92	50.26	
	605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654	605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654	GPS-SimCLR (ours)	<b>85.83</b>	<b>18.48</b>	<b>88.62</b>	<b>55.98</b>
		GPS-BYOL (ours)	84.56	14.79	81.66	<b>56.21</b>	
		GPS-Barlow (ours)	<b>84.83</b>	<b>18.12</b>	<b>87.79</b>	<b>55.86</b>	
		GPS-VICReg (ours)	<b>85.38</b>	<b>20.16</b>	<b>87.83</b>	55.26	
	605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654	605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654	SimCLR	88.26	<b>39.87</b>	91.56	61.51
		BYOL	86.90	27.33	91.24	60.73	
		NNCLR	87.95	39.12	91.14	52.42	
		Barlow Twins	88.89	25.71	<b>92.23</b>	60.06	
		VICReg	<b>89.34</b>	33.21	<b>92.27</b>	59.41	
	605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654	GPS-SimCLR (ours)	<b>89.57</b>	<b>50.08</b>	<b>92.19</b>	<b>62.76</b>	
		GPS-BYOL (ours)	88.46	32.07	91.05	54.05	
		GPS-Barlow (ours)	88.39	25.35	91.55	<b>62.93</b>	
		GPS-VICReg (ours)	<b>89.68</b>	<b>45.48</b>	91.88	<b>62.46</b>	

642 *Table 9. Results of ResNet50 with SimCLR and GPS-SimCLR on ImageNet100.*

643 644 645 646 647 648 649 650 651 652 653 654	643 644 645 646 647 648 649 650 651 652 653 654	643 644 645 646 647 648 649 650 651 652 653 654	642 <b>ImageNet100</b>			
			643 644 645 646 647 648 649 650 651 652 653 654		643 644 645 646 647 648 649 650 651 652 653 654	
			643 644 645 646 647 648 649 650 651 652 653 654	643 644 645 646 647 648 649 650 651 652 653 654	643 644 645 646 647 648 649 650 651 652 653 654	643 644 645 646 647 648 649 650 651 652 653 654
651 652 653 654						

#### A.4.3. ABLATING DIFFERENT $k$ s

655 We additionally ablate different values for  $k$ , i.e., the number of nearest neighbors for each datapoint to consider for the  
 656 positive sampling selection while training a ResNet18 and ResNet50 on FGVCAircraft. Note that when  $k = 0$ , GPS-  
 657 SimCLR reduces to the original SimCLR method. As seen in Figure 5, there is an optimal number of nearest neighbors to  
 658 use and using more results in lower performance, yet still outperforming the original SimCLR.  
 659

Table 10. Results of ResNet18 with SimCLR and GPS-SimCLR on Food101 and Cifar100.

Method	GPS-BB	Food101		Cifar100	
		RHFlipAug	StrongAug	RHFlipAug	StrongAug
SimCLR	—	16.88	73.35	26.1	64.78
GPS-SimCLR	RN50 <sub>CLIP</sub>	<b>70.25</b>	<b>78.94</b>	<b>61.34</b>	<b>66.45</b>

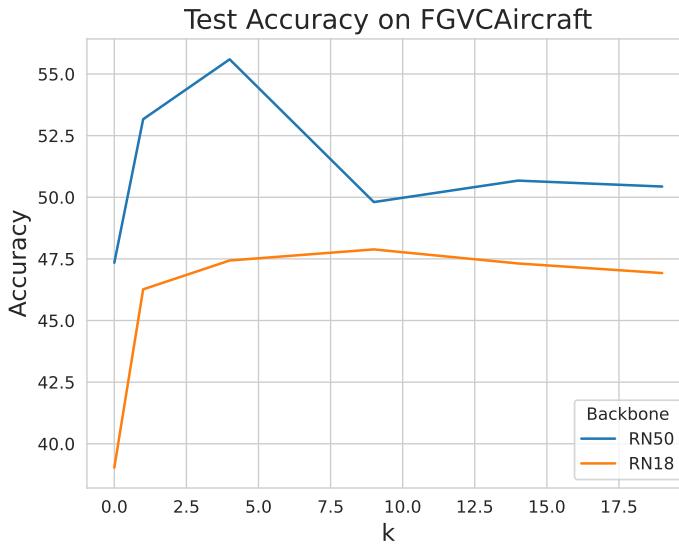
Figure 5. Ablation over different values of  $k$  with GPS-SimCLR on FGVC Aircraft.

Table 11. Results of ResNet18 with MoCo V3 and GPS-MoCoV3 on Cifar10 and FGVC Aircraft.

Method	GPS-BB	Cifar10		FGVC Aircraft	
		RHFlipAug	StrongAug	RHFlipAug	StrongAug
MoCo v3	—	49.7	87.9	8.37	32.7
GPS-MoCo v3	RN50 <sub>CLIP</sub>	<b>83.27</b>	<b>90.39</b>	<b>15.60</b>	<b>45.87</b>

#### A.4.4. MoCo v3

We also explore MoCo v3 (He et al., 2020; Chen et al., 2021), another SSL backbone based on momentum encoders, on FGVC Aircraft and to do so, we train a ResNet18 for 400 epochs with and without GPS on MoCo v3. As seen in Table 11, GPS improves the performance of this method as well.

#### A.4.5. FINETUNING FOR R-HID

We further try a trivial way of transferring knowledge from a pretrained network to other SSL baseline models and compare it to GPS-SimCLR; we initialize the base encoder in any SSL method, i.e., the ResNet18, to the pretrained network’s weights, as opposed to random initialization, and train it i.e., finetuning. Ultimately, we compare the results on R-HID in Table 12.

Although this might perform better if the pretrained network was trained on a visually similar dataset to the target dataset, Table 12 shows that it may harm the generalization on datasets that are different, e.g., ImageNet and R-HID, compared to being trained from scratch. However, GPS-SSL proves to be a stable method for transferring knowledge even if the

715 pretrained and target dataset are visually different (Table 7).

716  
 717  
 718 *Table 12.* Comparing the R@1 performance of SSL methods on R-HID when trained from scratch against being finetuned, i.e., being  
 719 initialized to a supervised ImageNet pretrained network, on a ResNet50. We highlight the difference in R@1 of the pretrained against the  
 720 scratch version with **green** when it improves and **red** when it worsens.

Method	Weight Init.	$\mathcal{D}_{SS}$	$\mathcal{D}_{SU}$	$\mathcal{D}_{UU}$	$\mathcal{D}_{??}$
<b>SimCLR</b>	RAND	3.23	16.10	19.62	15.12
	SUP	<b>-0.10</b>	<b>-0.21</b>	<b>-0.40</b>	<b>+0.27</b>
<b>BYOL</b>	RAND	3.27	16.25	20.20	15.91
	SUP	<b>-0.57</b>	<b>-1.75</b>	<b>-2.23</b>	<b>-1.50</b>
<b>NNCLR</b>	RAND	2.84	13.91	17.15	13.96
	SUP	<b>-0.54</b>	<b>-2.44</b>	<b>-3.18</b>	<b>-2.67</b>
<b>VICReg</b>	RAND	3.24	16.67	19.97	15.86
	SUP	<b>-0.43</b>	<b>-1.54</b>	<b>-2.45</b>	<b>-1.92</b>

735  
 736 **A.4.6. COMPARING TO LINEAR PROBING**

737 Finally, we compare the linear probing performance of the embeddings generated from different architectures, i.e. GPS  
 738 backbones (GPS-BB), pretrained on different datasets, i.e., GPS Datasets (GPS-DS), with the performance of GPS-SSL  
 739 using them. More specifically, in Tables 14 and 13, we compare the linear probe performance of the CLIP pretrained  
 740 ResNet50 on LAION-400M (Schuhmann et al., 2021) along with vision transformers (ViTs) pretrained on ImageNet  
 741 using Masked Auto Encoders (MAE) (He et al., 2022), a popular self-supervised method that also does not rely on strong  
 742 augmentations. We see our method outperforms the linear probe accuracy of CLIP embeddings for both Cifar10 and  
 743 FGVCAircraft and matches that of ViT-Base and ViT-Large for Cifar10 and ViT-Large for FGVCAircraft.

744 However, we further see that if we train the ViT-Large on the FGVCAircraft, using MAE with minimal augmentations, we  
 745 can use that as the positive sampler for GPS-SSL and beat the baseline SSL method on FGVCAircraft. This shows that  
 746 GPS-SSL does not entirely rely on huge pretrained models and that there is potential possibilities for training a positive  
 747 sampler prior to applying GPS-SSL to further boost the performance of baseline SSL methods.

748  
 749  
 750  
 751  
 752  
 753  
 754  
 755  
 756  
 757  
 758  
 759  
 760  
 761  
 762  
 763  
 764  
 765  
 766  
 767  
 768  
 769

770  
771  
772  
773  
774  
775  
776  
777

778 *Table 13.* Comparison of linear probing (LP) and GPS-VICReg's (with ResNet50) classification accuracy on FGVC Aircraft with different  
779 GPS backbones (GPS-BB) pretrained with CLIP and masked auto encoders (MAE) on different datasets without supervision (GPS-DS).  
780 The performance of the vanilla VICReg is also depicted for comparison. RN50 and ViT-L refer to ResNet50 and ViT-Large, respectively.

<b>GPS-BB</b>	<b>GPS-DS</b>	<b>LP</b>	<b>GPS-VICReg</b>	<b>VICReg</b>
RN50 <sub>CLIP</sub>	LAION-400M	44.55	46.44	
ViT-L <sub>MAE</sub>	ImageNet	37.32	38.44	39.99
ViT-L <sub>MAE</sub>	FGVCAircraft	17.01	42.87	

781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803

804 *Table 14.* Classification accuracy comparison of linear probing (LP) using embeddings with different GPS backbones (GPS-BB) pretrained  
805 with CLIP and masked autoencoders (MAE) on different upstream datasets, i.e., GPS-DS, and a trained ResNet50 with GPS-SimCLR  
806 on FGVC Aircraft and Cifar10 using the same GPS backbones and datasets. RN50, ViT-L, and Vit-B refer to ResNet50, ViT-Large, and  
807 ViT-Base, respectively.

<b>GPS-BB</b>	<b>GPS-DS</b>	<b>Cifar10</b>		<b>FGVCAircraft</b>	
		<b>LP</b>	<b>GPS-SimCLR</b>	<b>LP</b>	<b>GPS-SimCLR</b>
RN50 <sub>CLIP</sub>	LAION-400M	87.85	91.17	44.55	53.81
ViT-B <sub>MAE</sub>	ImageNet	85.78	87.35	27.96	29.55
ViT-L <sub>MAE</sub>	ImageNet	91.45	90.11	37.29	35.28
ViT-L <sub>MAE</sub>	FGVCAircraft	—	—	17.01	46.93

811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824