International Journal of Multimedia Information Retrieval
https://doi.org/10.1007/s13735-022-00245-6

TRENDS AND SURVEYS

Contrastive self-supervised learning: review, progress, challenges and future research directions

Pranjal Kumar[1] · **Piyush Rawat**[2] · **Siddhartha Chauhan**[1]

Received: 30 May 2022 / Revised: 7 July 2022 / Accepted: 14 July 2022

Abstract
In the last decade, deep supervised learning has had tremendous success. However, its flaws, such as its dependency on manual and costly annotations on large datasets and being exposed to attacks, have prompted researchers to look for alternative models. Incorporating contrastive learning (CL) for self-supervised learning (SSL) has turned out as an effective alternative. In this paper, a comprehensive review of CL methodology in terms of its approaches, encoding techniques and loss functions is provided. It discusses the applications of CL in various domains like Natural Language Processing (NLP), Computer Vision, speech and text recognition and prediction. The paper presents an overview and background about SSL for understanding the introductory ideas and concepts. A comparative study for all the works that use CL methods for various downstream tasks in each domain is performed. Finally, it discusses the limitations of current methods, as well as the need for additional techniques and future directions in order to make meaningful progress in this area.

Keywords Contrastive learning · Self-supervised learning · Unsupervised learning · Data augmentation · Survey

1 Introduction

Deep learning has advanced to the point where it is now an essential part of nearly all intelligent systems. Using the abundance of data available today, deep neural networks (DNNs) have become a compelling approach for a wide range of computer vision (CV) tasks including object detection, image classification [1–3], image segmentation [4,5], activity recognition, etc., and natural language processing (NLP) tasks such as sentiment analysis [6], pre-trained language models [7–10], question answering [11–14], etc. It is possible, however, that the labour-intensive process of manually annotating millions of data samples has exhausted the supervised approach to learning features. Most modern computer vision systems (that are supervised) attempt to learn some form of image representation in order to discover a pattern between data points and their respective annotations in large datasets. It has been suggested that providing visual explanations for decisions made by models can help to make them more transparent and understandable. [15].

On the other hand, supervised learning has hit a snag. It is prone to generalization errors, spurious correlations, and adversarial attacks because of its reliance on time-consuming and expensive manual labelling. We expect the neural networks to learn more quickly with fewer labels, samples, and trials. This paradigm has been adopted by many current models because it is data efficient and generalizable as an alternative that has received significant attention in the research community. In traditional supervised learning methods, the amount of available annotated training data is extremely important. A dearth of annotations has forced researchers to develop new methods for making use of the vast amount of data already available. With the help of self-supervised methods, deep learning progresses without expensive annotations and learns feature representation where data serve as supervision. Autoencoders and extensions, Deep Infomax, and Contrastive Coding, among other self-supervised learning models, will be thoroughly exam-

✉ Pranjal Kumar
pranjal@nith.ac.in

Piyush Rawat
psh.rawat@gmail.com

Siddhartha Chauhan
sid@nith.ac.in

[1] NIT Hamirpur, Hamirpur, Himachal Pradesh 177005, India

[2] Department of Systemics, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand 248007, India

Published online: 05 August 2022

ined in this review. These models will also be examined in terms of their theoretical soundness.

Learning a good representation explicitly, as opposed to learning a good representation implicitly, can be challenging. Firstly, it is not clear what constitutes a good depiction of the subject. Good representations have local smoothness and sparse activation for specific inputs, as well as temporal and spatial coherence, a variety of hierarchically organized explanatory factors that are shared across tasks, and simple dependencies among factors. According to [17], a good representation has these characteristics and more. In contrast to the generative approach, contrastive learning (CL) is a discriminative method that seeks to group samples that are alike together while keeping samples that are different from one another apart as is depicted in Fig. 1. When it comes to real-world applications, representation learning methods are proven to accurately represent the data [18]. In natural language processing and computer vision, contrastive learning has emerged as a popular method for training large pre-trained models. In spite of the fact that the underlying logic has not changed, the development process has involved numerous subfields and application domains, making it more difficult to grasp the current state of the art.

Recent interest in Contrastive Learning and its applications has grown, but without a proper framework for analysis, new methods' innovations and trade-offs are difficult to comprehend. It is clear that this sub-field has made significant progress, which can be categorized and explained in simple terms, ranging from the supervised to the self-supervised methodologies in a variety of application and input domains such as images, text and videos. No other paper has examined the recent evolution of contrastive approaches in as many fields as this one, to the best of our knowledge. Finally, here are the summaries of our contributions:

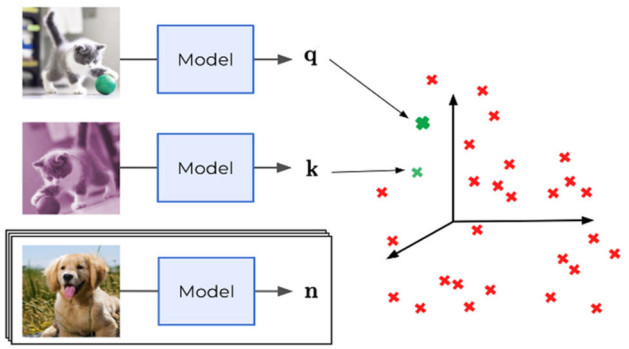

Fig. 1 The underlying principle of contrastive self-supervised learning: Utilising two transformed forms of the same entity, first there is a creation of a positive pair. In the next step, train a model that solves for the proxy task of bringing representations of the positive pair closer to each other, and further than the representations of any other entity from a set of negatives [16]

- An overview and background of self-supervised learning is presented in order to make the frontier ideas more understandable.
- Design of Contrastive learning in terms of its approaches, encoding schemes and loss functions are discussed in detail for better understanding about how the methodology functions.
- A comparative review of all the most recent works that use the contrastive learning paradigm for numerous downstream tasks across various domains is conducted.
- Some open issues in this field have been analysed, and the future direction for contrastive self-supervised representation learning is discussed.

1.1 Organisation of paper

The survey is set up in the following manner: Sects. 2 and 3 provide an overview of self-supervised learning and generative self-supervised learning, respectively, and how it relates to the development of representations. In Sect. 4, starting with a formal overview description of contrastive learning, we then present architectural details based on various approaches, encoding methodologies, and loss functions. Section 5 then covers a wide range of data domains and problem areas, including applications in vision, language, graph-structured data, audio, multimodal data, and others. To wrap up the paper, we discuss about a few hot button issues and how they might play out in the future in our concluding section.

2 Self-supervised learning

Self-supervised learning has led to significant advances in natural language processing [7,19–21], speech processing [22–24], and computer vision [25–29] because it builds representations of data without human annotated labels. There are three broad categories of mainstream self-supervised learning as shown in Fig. 2, and numerous subcategories within each of these. Machine translation [30] and speech recognition have all been made possible by self-supervised representations [31]. Algorithms that learn from their own mistakes have been developed by focusing on specific modalities. It is not possible to define a self-supervised learning task for speech units, like words in NLP, and so several prominent models have mechanisms for learning an inventory of speech units [24,32]. There is a similar issue in computer vision research in which researchers must choose between learning discrete visual tokens [28,33], regressing on the input [29], or learning representations that are invariant to data augmentation [27,34].

Learning biases can be helpful, but it is not always clear if they can be applied to other forms of learning. Further-

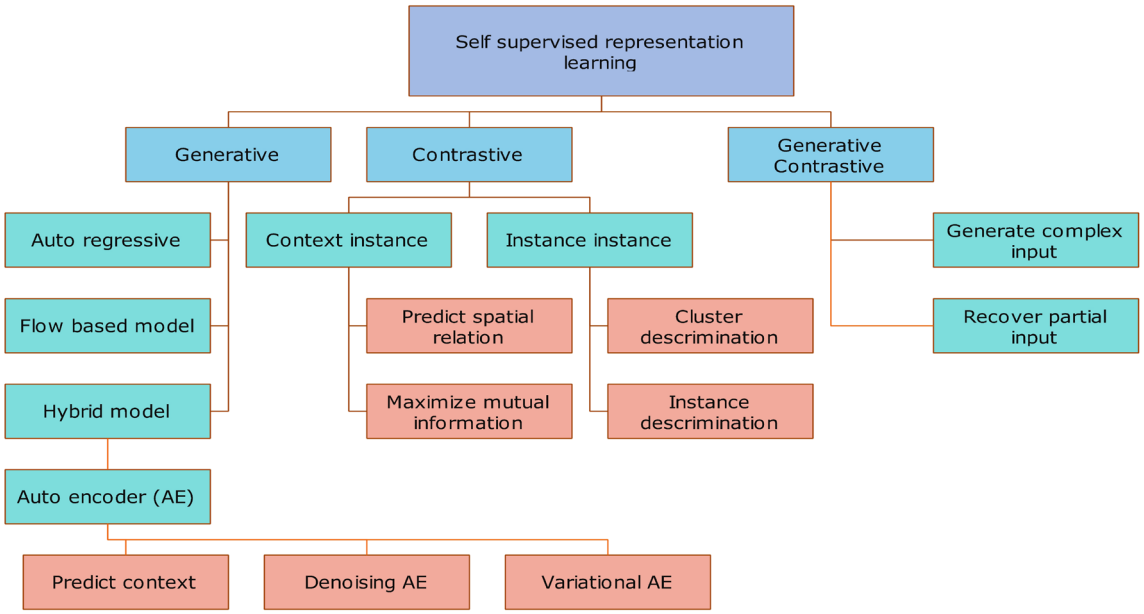

Fig. 2 Self-supervision learning classification into three general categories and various other subcategories

more, leading theories on the biology of learning suggest that humans use similar learning processes to comprehend the visual world as they do to comprehend language [35,36]. An interesting comparison has been made between modality-specific and general neural network architectures [37].

There are two types of SSL tasks that can be broken down into pretext and downstream tasks. Supervised learning is used for the former, while unsupervised learning is used for the latter, which generates labels based on the data it receives as input. When the model has finished learning the pretext task's representations, it can then apply and fine-tune those representations to the downstream task [38]. Figure 3 depicts the differences between the Transfer Learning (TL) and the Self-supervised Learning (SSL) work flows. The training process begins with a source task and ends with refinement of the downstream task, just like SSL. Pre-trained weights used to initialize the weights of the model in the target task, if the architecture of the source and target tasks are similar. To put it simply, TL uses pre-labelled data to train, whereas SSL relies on unlabelled data to learn features.

2.1 Self-supervised learning in vision

Self-supervised learning can take advantage of the variety of modalities (such as image, sound, and text) that videos offer over still images [39,40]. Self-supervised learning using 2D CNNs is a common early approach. For example, Lee et al. [41] and Misra et al. [42] use frame order as the supervision signal. It is proposed in Fernando et al. [43] to identify the odd or unrelated clips from a set of related clips using an odd one out network. The models must learn spatio-temporal features that can distinguish between similar clips in order to find the odd one. Proxy tasks may also include determining whether the video is moving forward or backward. According to Wei et al. [44], the arrow of time can be used to monitor activity. By extracting pixel-level geometry information, Gan et al. [45] used these data as auxiliary supervision. In order to learn the complex spatial and temporal representation, several methods for 3D CNN self-supervised learning have been proposed recently [46,47]. According to Kim et al. [48], 3D CNNs can be trained using a video representation learning method that is based on solving 3D video cubic jigsaw puzzles. There has been a lot of success recently with self-supervised learning combined with unsupervised contrastive learning [49–51]. Despite this, all methods of contrastive learning rely on instance discrimination. For each instance, a pair of augmented samples from that instance are used as predictions, and the augmented samples from other instances are repelled away. These methods, on the other hand, are inapplicable to untrimmed videos in which the instances are complex and composite.

2.2 Self-supervised learning in NLP

Methods for natural language processing (NLP) that use masked language modelling to improve the performance of various application tasks are self-supervised. Recurrence, adversarial or property masking, and contrastive learning objectives are frequently added to these pre-training methods [7,10,19–21,52,53]. Learning to compare two augmented images as similar or dissimilar using contrastive self-supervised training objectives has led to recent successes

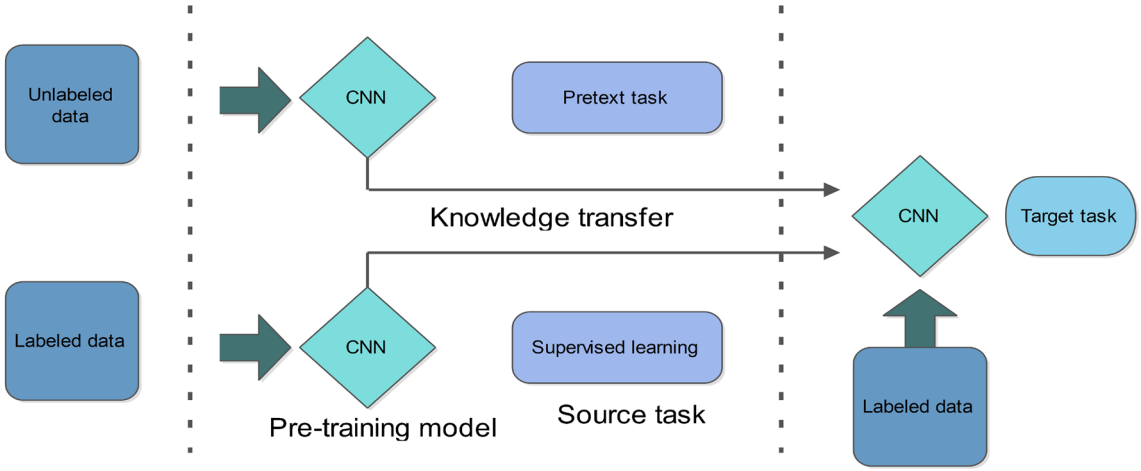

Fig. 3 Knowledge transfer for downstream tasks by training an encoder. In order to learn a feature, SSL uses unlabelled data, which is then further trained on the target task

in image representation pretraining. The automated creation of text input augmentations in NLP is still very difficult because a single token can invert the meaning of a sentence. Among the most popular models is BERT [7], which is used to solve a masked prediction task in which some of the input tokens are blanked out. For many languages, word boundaries are easy to identify, and most methods therefore predict word or sub-word units for pre-training purposes. As part of the pre-training and fine-tuning process, smaller BERT-style models are being developed [54].

2.3 Self-supervised learning in speech

Recently, the quality of speech recognition models [23,24,32,55–59] trained using self-supervised learning has improved dramatically. As a result of these methods, models can learn from unsupervised data and combine it with supervised learning to improve recognition accuracy. For languages and domains with a limited amount of supervised data, the ability to learn from unsupervised data is particularly beneficial. Learning representations are a common design principle in self-supervised learning for speech recognition. BERT-inspired algorithms are becoming increasingly popular in the speech community as a result of its success [7]. BERT-style self-supervised learning for speech faces the challenge of bridging the gap between continuous speech signals and discrete text tokens, and a solution for addressing this issue is to learn speech representation [23,53] or learn quantized representation [24,32,58]. It was found that the quantized results of speech representations learned using effective algorithms had a strong correlation with the phoneme of the utterances.

Combining self-supervised learning with representation learning has two drawbacks that can slow down research in the speech field: (1) The architecture of the model is restricted. Self-supervised learning often necessitates that a model serve as both a speech representation and a tool for the integration of representation learning and self-supervised learning; (2) a rise in the level of difficulty. Algorithms to design both algorithms and find a balance between them can stifle research. Additionally, this complexity can lead to the development of more complex algorithms rather than a simple and effective solution [60].

3 Generative self-supervised learning

Unsupervised pretext tasks are becoming increasingly popular as self-supervised learning methods that are becoming more and more effective. The pretext task is designed to generate its own supervisory signal rather than relying on human annotation. Here, we update the research presented in [61] with more recent findings.

3.1 Auto-regressive model

In terms of "Bayes net structure," auto-regressive (AR) models can be considered (directed graph model). Conditionals can be factored into a joint distribution if one variable's probability depends on the probability of the other.

$$\max_{\theta} p_{\theta}(\mathbf{x}) = \sum_{t=1}^{T} \log p_{\theta}\left(x_t \mid \mathbf{x}_{1:t-1}\right) \tag{1}$$

For many NLP tasks, large auto-regressive language models have attracted a lot of interest because of their zero-shot and few-shot capabilities, which allow them to be applied without task-specific fine-tuning or annotation information. [21,62]. Even though large language models like GPT-3 [21] have

been trained on multilingual corpora, there may still be significant differences in NLP performance between languages. The lack of human-curated benchmarks makes it difficult to evaluate zero-shot performance in non-English languages. With the exception of recent work in machine translation [63], multilingual models tend to perform worse than mono- or bilingual language models [64].

An auto-regressive model treats image pixels as a sequence and generates pixels one by one by modelling their conditional distributions. This is unlike VAE or GANs in image generation. As deep learning has progressed, researchers have looked at using deep auto-regressive models to produce image pixels in a sequential manner. PixelRNN [65] and PixelCNN [66] are examples of auto-regressive models being used in computer vision. The general idea is that images can be modelled pixel-by-pixel using auto-regressive methods. Looking at the lower (right) pixels as an example, we can see that they are generated by applying a condition to the upper (left). PixelRNN and PixelCNN use RNN and CNN to model their pixel distributions, respectively. 2D images can only use these models to factor in probabilities in specific directions, such as right and down. The emergence of flow models [67–69] and auto-regressive models [70] has ramped up recent progress in techniques for creating realistic content. For the generation of realistic graphs with deep auto-regressive models, You et al. [71] propose GraphRNN. Nodes and edges are generated sequentially in response to the graph that has already been created.

3.2 Flow-based model

An invertible transformation $f : X \rightarrow \mathrm{Z}$ to auxiliary base distribution $p_Z(z)$ in the latent space is modelled by flow-based models, which are a flexible type of deep generative model. To train flow models, the transformations f and f^{-1} must be differentiable, and z must have the same dimension as the input x in order for the function to be bijective. Flow models as a result, they are called diffeomorphisms [72]. The probability of a single input x can be expressed as follows using the theorem of variable change:

$$p_X(x) = p_Z(f(x)) \left| \det \frac{\partial f(x)}{\partial x} \right| \tag{2}$$

Determinants are represented by det operations, and the Jacobian matrices are partial derivatives of these det operations. It follows that as z neighbourhood shrinks, so does the density at x, and vice versa. This is because the total probability mass must be preserved. Neural networks are used to create the transformation f, and the latent space distribution $p_Z(z)$ is simple, such as a Gaussian.

3.3 Auto-encoder model

To rebuild data feature representations, a basic methodology relies on an assumption that enough information is available. Auto-encoder architecture has been extensively studied in the literature for learning representations in an unsupervised manner [73–75]. In addition to the variations auto-encoder, which explicitly incorporates a probabilistic assumption about the distribution of features extracted from data, there are numerous other encoders. By reconstructing the original inputs from noise-corrupted inputs, the denoising auto-encoder [75] begins learning more robust representations. Instead of using Auto-Encoding Data (AED), a new method of unsupervised representation learning based on Auto-Encoding Transformation (AET) is employed in this technique. Using encoded features at the output end, AET tries to predict the random transformation as accurately as possible. A decoder uses the encoder's extracted features to reconstruct the input data, while the encoder acts as an extractor of features that typically represent the most critical information about the input data. In theory, a feature representation should be able to accurately represent the input data. Networks $h = f_{enc}(x)$ and $x' = f_{dec}(h)$ make up the AE, which consists of encoder and decoder networks. The goal of AE is to get x and x' as close to each other as possible (such as through mean-square error). The linear autoencoder can be shown to be similar to the PCA method. When there are more hidden units than input units, it is possible to discover interesting structures by imposing sparsity constraints on the hidden units. Reconstructing original inputs from noise-corrupted inputs is the goal of the denoising auto-encoder [75]. Contrastive Auto-encoder [76] penalises abrupt changes in representation around given data, encouraging representation invariance to small perturbation on input data. Cross-channel features are concatenated as data representation in the auto-encoder proposed by Zhang et al. [77] by reconstructing a subset of data channels from another subset. A transforming auto-encoder proposed by Hinton et al. [78] is still trained in the AED fashion by minimising the difference between the reconstructed and target images.

4 Contrastive learning

Learning by comparing is a good description of contrastive representation learning. It is through comparison of input samples that contrastive learning, unlike discriminative modelling, is able to learn a mapping to some (pseudo-)labels and to reconstruct the input samples. Contrastive learning, on the other hand, learns by comparing multiple samples of the same signal. Negative pairs of inputs can be compared to positive pairs of inputs that are "similar" to each other. In order to learn representations that are invariant across differ-

ent perspectives on the same instance, contrastive learning methods are employed. Make positive pairs closer and separate negative ones to learn a general feature function that maps a raw entity into hypersphere space-based features as shown in Fig. 4. Convolutional neural networks with strong abstraction capabilities and heavy augmentations can help unsupervised contrastive models learn some semantic structures [2,80,81].

Softmax loss functions are used to distinguish between positive and negative samples in contrastive learning methods, which rely on the similarity of features to temperature. This is achieved by comparing the temperature of each feature. Unsupervised contrastive learning relies heavily on the contrastive loss. Nonparametric classification loss and its variants, such as InfoNCE and NT-Xent, are frequently employed, when dealing with a contrastive loss function such as the following:

$$\mathcal{L}_i = -\log \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_i'/\tau\right)}{\sum_{j=0}^{K} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_j'\right)/\tau\right)} \tag{3}$$

$\mathbf{z}_i^\top \mathbf{z}_i'$ is the product of the two vectors, and τ denotes a temperature hyperparameter impacting the product's sensitivity in terms of penalising the hard negative samples. Constantly penalising the hardest negative samples, the contrastive loss is a hardness-aware loss function [79]. Discriminative models for representation have recently made significant strides in contrastive learning with breakthroughs like Deep InfoMax, MoCo, and SimCLR. Figure 5 illustrates that the SimCLR architecture includes a base encoder that is based on a ResNet architecture and generates an embedding representation from augmented images. A Noise Contrastive Estimation (NCE) [82] objective is used to reflect how to learn and then compare as reflected in:

$$\mathcal{L} = \mathbb{E}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(x^-)}}\right)\right] \tag{4}$$

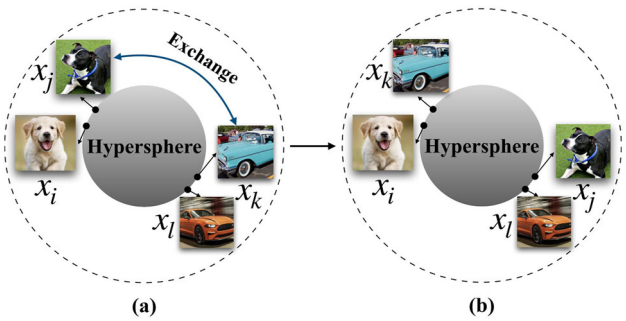

Fig. 4 Contrastive loss will not be affected if x_j and x_k are swapped, as well as their corresponding augmentations. It is important to note that the embedding distribution of **a** captures the semantic relations between instances, which makes it more useful for downstream tasks [79]

x^+ is similar to x, x^- is different from x, and "f" is an encoder (representation function). Although the encoder and similarity measure may be different for each task, the framework remains the same. For self-supervised learning methods, instead of using the pretext task to derive an arbitrary pseudo-label, contrastive learning methods use a set of multiple input pairs to build a discriminative model. To get around the problem of only having a finite number of label pairs available when using supervised learning, tasks like this one can be defined directly from the data. Instead of modifying the model architecture between training and fine-tuning, contrastive methods do not require any model architecture modification during learning (such as in [77]).

Adversarial methods preserve the encoder-decoder generator structure in contrastive learning. The contrastive, on the other hand, does away with the decoder altogether. Contrary to contrastive methods, the generator makes adversarial learning much more difficult than it would be without it, leading to unstable convergence when used in adversarial learning. First, the generator gives adversarial learning a unique representation. Due to the existence of the decoder in an adversarial setting, the representation is required to be "reconstructive," meaning it contains all the information required to construct the inputs. Only "distinguishable" information is needed in contrastive settings to distinguish between samples.

4.1 Approaches for contrastive learning

It is through the use of similarity distributions that contrastive learning is able to formalize the mapping from various viewpoints on a scene or context to the same region of a representation spatial [83]. Many different ideas of similarity and dissimilarity can be generated depending on the final objectives, which is why contrastive methods are so effective [84]. As depicted in Fig. 6, Contrastive Representation Learning (CRL) framework is utilised to learn the visual representations in the context of an instance discrimination task based on Chen et al. [25] work. However, there are a few general principles that underlie the construction of similarity and dissimilarity, which we will now examine.

4.1.1 Context based

A sample representation's context–instance relationship can be used to find other representations of the same scene. If we want to learn about a scene's context, we need a representation that includes all relevant information. There are usually smaller parts of a scene's context that contain a specific subset of the scene's information. Context-instance and context-context contrast are the two types of contrastive learning that aim to group similar samples closer and diverse samples farther away from each other in the embedding space

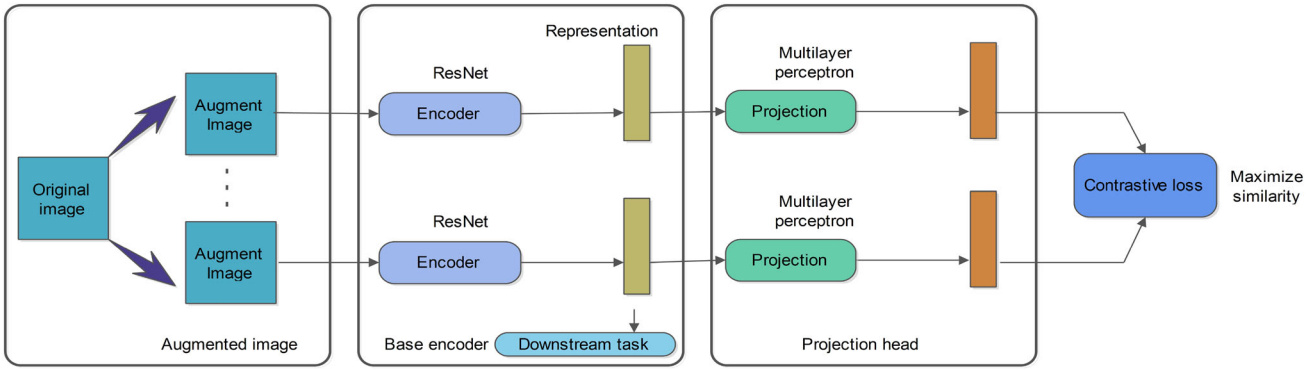

Fig. 5 Architectural representation for SimCLR. Two augmented images are created by applying data augmentation techniques to the original input image. An encoder network and projection head are trained to maximize the similarity between the augmented images by using contrastive loss. The results are used for downstream tasks after the training process has been completed

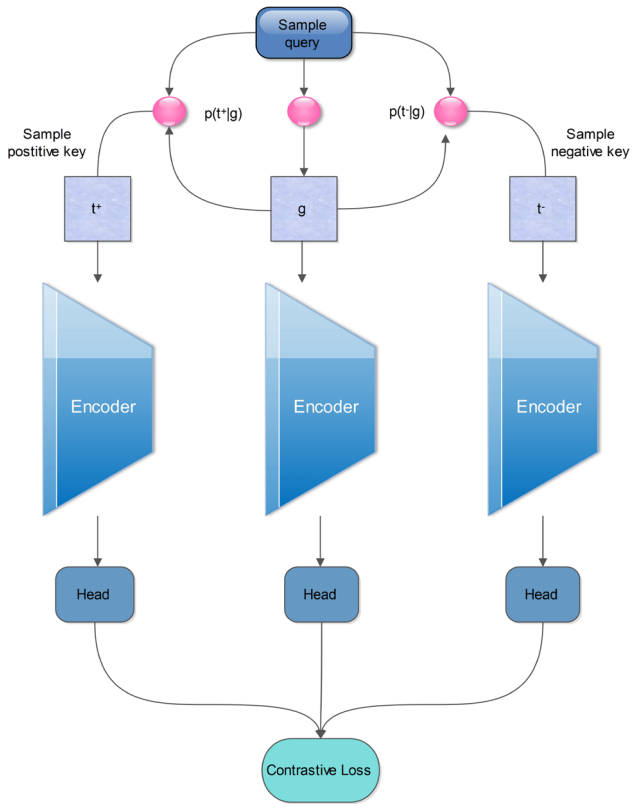

Fig. 6 In contrastive learning, there are the encoders and corresponding transform heads for individual data modality, along with a loss function for evaluating positive and negative data samples[83]

[82,85]. The local feature of a sample and its global context representation are modelled similarly in the context-instance contrast [86–88]. In the embedding space, we expect that a sentence's representation will be linked to that of the paragraph to which it belongs for natural language processing. With Deep InfoMax, the mutual information between a local patch and its global context can be explicitly modelled by distinguishing the negative image sample [87]. There is also Deep Graph InfoMax in graph learning, which uses the representation of each node as the local feature and the average of the node's representation to determine its context. Velickovic et al. [86] propose this method. In a similar vein, Hassani and Ahmadi [89] present a method for learning the graph's contrastive multi-view representation.

Tschannen et al. [90] argue that studying the relationships between global features of different samples directly can achieve rather good performance on representations [91,92], despite previous context-instance contrast having made significant progress. Contrastive Multiview Coding, proposed by Tian et al. [93], makes use of multiple views of an image as positive samples and a random one as a negative sample. For the first time in a fully supervised setting, Khosla et al. [94] extend the contrastive learning paradigm by allowing the model to use label information to group points belonging to the same class.

4.1.2 Consistency and coherence

In contrastive learning, it is also possible to use spatial or temporal coherence and consistency in an observation sequence to define similarity. This method uses semantically positive and negative image pairs as a basis for building discriminative representations through a process known as contrastive self-supervised learning (CSL). Semantic augmentations of the same image are used to create invariant representations, and representations of different images are used to create representations that are dissimilar. Because of this tendency to generate inconsistent local representations for the same spatial regions after image transformation, contrastive methods that use global representations on an image-by-image basis tend to be used for image classification. Previous global contrastive methods, for example, can produce a similar global representation even if the local feature of that object ends

up losing consistency [95], since they use global pooling to attend to other discriminative areas instead when an object in an image is geometrically shifted or scaled. As a result, the performance of localization tasks relying on spatial representations may be negatively affected. Previous contrastive approaches often use heavily cropped views from an image in order to make a positive pair, so the representations between semantically different regions are induced to match [96].

4.1.3 Data augmentation

There are two primary steps in contrastive learning: generating positive and negative samples for a given anchor data point, and comparing these samples. However, given the discrete nature of the input space, making transformations on the anchor to generate positive samples is a more difficult task in the text domain. The transformation or augmentation methods that are most commonly used in the discrete input space, in the latent representation space, etc., are discussed in this section.

- *Input-space transformations*: Text transformations performed in the discrete input space, also referred to as instance-based transformation, are the most straightforward to implement. Many different approaches have been tested in literature, with varying degrees of success, despite the fact that these methods are less intuitive than similar image transformations (such as cropping, flipping, or rotating). Giorgi et al. [97] used a span sampling approach in their DeCLUTR technique, and considered segments that are next to the original text segment, overlapped with it or subsumed it, as positive samples.
- *Latent-space transformations*: Contrastive representation learning can benefit from the same data augmentation techniques proposed for low-resource learning environments. Back-translation using another intermediate language [98,99] and language models to replace selected words from the text with nearest neighbour words [54] such as word2vec [100] or GloVe are two methods that generally preserve the semantic meaning of the original text.
- *Transformations via architecture and combined methods*: Using slightly different architectures or modifying some aspect of the architecture in a certain way, positive pairs for text can also be generated. Dropout noise is one example of an architecture-based method for text augmentation in contrastive learning. Gao et al. [101] feed the sample input twice to the encoder, they get two embeddings with different dropout masks, creating the positive pairs. Using adversarial training, a perturbed version of the input can be created and tagged as a positive example. In addition to lexical transformation and dropout approaches, Yan et al. [102] used Fast Gradient Value (FSV) [103] as an adversarial attack method to perturb the input data.

4.1.4 Multisensory signals

Multisensory signals (such as vision, sound, and touch) provide humans with a wealth of data with which to make sense of their surroundings. As a result of these cues from various modalities, we are able to perform complex tasks in our daily lives. Cross-modal representations of visual and textual data have been learned through the use of contrastive methods [104,105]. Considering vision and audio for instance, it is not hard to picture lightning storms, identify and converse with friends in a packed cocktail party, or associate multiple objects with their sources when you hear noise.

To address the problem of limited data, semi-supervised [106,107], weakly supervised [108], and self-supervised learning frameworks [109–111] are proposed. Existing self-supervised methods rely on a predefined number of clusters [109–111] or require videos of single sound sources [110], while weakly supervised methods require audio-visual event labels. In addition, semi-supervised methods [106,107] using audio visual correspondences alone as the supervision are less effective because a scene may contain non-sounding or ambient regions, which leads to the association between the incorrectly sounding regions and reference audio signals. If the number of sound sources is unknown and there are objects that cannot be seen during training, the performance of sound localization in unconstrained scenarios is negatively affected.

4.1.5 Clustering

There are many ways to learn discriminative representations, but contrastive learning aims to maximise the contrastive loss by encouraging closeness between representations from the same set of images. A distance measure in the embedding space can be used to derive high-level semantics for groups of instances. There is a correlation between the distance between clusters and the similarity of the categories they represent. Many contrastive learning-based deep image clustering methods have recently been proposed. Contrastive learning is combined with clustering in the first group of methods. There are many ways to combine contrastive loss and cluster loss to learn representations as well as assigning clusters. Two independent subspaces have been created for instance-contrastive and cluster-contrastive representations. Zhan et al. [112] developed an online deep clustering framework that used contrastive learning to initialise centroids and sample labels and then performed an update at the same time. Using instance discrimination and feature decorrelation, Tao et al. [113] learned representations that were clustering-friendly. Tsai et al. [114] optimised contrastive loss and the mixture of experts formulation to learn

the semantic representations. Using contrastive learning, the cluster centroids are iteratively optimised by the loss of clustering.

4.2 Encoding methodologies

In contrastive representation learning, two objectives must be met: Effective and efficient embedding that permits distance measurements between samples and mapping to a powerful representation of input data. To train representations unsupervised, contrastive learning has proven to be an extremely successful method. It teaches an encoder to distinguish between positive and negative samples by using query anchors as training data for the algorithm. Positive and negative samples are critical in defining the goal of learning a discriminative encoder, preventing it from learning trivial features. The goal of an encoder is to learn a good representation space mapping from inputs. The base encoder must be able to distinguish between learning a good representation and learning an embedding that is effective and efficient in computing similarity metrics. It is necessary to discard potentially useful information in order to maximize similarity between positive samples when the primary task of learning a representation and the pretext task of learning a similarity measure become intertwined in order to maximize the similarity between positive samples.

There are two main parts to the typical contrastive learning pipeline: pre-training encoders and building downstream classification. Pre-trained encoders are typically generated with a large number of unlabelled data and a large number of training costs as the most important component (e.g., GPU resources). In addition, gathering essential training data is an expensive process for specialised tasks like medical diagnosis. Due to the wide variety of downstream classification tasks that can be performed with just a small number of labelled data and a well-trained encoder, this skill is in high demand. Transform heads above encoders are a relatively new development. Before the recent advances in transform heads, many methods compared which layers were best suited for transfer learning after a standard encoder was trained [83]. Although specific encoders and the interactions among them allow for different downstream transform heads and contrastive losses, contrastive learning is not limited to any particular encoder.

Contrary to popular belief, adversarial contrastive learning [115] shows that negative samples can be directly learned by treating them as a part of network weights, resulting in hard negatives being pushed toward query anchors. A more principled approach to learning than constructing discriminative negatives heuristically is provided. The AdCo outperforms the SimCLR [25] and MoCo [91] in terms of accuracy and efficiency by treating the negatives and the encoder as two adversarial players. There are numerous variations on the contrastive learning strategy based on the various ways negatives can be formed. Both the SimCLR [25] (negatives come from the other samples in that batch) and MoCo [91,116] (negatives from previous batches are stored in memory) are examples of these. The former relies on a large number of mini batches to train the encoder, which requires a large amount of computing and memory resources. Instead of having to re-compute past samples for each mini-batch, MoCo utilises the representations that were previously obtained.

Positive samples are more easily distinguished from their negative counterparts, so the harder negative samples appear to be more important for contrastive learning. A trivial feature that could be used to shorten the aforementioned instance discrimination task is prevented from being learned by the encoder. There are methods that use heuristic sampling to select samples that are near positive anchors. Using a concentration parameter, a sampling method known as a "hard negative sampling method" [117,118] finds harder negatives.

4.3 Contrastive loss

There should be some tolerance for samples that are semantically similar when designing a contrastive loss [79]. A softmax function of feature similarities with a temperature(τ) is used as the loss function in contrastive learning methods in order to help differentiate between positive and negative samples. Contrastive loss is essential to the success of unsupervised contrastive learning. As a result, both explicit and implicit properties of the hard contrastive loss take into account the hardness of negative samples when selecting the "K" closest negative samples. The positive features can be aligned, and the embeddings can match a uniform distribution in a hypersphere by separating the contrastive loss into two parts. The distribution of negative samples is not constrained by the contrastive loss. Deep convolutional neural networks have strong abstraction and augmentation capabilities, which allow some semantics to be discerned in the negative distribution.

The latent embedding distance between positive and negative pairs is minimized by contrastive loss functions, while the distance between the two is maximized. A variety of functions, including NCE loss [82], InfoNCE loss [22], and NT-Xent loss [119], have been used for contrastive learning. The NCE method is typically used to learn datasets in these systems. This technique aids the model in bringing together similar images and separating those that are not. NCE uses nonlinear logistic regression for this [22], which aids the model in distinguishing real data from noise. If you want to find positive pairs, you can use SimCLR [25] to normalize your loss (NT-Xent). For each unlabelled sample $(x_1, x_2, \ldots, x_N)$, to get the embedding vector extract, i.e., $\left(z_i, z^+\right)$ from the base header $g(.)$, a stochastic augmentation $T(.)$ generates two different views x_i^+ and x_j^+ of the given sample x and feeds them through the encoder $f(.)$.

The following is the mathematical representation:

$$\mathcal{L}_{NCE} = -\log \frac{\exp\left(\text{sim}\left(z_i, z_j\right)/\tau\right))}{\sum_{k=1}^{2N} \mathbb{I}_{[k\neq i]} \exp\left(\text{sim}\left(z_i, z_k\right)/\tau\right)} \tag{5}$$

If $k \neq i$ and τ is the corresponding temperature hyperparameter that help in regulating the penalties on the hard negative samples [79]), then $\mathbb{I}_{[k\neq i]} \in [0, 1]$ denotes an indicator function which is equal to 1, and N is the total number of respective examples where two augmented views (i.e., positive pair denoted by x_1^+ and x_2^+) are getting generated from each of the given example x. Therefore, the total number of augmented pairs is calculated as $2N$, and correspondingly there are $2(N-1)$ negative augmented examples from the required dataset. The function that measures the similarity between z_i and z_j in terms of their embedding representation is sim $\left(z_i, z_j\right)$. The most common function is the cosine function, which is defined as following:

$$\text{sim}\left(z_i, z_j\right) = \frac{z_i \cdot z_j}{\|z_i\| \cdot \left\|z_j\right\|} \tag{6}$$

$\|\cdot\|$ represents the vector's Euclidean norm. The cosine function can be used to calculate the angle between any two nonzero vectors in a d-dimensional space. At zero degrees, the cosine similarity is equal to one. When viewed from different angles, the cosine similarity ranges from one to a negative one. There are many different ways to use InfoNCE [22] in contrastive model training. Because cross-entropy loss estimates the information shared between two images, InfoNCE is commonly used for contrastive loss in these cases. It distinguishes $\left(z_i, z^+\right)$ from its k negative pairs, which are written in the form $\left(z_i, z_1^-\right), \left(z_i, z_2^-\right), \ldots, \left(z_i, z_k^-\right)$ [62]. The formula for InfoNCE is as follows:

$$\mathcal{L}_{\text{infoNCE}} = -\log \frac{\exp\left(\text{sim}\left(z_i, z^+\right)/\tau\right)}{\exp\left(\text{sim}\left(z_i, z^+\right)/\tau\right) + \sum_{j=1}^{k} \exp\left(\text{sim}\left(z_i, z_j^-\right)/\tau\right)} \tag{7}$$

Mutual information estimates have higher lower bounds when the labels are clean and a sufficient number of negative samples are used. In most cases, this leads to a higher level of performance [120]. This lower bound between z_i and z_+ is calculated as follows:

$$I\left(z_i, z^+\right) \geq \log(k+1) - \mathcal{L}_{\text{infoNCE}} \tag{8}$$

An embedding space (embedding space) and loss functions designed to attract/repel similar/dissimilar samples are used in contrastive learning to establish similarities between samples. Only the features of the sample are taken into account when calculating the distance. It is possible to achieve separation by employing loss functions that are based on various combinations and variations of quadratic, exponential, logloss and maximum-margin (hinge) losses.

4.3.1 Maximum margin

A max-margin separation for the distances between an anchor point and a selection of its negative and positive samples was presented in [121] for the triplet loss. A single positive and negative sample is used for each comparison, and the samples are collected in a self-supervised or unsupervised fashion. In [122], a ranked list loss based on weighting positive and negative examples in relation to their distance from the anchor point using class information is used with a max-margin loss (supervised).

4.3.2 Maximum margin with minimum separation

K-nearest neighbour framework is presented in [123] to minimise Mahalanobis distance for positive samples and maximise distance for negative samples with a margin. Contrastive losses are computed using a subset of positive samples [124]. For the distance to positive samples, a quadratic loss is presented in [124,125], while for the distance to negative samples, a quadratic max-margin loss is presented. Images and faces can be recognised using label information [125] and a Siamese network [124] in a self-supervised fashion. There is only one positive and one negative sample to compare in all cases. Alternative to [126] is a modified loss that includes an exponential loss with a margin for the distance to all negative samples, as well as the direct distance to the positive sample. Self-monitored experiments are conducted.

4.3.3 Maximum separation with minimum separation

Neighbour Component Analysis (NCA), proposed by [127] in the context of a supervised learning framework, is designed to maximise the likelihood that a sample will be correctly classified. A connection to linear discriminant analysis (LDA) [128] can be made; however, all class distributions need not have equal covariance in order to apply this technique to discriminant analysis. One positive and one negative samples are used for each comparison. Each negative class is represented by a positive sample and a randomly chosen element in [119]. (multi-class N-pair). The goal is to push or pull the negative and positive points as far apart or as close together as possible to the reference sample. For NT-Xent/SimCLR, the authors in [25] present a linear loss over a distance function between similar points normalised by the sum of distances for other points in a training batch (assumed as negative samples).

New loss functions can be used to reduce training costs. A Self-adversarial Negative Sampling loss has been proposed by Sun et al. [129] to speed up model convergence by using the Softmax normalized triple score as the weight of each negative sample, as the Negative Sampling loss has been shown to be time consuming and unstable. Apart from that, recent studies conducted by the community have proposed a number of non-sampling training strategies in either an entirely negative or completely negative way [130,131].

5 Progress in contrastive learning

There have been a number of significant advancements in contrastive methods over the years, spanning a wide range of fields and domains. Because of its use of a contrastive cosine similarity loss in the latent space, SimCLR's visual representations [25] have been shown for the first time to perform on a par with supervised models, thanks to the two-network training method SimCLR devised. According to BYOL's recent results, the two-network approach is even more effective when negative pairs are removed from training [34]. Models like generative adversarial networks (GANs), where the goal is to use game-theoretic principles and two networks compete against each other in a game to generate more realistic data [61], sound eerily similar to this new model family. Models classified as generative (e.g. autoencoders), contrastive (e.g., SimCLR), or generative-contrastive SSL are introduced in a useful taxonomy (e.g., GANs or adversarial autoencoders). Each category has a distinct objective, ranging from restoration and contrastive losses, to distributional divergences. In the future, we expect generative, adversarial, and contrastive training to have more in common with one another [132].

In robotics, the term "self-supervised learning" is first used, where training data are automatically labelled by leveraging the relationships between different input sensor signals. Machine learning experts then build on the concept. In 1992, authors in [133,134], respectively, first proposed the idea of learning by comparing data points that are related but unsupervised. The authors of [133] formulate the problem of learning invariant representations by maximizing the mutual information among various views of the same scene using the "Siamese Network," which consists of two identical weight-sharing networks in a metric learning setup. Comparing training samples is the first step in applying the general principle of comparison-based teaching. There are three main categories of self-supervised models based on the objective function of deep neural networks: generative, contrastive, and generative contrastive (or adversarial). The Contrastive learning representations are being applied to a wide array of data and problem domains, most of which have been developed in the last few years or so. In this section, we will take a look at a number of areas where contrastive learning has recently made significant progress [83].

5.1 Natural language

In representation learning, contrastive self-supervised learning has become a popular technique. These methods rely on comparing samples that are semantically similar and those that are semantically dissimilar. But in Natural Language, the augmentation methods used in creating similar pairs with regard to contrastive learning assumptions are difficult. This is due to the fact that even a simple change in a word in the input could alter the sentence's semantic meaning, which would contradict the distributional hypothesis. There are two types of examples used in contrastive learning: positive samples and negative samples. The positive samples act as an anchor, and the negative samples are the augmented versions. This is a widely used contrastive learning setup for natural language. There are no well-defined transformation functions for texts, making it more difficult to create semantically similar pairs.

Recent years have seen a rise in the use of Contrastive Learning in the field of Natural Language Processing (NLP). NLP downstream tasks such as language understanding [98], cross-lingual pretraining [135], and textual representations learning [97] have shown significant improvement with the use of this technology. In order to learn how to distinguish machine translation of input sequences, cross-lingual pretraining task based on maximization of mutual information between two input sequences has been proposed by authors in INFOXLM [135]. To the contrary of the TLM [136], this model aims to maximize the exchange of information between machine translation pairs on a multilingual platform, and this helps in a variety of downstream tasks, such as cross-lingual classification and answering questions. There are two primary steps in contrastive learning: generating positive and negative samples for a given anchor data point, and comparing these samples. However, due to the discrete nature of the input space, transforming the anchor to generate such positive samples is a more difficult task in the text domain. Text transformations based on discrete inputs, or "instance-based" transformations, are the most straightforward. Many different approaches have been tested in literature, with varying degrees of success, even though these methods are not as intuitive as similar transformations in images (such as cropping, flipping, or rotating). Giorgi et al. [97] used a span sampling approach in their DeCLUTR technique, and considered segments that are next to the original text segment, overlapped with it or subsumed it, as positive samples as shown in Fig. 7. Additionally, lexical and sentence transformations can be used to augment the samples. In their CLEAR contrastive learning method, Wu et al. [137] used strategies like word deletion, span deletion, token reordering, and syn-

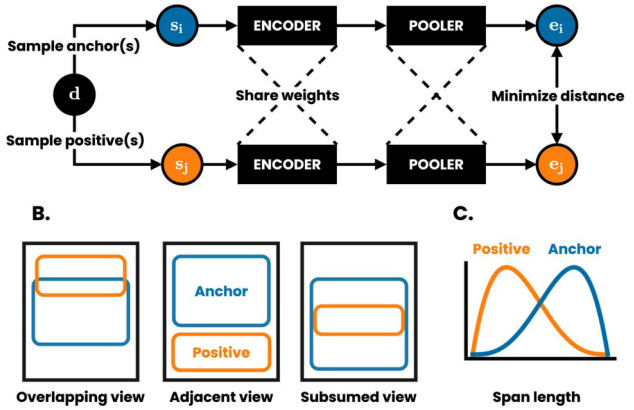

Fig. 7 A contrastive prediction task is used to train the encoder and pooler to reduce the distance between embeddings in DeCLUTR methodology [97]

onym substitution for sentence augmentation. The standard data augmentation methods [138], such as random insertion, synonym replacement, random deletion, and random swap, can also be used to generate the positive pairs. Natural Language Inference datasets can be used as a starting point for creating the desired data. Each of these datasets is made up of an entailment, neutral, or contradiction relationship between the premise and hypothesis [139]. If the relationship is one of entailment, the hypothesis would be labelled as positive; otherwise, the hypothesis would be labelled as negative.

The quality of the final representation depends on the quality of the negative data points sampled. Using an efficient sampling function for these negative examples can also speed up the learning process by correcting the model's mistake more quickly. Samples that are close to the anchor and have a high likelihood of being labelled with the same name can significantly improve the representations. Hard negative samples are what they are called. Hard negatives can be easily identified when latent classes are known (i.e., in the supervised case). However, mining the hard negatives in an unsupervised setting is more difficult. To ensure that a wide range of negative samples [25,91] are included in a single loss function, researchers typically increase the number of samples in the batch. But Arora et al. [140] find that, due to the inherent nature of contrastive learning, larger numbers of negative samples in some cases may even decrease performance on the subsequent task.

As a contrastive method, sentence prediction is widely used. NSP and SSP take inspiration from the skip-gram model [141], in which surrounding and non-surrounding words are contrastively predicted given a central word to learn word embeddings using an NCE variant [141] for NSP. Negative sampling strategies, such as undersampling frequent words, in [100], are critical in the majority of methods. For language models, better discourse performance can be achieved by predicating surrounding sentences in a k-sized window around an anchor sentence. [140] show that "increased negative sampling only helps if negatives are taken from the original texts' context or block of information," that is, the same document, paragraph, or sentence, theoretically and empirically, that contextual negative sampling is useful. [142] investigate contrastive sentence structure learning, while [143] investigate how to combine different variants of the NSP pretraining tasks with non-contrastive, auxiliary self-supervision signals.

It is proposed by [144] that the CoDIR method can be used to distil an already pretrained larger teacher model like the Masked Transformer Language Model into a smaller student model that can then be pretrained. There are many subtleties that can be lost when compressing a pre-trained language model, including interactions between the original layer representations. The large teacher network and the small student network are used to extract layer representations from the input texts in order to create a student and teacher view of those texts for distillation.

Table 1 provides a comparison of the Contrastive learning methods for learning language representations. An auto-regressive language model and a contrastive text continuation EBM model are combined by [145] to improve text generation. Conditional NCE is a technique they learn during pretraining to use when comparing real data and language model-generated text continuations. In order to improve model perplexity, they sample the best text completions from the auto-regressive language model and score the best continuation via the trained EBM.

5.2 Computer vision

There are numerous computer vision applications that benefit from strong and discriminative representations. These include computer vision tasks ranging from object detection to video classification to medical imaging. Self-supervised image representation learning has recently been demonstrated by Contrastive Self-supervised Learning (CSL)-based methods [25,91], which have closed the gap between unsupervised and supervised representation learning on various downstream tasks. They can learn general purpose visual representations without labels and perform well in linear classification and transferability to other tasks or datasets. In particular, the concept of contrastive learning has played a significant role in the development of the recent self-supervised representation learning framework. As a pretext task, a typical contrastive learning method uses noise contrastive estimation (NCE) [146] to perform nonparametric instance discrimination [85]. This method encourages the two augmented views of the same image to be closer together on the embedding space, but pushes apart all the other pictures. Recently, researchers have focused on

Table 1 An overview of approaches in Natural Language Processing (NLP) that used contrastive methods

Paper	Methodology	Loss	Query	Key features	Application
Chi et al. [135]	INFOXLM	NT-Xent	Monolingual text context	Joint training of the pretext tasks	A cross-lingual language model for pre-training in order to maximize mutual information between multilingual-multigranularity texts
Fang et al. [98]	CERT	InfoNCE	Augmented sentence	Pretrains language representation models	Creates augmentations of original sentences using back-translation
Giorgi et al. [97]	DeCLUTR	InfoNCE	Query text	Learning sentence embeddings	Learning universal sentence embeddings that does not require labelled training data
Wu et al. [137]	CLEAR	InfoNCE	Sentence token	Multiple sentence-level augmentation strategies	To learn a noise-invariant sentence representation
Liao [139]	EDA	NT-Xent + InfoNCE	Sentence pair	fine-tuning pre-trained BERT on SNLI data	To build sentence embeddings
Deng et al. [145]	EBM	NT-Xent	Text sequence	Sequence level operation instead of token level	Text summarisation, dialogue and machine translation
Sun et al. [144]	CODIR	L2 Loss + KLD Loss + InfoNCE	Text sequence	Facilitating the exploitation of rich information in hidden layers	To compress large-scale language models in both pretraining and finetuning stages
Stephane et al. [218]	NSP	NT-Xent	Ordered sentence	Context splitting and shallow semantic signal	Enhancing the performance using multiple tasks in a multi-task pre-training framework

improving contrastive learning through image augmentation and exploration of negative examples. Although instance discrimination-based methods inevitably lead to class collisions, even very similar instances still need to be separated. As a result, the representation quality is likely to suffer [140]. When it comes to learning representations, it is critical to identify and even exploit similar examples.

The problem of class collisions appears to receive much less attention in contrastive learning, which is surprising. Identification of samples that are similar has received scant attention. All positive samples in AdpCLR [147] come from the embedding space's nearest neighbours. In order to get the best performance, this method must first pre-train using SimCLR [25] and then switch to AdpCLR. This is because the model cannot effectively extract semantic information from the images at this stage of training. For each sample, FNCancel [148] generates a support set that contains different augmented views from the same image and then uses mean or max aggregation strategy over the cosine similarity score between the augmented views in the support set and finally identify the top-K similar samples, which is a very different approach than that proposed in the previous section. However, in their experiments, the optimal support size is 8, which necessitates 8 additional forwarding passes to generate the embedding vectors.

It has been proposed by Ermolov et al. [149] that the MSE distance between augmented instances should be minimized in order to ensure that in many image processing pipelines, the whitening operation can be applied in batch. The MSE objective tries to reduce the distance (separation) between positive pairs, which means that the distance between positive pairs is reduced, but the corresponding representation space do not get collapsed into a single cluster when whitened vectors are distributed.

Time-varying features are commonly used in video understanding models, and they are typically aggregated to produce a video level prediction. The temporally varying properties may play an important role in further improvements on video understanding tasks, despite the fact that modelling only the time-invariant properties can lead to significant success on many video comprehension tasks. Literature is split on whether video representations should be consistent across the temporal dimension or distinct. However, instance contrastive pre-training does not force the model to learn similar features for clips that are time-separated to represent the video as a whole. The invariance constraint, however, has recently been alleviated by various methods, such as using a weighted temporal sampling strategy [51], cross-modal mining of positive samples from across video instances [49], or introducing additional pretext tasks that require learning temporal features [150,151]. Studying the importance of time augmentation, CVRL [51] has developed an algorithm to avoid enforcing too much temporal invariance when learning video representations. As a result of encouraging encoder temporal robustness and simulating key decay, VideoMoCo [152] enhances the image-based MoCo framework for video representation. Contrastive loss is used as the positive pair in VTHCL's [153] SlowFast architecture [154], with slow and fast pathway representations as the negative pair. An unsupervised method for learning video representations using deep neural embeddings is proposed as VIE [155], which combines the static image representation of a two-dimensional neural network with the dynamic motion representation of a three-dimensional neural network. There have also been studies in the literature looking at approaches based on generative contrastive learning such as the prediction of the dense representation of the next video block [156,157] or Contrastive Predictive Coding (CPC) [22] for videos [158]. In the past, unsupervised representations have been learned by utilizing natural clusters formed by visual similarity [159]. According to [160], this goal can be achieved through an approach known as contrastive learning, in which close neighbours are combined with those in the background to form an aggregate. The background neighbours are a random sample of points in the embedding space that are close to the query image. As part of an unsupervised clustering algorithm, a set of positive samples and their background neighbours are used to feed data into a clustering algorithm, which uses the data from these samples. NCE loss is used to distinguish between close neighbours and background neighbours during iterative embedding learning. While PCL [161] preserved the smoothness around each instance in a cluster, through the cluster's centroid, it encoded its higher semantic structure as well. In the case of latent class variables, PCL uses the Expectation Maximisation (EM) framework to learn the class's prototype and ensure that points in a cluster are kept close together. In the E-step, the momentum encoder's features are used to create k-clusters using k-means, and the InfoNCE loss is used in the M-step to minimize the distance between each point and the prototype of its cluster. Local views (spatial slices of the feature map taken from a middle layer) and global views (complete feature maps) of different enhanced versions of the same image are considered as a positive pair for image representation learning using AMDIM [88], another CSL approach. Global views of other images form a negative pair for the contrastive loss when using global views of other images. AMDIM uses this approach. Spatio-temporal features are used to generate local views for use in the video domain [162,163]. While other methods of this type aim to maximize the agreement between features encoded at various levels, the Global Local Loss method instead aims to learn distinct features across time slices of the feature map rather than maximize their agreement. An alternative to learning maximum agreement between features can be envisioned in this way. Recent research has used pretext tasks and contrastive learning in a multi-tasking setting to learn

video representation features that change over time. Playback rates can be predicted using temporal transforms [151] or by using clips with varying playback rates [164] as positive pairs for contrastive loss. Pretext tasks such as frame rotation prediction [165] and frame-tuple order verification [166] and frame-based contrastive learning are other approaches to contrastive learning. To avoid the use of a pretext task, the TCLR approach adds explicit temporal contrastive losses to encourage the learned features to be more diverse in time. An interesting comparison can be drawn between the work that attempts to capture intra-video variance using optical flow and others. To generate these "hard" negatives, IIC [167] uses frame repeating and shuffling rather than learning distinct features along the temporal axis, as is the case with intra instance negatives. For this purpose, DSM [168] uses an intra-instance triplet loss based on negatives generated by optical flow scaling and spatial warping. Video representations can be learned in a self-supervised manner by employing additional supervisory signals in addition to RGB video data. Although these methods require cross-modal data such as audio [169] and text narration [170], or the costly and time-consuming computation of hand-crafted visual priors (e.g. optical flow [167] or dense trajectories [171]), these methods are the only ones available. Table 2 provides a comparison of the Contrastive learning methods for learning visual representations.

5.3 Audio and speech

Using past speech segments to predict future segments, Contrastive Predictive Coding (CPC) [22] is emerging as a potent algorithm for representing speech signals. Unsupervised evaluation benchmarks, on the other hand, show that it falls short of competing methods. A standard Automatic Speech Recognition (ASR) system relies on a large amount of annotated data for training acoustic models and even larger amounts of text for training language models. Costly resources may only be available for a fraction of the world's languages, leaving the rest of the languages unaddressed. This is not a foregone conclusion, however. To build a working acoustic and language model, it is not necessary to have a large amount of textual resources and curated datasets.

The radical goal of building the foundations of speech technology from raw audio, with no labels or text [172,173], is addressed by the Zero Resource setting. One of these bricks is the unsupervised learning of phonetically contrastive speech embeddings or discrete units. Distance-based metrics like the ABX score [174,175] are commonly used to evaluate the intrinsic quality of embeddings in this context. However, CPC [22] has proven useful as a pretraining method in low resource settings, as it produces good speech features that can be fine-tuned for subsequent supervised tasks with limited labels [23,176,177]. Aside from the pure zero resource setting, CPC has yet to reach its potential and is outperformed by simpler systems like DPGMM [178] that cannot be fine-tuned for additional downstream tasks.

A different approach was taken by the authors of [179] who used mel spectrogram images to train their speech model. In order to create a learning representation that is language agnostic and also can be transferred well into, for instance, an emotion classification task, many image instance discrimination methods (e.g., frequency masking and time [180]) are used.

Table 3 compares the Contrastive learning methods for learning an audio representation. For solving paralinguistic classification tasks, contrastive learning is one of the most popular paradigms in speech self-supervised representation learning [181,182]. A version of these models known as COLA (COntrastive Learning for Audio) [183] makes use of audio instead of visuals. It is based on assigning high similarity to segments extracted from a single audio file and low similarity to segments extracted from different audio files in order to learn representations. It is then fed into other models that help solve the problem. When compared with other approaches, however, COLA does not look into data augmentation to enforce additional representational invariances, as other approaches have done [25].

5.4 Graphs

For learning graph representations, Graph Contrastive Learning (GCL) has emerged as the dominant technique for maximizing the mutual information between pairs of graph enhancements that share the same semantics. Due to the complex nature of graph data, it is difficult to preserve semantics during augmentations. Right now there are three problems with GCL data enhancements that are meant to preserve semantics. An initial step is to perform trial-and-error selection of augmentations for each individual dataset. An additional option is a laborious search for the various augmentations. When expensive domain-specific knowledge is introduced as guidance, the enhancements can be obtained. The effectiveness and generalizability of current GCL methods are hindered by a number of factors [184].

The general framework of contrastive learning in the computer vision domain [25,185] is followed by Graph contrastive Learning (GCL) [186–188], which generates two augmentations for each graph and then maximizes the mutual information between these two augmented views as shown in Fig. 8.

GCL is based on this general framework. As a result, the model can develop representations that are impervious to perturbations of any kind. Examples include GraphCL [188], which identifies four general enhancement types (node dropping, edge perturbation, attribute masking and subgraph) for GCL before moving on to more specific ones. There is a

Table 2 An overview of approaches in Computer Vision (CV) that used contrastive methods

Paper	Methodology	Loss	Query	Key Features	Application
Wu et al. [85]	Neural net classifiers	Non-parametric classification	Augmented image	Noise contrastive estimation	Improving test performance with more training data and better network architectures.
Zhuang et al. [160]	Local aggregation	Instance recognition (IR) loss	Query image	Maximizing a metric of local aggregation	Allowing similar data instances to move together in embedding space and dissimilar instances to separate.
Bachman et al. [88]	AMDIM	Log-softmax loss	Global features	Maximizing mutual information between features	To learn image representations effectively
Li et al. [161]	PCL	InfoNCE	Query image	Encoding semantic structures into the learned embedding space	Low-resource transfer learning
Zhang et al. [59]	ADACLR	InfoNCE	Sampled image	Nearest positive samples mining	To explore the samples (data) comparable to supervised contrastive learning.
Huynh et al. [148]	FNCancel	InfoNCE	Augmented image	Negative elimination and attraction by false negatives.	Learning transformations that keep positive input pairs close together while pushing negative input pairs farther apart.
Ermolov et al. [149]	Whitening MSE	MSE	Query image	"Whitening" of the latent space features	To avoid degenerate solutions for representations collapsing to a single point.
Qian et al. [51]	CVRL	InfoNCE	Augmented video clip	Data augmentations involving spatial and temporal cues	To learn spatio-temporal visual representations from unlabelled videos.
Han et al. [49]	CoCLR	InfoNCE	32-frame RGB (or flow) clip	Self-supervised co-training scheme	To exploit the complementary information from different views, of the same data source
Yao et al. [150]	PRP	NT-Xent	Query video clip	Self-supervised signals about video playback rates	To learn spatio-temporal representation in a simple yet effective way.
Pan et al. [152]	VideoMoCo	L1-norm + InfoNCE	Query video clip	Adaptively dropping out different frames during training	To learn video representations without empirically designing pretext tasks.
Yang et al. [204]	VTHCL	InfoNCE	Query video clip	Maximizing the mutual information between slow and fast videos	Learned representations are generalized well to other downstream tasks.
Zhuang et al. [155]	VIE	Instance recognition (IR) loss	Video frames	Training deep nonlinear embeddings on video sequence inputs	Capturing the strong statistical structure inherent in videos without annotation labels
Han et al. [49]	MemDPC	InfoNCE	Optical flow frames	Predictive attention mechanism	Self-supervised learning representations for action recognition.
Tao et al. [167]	IIC	Triplet loss	Query video clip	Extending the negative samples by intra-negative samples	To learn video representations.
Wang et al. [168]	DSM	Triplet loss	Query video clip	Attention towards the motion information	To enhance temporal sensitivity of the network

Table 3 An overview of approaches in Audio and speech that used contrastive methods

Paper	Methodology	Loss	Query	Key Features	Application
Oord et al. [22]	CPC	InfoNCE	Text context	In latent space, predictions about the future are made using autoregressive models	To learn useful representations achieving strong performance in speech predictions
Dunbar et al. [172]	Zero Resource setting	N/A	Speech features (series)	To discover low bit-rate subword representations	Learning speech representations from raw audio signals without any labels
Schneider et al. [23]	wav2vec	InfoNCE	Text context	Multi-layer convolutional neural network	Speech recognition by learning representations of raw audio
Kawakami et al. [176]	TDNN	InfoNCE	CPC representation	Representations from up to 8000 hours of diverse and noisy speech data	Improvements in out-of-domain transfer in 25 phonetically diverse languages
Nandan and Vepa [179]	Convolutional encoder model	InfoNCE	Augmented mel spectrograms	Augmented Log Mel spectrograms	Emotion classification via speech representations or embeddings
Saeed et al. [183]	COLA	Multi-class NT-Xent	Audio sequence	High similarity to audio segments from the same recording, lower similarity to segments from different recordings	To learn a general-purpose representation of audio

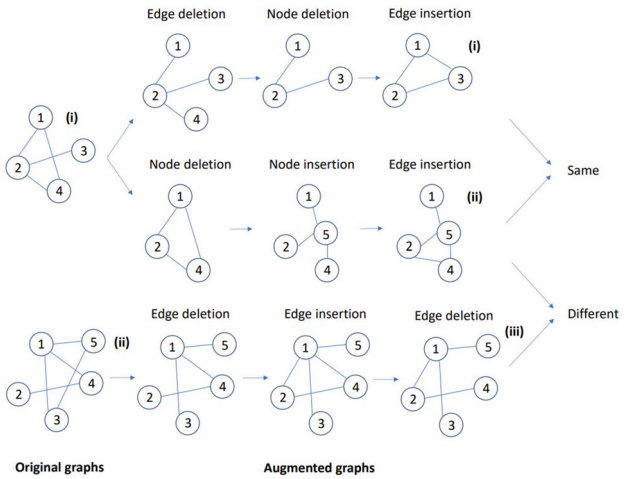

Fig. 8 Data augmentation illustration in Graph contrastive Learning (GCL) [186]

wide range of structural and semantic information in graphs, so these enhancements are not appropriate for every situation. According to GraphCL [188], social networks benefit from edge perturbation, but some biochemical molecules are negatively affected. Even if the perturbation is small, these augmentations can completely alter the graph's semantics. It is possible to change the molecular structure by removing one carbon atom from the phenyl ring, which will result in an alkene chain [189].

There are two types of graph contrastive learning. By comparing local and global representations, a single group can encode useful information. Initial proposals for expressive representations for graphs or nodes based on maximizing mutual information between graph-level representations and substructure-level representations of varying granularity were made by DGI [86] and InfoGraph [190]. To learn both node-level and graph-level representations at the node level, the MVGRL [89] model proposes node diffusion and contrast of node representations with augmented graph representations. Another group is geared toward learning representations that are more tolerant of data transformations. This is how it works: they first augment the graph data, and then they feed the enhanced graphs into a shared encoder and projection head. Adaptive augmentations that only perturb unimportant components are proposed by GCA [191] for node-level tasks, in order to preserve the graph's intrinsic structure and attributes. GCL's false negative problem is addressed in DGCL [192] by introducing a new probabilistic method. GraphCL [188] proposes four types of augmentations for general graphs and shows that the learned representations can help downstream tasks for graph-level tasks.

It is common practice in GraphCL [188] to manually select data augmentations for each dataset, which significantly restricts the framework's generalizability and usability. JOAO [186] proposes to automate GraphCL's selection of augmentation pairs in order to eliminate the tedious process of manually tuning the algorithm for each dataset. However, the search for appropriate augmentations incurs a higher computational cost, and the augmentation pool is still built and configured using human prior knowledge. Bioisosteres with similar properties can be substituted for valid molecular graph substructures to avoid altering the semantics of GraphCL and JOAO(v2), according to MoCL [189]. However, it necessitates in-depth domain expertise and cannot be used in other contexts, such as social networks.

Table 4 provides a comparison of the Contrastive learning methods for learning graph representations.

5.5 Miscellaneous

At the end of this section, we will look at some unrelated works that apply contrastive learning in different contexts than those discussed thus far.

5.5.1 Multi-modal

Contrastive learning for multimodal representation learning has recently been examined in a number of studies [93,193,194]. Cross-modal embedding space is learned by the majority of them [93,193]. Their goal is to capture the information that is shared between different modalities. The combined representation of multiple modalities is not directly examined, so the multimodal synergies are not fully leveraged. In order to address this problem, the authors in [194] propose an RGB-D representation learning framework that compares directly pairs of point-and-pixel pairs. However, it can only be used in two ways.

Contrastive learning can be divided into single and multimodality-based CL based on the type of data used. In order to construct a negative pair, CL must select hard negative samples [117,118,195,196]. The majority of currently available methods either increase batch size or maintain large memory banks, resulting in high memory requirements [91]. Recently, a number of studies have examined CL from a mutual information perspective (MI). According to the authors in [197], the MI between views ought to be reduced as much as possible by augmenting the data while keeping any information that is relevant to the task intact. [198] demonstrates that the CL algorithm family maximizes the lower bound on MI between multiple "views," with the choice of negative samples and views being crucial to these algorithms. "Views" in this context refer to augmented images.

Table 4 An overview of approaches in Graphs that used contrastive methods

Paper	Methodology	Loss	Query	Key Features	Application
Xia et al. [184]	SimGRACE	NT-Xent	Global graph	Increase mutual information between graph augmentations by as much as possible	To enhance the robustness of graph contrastive learning
You et al. [186]	JOAO	NT-Xent	Sub-graph structure	Min-max optimization	Learning generalizable, transferable and robust representations from unlabelled graphs
Zeng and Xie [187]	CSSL	InfoNCE	Global graph	Creating augmented graphs from the original graphs using data augmentation	To pretrain graph encoders on widely-available unlabelled graphs
You et al. [186]	GraphCL	InfoNCE	Global graph	Designing of four types of graph augmentations	Graph representations of similar or better generalizability, transferability, and robustness
Sun et al. [105]	InfoGraph	InfoGraph loss	Global graph	Mutual information between graph-level representations is maximized	To learn from unlabelled data while preserving the semantic space of the data itself
Hassani and Ahmadi [89]	MVGRL	InfoNCE	Graph Augmentations	Contrasting encodings from first-order neighbours and a graph diffusion	To learn node and graph level representations
Zhu et al. [191]	GCA	Triplet loss	Global graph	Augmentation schemes based on node centrality measures to highlight important connective structures	To preserve intrinsic structures and attributes of graphs
Xia et al. [192]	DGCL	NT-Xent	Query node/edge	DGCL-weight and DGCL-mix to boost the performance of GCL	To estimate the probability whether each negative sample is true or not

5.5.2 Audio-visual

There has been a great deal of research into how to learn video representations from natural audio-visual correspondence. Most existing approaches concentrate on capturing high-level semantic information for sequence-level (global) discrimination tasks like audio/visual video classification [199–201]. In this line of research, the model is encouraged to capture fine-grained temporal information by using various temporal scales for audio and visual data [202]. Classification tasks remain the primary evaluation method for only global representations induced by changes in sampling rates.

For "local" tasks like sound source separation and localization [203–205], spatial-temporal local information is another area of research. Lin et al. [205] were able to learn patch-level correspondence between audio and visual input by drawing positive/negative patches from sounding/non-sounding regions as a result of audio-visual feature correlation. In contrast, their approach is limited to the learning of spatial and temporal global representations and is tested on localization tasks. This approach differs from previous ones in that it explicitly optimizes for both global and local spatio-temporal representations and evaluates on both classification and localization downstream tasks. This is a significant advance. As for the learning of audio-visual representation, there is some work that focuses on specific tasks, such as speaker recognition [206]. Video representation is learned and demonstrated on a variety of downstream tasks as part of our work.

5.5.3 Code representation

According to previous studies [207,208], BERT native sentence representations are heavily weighted toward tokens with high frequency. It is even more serious in codes because of the token imbalance problem. As an example, the "def" token is used in almost all Python functions. Converse learning encourages the original sequence to be more like the "positive" augmented sequence, while staying away from the "negative" sequences, making the model learn a more even decision boundary across different data points to reconcile the token imbalance representation bias [102]. There have been a number of recent works that attempt to compare code snippets that are similar and different [209,210]. However, they only deal with a single mode of code, ignoring the multi-modality of programming languages. In order to acquire more comprehensive code representations, these two semantically equivalent modalities can offer complementary information.

5.5.4 Multi-view learning

There are two types of information in multi-view data: the common semantics across all views and the view-specific information for each view. For example, the unrelated context in the text and the background pixels in the image are useless view-private information for learning common semantics when they are combined with a text. In multi-view learning, common semantics and the avoidance of meaningless view-private information is a constant concern. Yunfan et al. [211] proposed a one-stage online image clustering method that explicitly conducted contrastive learning at the instance and cluster level. Contrastive learning has also been used for multi-view learning in some cases [89,93,212]. Tian et al. [93] proposed a contrastive multi-view coding framework to capture the underlying scene semantics, for instance. Contrastive learning was used to develop a multi-view representation learning method in [89]. Multiple contrastive learning frameworks for multi-view clustering have recently been investigated [212–214].

Table 5 compares the performance gains across all domains before and after the implementation of Contrastive Learning (CL).

6 Challenges and future research directions

As a result of contrastive learning, the performance gap between supervised and unsupervised models has narrowed, but more theoretical analysis is needed to provide a strong justification. Researchers are divided on the generalizability of representations learned using such methods, as well as on the appropriateness of transformations. We address some of the most pressing issues and suggest future research directions in this section.

6.1 Negative sampling

Contrastive learning works best when a large number of negative examples are analysed. Sampling negatives that are difficult to contrast, or relevant negatives, has the potential to improve sampling efficiency. sampling a variety of negatives can play an important role, depending on the task at hand. Easy-to-contrast negative samples have received little attention in research to date [61]. An understanding of how positive and negative samples are generated, along with bias in the data, is required before contrastive learning can be effectively applied to other datasets or problems [84,215].

The unsupervised nature of contrastive learning and lack of access to the labels may lead us to accept examples that are semantically similar to the anchor. This sampling bias can be reduced without relying on the actual data labels in future research.

Table 5 A comparison for performance enhancement after incorporation of contrastive learning (CL) across all domains

Paper	Objective	Dataset	Evaluation metric	State-of-the-art performance (without CL)	State-of-the-art performance (with CL)	Enhancement
Alayrac et al. [193]	Image classification	PASCAL	mAP	77.4	80.5	3.1
		ImageNet	Accuracy (%)	57.4	69.3	11.4
Kalantidis et al. [117]	Object detection and instance segmentation	COCO	AP	38.2	39.4	1.2
He et al. [91]	Linear image classification	ImageNet	Accuracy (%)	61.3	68.6	7.3
	Object detection	PASCAL VOC	AP	81.4	81.6	0.2
Asano et al. [199]	Multi-modal understanding	VGG-sound, kinetics-400	NMI	46.5	55.9	9.4
	Retrieval via various number of nearest neighbours	HMDB, UCF	Recall@20	54.6	75.5	20.9
Patrick et al. [201]	Video action recognition	HMDB	Accuracy (%)	61.9	62.3	0.4
		UCF	Accuracy (%)	91.2	90.9	n/a
Xiao et al. [202]	Video recognition	Kinetics-400	mAP	42.5	43.7	1.2
Lin et al. [212]	Sound localization	SoundNet-Flickr 10K	cIoU@0.5	56.6	71.1	14.5
		SoundNet-Flickr 10K	AUC	51.5	58.0	6.5
		MUSIC-Synthetic	cIoU@0.5	15.4	25.1	9.7
		MUSIC-Synthetic	AUC	17.0	21.9	4.9
Jain et al. [209]	Zero-shot code clone detection	Natural code	AUROC	74.04	79.39	5.35
		Natural code	AP	77.65	81.47	3.82
Bui et al. [210]	Code-to-code retrieval	JavaMed	MAP (%)	71.9	82.5	10.6
		JavaMed	MRR (%)	69.3	84.1	14.8
		JavaSmall	MAP (%)	58.0	64.0	6.0
Li et al. [130]	Clustering	STL-10	NMI (%)	61.1	76.4	15.3
		ImageNet-10	NMI (%)	80.2	85.9	5.7

6.2 Appropriate transformation function

In theory, there is a clear distinction between the transform heads and the base encoder, but in practice, there is not much of a difference between the two. This serves to emphasize the importance of transfer learning to downstream tasks. It is not clear which projection and transform heads are the best choices, but the basic encoders are, for the most part, taken directly from supervised learning; however, there have been some modifications made, such as making the layers wider so that more features can be captured [216]. Self-supervised contrastive representation learning assumes that the data transformations that are done on the data points are semantically invariant, and thus are simply two views of the same object. It is a common assumption in contrastive representation learning schemes that downstream tasks will be unaffected by the learning process. There should be no change in the semantic meaning of a data point after it has been transformed or augmented in any way.

6.3 Data augmentation

A variety of data augmentation strategies, such as resizing, rotating, and colouring, are largely responsible for recent advances in the field of visual representation learning [217]. It is, however, difficult to directly apply existing image-based augmentation to some data (e.g., graphs). For an accurate representation, it is necessary to take into account the dataset's inherent bias when looking at augmentations [61,215]. It is a promising direction to design more efficient augmentation strategies for SSL in order to further improve its performance.

6.4 Theoretical inadequacy

When investigating the generalizability of a contrastive objective function, architecture design and sampling techniques have a major impact. Using self-supervised learning representations, it is possible to keep only the data necessary for the task at hand and discard everything else (with a fixed gap) [61]. To summarize, these findings show that the effectiveness of these methods relies heavily on the pretext task selected during training. More theoretical analysis is needed on different pipeline modules in order to better understand how they all work together [83,84].

6.5 Selection of scoring function

For retrieval and ranking applications, prior to contrastive learning, the learned similarity score was used. Also as a proxy task for representation learning, it is most commonly used these days. Researchers can look into whether it is possible to use the learned similarity score in new and novel ways or not. An asymmetric scoring function is an interesting addition to the scoring function [83]. However, not all kinds of similarities are the same; for instance, "dog-cat" similarity metric should be different from "dog-animal." Current literature on contrastive methods assumes a simple symmetric relationship between distance and similarity. How could a non-transitive similarity relationship be used to develop a contrastive loss is a plausible research direction.

7 Conclusion

Recent Self-supervised Learning (SSL) advancements have provided novel insights into reducing the dependency on annotated labels and enabling training on massive unlabelled data. In this paper, a comprehensive review and comparative analysis of the literature on contrastive self-supervised learning methods are provided in a variety of applications and various input domains that include video, image, audio, and text, among others. Also, a thorough examination of contrastive learning approaches, encoding methods, and loss functions and overall methodology is presented. Self-supervised learning is the present and future of deep learning in many ways because of its superior ability to use Web-scale unlabelled data to train feature extractors and context generators efficiently. Finally, the paper discusses the technical limitations in current research and suggest potential future research directions. This paper will be of interest to both those who are not familiar with the contrastive learning methodology and wish to learn more about its operation, as well as those who are already familiar with the topic. This is due to the fact that contributions to the development of this field have come from a wide variety of sources.

Acknowledgements Not applicable.

Author Contributions PK has written and organised the paper. PR has formatted and presented the digrammatic representations. SC has supervised the work.

Funding Not applicable.

Data availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval and consent to participate Not applicable.

Consent for publication Yes.

Code availability Not applicable.

References

1. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 248–255
2. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
3. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
4. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
5. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
6. Liu B (2012) Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol 5(1):1–167
7. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
8. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) Albert: A lite Bert for self-supervised learning of language representations. arXiv:1909.11942
9. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized Bert pretraining approach. arXiv:1907.11692
10. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) Xlnet: Generalized autoregressive pretraining for language understanding. Adv Neural Inf Process Syst 32
11. Asai A, Hashimoto K, Hajishirzi H, Socher R, Xiong C (2019) Learning to retrieve reasoning paths over wikipedia graph for question answering. arXiv:1911.10470
12. Ding M, Zhou C, Chen Q, Yang H, Tang J (2019) Cognitive graph for multi-hop reading comprehension at scale. arXiv:1905.05460
13. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) Squad: 100,000+ questions for machine comprehension of text. arXiv:1606.05250
14. Yang Z, Qi P, Zhang S, Bengio Y, Cohen WW, Salakhutdinov R, Manning CD (2018) Hotpotqa: a dataset for diverse, explainable multi-hop question answering. arXiv:1809.09600
15. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
16. Kalantidis Y, Sariyildiz M, Weinzaepfel P, Larlus D (2020) Improving self-supervised representation learning by synthesizing challenging negatives. Naver Labs Europe
17. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828
18. Zimmermann RS, Sharma Y, Schneider S, Bethge M, Brendel W (2021) Contrastive learning inverts the data generating process. In: International conference on machine learning. PMLR, pp 12979–12990
19. Ilić S, Marrese-Taylor E, Balazs JA, Matsuo Y (2018) Deep contextualized word representations for detecting sarcasm and irony. arXiv:1809.09795
20. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training
21. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
22. Van den Oord A, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv:1807.03748
23. Schneider S, Baevski A, Collobert R, Auli M (2019) wav2vec: Unsupervised pre-training for speech recognition. arXiv:1904.05862
24. Baevski A, Zhou Y, Mohamed A, Auli M (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv Neural Inf Process Syst 33:12449–12460
25. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR, pp 1597–1607
26. Chen X, Xie S, He K (2021) An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9640–9649
27. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A (2021) Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9650–9660
28. Bao H, Dong L, Wei F (2021) Beit: Bert pre-training of image transformers. arXiv:2106.08254
29. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2021) Masked autoencoders are scalable vision learners. arXiv:2111.06377
30. Lample G, Conneau A, Denoyer L, Ranzato M (2017) Unsupervised machine translation using monolingual corpora only. arXiv:1711.00043
31. Baevski A, Hsu W-N, Conneau A, Auli M (2021) Unsupervised speech recognition. Adv Neural Inf Process Syst 34
32. Hsu W-N, Tsai Y-HH, Bolte B, Salakhutdinov R, Mohamed A (2021) Hubert: how much can a bad teacher benefit ASR pre-training?. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6533–6537
33. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR, pp 8748–8763
34. Grill J-B, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, Doersch C, Avila Pires B, Guo Z, Gheshlaghi Azar M et al (2020) Bootstrap your own latent-a new approach to self-supervised learning. Adv Neural Inf Process Syst 33:21271–21284
35. Friston K, Kiebel S (2009) Predictive coding under the free-energy principle. Philos Trans R Soc B Bioll Sci 364(1521):1211–1221
36. Friston K (2010) The free-energy principle: A unified brain theory? Nat Rev Neurosci 11(2):127–138
37. Jaegle A, Gimeno F, Brock A, Vinyals O, Zisserman A, Carreira J (2021) Perceiver: general perception with iterative attention. In: International conference on machine learning. PMLR, pp 4651–4664
38. Holmberg OG, Köhler ND, Martins T, Siedlecki J, Herold T, Keidel L, Asani B, Schiefelbein J, Priglinger S, Kortuem KU et al (2020) Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy. Nat Mach Intell 2(11):719–726
39. Arandjelovic R, Zisserman A (2017) Look, listen and learn. In: Proceedings of the IEEE international conference on computer vision, pp 609–617
40. Arandjelovic R, Zisserman A (2018) Objects that sound. In: Proceedings of the European conference on computer vision (ECCV), pp 435–451

41. Lee H-Y, Huang J-B, Singh M, Yang M-H (2017) Unsupervised representation learning by sorting sequences. In: Proceedings of the IEEE international conference on computer vision, pp 667–676
42. Misra I, van der Maaten L (2020) Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6707–6717
43. Fernando B, Bilen H, Gavves E, Gould S (2017) Self-supervised video representation learning with odd-one-out networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3636–3645
44. Wei D, Lim JJ, Zisserman A, Freeman WT (2018) Learning and using the arrow of time. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8052–8060
45. Gan C, Gong B, Liu K, Su H, Guibas LJ (2018) Geometry guided convolutional neural networks for self-supervised video representation learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5589–5597
46. Vondrick C, Pirsiavash H, Torralba A (2016) Generating videos with scene dynamics. Adv Neural Inf Process Syst 29
47. Zhao Y, Deng B, Shen C, Liu Y, Lu H, Hua X-S (2017) Spatio-temporal autoencoder for video anomaly detection. In: Proceedings of the 25th ACM international conference on multimedia, pp 1933–1941
48. Kim D, Cho D, Kweon IS (2019) Self-supervised video representation learning with space-time cubic puzzles. Proc AAAI Conf Artif Intell 33(01):8545–8552
49. Han T, Xie W, Zisserman A (2020) Self-supervised co-training for video representation learning. Adv Neural Inf Process Syst 33:5679–5690
50. Kong Q, Wei W, Deng Z, Yoshinaga T, Murakami T (2020) Cycle-contrast for self-supervised video representation learning. Adv Neural Inf Process Syst 33:8089–8100
51. Qian R, Meng T, Gong B, Yang M-H, Wang H, Belongie S, Cui Y (2021) Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6964–6974
52. McCann B, Bradbury J, Xiong C, Socher R (2017) Learned in translation: contextualized word vectors. Adv Neural Inf Process Syst 30
53. Baevski A, Edunov S, Liu Y, Zettlemoyer L, Auli M (2019) Cloze-driven pretraining of self-attention networks. arXiv:1903.07785
54. Jiao X, Yin Y, Shang L, Jiang X, Chen X, Li L, Wang F, Liu Q (2019) Tinybert: distilling Bert for natural language understanding. arXiv:1909.10351
55. Baevski A, Auli M, Mohamed A (2019) Effectiveness of self-supervised pre-training for speech recognition. arXiv:1911.03912
56. Baevski A, Schneider S, Auli M (2019) vq-wav2vec: Self-supervised learning of discrete speech representations. arXiv:1910.05453
57. Zhang Y, Qin J, Park DS, Han W, Chiu C-C, Pang R, Le QV, Wu Y (2020) Pushing the limits of semi-supervised learning for automatic speech recognition. arXiv:2010.10504
58. Chung Y-A, Zhang Y, Han W, Chiu C-C, Qin J, Pang R, Wu Y (2021) W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. arXiv:2108.06209
59. Zhang Y, Park DS, Han W, Qin J, Gulati A, Shor J, Jansen A, Xu Y, Huang Y, Wang S et al (2021) Bigssl: exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. arXiv:2109.13226
60. Chiu C-C, Qin J, Zhang Y, Yu J, Wu Y (2022) Self-supervised learning with random-projection quantizer for speech recognition. arXiv:2202.01855
61. Liu X, Zhang F, Hou Z, Mian L, Wang Z, Zhang J, Tang J (2021) Self-supervised learning: Generative or contrastive. IEEE Trans Knowl Data Eng
62. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I et al (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
63. Tran C, Bhosale S, Cross J, Koehn P, Edunov S, Fan A (2021) Facebook ai wmt21 news translation task submission. arXiv:2108.03265
64. Arivazhagan N, Bapna A, Firat O, Lepikhin D, Johnson M, Krikun M, Chen MX, Cao Y, Foster G, Cherry C et al (2019) Massively multilingual neural machine translation in the wild: findings and challenges. arXiv:1907.05019
65. Van Oord A, Kalchbrenner N, Kavukcuoglu K (2016) Pixel recurrent neural networks. In: International conference on machine learning. PMLR, pp 1747–1756
66. Van den Oord A, Kalchbrenner N, Espeholt L, Vinyals O, Graves A et al (2016) Conditional image generation with Pixelcnn decoders. Adv Neural Inf Process Syst 29
67. Rezende D, Mohamed S (2015) Variational inference with normalizing flows. In: International conference on machine learning. PMLR, pp 1530–1538
68. Yang G, Huang X, Hao Z, Liu M-Y, Belongie S, Hariharan B (2019) Pointflow: 3d point cloud generation with continuous normalizing flows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4541–4550
69. Vahdat A, Kautz J (2020) Nvae: a deep hierarchical variational autoencoder. Adv Neural Inf Process Syst 33:19667–19679
70. Chen M, Radford A, Child R, Wu J, Jun H, Luan D, Sutskever I (2020) Generative pretraining from pixels. In: International conference on machine learning. PMLR, pp 1691–1703
71. You J, Ying R, Ren X, Hamilton W, Leskovec J (2018) Graphrnn: generating realistic graphs with deep auto-regressive models. In: International conference on machine learning. PMLR, pp 5708–5717
72. Zhang L, Lin J, Shao H, Zhang Z, Yan X, Long J (2021) End-to-end unsupervised fault detection using a flow-based model. Reliab Eng Syst Saf 215:107805
73. Hinton GE, Zemel R (1993) Autoencoders, minimum description length and helmholtz free energy. Adv Neural Inf Process Syst 6
74. Japkowicz N, Hanson SJ, Gluck MA (2000) Nonlinear autoassociation is not equivalent to PCA. Neural Comput 12(3):531–545
75. Vincent P, Larochelle H, Bengio Y, Manzagol P-A (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on machine learning, pp 1096–1103
76. Rifai S, Vincent P, Muller X, Glorot X, Bengio Y (2011) Contractive auto-encoders: explicit invariance during feature extraction. In: ICML
77. Zhang R, Isola P, Efros AA (2017) Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1058–1067
78. Hinton GE, Krizhevsky A, Wang SD (2011) Transforming auto-encoders. In: International conference on artificial neural networks. Springer, pp 44–51
79. Wang F, Liu H (2021) Understanding the behaviour of contrastive loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2495–2504
80. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25
81. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
82. Gutmann M, Hyvärinen A (2010) Noise-contrastive estimation: A new estimation principle for unnormalized statistical models.

In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, pp 297–304
83. Le-Khac PH, Healy G, Smeaton AF (2020) Contrastive representation learning: a framework and review. IEEE Access 8:193907–193934
84. Jaiswal A, Babu AR, Zadeh MZ, Banerjee D, Makedon F (2020) A survey on contrastive self-supervised learning. Technologies 9(1):2
85. Wu Z, Xiong Y, Yu SX, Lin D (2018) Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3733–3742
86. Velickovic P, Fedus W, Hamilton WL, Liò P, Bengio Y, Hjelm RD (2019) Deep graph infomax. ICLR (Poster) 2(3):4
87. Hjelm RD, Fedorov A, Lavoie-Marchildon S, Grewal K, Bachman P, Trischler A, Bengio Y (2018) Learning deep representations by mutual information estimation and maximization. arXiv:1808.06670
88. Bachman P, Hjelm RD, Buchwalter W (2019) Learning representations by maximizing mutual information across views. Adv Neural Inf Process Syst 32
89. Hassani K, Khasahmadi AH (2020) Contrastive multi-view representation learning on graphs. In: International conference on machine learning. PMLR, pp 4116–4126
90. Tschannen M, Djolonga J, Rubenstein PK, Gelly S, Lucic M (2019) On mutual information maximization for representation learning. arXiv:1907.13625
91. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9729–9738
92. Noroozi M, Vinjimoor A, Favaro P, Pirsiavash H (2018) Boosting self-supervised learning via knowledge transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9359–9367
93. Tian Y, Krishnan D, Isola P (2020) Contrastive multiview coding. In: European conference on computer vision. Springer, pp 776–794
94. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C, Krishnan D (2020) Supervised contrastive learning. Adv Neural Inf Process Syst 33:18661–18673
95. Singh B, Davis LS (2018) An analysis of scale invariance in object detection snip. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3578–3587
96. Purushwalkam S, Gupta A (2020) Demystifying contrastive self-supervised learning: invariances, augmentations and dataset biases. Adv Neural Inf Process Syst 33:3407–3418
97. Giorgi J, Nitski O, Wang B, Bader G (2020) Declutr: deep contrastive learning for unsupervised textual representations. arXiv:2006.03659
98. Fang H, Wang S, Zhou M, Ding J, Xie P (2020) Cert: contrastive self-supervised learning for language understanding. arXiv:2005.12766
99. Xie Q, Dai Z, Hovy E, Luong T, Le Q (2020) Unsupervised data augmentation for consistency training. Adv Neural Inf Process Syst 33:6256–6268
100. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781
101. Gao T, Yao X, Chen D (2021) Simcse: simple contrastive learning of sentence embeddings. arXiv:2104.08821
102. Yan Y, Li R, Wang S, Zhang F, Wu W, Xu W (2021) Consert: a contrastive framework for self-supervised sentence representation transfer. arXiv:2105.11741
103. Rozsa A, Rudd EM, Boult TE (2016) Adversarial diversity and hard positive generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 25–32
104. Ilharco G, Zellers R, Farhadi A, Hajishirzi H (2020) Probing Contextual Language Models for Common Ground with Visual Representations. https://doi.org/10.48550/arxiv.2005.00619
105. Sun C, Baradel F, Murphy K, Schmid C (2019) Learning video representations using contrastive bidirectional transformer. arXiv:1906.05743
106. Senocak A, Oh T-H, Kim J, Yang M-H, Kweon IS (2018) Learning to localize sound source in visual scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4358–4366
107. Senocak A, Oh T-H, Kim J, Yang M-H, Kweon IS (2019) Learning to localize sound sources in visual scenes: analysis and applications. IEEE Trans Pattern Anal Mach Intell 43(5):1605–1619
108. Qian R, Hu D, Dinkel H, Wu M, Xu N, Lin W (2020) Multiple sound sources localization from coarse to fine. In: European conference on computer vision. Springer, pp 292–308
109. Hu D, Nie F, Li X (2019) Deep multimodal clustering for unsupervised audiovisual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9248–9257
110. Hu D, Qian R, Jiang M, Tan X, Wen S, Ding E, Lin W, Dou D (2020) Discriminative sounding objects localization via self-supervised audiovisual matching. Adv Neural Inf Process Syst 33:10077–10087
111. Hu D, Wang Z, Xiong H, Wang D, Nie F, Dou D (2020) Curriculum audiovisual learning. arXiv:2001.09414
112. Zhan X, Xie J, Liu Z, Ong Y-S, Loy CC (2020) Online deep clustering for unsupervised representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6688–6697
113. Tao Y, Takagi K, Nakata K (2021) Clustering-friendly representation learning via instance discrimination and feature decorrelation. arXiv:2106.00131
114. Tsai TW, Li C, Zhu J (2020) Mice: mixture of contrastive experts for unsupervised image clustering. In: International conference on learning representations
115. Hu Q, Wang X, Hu W, Qi G-J (2021) Adco: adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1074–1083
116. Chen X, Fan H, Girshick R, He K (2020) Improved baselines with momentum contrastive learning. arXiv:2003.04297
117. Kalantidis Y, Sariyildiz MB, Pion N, Weinzaepfel P, Larlus D (2020) Hard negative mixing for contrastive learning. Adv Neural Inf Process Syst 33:21798–21809
118. Robinson J, Chuang C-Y, Sra S, Jegelka S (2020) Contrastive learning with hard negative samples. arXiv:2010.04592
119. Sohn K (2016) Improved deep metric learning with multi-class n-pair loss objective. Adv Neural Inf Process Syst 29
120. Wu C, Wu F, Huang Y (2021) Rethinking infonce: How many negative samples do you need? arXiv:2105.13003
121. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
122. Wang X, Hua Y, Kodirov E, Hu G, Garnier R, Robertson NM (2019) Ranked list loss for deep metric learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5207–5216
123. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10(2)

124. Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 1. IEEE, pp 539–546
125. Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), vol 2. IEEE, pp 1735–1742
126. Oh Song H, Xiang Y, Jegelka S, Savarese S (2016) Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4004–4012
127. Goldberger J, Hinton G E, Roweis S, Salakhutdinov R R, "Neighbourhood components analysis," *Advances in neural information processing systems*, vol. 17, (2004)
128. Ghojogh B, Karray F, Crowley M (2019) Fisher and kernel fisher discriminant analysis: tutorial. arXiv:1906.09436
129. Sun Z, Deng Z-H, Nie J-Y, Tang J (2019) Rotate: knowledge graph embedding by relational rotation in complex space. arXiv:1902.10197
130. Li Z, Ji J, Fu Z, Ge Y, Xu S, Chen C, Zhang Y (2021) Efficient non-sampling knowledge graph embedding. Proc Web Conf 2021:1727–1736
131. Peng X, Chen G, Lin C, Stevenson M (2021) Highly efficient knowledge graph embedding learning with orthogonal procrustes analysis. arXiv:2104.04676
132. Cheng JY, Goh H, Dogrusoz K, Tuzel O, Azemi E (2020) Subject-aware contrastive learning for biosignals. arXiv:2007.04871
133. Becker S, Hinton GE (1992) Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature 355(6356):161–163
134. Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R (1993) Signature verification using a "siamese" time delay neural network. Adv Neural Inf Process Syst 6
135. Chi Z, Dong L, Wei F, Yang N, Singhal S, Wang W, Song X, Mao X-L, Huang H, Zhou M (2020) Infoxlm: an information-theoretic framework for cross-lingual language model pre-training. arXiv:2007.07834
136. Lample G, Conneau A (2019) Cross-lingual language model pretraining. arXiv:1901.07291
137. Wu Z, Wang S, Gu J, Khabsa M, Sun F, Ma H (2020) Clear: contrastive learning for sentence representation. arXiv:2012.15466
138. Wei J, Zou K (2019) Eda: easy data augmentation techniques for boosting performance on text classification tasks. arXiv:1901.11196
139. Liao D (2021) Sentence embeddings using supervised contrastive learning. arXiv:2106.04791
140. Arora S, Khandeparkar H, Khodak M, Plevrakis O, Saunshi N (2019) A theoretical analysis of contrastive unsupervised representation learning. arXiv:1902.09229
141. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 26
142. Simoulin A, Crabbé B (2021) Contrasting distinct structured views to learn sentence embeddings. In: European chapter of the association of computational linguistics (student)
143. Aroca-Ouellette S, Rudzicz F (2020) On losses for modern language models. arXiv:2010.01694
144. Sun S, Gan Z, Cheng Y, Fang Y, Wang S, Liu J (2020) Contrastive distillation on intermediate representations for language model compression. arXiv:2009.14167
145. Deng Y, Bakhtin A, Ott M, Szlam A, Ranzato M (2020) Residual energy-based models for text generation. arXiv:2004.11714
146. Lai C-I (2019) Contrastive predictive coding based feature for automatic speaker verification. arXiv:1904.01575
147. Zhang S, Yan J, Yang X (2020) Self-supervised representation learning via adaptive hard-positive mining
148. Huynh T, Kornblith S, Walter MR, Maire M, Khademi M (2022) Boosting contrastive self-supervised learning with false negative cancellation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2785–2795
149. Ermolov A, Siarohin A, Sangineto E, Sebe N (2021) Whitening for self-supervised representation learning. In: International conference on machine learning. PMLR, pp 3015–3024
150. Yao Y, Liu C, Luo D, Zhou Y, Ye Q (2020) Video playback rate perception for self-supervised spatio-temporal representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6548–6557
151. Bai Y, Fan H, Misra I, Venkatesh G, Lu Y, Zhou Y, Yu Q, Chandra V, Yuille A (2020) Can temporal information help with contrastive self-supervised learning? arXiv:2011.13046
152. Pan T, Song Y, Yang T, Jiang W, Liu W (2021) Videomoco: contrastive video representation learning with temporally adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11205–11214
153. Yang C, Xu Y, Dai B, Zhou B (2020) Video representation learning with visual tempo consistency. arXiv:2006.15489
154. Feichtenhofer C, Fan H, Malik J, He K (2019) Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6202–6211
155. Zhuang C, She T, Andonian A, Mark M S, Yamins D (2020) Unsupervised learning from video with deep neural embeddings. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9563–9572
156. Han T, Xie W, Zisserman A (2019) Video representation learning by dense predictive coding. In: Proceedings of the IEEE/CVF international conference on computer vision workshops
157. Han T, Xie W, Zisserman A (2020) Memory-augmented dense predictive coding for video representation learning. In: European conference on computer vision. Springer, pp 312–329
158. Lorre G, Rabarisoa J, Orcesi A, Ainouz S, Canu S (2020) Temporal contrastive pretraining for video action recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 662–670
159. Caron M, Bojanowski P, Joulin A, Douze M (2018) Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV), pp 132–149
160. Zhuang C, Zhai A L, Yamins D (2019) Local aggregation for unsupervised learning of visual embeddings. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6002–6012
161. Li J, Zhou P, Xiong C, Hoi SC (2020) Prototypical contrastive learning of unsupervised representations. arXiv:2005.04966
162. Hjelm RD, Bachman P (2020) Representation learning with video deep infomax. arXiv:2007.13278
163. Xue F, Ji H, Zhang W, Cao Y (2020) Self-supervised video representation learning by maximizing mutual information. Signal Process Image Commun 88:115967
164. Wang J, Jiao J, Liu Y-H (2020) Self-supervised video representation learning by pace prediction. In: European conference on computer vision. Springer, pp 504–521
165. Knights J, Harwood B, Ward D, Vanderkop A, Mackenzie-Ross O, Moghadam P (2021) Temporally coherent embeddings for self-supervised video representation learning. In: 2020 25th international conference on pattern recognition (ICPR). IEEE, pp 8914–8921
166. Yao T, Zhang Y, Qiu Z, Pan Y, Mei T (2021) Seco: exploring sequence supervision for unsupervised representation learning. In: AAAI, vol 2, p 7

167. Tao L, Wang X, Yamasaki T (2020) Self-supervised video representation learning using inter-intra contrastive framework. In: Proceedings of the 28th ACM international conference on multimedia, pp 2193–2201
168. Wang J, Gao Y, Li K, Jiang X, Guo X, Ji R, Sun X (2021) Enhancing unsupervised video representation learning by decoupling the scene and the motion. In: AAAI, vol 1, no. 2, p 7
169. Afouras T, Owens A, Chung JS, Zisserman A (2020) Self-supervised learning of audio-visual objects from video. In: European conference on computer vision. Springer, pp 208–224
170. Miech A, Alayrac J-B, Smaira L, Laptev I, Sivic J, Zisserman A (2020) End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9879–9889
171. Tokmakov P, Hebert M, Schmid C (2020) Unsupervised learning of video representations via dense trajectory clustering. In: European conference on computer vision. Springer, pp 404–421
172. Dunbar E, Karadayi J, Bernard M, Cao X-N, Algayres R, Ondel L, Besacier L, Sakti S, Dupoux E (2020) The zero resource speech challenge 2020: discovering discrete subword and word units. arXiv:2010.05967
173. Glass J (2012) Towards unsupervised speech processing. In: 2012 11th international conference on information science, signal processing and their applications (ISSPA). IEEE, pp 1–4
174. Schatz T (2016) Abx-discriminability measures and applications. Ph.D. Dissertation, Université Paris 6 (UPMC)
175. Dunbar E, Cao XN, Benjumea J, Karadayi J, Bernard M, Besacier L, Anguera X, Dupoux E (2017) The zero resource speech challenge 2017. In: 2017 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE, pp 323–330
176. Kawakami K, Wang L, Dyer C, Blunsom P, van der Oord A: Learning robust and multilingual speech representations. arXiv:2001.11128
177. Wang W, Tang Q, Livescu K (2020) Unsupervised pre-training of bidirectional speech encoders via masked reconstruction. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6889–6893
178. Heck M, Sakti S, Nakamura S (2017) Feature optimized DPGMM clustering for unsupervised subword modeling: A contribution to zerospeech 2017. In: 2017 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE, pp 740–746
179. Nandan A, Vepa J (2020) Language agnostic speech embeddings for emotion classification
180. Park DS, Chan W, Zhang Y, Chiu C-C, Zoph B, Cubuk ED, Le QV (2019) Specaugment: a simple data augmentation method for automatic speech recognition. arXiv:1904.08779
181. Shor J, Jansen A, Han W, Park D, Zhang Y (2021) Universal paralinguistic speech representations using self-supervised conformers. arXiv:2110.04621
182. Al-Tahan H, Mohsenzadeh Y (2021) Clar: contrastive learning of auditory representations. In: International conference on artificial intelligence and statistics. PMLR, pp 2530–2538
183. Saeed A, Grangier D, Zeghidour N (2021) Contrastive learning of general-purpose audio representations. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 3875–3879
184. Xia J, Wu L, Chen J, Hu B, Li SZ (2022) Simgrace: a simple framework for graph contrastive learning without data augmentation. arXiv:2202.03104
185. Wang T, Isola P (2020) Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. PMLR, pp 9929–9939
186. You Y, Chen T, Shen Y, Wang Z (2021) Graph contrastive learning automated. In: International conference on machine learning. PMLR, pp 12121–12132
187. Zeng J, Xie P (2020) Contrastive self-supervised learning for graph classification. arXiv:2009.05923
188. You Y, Chen T, Sui Y, Chen T, Wang Z, Shen Y (2020) Graph contrastive learning with augmentations. Adv Neural Inf Process Syst 33:5812–5823
189. Sun M, Xing J, Wang H, Chen B, Zhou J, "Mocl: Contrastive learning on molecular graphs with multi-level domain knowledge," arXiv preprint arXiv:2106.04509, (2021)
190. Sun F-Y, Hoffmann J, Verma V, Tang J (2019) Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. arXiv:1908.01000
191. Zhu Y, Xu Y, Yu F, Liu Q, Wu S, Wang L (2021) Graph contrastive learning with adaptive augmentation. Proc Web Conf 2021:2069–2080
192. Xia J, Wu L, Chen J, Wang G, Li SZ (2021) Debiased graph contrastive learning. arXiv:2110.02027
193. Alayrac J-B, Recasens A, Schneider R, Arandjelović R, Ramapuram J, De Fauw J, Smaira L, Dieleman S, Zisserman A (2020) Self-supervised multimodal versatile networks. Adv Neural Inf Process Syst 33:25–37
194. Liu Y, Yi L, Zhang S, Fan Q, Funkhouser T, Dong H (2020) P4contrast: contrastive learning with pairs of point-pixel pairs for RGB-D scene understanding. arXiv:2012.13089
195. Chuang C-Y, Robinson J, Lin Y-C, Torralba A, Jegelka S (2020) Debiased contrastive learning. Adv Neural Inf Process Syst 33:8765–8775
196. Ho C-H, Nvasconcelos N (2020) Contrastive learning with adversarial examples. Adv Neural Inf Process Syst 33:17081–17093
197. Tian Y, Sun C, Poole B, Krishnan D, Schmid C, Isola P (2020) What makes for good views for contrastive learning? Adv Neural Inf Process Syst 33:6827–6839
198. Wu M, Zhuang C, Mosse M, Yamins D, Goodman N (2020) On mutual information in contrastive learning for visual representations. arXiv:2005.13149
199. Asano Y, Patrick M, Rupprecht C, Vedaldi A (2020) Labelling unlabelled videos from scratch with multi-modal self-supervision. Adv Neural Inf Process Syst 33:4660–4671
200. Morgado P, Vasconcelos N, Misra I (2021) Audio-visual instance discrimination with cross-modal agreement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12475–12486
201. Patrick M, Asano YM, Kuznetsova P, Fong R, Henriques JF, Zweig G, Vedaldi A (2020) Multi-modal self-supervision from generalized data transformations. arXiv:2003.04298
202. Xiao F, Lee YJ, Grauman K, Malik J, Feichtenhofer C (2020) Audiovisual slowfast networks for video recognition. arXiv:2001.08740
203. Gan C, Huang D, Zhao H, Tenenbaum JB, Torralba A (2020) Music gesture for visual sound separation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10478–10487
204. Yang K, Russell B, Salamon J (2020) Telling left from right: learning spatial correspondence of sight and sound. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9932–9941
205. Lin Y-B, Tseng H-Y, Lee H-Y, Lin Y-Y, Yang M-H (2021) Unsupervised sound localization via iterative contrastive learning. arXiv:2104.00315
206. Nagrani A, Chung JS, Albanie S, Zisserman A (2020) Disentangled speech embeddings using cross-modal self-supervision. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6829–6833

207. Li B, Zhou H, He J, Wang M, Yang Y, Li L (2020) On the sentence embeddings from pre-trained language models. arXiv:2011.05864
208. Reimers N, Gurevych I (2019) Sentence-Bert: sentence embeddings using Siamese Bert-networks. arXiv:1908.10084
209. Jain P, Jain A, Zhang T, Abbeel P, Gonzalez JE, Stoica I (2020) Contrastive code representation learning. arXiv:2007.04973
210. Bui N D, Yu Y, Jiang L (2021) Self-supervised contrastive learning for code retrieval and summarization via semantic-preserving transformations. In: Proceedings of the 44th International ACM SIGIR conference on research and development in information retrieval, pp 511–521
211. Li Y, Hu P, Liu Z, Peng D, Zhou JT, Peng X (2021) Contrastive clustering. In: 2021 AAAI conference on artificial intelligence (AAAI)
212. Lin Y, Gou Y, Liu Z, Li B, Lv J, Peng X (2021) Completer: incomplete multi-view clustering via contrastive prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11174–11183
213. Pan E, Kang Z (2021) Multi-view contrastive graph clustering. Adv Neural Inf Process Syst 34
214. Trosten DJ, Lokse S, Jenssen R, Kampffmeyer M (2021) Reconsidering representation alignment for multi-view clustering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1255–1265
215. Wu L, Lin H, Tan C, Gao Z, Li SZ (2021) Self-supervised learning on graphs: contrastive, generative, or predictive. IEEE Trans Knowl Data Eng
216. Bhattacharjee A, Karami M, Liu H (2022) Text transformations in contrastive self-supervised learning: a review. arXiv:2203.12000
217. Albelwi S (2022) Survey on self-supervised learning: auxiliary pretext tasks and contrastive learning methods in imaging. Entropy 24(4):551
218. Stephane A-O, Frank R (2020) On losses for modern language models. arXiv:2010.01694