FACULTY OF INFORMATICS AND COMPUTING

MASTER OF DATA SCIENCE

Research topic: **Computer Vision-Aided Recyclable Waste Item Detection**

A Comparative Analysis of Self-Supervised Learning Algorithms in computer vision object detection.

Abstract: The field of Self-Supervised Learning (SSL) has rapidly evolved, moving from instance discrimination-based methods to more complex paradigms inspired by masked modelling and latent prediction. This study provides a theoretical comparison of five prominent SSL algorithms: SimCLR, Bootstrap Your Own Latent (BYOL), Self-Distillation with No Labels (DINO) and Yann LeCun's proposed frameworks, **JEPA (**Joint Embedding Predictive Architecture), **IJEPA (**Image-based Joint Embedding Predictive Architecture**)**, and **LeJEPA (**Language-enhanced Joint Embedding Predictive Architecture**)**. The contenders are analysed based on a clear set of criteria: underlying principle, architectural requirements, training stability, computational cost, and representational properties. The analysis reveals a clear trajectory from contrastive to generative and finally to energy-based models, with a growing emphasis on learning abstract, hierarchical representations while improving computational and data efficiency. We conclude that while no single algorithm is superior in all aspects, the JEPA family represents a significant conceptual shift towards more efficient and scalable world models.

Student Name	Hambeleleni P Shaningwa
Student Number	213091704

Table of Contents

1. Introduction

Self-supervised learning (SSL) has emerged as a dominant paradigm in representation learning. It enables models to learn from unlabelled data by creating supervisory signals from the data itself. Unlike supervised methods that rely on large annotated datasets, SSL methods exploit intrinsic structure in data to learn robust representations, particularly in the visual domain. The past decade has produced several landmark SSL algorithms contrastive, predictive, cluster-based, and reconstruction-based models each with distinct learning principles and architectural designs. This report compares the contenders of SSL algorithms: SimCLR, BYOL, DINO, JEPA, and I-JEPA, highlighting their methodological differences, theoretical motivations, and empirical performance trends.

2. The importance of Self- Supervised Learning in Waste Item Detection

Waste images vary dramatically due to crushing, occlusion, dirt, lighting differences, similarities and deformation. SSL helps to address these challenges by:

- Learning invariant representations from unlabelled waste streams
- Reducing the annotation costs,
- Learns features transferable to object detectors and
- Adapts continuously to new waste types and packaging redesigns

Thus, SSL is ideal for large-scale recycling facilities and real-time robotic sorting applications.

3. Self- Supervised Learning Algorithm and relevance to Waste Detection

3.1 Contrastive Methods SimCLR (Simple Framework for Contrastive Learning of Visual Representations)

SimCLR is a self-supervised learning method aimed at learning high-quality image representations without human-annotated labels. It improves over previous contrastive learning methods by simplifying the architecture while demonstrating that data augmentation, projection heads, contrastive loss, and large batch sizes are fundamental to performance.

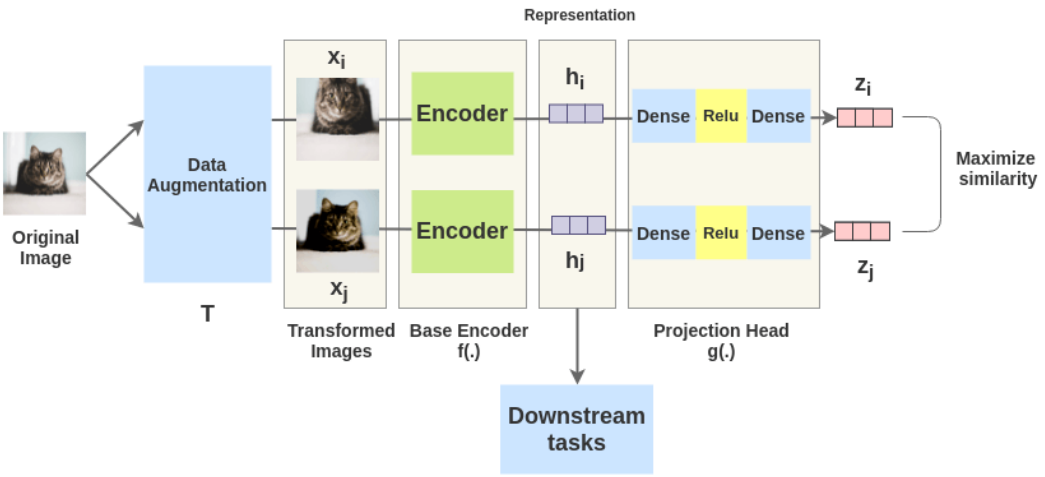
Core Principle: SimCLR learns by maximizing agreement between differently augmented views of the same image (positive pairs) while minimizing agreement with views from other images in the batch (negative pairs). It directly contrasts samples in the embedding space.

SimCLR Architecture:

It requires a Siamese network and uses four major components to learn robust visual representations without human-provided labels: data augmentation, a base encoder, a projection head, and a contrastive loss function see figure 1.

- Data Augmentation: This component applies a sequence of random transformations to an input image to create two correlated but distinct views of the same example (x_i and x_j). This specific composition of augmentations is critical for performance, as it prevents the model from using simple shortcuts to solve the task.
- Base Encoder: A neural network takes the augmented images (x_i and x_j) as input and extracts feature representations. The encoder's weights are the main output of the pre-training process and are used for downstream tasks.
- Projection Head: A small, non-linear Multi-Layer Perceptron (MLP) maps the representations from the base encoder into a lower-dimensional latent space. The contrastive loss is applied in this latent space, which improves the quality of the representations learned by the encoder itself. The projection head is discarded after pre-training.
- Contrastive Loss Function: The Normalized Temperature-Scaled Cross-Entropy Loss (NT-Xent) is used to compare the representations. It encourages the latent representations of the positive pairs (two views of the same original image) to be similar and those of all other images in the mini-batch (negative pairs) to be dissimilar.

Figure 1:Simple Framework for Contrastive Learning of Visual Representations framework



Representational Properties: Tends to learn invariant features to the applied augmentations. Creates a uniformly distributed embedding space where semantically similar samples are clustered.

Relevance to Waste Detection: It can learn to be invariant to colour changes (from dirt), slight rotations, and occlusions. However, its need for a large batch size to provide negative examples is a limitation when dealing with the long-tailed distribution of waste types (e.g., many plastic bottles).

3.2 Non-Contrastive Siamese Method, Bootstrap Your Own Latent (BYOL)

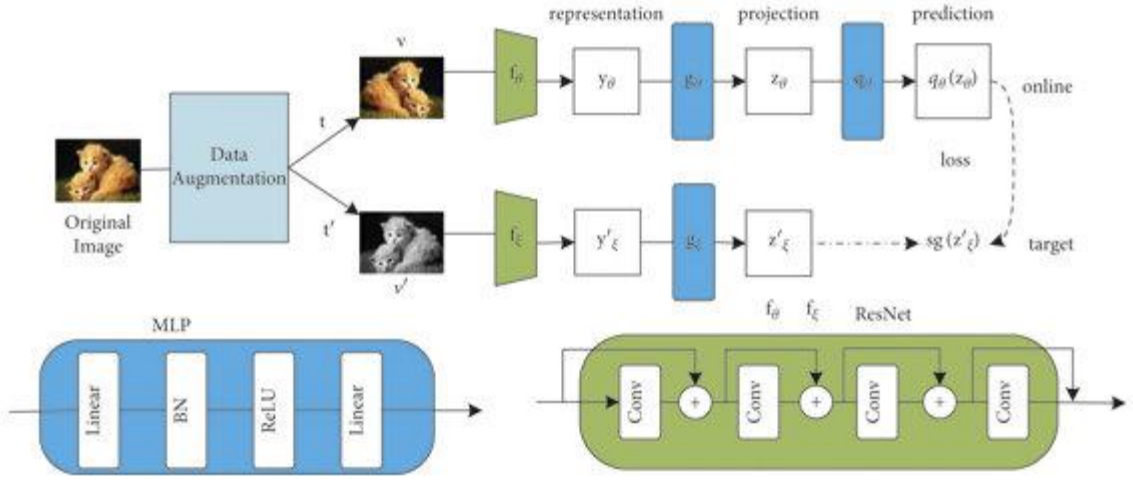
According to (Jean-Bastien, et al., 2020) BYOL is a new approach to self-supervised image representation learning. It relies on two neural networks, referred to as online and target networks, that interact and learn from each other. From an augmented view of an image, the online network is trained to predict the target network representation of the same image under a different augmented view. The Non-Contrastive Siamese methods avoid negative pairs and instead they use architectural asymmetry to avoid representation collapse.

Core Principle: Eliminates negative pairs. It uses two networks (online and target) where the online network learns to predict the target network's representation of the same image under a different augmentation. The target network is a slow-moving average of the online network.

BYOL Architecture:

The BYOL architecture consists of two identical Siamese networks, each with an encoder and a projector, but the online network also includes a predictor. Online Network: This network has trainable weights and comprises three stages:

- Encoder: A standard CNN (e.g., ResNet) that extracts features from the input image to produce a high-dimensional representation.
- Projector: An MLP that maps the representation to a lower-dimensional latent space (e.g., 256 dimensions).
- Predictor: Another MLP, exclusive to the online branch, which takes the projected representation and predicts the target network's output.
- Target Network: This network has the same architecture as the online network's encoder and projector but uses different weights and has no predictor. The target network's weights are not updated via backpropagation but through an exponential moving average (EMA) of the online network's weights, ensuring a more stable and slowly evolving target.



In a waste item detection context, data augmentation is a crucial pre-processing technique used during BYOL's self-supervised training phase to generate different views of the same waste image, enhancing dataset diversity and improving the model's robustness and generalization capabilities.

The entailment of augmentation generally involves applying a series of random transformations to an input image, including:

- Random resized crops: Taking random section of the image and resizing them to a consistent input size,
- Flipping the image along the vertical axis and randomly rotating the images within a specific range to simulate various object orientations.
- Scaling to ensure the waste of various sizes are represented.
- Random adjustments the image brightness, contract, saturation and hue and
- Converting the images to black and white and applying a blur effect

Representational Properties: Similar to SimCLR, it focuses on invariance. It may learn a strong representation of a "bottle" but might discard the fine-grained texture information needed to distinguish plastic from glass.

3.3 Clustering-Based SSL (Self-Distillation with No Labels (DINO))

DINO uses teacher–student Vision Transformers (Caron et al., 2020; Caron et al., 2021). It is particularly effective for recyclable waste segmentation, where attention heads naturally highlight object boundaries.

Core Principle: It is a form of self-distillation where a student network learns to match the output of a teacher network across different augmented views of an image. The teacher's weights are an exponential moving average (EMA) of the student's. It uses a centering and sharpening trick on the output distribution to avoid collapse.

DINO Architecture:

The DINO architecture is a Siamese network that is composed of a student and a teacher network, which share the same architecture (ViT or ResNet) but use different sets of weights.

- Self-Distillation: The core method involves training the student network to match the output distribution of the teacher network. A cross-entropy loss function is used for this comparison.
- Momentum Encoder: The teacher's weights are not updated via backpropagation; instead, they are an exponential moving average (EMA) of the student's weights. This provides a stable and consistent target for the student to learn from, acting as a moving ensemble and preventing training instabilities.
- Asymmetric Inputs: The student and teacher receive different augmented "views" (crops) of the same original image. The student processes all crops (both global and local views), while the teacher only processes global crops. This asymmetry encourages the student to align local features with the global context.
- Avoiding Collapse: To prevent the model from outputting trivial, constant representations, DINO employs two mechanisms on the teacher's output: **centering** (subtracting a running mean of the batch outputs) and **sharpening** (using a low teacher temperature in the SoftMax).

Representational Properties: Produces exceptionally strong semantic segmentations and object boundaries without any fine-tuning. The [CLS] token learns to capture scene-level semantics, while patch tokens retain local information. Creates well-structured, semantically clustered spaces.

3.4 Predictive Self Supervised Learning (LeJEPA, I-JEPA)

The JEPA (Joint-Embedded Predictive Architecture) family represents a conceptual shift from invariance-based learning to prediction in the latent space. It consists of JEPA, IJEPA and LeJEPA

3.4.1 LeJEPA

LeJEPA (Latent-Euclidean Joint Embedding Predictive Architecture) self-supervised learning (SSL) framework aims to learn rich visual representations by predicting the embeddings of masked data patches while enforcing a specific geometric constraint on the entire embedding space.

Core Principle (**LeJEPA):** An extension where the predictor operates across multiple layers of abstraction, enabling hierarchical prediction from low-level to high-level features. According to (Randall & Yann, 2025), the combination of the JEPA predictive loss with Sketched Isotropic Gaussian Regularization **(**SIGReg) yields LeJEPA with numerous theoretical and practical benefits: single trade-off hyper parameter, linear time and memory complexity, stability across hyper-parameters, architectures (ResNets)(Residual Networks), ViTs, ConvNets) and domains, heuristics-free, e.g., no

stop-gradient, no teacher–student, no hyper-parameter schedulers, and distributed training-friendly implementation requiring only ≈50 lines of code.

Le- JEPA Architecture

The LeJEPA method combines two main components in its loss function:

- JEPA Prediction Loss: This is the standard task for Joint Embedding Predictive Architectures (JEPAs). It involves training an encoder to make the embeddings of different augmented "views" (e.g., different crops of the same image) predict each other. The model learns to focus on the semantic content rather than trivial details.
- Sketched Isotropic Gaussian Regularization (SIGReg) Loss: This is the novel component introduced by LeJEPA. It enforces the desired isotropic Gaussian distribution on the model's embeddings. SIGReg uses random projections to reduce the high-dimensional embedding distribution to multiple 1D distributions, then applies a statistical test to ensure these 1D distributions match a standard Gaussian.

The underlying representation is a latent space where the learned embeddings are constrained to have an isotropic Gaussian distribution.

- Isotropic means the embeddings have the same variance in every direction, ensuring no single dimension or combination of dimensions is disproportionately important.
- This specific geometric structure is theoretically proven to minimize the bias and variance of downstream prediction tasks (both linear and non-linear), resulting in highly transferable and informative representations.

In the context of general waste detection and classification models, data augmentation typically involves techniques to increase the dataset size and improve the model's robustness and generalization to real-world conditions. The specific augmentation in a related method, called BackRep (Background Replacement), entails:

- Cropping waste items from images where they appear on a uniform background.
- Superimposing the cropped waste items onto more realistic and varied backgrounds (representing places where waste is typically littered).

This augmentation process helps the waste detection model to focus on the waste object itself and its salient visual attributes, rather than the background context, making it more effective in real-world scenarios.

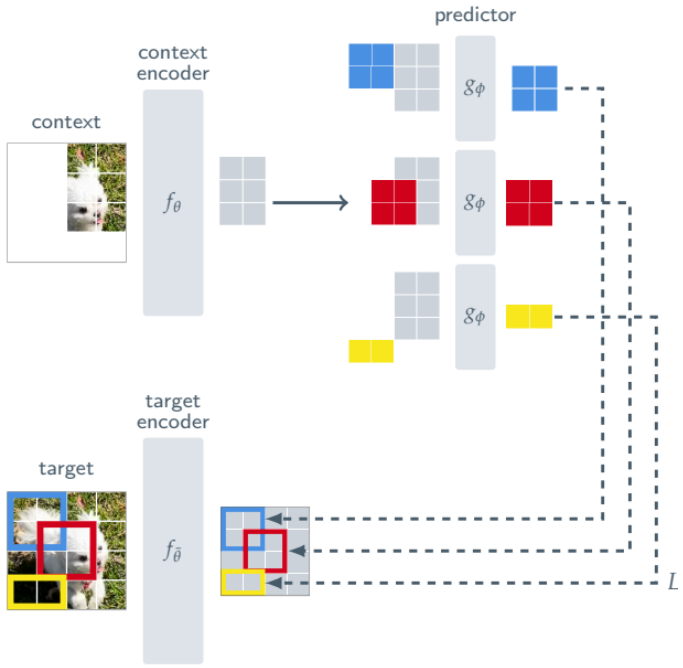
3.4.2 I-JEPA

The Image-based Joint-Embedding Predictive Architecture (I-JEPA) is a self-supervised learning method that learns highly semantic image representations by predicting abstract representations of target blocks from a single context block within the same image. It consists of three main components that are all based on Vision Transformers (ViTs).

Core Principle: The image-based implementation. It uses multi-block masking, where the context is a set of image blocks, and the target is a different set of blocks at a different spatial location and scale.

I-JEPA Architecture

- Context Encoder: The aim of this component is to process the visible patches of the part of image given to the model (context block) as a clue.
- Target Encoder: It is responsible for computing the representations of the Target Block (part of the image that the model needs to predict). To prevent the model from finding trivial solutions, the weights of the target encoder are an exponential moving average (EMA) of the Context encoder.
- Predictor: This is a light-weight ViT that takes two things as input; the output representation from the context encoder (what the model sees from the context block) and positional mask tokens that tells the predictor where in the image the target block is located. Based on this these inputs the predictor outputs a predicted representation for that specific target block.



Representational Properties: An input image is divided into patches and single, sufficiently informative context block is randomly sampled, and several large target blocks are also randomly sampled. The context encoder than processes only the visible context patches and the target encoder processes the full image's patches to generate the ground truth target representations. The predictor uses the encoded context and mask tokens (indicating where the target blocks are located) to predict the representations of the target blocks.

In the context waste item detection, data augmentation techniques are still widely and rigorously implemented to maximize the generalizability of the final model and counteract overfitting on bespoke datasets.

The augmentation for a waste detection system typically entails techniques designed to introduce variability that reflects real-world conditions, for example:

- Geometric Transformations: Random rotation, scaling, translation, and flips to make the model robust to different object orientations and sizes.
- Colour and Brightness Adjustments: Modifying hue, saturation, brightness, and contrast to simulate various lighting conditions.
- Masking and Cropping: Used to ensure the model can identify waste objects even when partially obscured or only a portion is visible.
- Synthetic Image Generation: Advanced methods like using Generative Adversarial Networks (GANs) to create synthetic images of waste, which helps increase dataset size and realism, a crucial factor given the challenges in collecting diverse real-world waste images from material recovery facilities.

The goal of these augmentations in the waste detection context is to ensure the trained model is robust enough to perform accurately in varied, real-time, real-world scenarios.

4. Conclusion

Modern SSL methods have rapidly evolved from contrastive paradigms requiring negative pairs to predictive architectures modelling latent variables. Contrastive methods such as SimCLR provides a robust theoretical footing but require heavy computational resources. Non-contrastive approaches like BYOL demonstrate that collapse can be prevented through architectural choices, offering simplicity and efficiency. Clustering-based methods such as DINO enable strong semantic feature learning, particularly when paired with ViTs. Finally, predictive frameworks such as JEPA, and I-JEPA represent a shift toward scalable, reconstruction-free representation learning, achieving state-of-the-art performance while aligning more closely with theoretical models of perception. The choice of SSL algorithm depends heavily on computational constraints, dataset scale, and desired representation properties, but the JEPA family currently leads in scalability and performance on large-scale vision benchmarks.

5. Characteristics of the images in different types of datasets

ZeroWaste dataset

The images in the ZeroWaste dataset are taken from real environments**,** making them complex as one image might contain multiple waste items. The same object within the dataset has different shapes, size and textures as it might appear torn or crumpled in the case of plastic or paper. Some pictures were taken from far so data augmentations will be seriously being required, see picture below.

TrashNet Dataset

The characteristics of images in the TrashNet dataset are the opposite of the ZeroWaste images. The images in this dataset possess several distinct visual and structural characteristics that make them suitable for evaluating computer vision and self-supervised learning models. Only one waste object is found per picture and the picture was taken in a plain or uniform backgrounds such as on tables with white surfaces and they have a fixed image resolution hence nor much of data augmentation is required as depicted in the picture below.

6. Bibliography

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple Framework for Contrastive learning Of Visual Representation (SimCLR). *International Conterence on Machine Learning (ICML)*.

Jean-Bastien, G., Florian, S., Florent, A., Corentin, T., Pierre H, R., Elena, B., . . . Rémi, M. (2020). *Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning.* Retrieved from NeurIPS.

Mahmoud, A., Quentin, D., Ishan, M., Piotr, B., Pascal, V., Micheal, R., . . . Nicolas, B. (2023, April 13). *Self-Supervised Learning from Images with a Joint-embedding Predictive Architecture.* Retrieved from arXiv Web site.

Mathilde, C., Hugo, T., Ishan, M., Herve, J., Julien, M., Piotr, B., & Armand, J. (2021, May 24). *Emerging Properties in Self-Supervised Vision Transformers.* Retrieved from arXiv: arXiv:2104.14294v2 [cs.CV] 24 May 2021

Randall, B., & Yann, L. (2025, November 12). *LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics.* Retrieved from arXiv: https://arxiv.org/abs/2511.08544

Tahira, S., Khurram, A. H., Didier, S., & Muhammad, Z. A. (2023, July 10). *Object Detection with Transformers: A Review.* Retrieved from arXiv Web site.