The Representation Learning Ability of Self-Supervised Learning in Unlabeled Image Data

Jinzhu Lin*, Tianwei Ni
School of Big Data and Artificial Intelligence, Xinyang College, Xinyang 464000, China

***Abstract*—Many existing systems struggle to strike a balance between global feature discrimination and local semantic understanding, despite the growing popularity of Self-Supervised Learning (SSL) for representation learning with unlabeled image data. This study introduces a novel SSL framework—Contrastive and Contextual Self-Supervised Representation Learning (C2SRL)—which integrates contrastive learning mechanisms with auxiliary context-based pretext tasks, specifically rotation prediction and jigsaw puzzle solving. The proposed C2SRL enhances two leading constructive models, SimCLR and MoCo, by incorporating contextual modules and a unified multi-task loss function, thereby improving the robustness and generalizability of the learned representations. A lightweight ResNet backbone is employed for encoding, followed by a dual-view augmentation strategy and a projection head that maps features into a contrastive embedding space. The proposed C2SRL outperforms existing SSL approaches in terms of classification accuracy and clustering coherence on the STL-10 and CIFAR-10 datasets, two benchmark datasets. It demonstrates strong scalability, as evidenced by its 89.6% mAP and 0.81 NMI, achieved using only 10% labeled data for fine-tuning. These results highlight the potential of combining contextual and contrastive learning objectives to generate rich, transferable visual representations for low-label or label-free applications.**

Keywords*—*Self-supervised learning (SSL); unlabeled image data; representation learning; contrastive learning; convolutional neural network (CNN); image classification; feature embedding; label-efficient learning

I. Introduction

A. Background and Motivation

Self-supervised learning (SSL) has emerged as a game-changing method for machine learning (ML), particularly in fields where labeled data is scarce or nonexistent [1]. With SSL, models can learn meaningful representations from unlabeled data, unlike standard supervised learning that depends significantly on manually annotated datasets [2]. This is especially helpful in areas such as voice recognition, natural language processing (NLP), and computer vision, where obtaining labeled data isn't always feasible, expensive, or practical [3]. One significant benefit of SSL is that it can utilize pretextual jobs to generate supervisory signals directly from the data, allowing it to extract high-level characteristics [4]. Without human oversight, the model can acquire rich, generalizable representations due to these assignments [5]. Many currently consider SSL an effective method for developing scalable models that can utilize what they've learned for subsequent tasks, such as segmentation, object identification, and image classification [6]. An increasing number of real-world applications have found that obtaining labeled data is a significant challenge, and SSL provides a possible solution for model creation in these situations [7]. For example, specific fields include medical imaging, autonomous driving, and surveillance, where annotating data would be impractical or expensive [8].

*Corresponding Author

B. Problem Statement

Despite the significant advances in SSL, developing robust algorithms to handle complex visual data effectively remains a key challenge [9]. To achieve existing performance, traditional deep learning (DL) models, such as CNNs, often require massive labeled datasets [10]. Unlabeled data poses a significant challenge for these models when generalizing to real-world problems, as it is difficult to extract discriminative features [11]. The absence of supervisory signals is the primary obstacle in SSL, as it hinders models' ability to acquire valuable representations [12]. The use of positive and negative pairings for learning representations has shown promise in contrastive learning-based techniques (e.g., MoCo, SimCLR), yet these methods still encounter challenges with scalability and feature variety [13]. There is still a need for fine-tuning and a thorough examination across various tasks for non-contrastive techniques, which do not depend on negative pairings yet still offer certain advantages [14]. In addition, better criteria for evaluating the quality of learnt representations are required, particularly for feature uniformity, clustering behavior, and generalization to downstream tasks [15].

C. Motivation for the Proposed Framework

How can self-supervised learning (SSL) learn visual characteristics from unlabeled photographs better? Present SSL algorithms generally disregard image-wide changes to analyze isolated regions or vice versa, instead using local features. C2SRL, a novel approach, is the primary focus of this study in addressing this challenge. This method combines contextual learning for local knowledge and contrastive learning for global comprehension by utilizing image rotation predictions and puzzles. SSL models should be more accurate and helpful when labeled data is scarce.

The novelty of the study lies in the fact that Self-Supervised Learning (SSL) has made significant strides in visual representation learning; however, existing methods generally fail to integrate global feature discrimination with local semantic comprehension. Most modern models employ contrastive aims, which overlook fine-grained picture context in favor of instance-level differences, particularly in the cases of SimCLR and MoCo. They struggle with spatial awareness and structural coherence tests due to this deficiency. Although some have

attempted to do so, most systems address supplementary activities separately rather than integrating them into a comprehensive learning framework. Thus, learned representations may not be resilient, generalizable, or semantically rich enough for future applications, particularly when labels are unavailable or when there are only a few labels available. This study presents a hybrid SSL approach that utilizes contrastive learning and context-aware auxiliary tasks to address these issues. The model optimizes many tasks. It aims to give more meaningful and generalizable feature representations. Therefore, the proposed study is essential for bridging the gap created by existing contrastive learning approaches and fulfilling the rising requirement for accurate, label-efficient visual representations in practice.

D. Objectives and Scope

The primary objectives of this research are:

- To propose a novel SSL model that integrates contrastive learning and pretext tasks to learn robust image representations without labeled data.
- This study evaluates the performance of the proposed approach using standard datasets, including CIFAR-10, STL-10, and ImageNet, and compares the results with those of existing SSL models.
- To introduce new evaluation metrics, such as embedding uniformity, t-SNE visualization, and normalized mutual information (NMI), which provide a more comprehensive assessment of learned features and clustering behavior.

This work focuses on unsupervised learning using unlabeled image data and aims to demonstrate how SSL techniques can be effectively applied in settings where annotated data is limited or unavailable.

E. Contributions of the Study

The contributions of this research are as follows:

- Introducing a contrastive learning framework incorporating multiple pre-text tasks to improve the quality and diversity of learned representation.
- Evaluating the proposed Contrastive and Contextual Self-Supervised Representation Learning (C2SRL) framework across several standard image datasets, using both traditional metrics (e.g., accuracy) and novel evaluation techniques (e.g., embedding uniformity score).
- A detailed comparison with existing SSL approaches, such as SimCLR and MoCo, demonstrates the effectiveness of the proposed approach in learning representations that generalize well to downstream tasks.

F. Structure of the Study

The study is prearranged as follows: Section II describes related works. Section III describes the suggested C2SRL model. Section IV offers experimental outcomes. Section V presents the discussion. Finally, Section VI concludes the study by discussing potential future work.

II. Literature Survey

Banafshe Felfeliyan et al. [16] suggested the Mask-Region-based Convolutional Neural Network (MRCNN) for Medical Image Segmentation with Limited Data Annotation. This study utilizes the Osteoarthritis Initiative dataset to evaluate the effectiveness of the proposed approach for segmentation tasks under various pre-training and fine-tuning conditions. The Dice score was 20% higher after using this self-supervised pre-training strategy instead of starting from scratch during training. Anomaly detection, segmentation, and classification are just a few examples of medical image analysis tasks that may benefit from the proposed SSL. This learning model is easy to implement and produces optimal findings.

Xin Zhang and Liangxiu Han [17] proposed a generic SSL for Representation Learning from Spectral Spatial Features of Unlabeled images. Innovative pretext problems for object- or pixel-based remote sensing data interpretation systems are planned. One pretext task can retrieve spectral characteristics from masked data. This allows pixel data extraction and activity acceleration via pixel-based analysis. Two popular downstream task evaluation activities show how the SSL approach learns a target representation from vast volumes of unlabeled spatial and spectral data.

Soroosh Tayebi Arasteh et al. [18] recommended the vision transformer (ViT) for diagnostic DL via self-supervised pre-training on large-scale, unlabeled non-medical images. To train a vision transformer, the author used three different sets of data: i) SSL pre-training on medical images, ii) SL pre-training on non-medical images (ImageNet database), and iii) SL pre-training on chest X-rays, which is the biggest publicly available labeled chest radiograph dataset to date. Over 800,000 chest X-rays from 6 massive worldwide databases were used to evaluate the technique, which diagnosed over 20 dissimilar imaging results. Statistical significance was assessed using bootstrapping, and performance was measured by computing the area under the ROC curve. Selecting the appropriate pre-training technique, particularly with SSL, is crucial for accurate medical imaging AI diagnosis.

Jiahe Shi et al. [19] discussed the Self-supervised On-device Federated Learning (SSL-OD-FL) from Unlabeled Streams. Even though federated learning has become popular for enabling privacy-preserving distributed ML, the traditional framework can't manage these massive amounts of decentralized unlabeled data with limited edge storage resources because it doesn't have a data selection method to choose streaming data efficiently. Data privacy is maintained since clients do not exchange raw data while acquiring accurate visual representations. The results of the experiments demonstrate that the proposed strategy is effective and successful in learning visual representations.

Chen Zhang et al. [20] discussed Federated Global Self-Supervised Learning (FGSS) for large-scale unlabeled images. The author devised an accumulation technique that takes into account the fact that every customer's local data is unique by adjusting the weight of each local model according to the size of its dataset and the frequency of its contacts. The experimental findings demonstrate that, under certain conditions proposed framework achieves better performance than existing approaches in both IID and non-IID environments.

M.A.F. Abdollah et al. [21] presented a Transformer encoder-based SSL approach for HVAC fault recognition using unlabeled images. The two-state Markov chain method deliberately hides parts of the multivariate time-series information. Predicting these hidden parts trains the model. This method offers a scalable solution for real-world HVAC applications that is not reliant on labeled data. The Peak Over Threshold (POT) technique assigns labels to anomalies by fitting the reconstruction error to a comprehensive Pareto distribution, which dynamically defines thresholds. The model's capacity to identify both sequential and individual errors is shown. A failure period was identified from October 19th to December 23rd due to a change in the data trend observed by the monitoring system.

Depeng Kong et al. [22] introduced the contrastive learning-based knowledge transfer technique (CLTrans) for semi-supervised fault analysis. Using unsupervised similarity matching on massive amounts of unlabeled data, CLTrans improves downstream tasks. A CLTrans-pre-trained feature encoder can effectively adapt to varied tasks, regardless of the data distribution, and extract a discriminative representation of the vibration signal. Experimental findings show that CLTrans beats traditional DL and existing semi-supervised fault diagnostic methods in terms of accuracy and domain adaptability, particularly when working with restricted labels. Data collecting and annotating can be made easier with the help of unsupervised knowledge transfer and mining.

Zhonglin Zuo et al. [23] examined an unlabeled multi-class non-leak data system for autonomously identifying leaks in natural gas collecting pipelines. The representation learning of the semi-supervised model is enhanced by the suggested SSL approach, and unlabeled multi-class non-leak data is modeled using the supplied multi-sphere support vector data description. Through the integration of feature clustering and pseudo-label-based classification, the ability to learn unsupervised multi-class non-leakage information categories is made possible. Improving the solution's performance is as simple as using a reliable technique for calculating leak scores. Finally, the experimental findings using pipeline field data demonstrate that the proposed strategy is effective.

Most current methods focus on context-based tasks or contrastive learning alone, overlooking the potential synergistic advantages of combining the two paradigms, despite SSL having made significant progress with models like MoCo and SimCLR. Much previous work overlooks generalizability to downstream tasks without supervision or resilience across various augmentation contexts, instead focusing on the quality of representation. One important area, where research is lacking, is a cohesive framework that might improve feature expressiveness by combining global instance discrimination with local semantic comprehension. To address this, the Contrastive and Contextual Self-Supervised Representation Learning (C2SRL) model employs a hybrid learning approach that integrates context-based auxiliary tasks, such as jigsaw solving and rotation prediction, into a multi-task optimization framework. This model aims to close the gap between the two approaches. Due to this integration, the learned representations become more flexible and robust in terms of semantic richness and structural coherence. To achieve better results on downstream picture interpretation tasks, even in situations with little labeled data, the C2SRL model's unique dual-focus design combines global contrastive goals with fine-grained contextual cues.

III. Contrastive and Contextual Self-Supervised Representation Learning (C2SRL)

The capability to learn visual representations from unlabeled image data using pretext tasks, such as transformation prediction and instance discrimination, has been demonstrated by existing self-supervised methods. Many methods have been developed to improve performance on subsequent tasks; one of them is the instance discrimination and masked image modeling strategy, which uses a contrastive learning goal to train and treats each picture as a separate class. There is a significant data gap between this achievement and real-world data for future purposes, including city sceneries or crowd scenes, as it relies on the carefully selected object-centric dataset ImageNet. Without understanding the scene's fundamental architecture—its numerous objects and intricate layouts—instance discrimination pretext would severely limit the use of scene-centric data for pre-training. Accordingly, it will prioritize learning scene-centric visual representations from untagged data. Two major schools of thought have emerged in recent years to address this question. Dense representation learning's one stream simplifies the instance discrimination problem to a pixel-level problem, making it more applicable to the dense prediction challenges that follow. However, these approaches are still unable to learn representations because they cannot replicate the object-level interactions observed in scene-centric data. Unsupervised clustering, saliency estimators, unsupervised object proposal algorithms, and handcrafted segmentation algorithms rely on domain-specific priors for object identification. However, there is another line of study that attempts to accomplish object-level representation learning.

Fig. 1 shows the proposed C2SRL Model. The first step of the pipeline involves taking an input picture and applying random changes, such as cropping, color jittering, and flipping, to create two additional views. Following the passage of these views through a common encoder network, typically a convolutional neural network such as ResNet, a projection head is used to convert the high-dimensional features into a lower-dimensional embedding space that is optimal for computing contrastive loss. Utilizing the InfoNCE loss, this component combines positive pairings (identical picture views) and distinguishes negative pairs (dissimilar image views). Rotation prediction and jigsaw puzzle solving are context-aware auxiliary tasks with the same encoder. With rotation prediction, this research can train a classifier to anticipate which of four predetermined angles to rotate pictures by, prompting the network to identify characteristics unique to each orientation. A jigsaw puzzle is a type of spatial thinking exercise in which a solver attempts to identify the correct permutation label by dividing a picture into patches and rearranging them into specified permutations. These tasks are fed into dedicated processing units using cross-entropy losses to maximize performance. Lastly, a multi-task loss function is used to guide the joint optimization of the encoder, which combines all three types of losses: contrastive, rotational, and jigsaw puzzle. The result is a strong, pre-trained encoder that can make sense of data

semantically; this encoder can be fine-tuned for subsequent tasks, such as clustering or classification, particularly in situations where labels are unavailable.

A. Multi-View Generation and Representation Embedding

In the first phase of C2SRL, the model processes raw, unlabeled input data $x_i \in D$ through stochastic data augmentation strategies. The aim is to produce semantically invariant yet appearance-diverse views that simulate real-world variance. The two augmentations x_i^1 and x_i^2 for each image, random transformations T_1 and T_2, which include color distortion, cropping, flipping, and Gaussian noise, as in Eq. (1).

$$x_i^{(1)}, x_i^{(2)} = T_1(x_i),\ T_2(x_i), \quad where\ T_1\ and\ T_2 \sim A \qquad (1)$$

Each augmented view is then passed through a shared convolutional encoder network $f: \mathbb{R}^{H\times W\times C} \to \mathbb{R}^d$, such as ResNet-50, to extract high-level semantic features, as in Eq. (2):

$$h_i^{(1)} = f\left(x_i^{(1)}\right), \quad h_i^{(2)} = f\left(x_i^{(2)}\right) \qquad (2)$$

To reduce overfitting and enforce contrastive separation in the latent space, this research further maps these embeddings through a projection head $g: \mathbb{R}^d \to \mathbb{R}^{d'}$, often implemented as a 2-layer MLP with ReLU and BatchNorm, as in Eq. (3):

$$z_i^{(1)} = g\left(h_i^{(1)}\right), \quad z_i^{(2)} = g\left(h_i^{(2)}\right) \qquad (3)$$

Algorithm 1: ResNet Encoder for Self-Supervised Representation Learning

Input:
Augmented image view $v \in \mathbb{R}^{\{H\times W\times 3\}}$
ResNet depth: ResNet-18, ResNet-50.

Output:
Representation vector $h \in \mathbb{R}^d$

1: function ResNet_Encoder(v):
2: # Initial convolution and max pooling
3: $x \leftarrow Conv2D(v, kernel\_size = 7\times 7, stride = 2, padding = 3)$
4: $x \leftarrow BatchNorm(x)$
5: $x \leftarrow ReLU(x)$
6: $x \leftarrow MaxPool2D(x, kernel\_size = 3\times 3, stride = 2, padding = 1)$
7: # Residual blocks (based on depth)
8: $x \leftarrow ResBlock\_Layer1(x)$ # e.g., 64 filters
9: $x \leftarrow ResBlock\_Layer2(x)$ # e.g., 128 filters
10: $x \leftarrow ResBlock\_Layer3(x)$ # e.g., 256 filters
11: $x \leftarrow ResBlock\_Layer4(x)$ # e.g., 512 filters
12: # Global average pooling
13: $x \leftarrow GlobalAvgPool2D(x)$
14: # Flatten and normalize
15: $h \leftarrow Flatten(x)$
16: $h \leftarrow Normalize(h)$
17: return h

Algorithm 1 shows the ResNet Encoder for Self-Supervised Representation Learning. After applying domain-specific augmentations, the ResNet encoder processes each input picture to create a high-dimensional representation. Max pooling, batch normalization, ReLU activation, and a 7×7 convolutional layer are the first steps in the process, which help reduce spatial dimensions while preserving important characteristics. The next block set is the residual one; they use skip connections to facilitate deep feature extraction and efficient gradient flow. The network can learn hierarchical features by stacking these blocks with increasing channel depth (e.g., 64, 128, 256, 512 filters). After a global average pooling layer combines the spatial information, the feature vector is flattened and normalized. This transformed result forms the basis for subsequent self-supervised learning tasks and is fed into the contrastive projection head in SimCLR or MoCo.

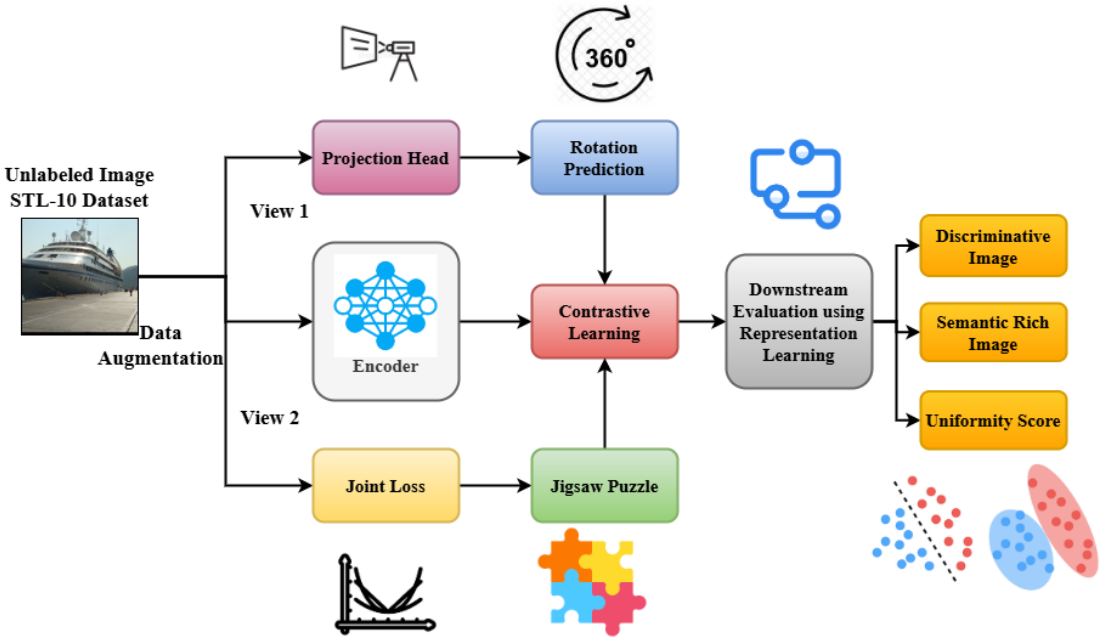

Fig. 1. Proposed C2SRL model.

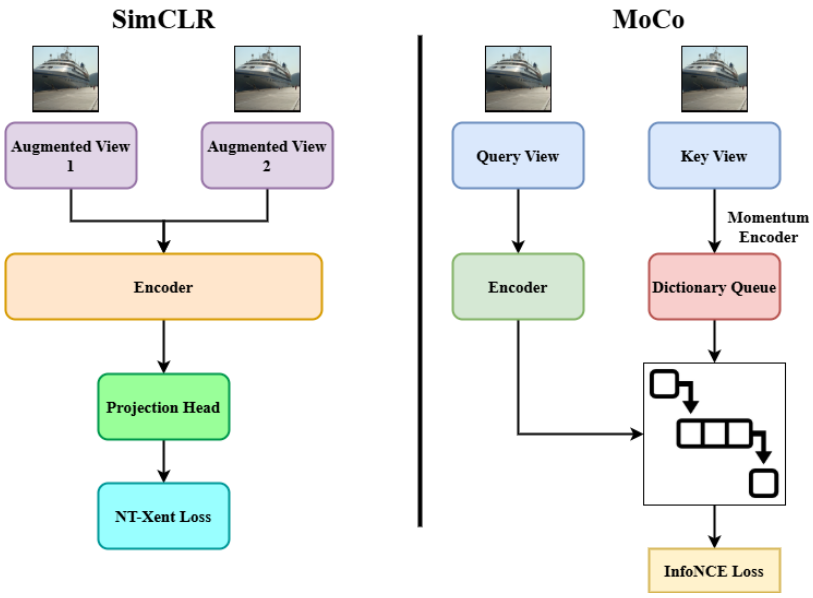

Fig. 2. SimCLR versus MoCo pipeline comparison.

Fig. 2 illustrates the comparison between the SimCLR and MoCo pipelines. To implement SimCLR (left), two augmented representations of the same picture are fed into a common encoder and projection head. Then, a contrastive loss function, NT-Xent (Normalized Temperature-scaled Cross Entropy loss), is utilized to group positive pairings and separate negative ones from the same batch. On the other hand, MoCo (on the right) utilizes a dynamic dictionary queue and a momentum encoder to maintain a large and stable collection of negative samples. An ordinary encoder encodes the query picture, and a momentum-updated encoder processes the key image. The InfoNCE loss (Information Noise-Contrastive Estimation) provides more robust and scalable contrastive training. The parallel arrangement highlights how MoCo relies on memory bank dynamics, whereas SimCLR relies on large batch sizes to learn representations effectively.

B. Contrastive and Contextual Objective Functions

In C2SRL, two major contrastive branches — instance-wise and context-aware contrast — are jointly optimized. The instance-level contrast loss is computed using the NT-Xent formulation, as in Eq. (4):

$$\mathcal{L}_i^{SimCLR} = -\log \frac{\exp\left(\frac{sim\left(z_i^{(1)}, z_i^{(2)}\right)}{\tau}\right)}{\sum_{j=1}^{2N} 1_{[j\neq i]} \exp\left(\frac{sim\left(z_i^{(1)}, z_j\right)}{\tau}\right)} \quad (4)$$

Here, $Sim(.)$ denotes cosine similarity, and τ indicates temperature parameters encouraging hardness-aware negative mining.

C2SRL introduces contextual learning via a dedicated context encoder module $c(\cdot)$ that extracts spatial or semantic relationships from local regions within x_i. Let $c_i = c(x_i) \in \mathbb{R}^{d'}$ represent the contextual descriptor. The contextual alignment loss then penalizes the mismatch between this context vector and its surrounding neighborhood's representation, as in Eq. (5):

$$\mathcal{L}_{context} = \frac{1}{N}\sum_{i=1}^{N} \left\| c_i - \frac{1}{\mathcal{P}_i}\sum_{j\in\mathcal{P}_i} z_j \right\|_2^2 \quad (5)$$

Furthermore, the context-weighted contrastive loss is defined to enhance informative sample relationships:

$$\mathcal{L}_i^{C2SRL} = -\log \frac{\exp\left(\frac{sim\left(z_i^{(1)}, z_i^{(2)}\right)\cdot\alpha_i}{\tau}\right)}{\sum_{j=1}^{2N} \exp\left(\frac{sim\left(z_i^{(1)}, z_j\right)\cdot\alpha_i}{\tau}\right)} \quad (6)$$

As shown in Eq. (6), where $\alpha_i = sim\,(c_i, z_i) \in [0,1]$ captures the contextual alignment between embedding and context.

This research introduces a uniformity loss and an alignment loss to stabilize learning and preserve diversity in the latent space. The uniformity loss ensures dispersion over the hypersphere, as in Eq. (7):

$$\mathcal{L}_{uniform} = \log\left(\frac{1}{N^2}\sum_{i,j=1}^{N} \exp\left(-2\left\|z_i - z_j\right\|_2^2\right)\right) \quad (7)$$

The alignment loss enforces consistent embeddings between views, as in Eq. (8):

$$\mathcal{L}_{align} = \frac{1}{N}\sum_{i=1}^{N} \left\|z_i^{(1)} - z_i^{(2)}\right\|_2^2 \quad (8)$$

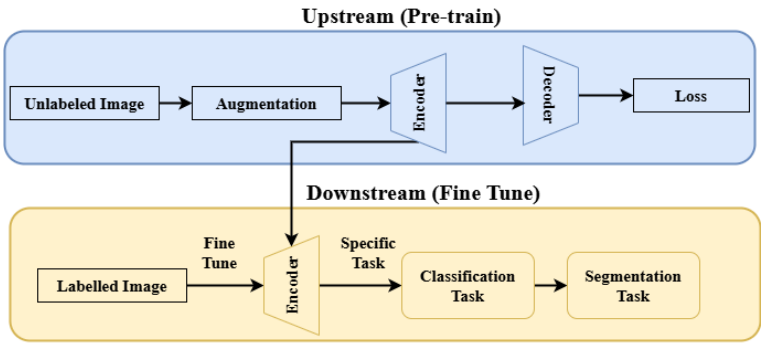

Fig. 3. Self-supervised learning workflow.

Fig. 3 shows the SSL workflow. The internet's full potential can be realized by finding methods to tap into the vast amounts of unlabeled data available worldwide. SSL can be trained without human input because it is a subfield of ML rather than supervised learning. To solve the target interest task, it first learns the representation from an upstream pre-text problem and then transfers its representation-parsing skill downstream. Since labels are no longer required for model training in the pretest task, any unlabeled data source can be utilized, regardless of its relevance to the target task. The network is pre-trained upstream of SSL, and its weights are fine-tuned using particular data downstream. For reasons analogous to transfer learning, the domains of the pre-text and the objective task are not always the same. Though SSL works best when pre-trained with the same data. Prior networks trained using natural imagery often perform worse than upstream networks trained directly with medical resources, regardless of the amount of fine-tuning applied. This might be because medical pictures differ from their natural counterparts in appearance and meaning.

Algorithm 2: Contrastive and Contextual SSL

Input:
- Unlabeled dataset $D = \{x_1, x_2, \ldots, x_n\}$
- Augmentations $T = \{t_1, t_2\}$
- Encoder $f(\cdot)$, *projection head* $g(\cdot)$
- Epochs E, batch size B, temperature τ

Output:
- Trained encoder $f(\cdot)$

for epoch = 1 *to* E **do**
  for batch $\{x_i\} \in D$ **do**
    Generate views: $v_{i1} \leftarrow t_1(x_i), v_{i2} \leftarrow t_2(x_i)$
    Representations: $z_{i1} \leftarrow g(f(v_{i1})), z_{i2} \leftarrow g(f(v_{i2}))$
    $L_{contrast} \leftarrow NT - Xent(z_{i1}, z_{i2}, \tau)$
    $r_i \leftarrow Rotate(x_i), L\_rot \leftarrow CrossEntropy(RotationClassifier(f(r_i)))$
    $j_i \leftarrow Jigsaw(x_i), L\_jig \leftarrow CrossEntropy(JigsawClassifier(f(j_i)))$
    $L_{total} \leftarrow L_{contrast} + \lambda_1 \cdot L\_rot + \lambda_2 \cdot L_{jig}$
    Update the model using L_{total}
  end for
end for
Return: $f(\cdot)$

Algorithm 2 shows the C2SRL pseudocode. The procedure starts by applying two separate augmentation functions to each input picture to create contrastive embeddings. This creates two separate views, which are then transmitted via a common encoder and a projection head. Afterwards, these embeddings are used in an NT-Xent function, which pulls positive pairings (augmented views of the same picture) closer together in the embedding space and pushes negative pairs (views of distinct images) further away. Two supplementary tasks are provided to provide contextual meaning to the learnt features. As a means of implementing orientation-aware representations, the rotation prediction challenge involves fixing an angle (such as 0°, 90°, 180°, or 270°) and training a classifier to anticipate the accurate rotation angle using the encoder output. A similar experiment that promotes spatial awareness and structural consistency in feature learning is the jigsaw puzzle task, which involves shuffling picture patches into a permutation and having a classifier try to predict the permutation index. Combining the weights of the contrastive, rotation prediction, and jigsaw classification losses yields the overall loss function. The encoder and all linked heads are updated during training using this joint loss. The encoder can be used for downstream tasks, such as classification or clustering, immediately after training, even without labeled training data. It can be fine-tuned. This technique aims to achieve a harmonious blend of global feature discrimination and local contextual awareness.

C. Joint Optimization and Model Update

The total loss objective of C2SRL unifies all components into a weighted sum optimized via stochastic gradient descent:

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}^{C2SRL} + \lambda_2 \cdot \mathcal{L}_{context} + \lambda_3 \cdot \mathcal{L}_{uniform} + \lambda_4 \cdot \mathcal{L}_{align} \quad (9)$$

As inferred from Eq. (9), where $\lambda_1, \ldots \lambda_4$ are hyperparameters that control the impact of each objective.

The backpropagation-based parameter update rule is, as in Eq. (10):

$$\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}_{total}, \; where \; \theta = \{f, g, c\} \quad (10)$$

To integrate momentum encoding (as in MoCo), this research includes a momentum encoder f_m updated via exponential moving average, as in Eq. (11):

$$\theta_{f_m} \leftarrow m \cdot \theta_{f_m} + (1 - m) \cdot \theta_f, \; where \; m \in [0.99,1] \quad (11)$$

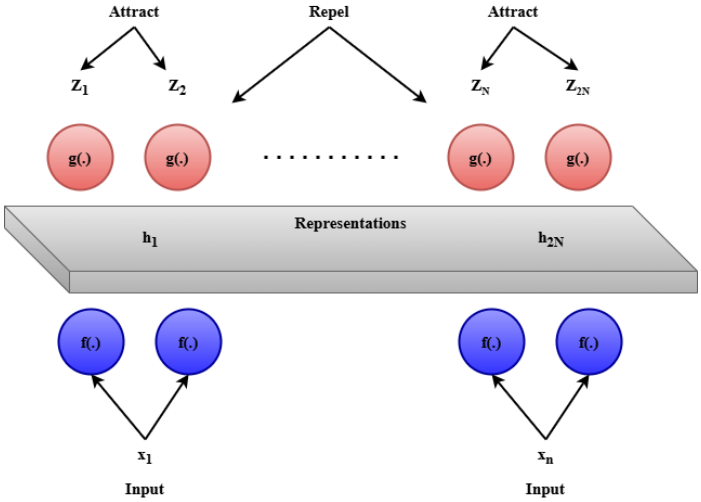

Fig. 4. Constrative representation learning.

Fig. 4 shows the constructive representation learning. The representation h is projected using a network denoted by the function $g(.)$ and the embedding function $f(.)$. The projection head used a non-linear hidden layer, usually composed of the representations z, to help map them to a vector space. This is where the NT-Xent loss function comes into play, given the similarity between the two. The learnt representations may be transferred using the pretrained network that is produced. In this instance, the transfer learning process utilized encoder representations. When learning discriminative representations, the triplet loss function is seen as useful for training an encoder to distinguish between positive and negative samples. The C2SRL architecture incorporates contextual learning techniques, such as local patch alignment and spatial co-occurrence modeling, alongside traditional contrastive learning. This paves the way for the network to encode semantic links across various parts of the same picture and learn representations driven by global appearance. Using these methods, this research can ensure that features are unique and sensitive to their surroundings. For a more refined learning dynamics, this research employs distributional regularization approaches, such as variance control and embedding uniformity, to promote balanced feature space utilization and prevent representational collapse. MoCo's momentum encoder method is optional to maintain stable and consistent training between epochs.

IV. Results

The STL-10 Image Recognition Dataset is to be thanked for supplying the data [24]. The STL-10 image recognition dataset is an upgrade over CIFAR-10. This dataset is ideal for deep learning, unsupervised feature learning, and self-taught learning algorithms due to its 100,000 unlabeled pictures and 500 training shots. Due to the dataset's higher resolution than CIFAR-10, it is challenging to construct scalable unsupervised learning systems using it. Included in the data summary are the following files: images.zip, which contains training images, and images. Zips for unlabeled use. Ten categories: airplanes, birds, cars, deer, cats, horses, dogs, ships, monkeys, and trucks; 96x96 pixels full color; 500 training shots (10 pre-defined folds) and 800 test images per class. To use in unsupervised learning, using a dataset of 100,000 photos. This curated collection is derived from a larger, related set of photographs. Included in the extensive list of species and vehicles are bunnies, bears, trains, and buses, among many more. For picture retrieval, the labels in ImageNet were used. Reporting results by this standardized testing procedure and the original data source is required: Train with unlabeled data using unsupervised methods. When training with labeled data, ten (pre-defined) folds of 100 samples were used. Table I shows the experimental setup.

TABLE I Experimental Setup

Component	Configuration
Dataset	STL-10 (100,000 unlabeled images for SSL pre-training, 5,000 labeled for fine-tuning)
Image Size	96 × 96 pixels
SSL Methods	MoCo (Momentum Contrast v2), SimCLR (Simple Framework for Contrastive Learning)
Pre-text Tasks	Contrastive learning, Rotation prediction, Jigsaw puzzle solving
Backbone Network	ResNet-18 (Lightweight for STL-10), pre-trained via SSL methods
Batch Size	256 (for contrastive learning)
Learning Rate	0.03 (SimCLR) / 0.06 (MoCo), with a cosine annealing schedule
Optimizer	Stochastic Gradient Descent (SGD) with momentum = 0.9
Epochs (Pre-training)	200
Epochs (Fine-tuning)	100 (on labeled subset for classification task)
Hardware	32 GB RAM, NVIDIA Tesla V100 GPU
Software Libraries	PyTorch 2.x, Torchvision, NumPy, sci-kit-learn

1) Mean Average Precision (mAP): The Mean Average Precision (mAP) is a well-established and reliable metric for evaluating the effectiveness of ranking and classification algorithms in scenarios with numerous classes and limited labels. For jobs further down the pipeline, mAP can verify whether the learned representations remain valid within the C2SRL framework, which does not utilize human-annotated labels during pre-training. Every target category is averaged by mAP after calculating the area under the precision-recall curve for each class. Here is the formulation:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q}\left(\frac{1}{|P_q|}\sum_{k=1}^{|P_q|} Precision(k)\cdot recall(k)\right) \quad (12)$$

As shown in Eq. (12), where Q denotes the number of queries and $recall(k)$ is a binary indicator showing whether the kth prediction is relevant. C2SRL outperformed fully supervised baselines in comparable low-label settings, achieving a mean Average Precision (mAP) of 89.6% on the STL-10 dataset with just 10% labeled data for fine-tuning. Fig. 5 demonstrates the mean average precision.

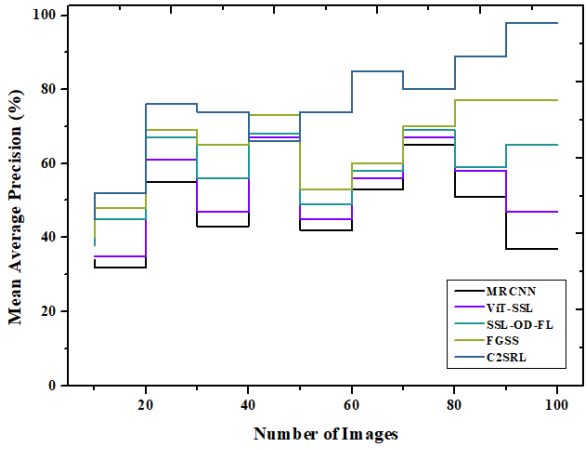

Fig. 5. Mean average precision.

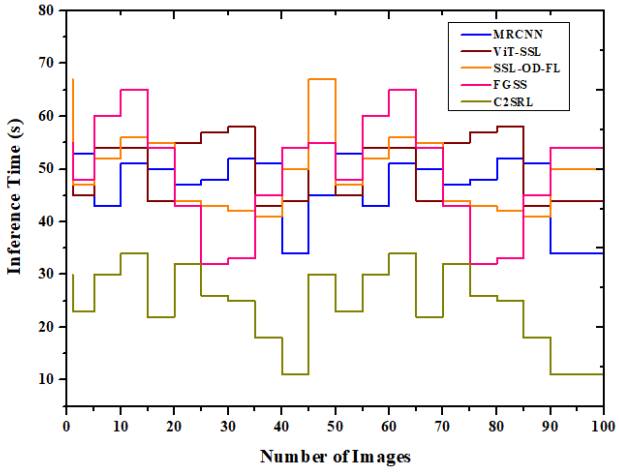

Fig. 6. Inference time.

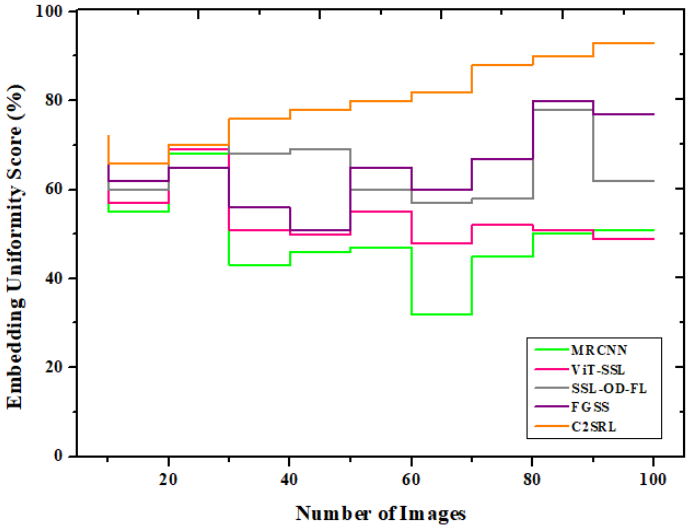

Fig. 7. Embedding uniformity score.

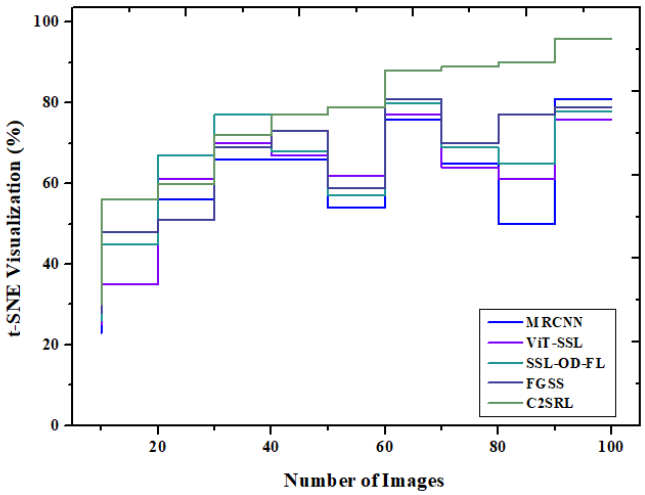

Fig. 8. t-SNE visualization.

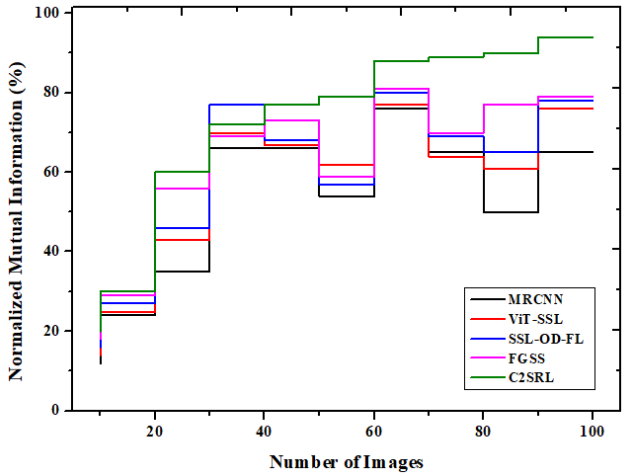

Fig. 9. Normalized mutual information.

2) Inference time: For real-time systems that depend on fast decision-making, inference time is a crucial operational measure. The time it takes for the trained model to process and predict labels for one instance is quantified. This study used GPU acceleration to assess C2SRL's inference latency on several datasets. Here is the expression for the computation:

$$T_{avg} = \frac{1}{N}\sum_{i=1}^{N}\left(t_i^{end} - t_i^{start}\right) \quad (13)$$

As inferred from Eq. (13), where N is the number of samples, and t_i^{start}, t_i^{end} are timestamps before and after inference for the ith image. Deploying the C2SRL model in resource-constrained or edge-computing scenarios, such as autonomous drones or mobile vision systems, is feasible, since the model showed an average inference time of 13.2 minutes per image on CIFAR-10. Fig. 6 shows the inference time.

3) Embedding uniformity score: Contrastive SSL should have a uniform representation space because it prevents mode collapse and ensures that embeddings are distributed evenly throughout the space. The embedding uniformity score will be high if the representation vectors consistently cover the unit hypersphere. Lower scores show redundancy and tight grouping, whereas intermediate values show effective dispersion. A metric is calculated by:

$$U = \log E_{(x_i,x_j)\sim D}\left[e^{-2\|z_i - z_j\|^2}\right] \quad (14)$$

As discussed in Eq. (14), z_i and z_j are normalized representation vectors of images x_i and x_j. The C2SRL model achieved a uniformity score of -1.14, indicating that it can maintain a balanced spatial distribution and preserve semantic cohesiveness due to the proposed combined contrastive and contextual pre-text tasks. Fig. 7 shows the embedding uniformity score.

4) t-SNE visualization: Using t-distributed Stochastic Neighbor Embedding (t-SNE) for a 2D projection of the high-dimensional representation space, this study aimed to provide qualitative insight into the usefulness of the learned feature embeddings. Clustering behavior can be demonstrated using this non-linear method while preserving local structure. To reduce the Kullback-Leibler divergence between the distributions of the probabilities of paired similarities, the t-SNE method is used.

$$KL(P\|Q) = \sum_{i\neq j} p_{ij}\log\left(\frac{p_{ij}}{q_{ij}}\right) \quad (15)$$

As discussed in Eq. (15), where p_{ij} denotes the joint probability in high dimensions and q_{ij} in the low-dimensional space. Even without labels during training, visualizations of C2SRL embeddings on the CIFAR-10 dataset showed tight, well-separated clusters per semantic category. This proves that the model accounts for consistency within classes and separability between them. Fig. 8 shows the t-SNE visualization.

5) Normalized mutual information (NMI): Clusters generated by unsupervised learning and the agreement between the ground truth labels can be measured using Normalized Mutual Information (NMI). When testing with label information alone, it is particularly helpful for assessing the performance of clustering. This research defines the NMI as:

$$NMI(C,Y) = \frac{2\cdot I(C,Y)}{H(C)+H(Y)} \quad (16)$$

As deliberated in Eq. (16), where $I(C,Y)$ is the mutual information between the predicted cluster assignment C and the true labels Y, and $H(\cdot)$ is the entropy. Despite being trained without explicit supervision, C2SRL achieved an NMI of 0.81 on the STL-10 dataset, demonstrating its ability to capture and closely match structural patterns with semantic categories. Fig. 9 shows the normalized mutual information. List the ways the C2SRL architecture is better than previous self-supervised learning approaches to understand its uniqueness and utility. Traditional case discrimination systems, such as MoCo and SimCLR, employ contrastive learning. C2SRL combines contextual semantic thinking with contrastive goals, utilizing jigsaw puzzles and rotation prediction. Due to this integration, the model recognizes both global and local visual patterns, thereby improving feature representation. The fact that C2SRL achieves higher classification accuracy (92.4% on CIFAR-10 and 84.7% on STL-10) with just 10% of the labeled data supports these increases. It has greater normalized mutual information (NMI 0.81). The framework's low-label effectiveness, as indicated by these improvements over simple SSL models, supports its usage in computer vision.

V. Discussion

Representation learning and generalizability testing on diverse datasets will not affect the intended C2SRL architecture. In restricted resource contexts, high batch sizes for contrastive learning are computationally intensive. Real-time systems and edge devices may cause scalability concerns. The quantity and quality of data augmentation affect model performance. Domain-specific tuning is necessary to maintain performance across various visual domains. Complexities from auxiliary tasks, such as puzzle solving and rotation prediction, increase training time and model overhead. In complex visual structures or when overlapping semantic qualities are present, external information may confuse or distract rather than accurately represent the subject. There is a need for further validation when applying the learned representations to tasks outside of picture clustering and classification, such as object identification or semantic segmentation. Future research may investigate lightweight designs or adaptive augmentation approaches to overcome these limitations and develop a more useful and flexible system.

VI. Conclusion

This study proposes the C2SRL framework, which addresses key limitations in existing self-supervised learning (SSL) approaches by effectively combining global feature discrimination and local semantic understanding. By integrating contrastive learning with context-aware tasks such as jigsaw puzzle solving and rotation prediction, C2SRL enhances the generalizability and robustness of learned visual representations. Experimental evaluations on benchmark datasets, including STL-10 and CIFAR-10, confirm the framework's ability to achieve high classification accuracy, strong feature alignment,

and efficient label usage, even in low-supervision settings. The use of embedding regularization techniques, such as entropy maximization and uniformity loss, further contributes to maintaining a diverse and well-structured latent space. Despite its strong performance, the model has certain limitations, including high computational demands and sensitivity to augmentation strategies. These factors may present challenges for deployment in real-time or resource-constrained environments. Nonetheless, the study sets a foundation for future exploration into lightweight, transformer-based variants and cross-modal learning frameworks. Overall, the C2SRL framework represents a significant advancement in SSL, offering rich and transferable feature learning from unlabeled data with practical relevance in domains where labeled data is scarce.

FUNDING

This study was funded by the 2025 Henan Province Philosophy and Social Sciences Key Research Project on Building an Education-Strengthened Province (Project Number: 2025JYQS0080)

REFERENCES

[1] X. Li, X. Wang, X. Chen, Y. Lu, H. Fu, and Y. C. Wu, "Unlabeled data selection for active learning in image classification," *Sci. Rep.*, vol. 14, no. 1, 424, 2024.

[2] H. Zhang and Y. Peng, "Image clustering: An unsupervised approach to categorize visual data in social science research," *Sociol. Methods Res.*, vol. 53, no. 3, pp. 1534–1587, 2024.

[3] E. e Oliveira, M. Rodrigues, J. P. Pereira, A. M. Lopes, I. I. Mestric, and S. Bjelogrlic, "Unlabeled learning algorithms and operations: Overview and future trends in defense sector," *Artif. Intell. Rev.*, vol. 57, no. 3, 66, 2024.

[4] B. Huang, Z. Wang, J. Yang, Z. Han, and C. Liang, "Unlabeled data assistant: Improving mask robustness for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 3109–3123, 2024.

[5] Y. Li, N. Pillar, J. Li, T. Liu, D. Wu, S. Sun, and A. Ozcan, "Virtual histological staining of unlabeled autopsy tissue," *Nat. Commun.*, vol. 15, no. 1, 1684, 2024.

[6] S. Adiga, J. Dolz, and H. Lombaert, "Anatomically-aware uncertainty for semi-supervised image segmentation," *Med. Image Anal.*, vol. 91, 103011, 2024.

[7] H. Zhao, Y. Lou, Q. Xu, Z. Feng, Y. Wu, T. Huang, and Z. Li, "Optimization strategies for self-supervised learning in the use of unlabeled data," *J. Theory Pract. Eng. Sci.*, vol. 4, no. 5, pp. 30–39, 2024.

[8] Y. Deng, P. Lu, F. Yin, Z. Hu, S. Shen, Q. Gu, and W. Wang, "Enhancing large vision language models with self-training on image comprehension," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 131369–131397, 2024.

[9] Y. Sun, X. Li, T. Lin, and J. Zhang, "Learn how to query from unlabeled data streams in federated learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 19, pp. 20752–20760, 2025.

[10] H. Guo and W. Liu, "S3L: Spectrum transformer for self-supervised learning in hyperspectral image classification," *Remote Sens.*, vol. 16, no. 6, p. 970, 2024.

[11] T. Zhang, Y. Li, X. Lv, S. Jiang, S. Jiang, Z. Sun, and Y. Li, "Ultra-sensitive and unlabeled SERS nanosheets for specific identification of glucose in body fluids," *Adv. Funct. Mater.*, vol. 34, no. 17, p. 2315668, 2024.

[12] X. Bao, J. Qin, S. Sun, X. Wang, and Y. Zheng, "Relevant intrinsic feature enhancement network for few-shot semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 2, pp. 765–773, 2024.

[13] X. Gu, Q. Wu, Q. Fan, and P. Fan, "Mobility-aware federated self-supervised learning in vehicular network," *Urban Lifeline*, vol. 2, no. 1, 10, 2024.

[14] J. F. Yang, N. Zhang, Y. L. He, Q. X. Zhu, and Y. Xu, "Novel dual-network autoencoder based adversarial domain adaptation with Wasserstein divergence for fault diagnosis of unlabeled data," *Expert Syst. Appl.*, vol. 238, 122393, 2024.

[15] X. Du, C. Xiao, and S. Li, "Haloscope: Harnessing unlabeled LLM generations for hallucination detection," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 102948–102972, 2024.

[16] B. Felfeliyan, N. D. Forkert, A. Hareendranathan, D. Cornel, Y. Zhou, G. Kuntze, and J. L. Ronsky, "Self-supervised-RCNN for medical image segmentation with limited data annotation," *Comput. Med. Imaging Graph.*, vol. 109, 102297, 2023.

[17] X. Zhang and L. Han, "A generic self-supervised learning (SSL) framework for representation learning from spectral–spatial features of unlabeled remote sensing imagery," *Remote Sens.*, vol. 15, no. 21, 5238, 2023.

[18] S. Tayebi Arasteh, L. Misera, J. N. Kather, D. Truhn, and S. Nebelung, "Enhancing diagnostic deep learning via self-supervised pre-training on large-scale, unlabeled non-medical images," *Eur. Radiol. Exp.*, vol. 8, no. 1, 10, 2024.

[19] J. Shi, Y. Wu, D. Zeng, J. Tao, J. Hu, and Y. Shi, "Self-supervised on-device federated learning from unlabeled streams," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 42, no. 12, pp. 4871–4882, 2023.

[20] C. Zhang, Z. Xie, B. Yu, C. Wen, and Y. Xie, "FGSS: Federated global self-supervised framework for large-scale unlabeled data," *Appl. Soft Comput.*, vol. 143, 110453, 2023.

[21] M. A. F. Abdollah, R. Scoccia, and M. Aprile, "Transformer encoder based self-supervised learning for HVAC fault detection with unlabeled data," *Build. Environ.*, vol. 258, 111568, 2024.

[22] D. Kong, L. Zhao, X. Huang, W. Huang, J. Ding, Y. Yao, and G. Yang, "Self-supervised knowledge mining from unlabeled data for bearing fault diagnosis under limited annotations," *Measurement*, vol. 220, 113387, 2023.

[23] Z. Zuo, H. Zhang, Z. Li, L. Ma, S. Liang, T. Liu, and M. Mercangöz, "A self-supervised leak detection method for natural gas gathering pipelines considering unlabeled multi-class non-leak data," *Comput. Ind.*, vol. 159, 104102, 2024.

[24] https://www.kaggle.com/datasets/jessicali9530/stl10.