Unsupervised Training of Vision Transformers with Synthetic Negatives

Nikolaos Giakoumoglou Andreas Floros Kleanthis Marios Papadopoulos Tania Stathaki
Imperial College London
{nikos, andreas.floros18, kleanthis-marios.papadopoulos18, t.stathaki}@imperial.ac.uk

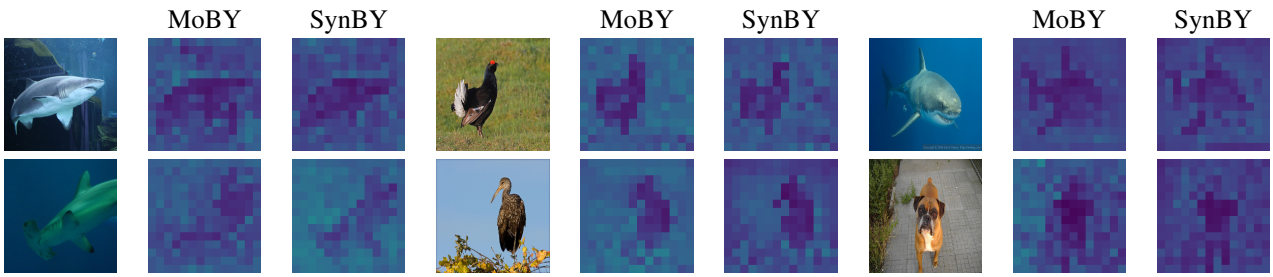


Figure 1. Self-attention patterns of the average attention head of DeiT-S from the last transformer layer for MoBY and our approach.

Abstract

This paper does not introduce a novel method per se. Instead, we address the neglected potential of hard negative samples in self-supervised learning. Previous works explored synthetic hard negatives but rarely in the context of vision transformers. We build on this observation and integrate synthetic hard negatives to improve vision transformer representation learning. This simple yet effective technique notably improves the discriminative power of learned representations. Our experiments show performance improvements for both DeiT-S and Swin-T architectures.[1]

1. Introduction

Computer vision has recently witnessed two major advances. Self-supervised contrastive learning [4, 11] has fundamentally transformed how machines learn from visual data without labels. Concurrently, vision transformer architectures [8, 23] have reshaped the field by applying attention mechanisms to image understanding tasks. Self-supervised methods have proven remarkably effective for building robust visual representations [15], often referred to as *"the dark matter of intelligence"* that underpins broader machine comprehension. As Yann LeCun aptly noted, *"if intelligence is a cake, the bulk of the cake is unsupervised learning"*, and the emergence of transformer models facilitates this by providing architectures capable of capturing complex relationships within visual data [8].

Despite their effectiveness, contrastive learning approaches face a persistent challenge regarding the quality of negative examples [13]. Standard techniques rely on randomly sampling negatives from a batch [4, 5] or memory bank [11, 24], but these negatives are often too easy to distinguish, limiting the discriminative power of learned representations [9, 13].

In this work, we attempt to overcome this limitation by integrating synthetic hard negatives into self-supervised vision transformer training. Building upon existing momentum-based frameworks [10, 11, 25], we generate challenging negative examples that force the model to learn more discriminative features [9, 13]. Inspired by recent advances in synthetic contrastive learning [9], our approach synthesizes hard negatives *"on-the-fly"* in the feature space, creating examples that improve representation quality while maintaining stability. The key insight of our approach is that synthetic negatives provide a controlled way to increase the difficulty of the learning task [9], pushing the model to develop more robust representations.

Our main **contributions** include exploring the previously uninvestigated application of synthetic negatives in vision transformers. Specifically:

- We demonstrate the potential of synthetic hard negatives in contrastive learning by integrating our approach with the MoBY framework and experimenting on the DeiT-S and Swin-T architectures.
- We experiment and benchmark on ImageNet, where we find that most configuration settings of our proposed method provide sufficient contrast for the models to learn highly discriminative features.

[1] https://github.com/giakoumoglou/synco-v2

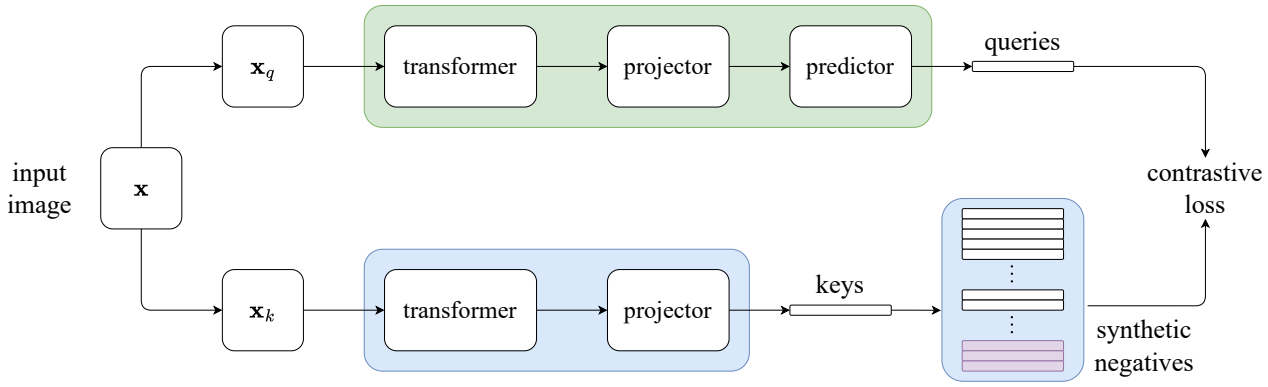


Figure 2. SynBY framework overview. Our approach incorporates synthetic hard negatives into the MoBY framework.

2. Related Work

Self-supervised visual representation learning. Unsupervised learning has emerged as a powerful approach to learn visual representations without manual annotations. Within this paradigm, contrastive learning has shown particular promise and has been widely adopted in various forms [4, 5, 11, 20]. SimCLR [4] demonstrated the effectiveness of a simple framework using data augmentation, large batch sizes, and nonlinear projection heads. MoCo [11] introduced a momentum encoder and queue-based mechanism, enabling contrastive learning with smaller batch sizes.

Hard negatives in contrastive learning. The quality of negative samples in contrastive learning has been a focus of extensive research [1, 6, 9, 13, 19, 24]. These studies aim to select informative negative samples and address false negatives in instance discrimination tasks. Recent work [13] explored mixing of hard negatives to create challenging contrasts, showing that harder examples lead to improved representations. Subsequent works developed this direction, with newer approaches [9] proposing systematic methods for generating synthetic hard negatives in the feature space.

Self-supervised transformers for vision. Self-supervised learning for vision transformers has rapidly evolved [2, 12]. Self-distillation methods operate without labels [3], while masked modeling draws inspiration from language processing [2, 12]. MoCo-v3 [5] adapted momentum-based frameworks for transformers, addressing instability through fixed patch projection and batch normalization. Other contrastive methods like MoBY [25] implemented asymmetric drop path rates and fewer stability "*tricks*".

3. Background

In this section, we introduce contrastive learning basics (Section 3.1) and our framework for generating synthetic hard negatives (Section 3.2), illustrated in Figure 2.

3.1. Contrastive Learning

Contrastive learning aims to learn representations by comparing similar and dissimilar samples. Given an image, $\mathbf{x}$, and two distributions of image augmentations, $\mathcal{T}$ and $\mathcal{T}'$, two augmented views of the same image are created via $\mathbf{x}_q = t_q(\mathbf{x})$ and $\mathbf{x}_k = t_k(\mathbf{x})$, with $t_q \sim \mathcal{T}$ and $t_k \sim \mathcal{T}'$. The views are encoded by *online* and *target* encoders, f_q and f_k, respectively, producing vectors $\mathbf{q} = f_q(\mathbf{x}_q)$ and $\mathbf{k} = f_k(\mathbf{x}_k)$. The learning objective is to minimize the InfoNCE loss [22]:

$$\mathcal{L}(\mathbf{q}, \mathbf{k}, \mathcal{Q}) = -\log \frac{\exp(\mathbf{q}^\top \cdot \mathbf{k}/\tau)}{\exp(\mathbf{q}^\top \cdot \mathbf{k}/\tau) + \sum_{\mathbf{n}\in\mathcal{Q}} \exp(\mathbf{q}^\top \cdot \mathbf{n}/\tau)}. \tag{1}$$

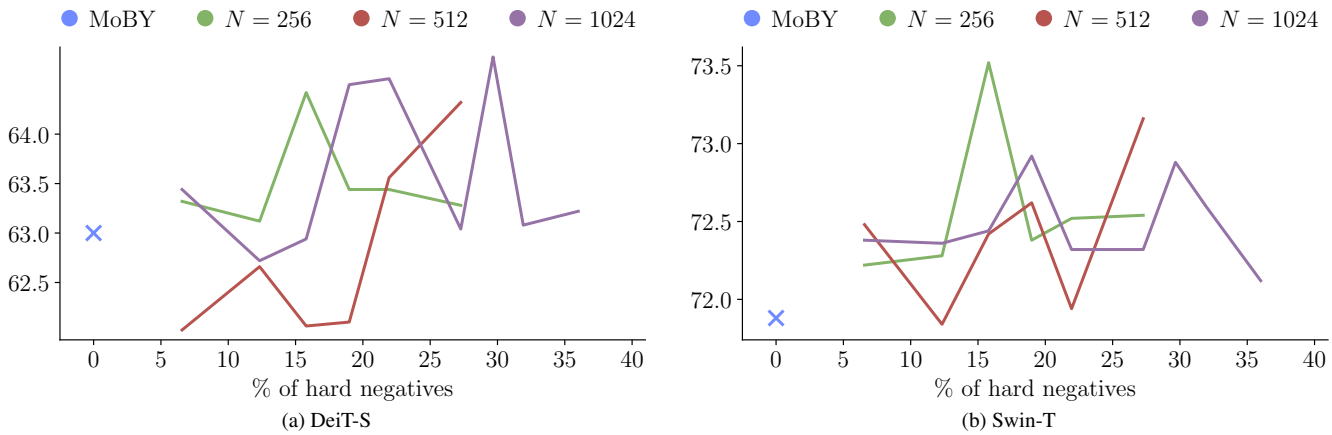
Here, $\mathcal{Q} = \{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_K\}$ is a set of K negative samples and τ is a temperature parameter. Negative samples are mined either from the batch [4, 5] or from a memory bank [11, 18, 20]. The online encoder is updated via gradient descent whereas the target encoder may be updated via momentum, $\theta_k \leftarrow m \cdot \theta_k + (1-m) \cdot \theta_q$, or through weight sharing in siamese networks (i.e., where $f_k \equiv f_q$).

3.2. Synthetic Hard Negatives

Synthetic negatives provide challenging examples that help models learn more discriminative features. Let $\mathcal{Q}^N = \text{TopK}(\{\text{sim}(\mathbf{q}, \mathbf{n}) \mid \mathbf{n} \in \mathcal{Q}\}, N)$ be the subset containing the $N < K$ hardest negatives, where $\text{sim}(\cdot,\cdot)$ is the cosine similarity. The synthetic hard negatives can be abstractly represented through a synthesis function, $\mathcal{F}$, as follows:

$$\mathbf{s} = \frac{\mathcal{F}(\mathbf{q}, \mathcal{Q}^N; \xi)}{\|\mathcal{F}(\mathbf{q}, \mathcal{Q}^N; \xi)\|_2}, \tag{2}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm, and ξ represents the parameters that control the synthesis process, described in [9]. The set of L synthetic negatives, $\mathcal{Q}_s = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_L\}$, is then combined with the existing queue of real negatives, $\mathcal{Q}$, effectively expanding the diversity of negative examples and exposing the model to more challenging contrasts.



(a) DeiT-S (b) Swin-T

Figure 3. Ablation study of different hardness selection values and synthetic negative percentages.

4. Experiments

We develop our approach in PyTorch, building upon the implementation of MoBY [25] and SynCo [9]. We refer to our approach as SynBY.

4.1. Setup

We pretrain SynBY on ImageNet ILSVRC-2012 [7] and its smaller ImageNet-100 subset [14] using a DeiT-Small [8, 21] or Swin-Tiny [16] encoder. Our implementation builds upon MoBY [25], where the online encoder, f_q, consists of the backbone, a projection head [4], and an extra prediction head [10]; the target encoder, f_k, has the backbone and projection head, but does not include the prediction head. For training, we use the AdamW optimizer [17] with a base learning rate of 0.03, weight decay of 10^{-4}, and batch size of 512. Unless otherwise stated, we use the following default hyperparameters. The momentum parameter starts at $m_{\text{start}} = 0.99$ and increases to 1 following a cosine schedule. For synthetic negatives, we select the top $N = 256$ hardest negatives. We use a temperature $\tau = 0.2$ for the contrastive loss of Equation (1) and a queue size $K = 4096$. For our experiments on ILSVRC-2012, we implement a cooldown period for the last 100 epochs where *no* synthetic negatives are generated. For linear evaluation, we train a fully-connected layer on frozen features for 100 epochs. We refer the reader to [25] for further implementation details.

4.2. Linear Evaluation on ImageNet

Table 1 shows top-1 accuracy of our method after pretraining for 300 epochs on ImageNet ILSVRC-2012. SynBY outperforms the MoBY baseline by 0.2% on both architectures and maintains superiority over DINO [3] and MoCo-v3 [5]. However, we note that there is a considerable gap in performance when one compares with supervised training.

Table 1. Top-1 classification accuracy on ImageNet for self-supervised methods with DeiT-S and Swin-T architectures.

Method	Arch.	Params (M)	Top-1 (%)	
Supervised	DeiT-S	22	79.8	
	Swin-T	29	81.3	
DINO [3]	DeiT-S	22	72.5	
MoCo-v3 [5]	DeiT-S	22	72.5	
MoBY [25]	DeiT-S	22	72.8	
	Swin-T	29	75.0	
SynBY (ours)	DeiT-S	22	**73.0**	↑0.2
	Swin-T	29	**75.2**	↑0.2

Visualizing attention. Figure 1 shows self-attention patterns of MoBY and SynBY. Though similar, ours tends to produce more focused attention maps with finer-grained patterns that highlight semantically meaningful regions. This suggests that synthetic negatives help develop more discriminative features, targeting relevant visual elements.

4.3. Ablation Study

We perform ablations studies of SynBY on ImageNet-100 classification, pretraining for 100 epochs.

Synthetic negatives. We observe architectural differences in how DeiT and Swin transformers respond to synthetic negatives, shown in Figure 3. DeiT benefits from mining negatives at either low (256) or high (1024) hardness levels, while Swin performs well across all hardness levels. Additionally, DeiT achieves better results with moderately hard negatives at medium or high proportions whereas Swin performs consistently well with all proportions. This likely stems from Swin's inductive biases requiring less aggressive negative samples than DeiT's pure transformer architecture.

Table 2. Ablation study on applying tricks of MoCo-v3.

Fixed Patch Embedding	Replace LN before MLP with BN	Top-1 (%) DeiT-S	Top-1 (%) Swin-T
		66.7	67.5
✓		66.4	67.2
	✓	67.2	67.9

Table 3. Ablation study on the drop path rates.

Online dpr	*Target* dpr	Top-1 (%) DeiT-S	Top-1 (%) Swin-T
0.1	0.1	61.9	74.3
0.05	0.0	65.0	75.3
0.1	0.0	65.0	75.4
0.2	0.0	64.7	72.7

Table 4. Ablation study on queue size K.

K	Top-1 (%) DeiT-S	Top-1 (%) Swin-T
1024	64.5	72.5
2048	64.5	72.5
4096	64.7	72.7
8192	63.6	72.3
16384	62.6	71.6

Table 5. Ablation study on temperature τ.

τ	Top-1 (%) DeiT-S	Top-1 (%) Swin-T
0.07	59.3	61.5
0.1	61.5	69.2
0.2	64.5	72.7
0.3	64.0	71.7

Table 6. Ablation study on momentum m_{start}.

m_{start}	Top-1 (%) DeiT-S	Top-1 (%) Swin-T
0.99	64.5	72.7
0.993	65.2	72.2
0.996	63.8	72.4
0.999	60.3	68.6

Applying MoCo-v3 tricks. Our experiments reveal that synthetic negatives eliminate the need for additional stabilization techniques from MoCo-v3. As shown in Table 2, fixing the patch embedding has minimal impact on performance, suggesting that our synthetic negatives already provide sufficient regularization. This allows for a simpler implementation without compromising performance. Notably, replacing Layer Normalization (LN) with Batch Normalization (BN) before MLP blocks yields improvements.

Asymmetric drop path rates. As shown in Table 3, the asymmetric configuration of drop path rates (dpr) significantly impacts model performance. Unlike MoBY which uses 0.2 for the online encoder, we find a smaller rate of 0.05 is optimal when combined with synthetic negatives. This is in agreement with our intuition that the synthetic negatives provide additional regularization and reduce the need for stablization tricks. Applying drop path only to the online encoder while keeping the target encoder stable yields the best balance.

Other hyperparameters. The default hyperparameters from MoBY work effectively with our synthetic negative approach. As shown in Tables 4 to 6, performance remains stable across different queue sizes (best at 4096), temperatures (optimal at 0.2), and momentum values (best at 0.99). This demonstrates that synthetic negatives can be incorporated without extensive re-tuning of existing parameters. Overall, our synthetic negative generation technique integrates seamlessly with established contrastive learning frameworks, requiring minimal adaptation effort.

5. Discussion

In this paper, we explored synthetic negatives in vision transformer pretraining. We found that synthetic negatives provide sufficient regularization for learning more robust representations while also reducing the need for stabilization tricks. Importantly, our approach requires minimal adjustments to current frameworks, working in a "*plug-and-play*" manner with existing architectures. The experimental results demonstrate that SynBY further improves unsupervised learning with minimal computational *overhead*, showing consistent gains across different transformer architectures.

Limitations. Our ablation studies were conducted on ImageNet-100, which may not fully capture the behavior on larger-scale datasets. Additionally, while we demonstrated improved performance on image classification, we did not evaluate on more complex downstream tasks.

Future work. Synthetic hard negatives have proven effective for vision transformers. Exploring their integration into vision-language frameworks represents a promising direction, with potential to enhance cross-modal contrastive learning through more challenging negative examples.

Acknowledgments

We acknowledge the computational resources and support provided by the Imperial College Research Computing Service (http://doi.org/10.14469/hpc/2232), which enabled our experiments.

References

[1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning, 2019. 2

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 2, 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 1, 2, 3

[5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. 1, 2, 3

[6] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, pages 8765–8775, 2020. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1, 3

[9] Nikolaos Giakoumoglou and Tania Stathaki. Synco: Synthetic hard negatives in contrastive learning for better unsupervised visual representations, 2024. 1, 2, 3

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 1, 3

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. 1, 2

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 2

[13] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning, 2020. 1, 2

[14] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. 3

[15] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Liming Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):21–40, 2021. 1

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 3

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 3

[18] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations, 2019. 2

[19] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021. 2

[20] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, pages 6827–6839. Curran Associates, Inc., 2020. 2

[21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2021. 3

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 2

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 1

[24] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018. 1, 2

[25] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers, 2021. 1, 2, 3