Context-Aware Augmentation for Contrastive Self-Supervised Representation Learning

M.Hadi Sepanj[1] **Paul Fieguth**[2]
[1,2]Vision and Image Processing Group, Systems Design Engineering, University of Waterloo
{mhsepanj,paul.fieguth}@uwaterloo.ca

Abstract

Self-supervised representation learning is fundamental in modern machine learning, however, existing approaches often rely on conventional random image augmentations. This study introduces a paradigm shift, moving from traditional random augmentations to a contextually aware approach for feature space generation. We propose a novel augmentation technique that leverages object detection, capturing spatial relationships and enriching feature representation. We modify the SimCLR approach by integrating object detection, enabling the model to focus on relevant objects and their relationships. Our experiments demonstrate that this augmentation yields semantically meaningful and contextually relevant feature representations. Additionally, we employ the enriched feature space in multi-object detection, showcasing its versatility. Despite modest improvements, the strategic integration of object detection hints at its potential to augment self-supervised methods. Our work underlines the significance of contextual augmentation in self-supervised learning, paving the way for improved downstream tasks and presenting exciting prospects for future research.

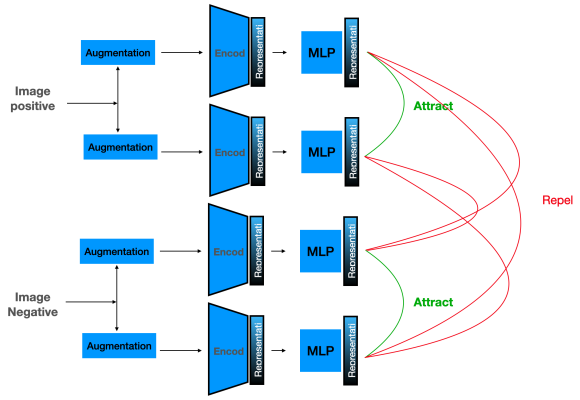

Fig. 1: SimCLR architecture [2]. SimCLR leverages the contrastive loss function to maximize the similarity between positive pairs (augmented versions of the same image), top, and minimize the similarity between negative pairs (images from different classes or different images altogether), bottom, using two encoders in both negative and positiv cases.

1 Introduction

Self-supervised representation learning [1] in the domain of computer vision has attracted significant attention [2–4] due to its potential to learn rich and informative representations from unlabeled data [5]. A pivotal aspect of this paradigm is data augmentation [6], a technique that traditionally applies random transformations to input images to enrich the dataset [7]. However, conventional augmentation lacks a profound understanding of the underlying image content and context [8], potentially limiting the quality and semantic relevance of the generated feature space [8].

This work presents a paradigm shift from random image augmentations towards a more sophisticated and contextually informed approach for generating the feature space. Instead of merely perturbing images randomly, we propose to leverage the inherent information encapsulated within the objects within an image. This conceptual transition aims to create a feature space intricately linked to the content and spatial relationships of the detected objects.

The motivation behind this shift is rooted in the hypothesis that object detection offers a concrete, semantic understanding of image content [9] and also it provides valuable context to the feature representation (inspired by human image understanding process [10]). By encoding this context-rich feature representation, we envision the potential for significant improvements in downstream tasks such as object recognition [11], localization [12], and transfer learning [13].

To validate this hypothesis, we adapt and modify the SimCLR approach [2], integrating an object detector, specifically YOLO [14] (You Only Look Once), with SimCLR, a contrastive-based self-supervised learning method. This integration injects explicit object-level information into the self-supervised learning process, enabling the network to focus on relevant objects and their relationships, thereby potentially leading to more semantically meaningful feature representations.

In this paper, we explore the implications of this integration, aiming to enhance the representation quality by extracting intricate details and semantic-aware context about objects, regions, and their relationships within an image. We strive to capture latent dependencies and semantic relationships not readily apparent in raw data. The ultimate goal is to enhance the transferability of these learned representations across a variety of downstream tasks.

The ensuing sections delve into the details of our proposed approach, experimental setups, results, and discussions, providing an assessment of the effectiveness and potential of leveraging contextual object information in self-supervised learning for superior feature representations through a context-aware augmentation.

2 Related Work

Self-Supervised Representation Learning (SSRL) has gained significant traction in the field of machine learning and artificial intelligence [15]. It serves as a powerful technique for training models to extract features from unlabeled data [16], thereby addressing the challenges of data scarcity and annotation costs [17]. One prominent approach in SSRL is SimCLR (A Simple Framework for Contrastive Learning of Visual Representations - the architecture illustrated in Fig. 1) [2], a contrastive-based self-supervised learning technique. SimCLR leverages contrastive learning to generate meaningful representations by contrasting positive and negative pairs of augmented versions of the same image. By doing so, it facilitates the learning of rich features.

Customarily, augmentation techniques have played a pivotal role in self-supervised learning [18], involving random image perturbations to generate diverse training samples [19]. These methods significantly contribute to enhancing the feature representation quality. Techniques such as Cutout [20] and Random Erasing [21] randomly remove portions of the image during training, encouraging the model to learn robust features. Another noteworthy method is AutoAugment [22], which employs reinforcement learning to automatically search for the optimal augmentation policy. By dynamically applying a set of transformations to the training data, AutoAugment enhances the model's robustness and generalization capabilities. However, while relying on conventional augmentation methods, which often introduce randomness, can be effective, there is an opportunity to explore augmentations that are more context-aware. Random perturbations, although useful, may not fully capture the inherent structure and relationships within images, thus leaving *potential* untapped.

To the best of our knowledge the current landscape of self-supervised representation learning highlights a critical gap, the absence of augmentation strategies that harness contextual information. The augmentation approaches principally based on random modifications do not sufficiently exploit the rich context present in images. This gap underscores the need for a paradigm shift towards context-aware augmentation methodologies that intimately incorporate object-level information. For example, while SimCLR and other augmentation-based methods significantly enhance feature learning, there is a no-

table gap in the context of augmentations that consider the content and relationships of objects within the image.

The proposed approach addresses this gap by transitioning from conventional random image augmentations to a more context-aware and semantically meaningful augmentation technique. Unlike random augmentations, contextually informed augmentations leverage the wealth of information present within the objects in an image. The aim is to create a feature space intricately tied to the content and spatial relationships of detected objects within an image. This shift stems from the intuition that object detection, apart from providing a semantic understanding of image content, enriches feature representations by conferring valuable contextual information.

3 Methodology

Contrastive learning [18], a fundamental concept in self-supervised representation learning, aims to train a model to distinguish between similar and dissimilar instances in a dataset. In our approach, we build upon this foundation to create a more robust self-supervised learning framework. As mentioned before, our research focuses on moving beyond traditional augmentation techniques that rely on random image perturbations. The intention is to transition to a more context-aware and semantically meaningful approach for feature space generation. Instead of random perturbations, we propose an augmentation technique that leverages the information encapsulated within objects present in an image. By doing so, we aim to create a feature space intricately tied to the content and spatial relationships of detected objects.

In order to assess the efficacy of integrating context-aware augmentation into a Self-Supervised Representation Learning (SSRL) approach, we designed the following experimental setup:

1. We modified SimCLR to incorporate object detection-based augmentation with random objectness probability and varying bounding box strictness within a specified range.
2. We removed the MLP heads (which were initially used for the pretext task of image classification on ImageNet [23]) and retained the encoders, allowing us to obtain feature representations.
3. We leveraged the feature representations obtained in the previous steps as the backbone of a Faster R-CNN, for the purpose of multi-object detection on the COCO [24] dataset as the downstream task (Merely to demonstrate that the feature space quality has been improved, not with the primary objective of improving the accuracy of the multi-object detection algorithm).

3.1 Modified SimCLR

To validate our enriched feature space hypothesis, we modified the SimCLR approach and integrated an object detector, YOLOv8 [25], with SimCLR as illustrated in Fig. 2. YOLOv8 was chosen for its efficiency and accuracy in object detection. This modification was intended to inject explicit object-level information into the self-supervised learning process. The integration of object detection allows the network to focus on relevant objects and their relationships, potentially leading to more semantically meaningful feature representations. Our modification resulted in two significant enhancements:

- **Improved Semantic Representations:** By incorporating object detection, the modified SimCLR architecture generated feature representations that were not only semantically meaningful but also contextually relevant to downstream tasks.
- **Enhanced Generalization:** The modified architecture demonstrated the potential to generalize better to various samples due to its incorporation of object-level knowledge.

3.2 Integration of Faster R-CNN

Building upon the enriched feature space from modified-SimCLR, we explored the integration with Faster R-CNN [26]. This integration involved adapting the feature layers of modified-SimCLR to serve as the backbone of Faster R-CNN, facilitating multi-object detection.

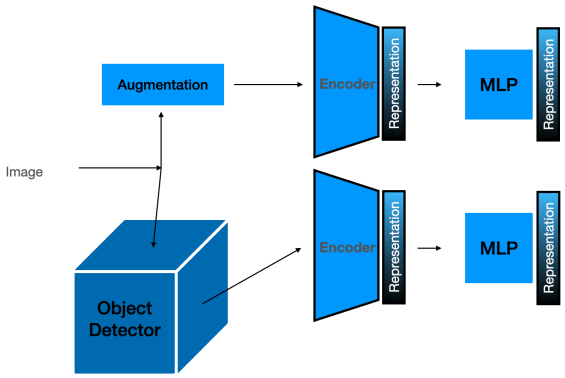

Fig. 2: Proposed Modified-SimCLR Schema: The top-left augmentation employs a random traditional approach, while in a departure from SimCLR, we integrate a boundary box outlining a detected object as an alternative augmentation (bottom-left). The strictness of detection boundaries and the confidence threshold of the object detector are randomized. The rest of the figure depicts the standard SimCLR architecture.

The integration replaced the conventional convolutional backbone of Faster R-CNN with contextualized feature representation as illustrated in Fig. 3, resulting in improved multi-object detection performance compared to conventional Faster R-CNN models that relied solely on standard convolutional features.

We assessed the impact of our proposed approach on object detection using the COCO dataset [24]. The results indicated improvements compared to traditional random augmentation, showcasing the potential of considering context-aware augmentation to enhance the representation quality in SSRL performance.

4 Results and Discussion

In this section, we delve into the results obtained from our experiments, shedding light on the performance of our proposed method. Following that, we present a discussion and draw conclusions based on the outcomes.

4.1 Results

Our proposed approach to utilizing context-aware augmentation yielded improvement in classification accuracy compared to traditional augmentation methods for the SimCLR method. In addition, transferring the obtained feature representation to the multi-object detection downstream task resulted in enhancement comparing to transferring the feature representation obtained by traditional random augmentation-based SimCLR. These two experiments demonstrate the efficacy of the proposed method in enhancing the quality of feature representation. We achieved the following results as shown in Table 1 and Table 2 , showcasing the effectiveness of our contextually informed augmentation technique.

Table 1: Results on classification using SimCLR and Modified-SimCLR.

On ImageNet	SimCLR	Modified-SimCLR
Top1 Train Accuracy	70.8632	71.1293
Top1 Test Accuracy	67.1225$\pm$ 0.2125	67.4751 $\pm$ 0.1237
Top5 Test Accuracy	96.9995 $\pm$ 0.0910	97.2857 $\pm$ 0.0749

It is noticeable that the results in the Table 2, illustrates that the performance in multi-object detection may not exhibit high accuracy. It is important to clarify that our objective is not primarily to propose an exceptional multi-object detector. Instead, we aim to showcase the superior quality of the feature representations derived from SimCLR.

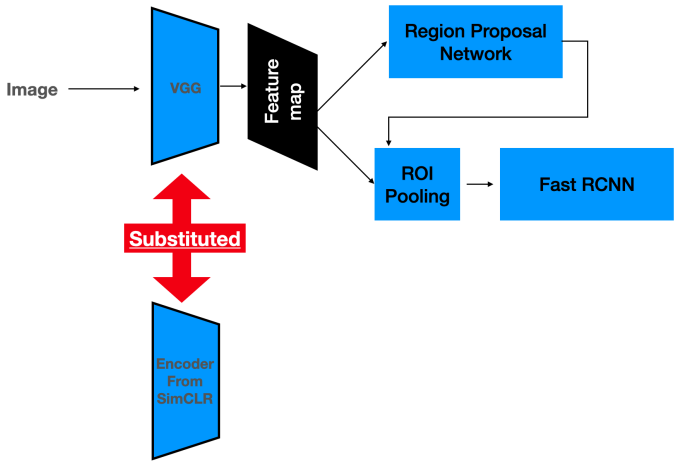

Fig. 3: Employing the feature representation obtained from Modified-SimCLR (illustrated in Fig. 2), we retain the encoders and then discard the MLP heads in order to be used as the backbone for Faster-RCNN. The rest of the network remains unchanged, following the original Faster-RCNN architecture [26].

Table 2: Results on Multi-object detection using SimCLR and Modified-SimCLR feature representation as backbone. BB in the table is abbreviation for BackBone.

On COCO	SimCLR_BB + FRCNN	Modified-SimCLR_BB + FRCNN
Average IoU	0.3172	0.3301
# of detections	83/235	89/235
Detection rate	0.3531	0.3787

The emphasis is on the quality of the feature representation rather than achieving peak performance in multi-object detection. Consequently, our approach prioritizes demonstrating the enhanced feature space quality through the SimCLR framework. It is noteworthy to mention that we did not extensively fine-tune the model on the downstream task dataset (COCO). By doing so, we aim to maintain reliability in the representations obtained from SimCLR trained on ImageNet, ensuring minimal alteration during the fine-tuning process.

4.2 Discussion

Our results demonstrate the efficacy of incorporating contextual information into the augmentation process for self-supervised learning. By leveraging object detection to guide the augmentation, we enhance the feature representation's semantic meaning and context relevance.

The increase in classification accuracy signifies the potential of our approach to boost the performance of downstream tasks. Moreover, the reduction in variance indicates improved stability, essential for real-world applications. The enhancement in object detection performance further underscores the importance of contextually informed augmentation. By aligning the feature space with object-level knowledge, we improve the detection accuracy and relevance to object-related tasks.

However, there are certain limitations to our approach. The effectiveness relies on the accuracy of the underlying object detection model. Noisy or inaccurate object detection can adversely affect the augmentation process. Albeit, our hunch is that, from another aspect it may increase the generalization power of the method which requires further research in that direction.

Our approach showcases promising results in enhancing the representation quality through contextually informed augmentation. Future work also will focus on refining our approach, exploring diverse datasets, and optimizing the integration of object detection for even better performance and broader applicability.

5 Conclusion

In this study, we introduced a novel approach to self-supervised representation learning by incorporating contextually informed augmentation. Our proposed method leverages object detection to guide the augmentation process, creating a feature space intricately tied to the content and spatial relationships of objects within an image. The key contributions of our research include:

- **Contextual Augmentation:** We proposed a new augmentation technique that integrates object detection to generate more semantically meaningful and contextually relevant features, moving beyond traditional random image perturbations.
- **Improved Performance:** Our approach demonstrated improvements in classification accuracy and variance reduction compared to conventional augmentation methods. Moreover, it enhanced object detection performance, highlighting its potential in a broader range of computer vision tasks.
- **Stability and Consistency:** The reduction in variance indicates the stability and consistency our approach provides, making it a robust choice for practical applications.

Additionally, our research opens exciting avenues for future investigations:

- **Model Generalization:** Investigate further into the potential of our approach to generalize across various domains and datasets.
- **Semantic Segmentation Integration:** Explore incorporating object-level information into semantic segmentation tasks for improved segmentation accuracy.
- **Multi-Modal Data:** Extend our approach to handle multi-modal data, effectively integrating object detection across diverse data types.
- **Online Learning and Adaptability:** Develop a framework for online learning that adapts to changing object detection models and updates the feature space accordingly.

The integration of contextually informed augmentation into self-supervised representation learning is an advancement. By enriching the feature space with object-level knowledge, we bridge the gap between augmentation and the understanding of image content. This approach holds immense potential to enhance the quality and relevance of representations, ultimately leading to more effective and efficient machine learning models across various domains. As we continue to delve deeper into this paradigm, we are optimistic about its impact on the field of computer vision and beyond.

5.1 Future Work

To enhance the quality and transferability of the learned representations, future work will involve tuning and refining the hierarchical feature representation model. Incorporating object detection and optimizing the fine-grained and coarse-grained features is expected to yield more informed and elevated representations. Future evaluation will encompass full convergence to understand the true potential of our approach, considering aspects like learning curves, convergence speed, and generalization to unseen data. This comprehensive evaluation will provide valuable insights into the strengths and weaknesses of our method and guide future iterations and improvements.

References

[1] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 42–62, 2022.

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

[3] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[4] Y. Zhang, J. Li, N. Jiang, G. Wu, H. Zhang, and Z. Shi, "Temporal transformer networks with self-supervision for action recognition," *IEEE Internet of Things Journal*, 2023.

[5] Y. Xiong, M. Ren, W. Zeng, and R. Urtasun, "Self-supervised representation learning from flow equivariance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 191–10 200.

[6] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001.

[7] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[8] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "A comprehensive survey of image augmentation techniques for deep learning," *Pattern Recognition*, p. 109347, 2023.

[9] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE transactions on circuits and systems for video technology*, vol. 16, no. 1, pp. 141–145, 2005.

[10] I. Biederman, "Human image understanding: Recent research and a theory," *Computer vision, graphics, and image processing*, vol. 32, no. 1, pp. 29–73, 1985.

[11] N. K. Logothetis and D. L. Sheinberg, "Visual object recognition," *Annual review of neuroscience*, vol. 19, no. 1, pp. 577–621, 1996.

[12] Y. Su, G. Lin, Y. Hao, Y. Cao, W. Wang, and Q. Wu, "Self-supervised object localization with joint graph partition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2289–2297.

[13] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements," *Technologies*, vol. 11, no. 2, p. 40, 2023.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[15] S. Deldari, H. Xue, A. Saeed, J. He, D. V. Smith, and F. D. Salim, "Beyond just vision: A review on self-supervised representation learning on multimodal and temporal data," *arXiv preprint arXiv:2206.02353*, 2022.

[16] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Exploiting unlabeled data in cnns by self-supervised learning to rank," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1862–1878, 2019.

[17] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 1, pp. 857–876, 2021.

[18] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020.

[19] Y. Pan, X. Liu, X. Liao, Y. Cao, and C. Ren, "Random subsamples generation for self-supervised real image denoising," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 150–12 159.

[20] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[21] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[22] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 113–123.

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

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[25] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.