Heterogeneous Contrastive Learning for Foundation Models and Beyond

Lecheng Zheng
University of Illinois at Urbana-Champaign
Champaign, IL, USA
lecheng4@illinois.edu

Baoyu Jing
University of Illinois at Urbana-Champaign
Champaign, IL, USA
baoyu2@illinois.edu

Zihao Li
University of Illinois at Urbana-Champaign
Champaign, IL, USA
zihaoli5@illinois.edu

Hanghang Tong
University of Illinois at Urbana-Champaign
Champaign, IL, USA
htong@illinois.edu

Jingrui He
University of Illinois at Urbana-Champaign
Champaign, IL, USA
jingrui@illinois.edu

ABSTRACT

In the era of big data and Artificial Intelligence, an emerging paradigm is to utilize contrastive self-supervised learning to model large-scale heterogeneous data. Many existing foundation models benefit from the generalization capability of contrastive self-supervised learning by learning compact and high-quality representations without relying on any label information. Amidst the explosive advancements in foundation models across multiple domains, including natural language processing and computer vision, a thorough survey on heterogeneous contrastive learning for the foundation model is urgently needed. In response, this survey critically evaluates the current landscape of heterogeneous contrastive learning for foundation models, highlighting the open challenges and future trends of contrastive learning. In particular, we first present how the recent advanced contrastive learning-based methods deal with view heterogeneity and how contrastive learning is applied to train and fine-tune the multi-view foundation models. Then, we move to contrastive learning methods for task heterogeneity, including pretraining tasks and downstream tasks, and show how different tasks are combined with contrastive learning loss for different purposes. Finally, we conclude this survey by discussing the open challenges and shedding light on the future directions of contrastive learning.

CCS CONCEPTS

• **Computing methodologies → Machine learning approaches**; **Multi-task learning**; **Supervised learning**; **Unsupervised learning**.

KEYWORDS

Foundation Model; Contrastive Learning; Multi-view Learning; Multi-task Learning, Heterogeneous Learning

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08.
https://doi.org/10.1145/3637528.3671454

ACM Reference Format:
Lecheng Zheng, Baoyu Jing, Zihao Li, Hanghang Tong, and Jingrui He. 2024. Heterogeneous Contrastive Learning for Foundation Models and Beyond. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3671454

1 INTRODUCTION

Recent years have witnessed the rapid growth of the volume of big data. A Forbes report shows that the amount of newly created data in the past several years had increased by more than two trillion gigabytes[1]. One major characteristic of big data is heterogeneity [152]. Specifically, big data are usually collected from multiple sources and associated with various tasks, exhibiting view or task heterogeneity. For instance, in a social media platform, such as Facebook or Twitter, a post usually consists of a mixture of multiple types of data, such as a recorded video or several photos along with the text description. In the financial domain, taking the stock market for instance, the collected data may include not only the numerical values (e.g., stock price, the statistics from a company quarter report) but also some textual data conveying important information (e.g., a piece of news about a pharmaceutical company receiving the approval from Food and Drug Administration for its new product).

In response to the challenges posed by the exponential growth of big data, a promising approach is emerging by leveraging contrastive self-supervised learning to pre-train foundational models tailored for large-scale heterogeneous datasets. Recently, Contrastive Learning (CL) has gained an increasing interest in training foundation models [12, 36, 55], due to its good generalization capability and the independence of labeled data. Amidst the explosive advancements in foundation models across multiple domains, including natural language processing and computer vision, there is an urgent need for a comprehensive survey on heterogeneous contrastive learning for foundational models.

However, existing surveys on this topic are limited in scope and fail to systematically evaluate the most advanced techniques. Previous survey papers [2, 56, 60, 77, 88, 96, 172, 176, 198] mainly focus on investigating single heterogeneity [88, 176, 198] (e.g., view heterogeneity, task heterogeneity), contrastive learning [2, 56, 77, 96]

[1] https://www.forbes.com/sites/gilpress/2020/01/06/6-predictions-about-data-in-2020-and-the-coming-decade/?sh=3214c68f4fc3

		Research Topics and Coverage			
Learning Paradigms	Survey Papers	View Heterogeneity	Task/Label Heterogeneity	Contrastive Learning	Foundation Model
Non-Contrastive Learning	Li *et al.*, 2018 [88]	✓	✗	✗	✗
	Zhang *et al.*, 2018 [198]	✗	✓	✗	✗
	Yan *et al.*, 2021 [176]	✓	✗	✗	✗
	Jin *et al.*, 2023 [60]	✓	✗	✗	✓
	Xu *et al.*, 2024 [172]	✓	✗	✗	✓
Contrastive Learning	Jaiswa *et al.* 2020 [56]	✗	✗	✓	✗
	Le-Khac *et al.* 2020 [77]	✗	✗	✓	✗
	Liu *et al.*, 2023 [96]	✗	✗	✓	✗
	Albelwi 2022 [2]	✗	✓	✓	✗
	This survey	✓	✓	✓	✓

Table 1: Comparison with the existing related survey papers.

or multi-modal foundation model [60, 172]. The comparison of these surveys is summarized in Table 1. Specifically, [88, 176, 198] solely focus on heterogeneous machine learning, (e.g., multi-view learning, multi-label learning) and they do not cover any topic about contrastive learning and foundation model. [56, 77] discuss some contrastive learning methods at the early stage and they fail to include the most recent advanced techniques; [2, 96] investigate the recent advances in contrastive learning, but these two papers are only limited to summarizing the traditional contrastive learning methods. [60, 172] introduce the multi-modal foundation models, but their topics are only limited to multi-modal large language models. This survey critically evaluates the current landscape of heterogeneous contrastive learning for foundation models from both view and task heterogeneities, highlighting the open challenges and future trends of contrastive learning.

Our contributions are summarized as follows:

- **Categorization of Contrastive Foundation Models.** We systematically review the contrastive foundation models and categorize the existing methods into two branches, including the contrastive foundation models for view heterogeneity and task heterogeneity.
- **Systematic Review of Techniques.** We provide a comprehensive review of heterogeneous contrastive learning for foundation models. For both view heterogeneity and task heterogeneity, we summarize the representative methods and make necessary comparisons.
- **Future Directions.** We summarize four possible research directions on heterogeneous contrastive foundation models for future exploration.

This paper is organized as follows. In Section 2, we briefly review the basic concept of contrastive learning, and in Section 3, we first introduce the traditional multi-view contrastive learning model as the basis and then present the multi-view contrastive learning for large foundation models. In Section 4, we summarize contrastive learning methods for task heterogeneity, including pretraining tasks and downstream tasks, and show how different tasks are combined with contrastive learning loss for different purposes. In Section 5, we present several open future directions in contrastive learning before we conclude this survey paper in Section 6.

2 BASIC CONCEPT OF CONTRASTIVE LEARNING

Contrastive learning (CL) aims at learning the compact representation by contrasting the embeddings with one negative sample, following the idea of Noise Contrastive Estimation (NCE) [42]. Its pipeline is comprised of three stages, including augmentation, contrastive pair construction, and loss function formulation. In the first stage, many existing works either augment the raw data (i.e., data augmentation) or the embedding (i.e., embedding augmentation) to get an augmented sample and enrich the negative sets. In the second stage, researchers design how to construct the positive and negative pairs based on the different purposes or different settings (e.g., unsupervised contrastive loss [12] vs supervised contrastive loss [73]). The most commonly used contrastive pair construction considers that two samples augmented from the same raw data can form a positive pair and the rest of the samples are treated as negative samples. In the third stage, various types of contrastive learning losses are formulated based on different contrastive pair constructions, e.g., instance-level contrastive loss [167] , cluster-level contrastive loss [85] , contrastive alignment [145, 146], inter-view contrastive loss [91, 171], etc. A detailed discussion of the various types of loss formulation is shown in the next several subsections. In addition to these stages, many existing works [121, 124, 180] design a variety of CL strategies for pre-training tasks and downstream tasks. During pre-training, various characteristics of the data are injected into the models by pre-training tasks, including pretext tasks [12, 34, 51], supervised tasks [73, 119], preference tasks [22, 50] and auxiliary tasks [117, 182]. After pre-training, the models are fine-tuned to learn task-specific patterns of the downstream tasks, including automated machine learning [61, 125, 139], prompt learning [1, 157, 175], multi-task learning [116, 192], task reformulation [85, 122, 124], etc.

$\mathcal{D}$	The training dataset
$\boldsymbol{x}$	The input feature
z_i^+ (z_i^-)	A positive (negative) sample for z_i
$sim(\cdot)$	The similarity measurement function
τ	The temperature to scale the similarity measurement

Table 2: Main symbols and notation

Formally, CL loss follows the idea of NCE loss [42] by including more negative samples as follows:

$$\mathcal{L} = -\mathbb{E}_{x_i \in \mathcal{D}} \log \frac{\exp(sim(z_i, z_i^+))}{\exp(sim(z_i, z_i^+)) + \sum_{k \neq i} \exp(sim(z_i, z_k^-))} \quad (1)$$

where z_i is the learned representation of a input sample $\boldsymbol{x}_i$ from the dataset $\mathcal{D}$, z_i^+ is the representation of a positive sample similar to $\boldsymbol{x}_i$ and z_k^- is the representation of a negative sample dissimilar to $\boldsymbol{x}_i$. $sim(\cdot)$ denotes the similarity measurement functions, (e.g., $sim(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{a})^T \boldsymbol{b}/\tau$, where τ is the temperature). It constructs

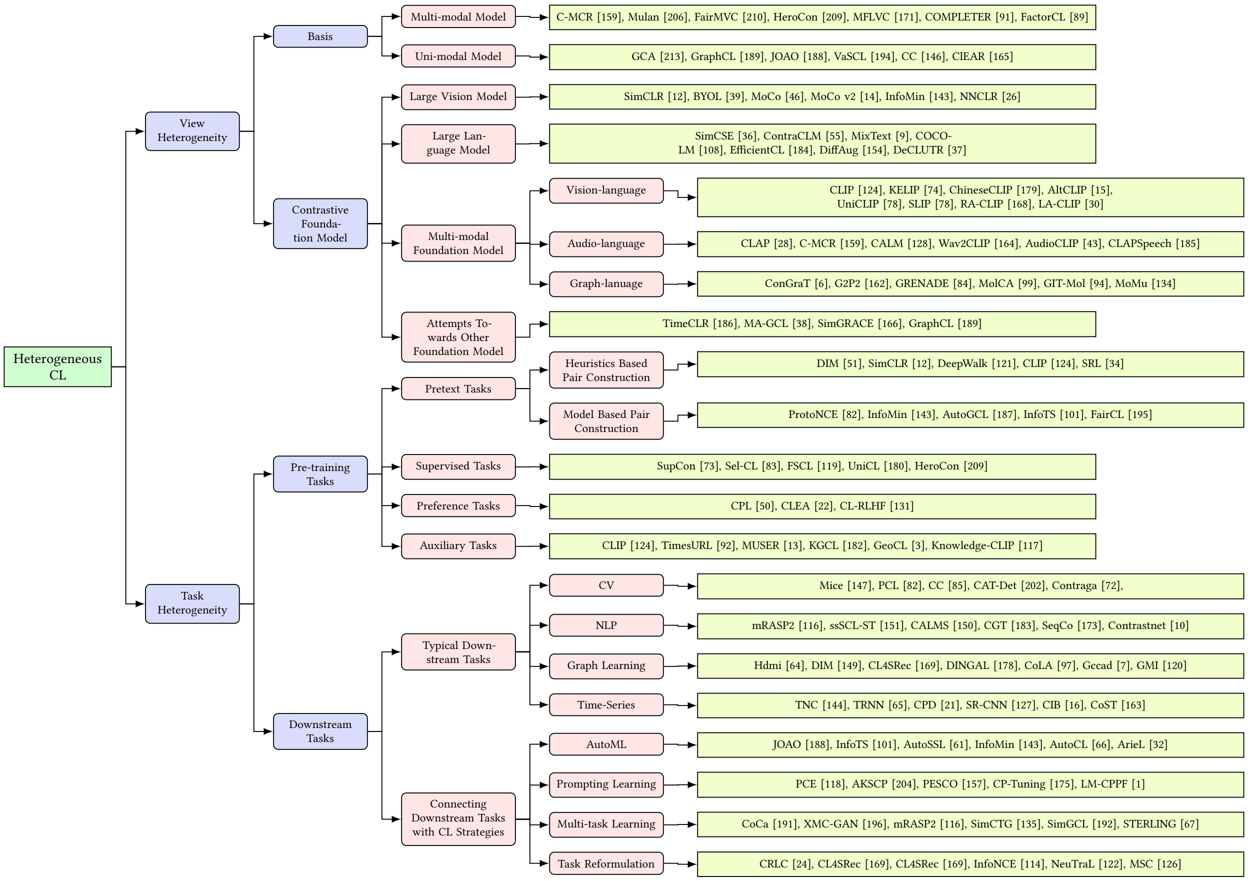

Figure 1: Taxonomy of this Survey with Representative Works.

a dataset $\mathcal{D}$ with n samples containing 1 positive samples and $n-1$ negative samples and then maximizes the similarities between z_i and z_i^+. Table 2 summarizes the symbols and their meanings. There are many other types of losses similar to CL loss, including InfoMax Based loss [51], triplet loss [129], etc. Specifically, [51] proposes to maximize the mutual information between the input feature $\boldsymbol{x}$ and the output of the encoder $\boldsymbol{z}$ (i.e., $\max \mathcal{I}(\boldsymbol{x}_i, z_i)$); [129] introduces the triplet loss to compare the representation of the anchor sample $\boldsymbol{x}_i$ with the positive and negative samples by $\mathcal{L} = \sum_i ||z_i - z_i^+||_2^2 - ||z_i - z_i^-||_2^2 + \alpha$, where α is a margin enforced between positive and negative pairs.

3 CONTRASTIVE LEARNING FOR VIEW HETEROGENEITY

In this section, we first present the basis of CL for view heterogeneity and then introduce traditional multi-view CL methods in different domains, including computer vision [89, 91, 159, 171] natural language processing [115, 194, 201], etc. Based on these traditional CL methods for view heterogeneity, we show how the researchers apply contrastive self-supervised learning to train the multi-view foundation models.

3.1 Basis of Contrastive Learning for View Heterogeneity

View heterogeneity refers to situations where data from different sources are available for training a model [207, 208, 212]. In CL, view heterogeneity can be categorized into two scenarios. In the first scenario, the raw data is unimodal or single-view (e.g., single-view image, text, or graph data), while in the second scenario, the raw data are collected from multiple data sources and the dataset naturally consists of multiple views (e.g., images with text descriptions in the social media). Different from InfoNCE [114] maximizing the input data $\boldsymbol{x}$ and its contextual information, CL for view heterogeneity aims to maximize the mutual information of multiple views of the same sample to extract the shared representations [142]. At the early stage, most CL methods tend to first use data augmentation methods to generate the augmented view and then apply CL [146, 167, 188, 189]. Differently, CMC [142] formally applies the idea of CL to handle the raw data with multiple views. Following CMC, various types of CL losses are proposed to model the multi-modal data, including inter-modality contrastive loss [89, 91, 159, 171, 205, 209, 210] , intra-modality contrastive loss [159, 209], contrastive alignment [115, 145, 146]. Unlike multi-modal data which naturally consists of various types of data, when handling single-view data such as images, text, and graphs, researchers often rely on data augmentation techniques to generate one or more augmented views for CL [146, 167, 188, 189]. Here, we characterize these view construction methods into two main categories: global and local. Global view construction involves augmenting samples globally, creating synthetic data akin to the original, commonly done through methods like random rotation and color jittering in computer vision [142, 146, 167], or graph-level augmentation in graph mining [189, 190, 213] , jitter-and-scale and permutation-and-jitter strategies for time-series data [27]. Conversely, local view construction focuses on augmenting samples locally or partially, often with specific purposes, such as image cropping in computer vision [142, 146, 167] or node-level augmentations and edge-level augmentations in graph mining [188, 189], sentence-level augmentation in NLP [165, 194], etc. For instance, DIM [51] proposes to maximize the mutual information between local representation and global representation. These approaches collectively enhance the diversity and richness of the data for effective CL.

3.2 Contrastive Learning for Foundation Model with View Heterogeneity

3.2.1 Large Vision Model. Different from many traditional CL methods, most of the large vision models [12, 14, 18, 20, 39, 46, 143] mainly implement the data augmentation methods to generate two augmented views and then apply CL to learn the representations. Specifically, SimCLR [12] achieves competitive performance on par with the supervised model after hundreds of iterations of fine-turning with 1% of the labeled data on the ImageNet dataset. Different from traditional CL methods (e.g., InfoNCE [114]), SimCLR introduces a learnable nonlinear transformation named *projection head* between the representation and the contrastive loss, and it contributes the success of the unsupervised CL large vision model to stronger data augmentation, normalized embeddings, an appropriately adjusted temperature parameter, larger batch sizes, longer training iterations, and deeper and wider networks. Despite the superiority of CL algorithms [12, 114] for many tasks, one major drawback of CL is its *high GPU memory requirement* as we need to increase the batch size to achieve better performance. To relax such a constraint, MoCo [46] stores the new encoded representations of the current batch in a dictionary and adopts a momentum-based moving average to maintain consistency for the newest and oldest representations. Chen *et al.* [14] verify the effectiveness of the projection head and stronger data augmentation proposed in SimCLR, showing further performance improvement. BYOL [39] trains the online network to predict the representation of the same image encoded by the target network under a different augmented view;

3.2.2 Large Language Model. CL with view heterogeneity is one of the most prevalent choices when pretraining large language models because of the scarcity of labeled data [199], by regarding augmented data as new views. A well-known pretraining technique that learns word embeddings through CL is word2vec [109]. Later, augmenting different views and conducting CL are used to pre-train state-of-the-art language models. SimCSE [36] and ContraCLM [55] adopt simple yet effective dropout-based augmentation. Shen *et al.* [130] propose to apply a cutoff for natural language augmentation to boost the model ability on both language understanding and generation. CERT [31] and MixText [9] create augmentations of original sentences using back-translation. DeCLUTR [37], which closely resembles QT [100], samples textual segments of the anchors up to paragraph length, allowing each sample to be as overlapping view, adjacent view or subsumed view with the anchor. CoDA [123] introduces contrast-enhanced and diversity-promoting data augmentation through a combination of back-translation, adversarial training, and label-preserving transformations. Additionally, CL with view heterogeneity has been widely used in other tasks. (a) *Fine-tuning.* Contrastive objectives have been used for language model fine-tuning [40, 104], and a recent work LM-CPPF [1] proposes to use few-shot paraphrasing for contrastive prompt-based fine-tuning. (b) *Machine Translation.* Different languages are naturally different views. mRASP2 [116] leverages CL with augmentations to align token representations and close the gap among representations of different languages. Li *et al.* [87] propose two-stage cross-lingual CL to improve word translation of language models. (c) *Other view heterogeneity-related tasks.* CALMS [150] leverages contrastive sentence ranking and sentence-aligned substitution to conduct multilingual text summarization.

3.2.3 Multi-modal Foundation Models. Multi-modal foundation models combine different modalities.

Vision-language model. The study of the vision-language models is evolving rapidly, and many surveys have provided comprehensive reviews from multiple perspectives. Mogadala *et al.* [111] survey common vision-language tasks, benchmark datasets, and seminal methods. Li *et al.* [80] first summarize the development of task-specific vision-language models, then review vision-language pretraining methods for general vision-language foundation models. Wang *et al.* [155] and Du *et al.* [25] share recent advances in vision-language model pretraining. Zhang *et al.* [197] review vision-language models specifically for various visual recognition tasks. In this work, we focus on leveraging heterogeneous CL in the pretraining phase of vision-language models. The seminal work in this group of models that uses heterogeneous CL, specifically cross-modal CL, is the well-known CLIP (Contrastive Language–Image Pre-training) [124]. By optimizing the contrastive loss from the language and vision views of over 400 million image-caption pairs, CLIP achieves strong performance in few-shot or zero-shot image classification settings. Another seminal work ALIGN [58] uses the same heterogeneous contrastive backbone over a larger noisy dataset. Following the impressive success of the image-text CL framework, many CLIP variants have been proposed. KELIP [74], ChineseCLIP [179] and AltCLIP [15] extend CLIP into other languages by leveraging the heterogeneous CL to fine-tune the CLIP model. DeCLIP [86] considers self-supervision together with image-text-pair supervision to achieve data-efficient training. RA-CLIP [168] and LA-CLIP [30] respectively introduce retrieval-augmented and LLM-augmented heterogeneous CL between images and texts.

Audio-language model. CLAP (Contrastive Language-Audio Pretraining) [28] imitates the process of CLIP to build an audio-language model that achieves state-of-the-art performance on multiple downstream tasks, even with much less training data compared to the vision-language domain. AudioCLIP [43] adds the audio modality to the two-modality CLIP through three two-view CL. Wav2CLIP [164] learns robust audio representations by projecting audio into a shared embedding space with images and text and distilling from CLIP through contrastive loss projection layers. C-MCR [159] offers a framework to efficiently train CLIP and CLAP by connecting the representation spaces. CALM [128] can efficiently bootstrap high-quality audio embedding by aligning audio representations to pretrained language representations and utilizing contrastive information between acoustic inputs.

Graph-language model. On text-attributed graphs, ConGraT [6] conducts CLIP-like contrastive pretraining for both language and graph tasks. G2P2 [162] extends the CLIP framework to text attributed graphs for zero-shot and few-shot classification. GRENADE [84] conducts graph-centric CL and knowledge alignment that considers neighborhood-level similarity to learn expressive and generalized representations. For text-paired graphs [60], MoleculeSTM [95] and MoMu [134] bridge molecular graphs and text data through contrastive learning. MolCA [99] adopts a cross-modal projector and uni-modal adapter to practically and efficiently understand molecular contents in both text and graph form. MolFM [103] leverages information from the input molecule structure, the input text description, and the auxiliary knowledge graph to build a multimodal molecular foundation model. GIT-Mol [94] incorporates cross-modal CL to build a multi-modal molecular foundation model with graphs, SMILES (Simplified Molecular Input Line Entry System), images, and text.

3.2.4 Other Foundation Models. Inspired by foundation models for language and vision data, recently, some attempts have been made to build foundation models for other data types. For time series, TimeCLR [186] develops a time series foundation model by leveraging CL to train unlabeled samples from multiple domains. For the graph data, due to the complex nature of graphs, we only find some initial attempts [189, 190] to develop graph foundation models and most of them follow the pertaining strategies of large language models [93, 107]. For instance, GraphCL [189] studies several intuitive augmentation strategies and proposes the initial framework for graph CL by maximizing the agreement of the augmented graphs in different views. Some later works [136, 190] follow this track and propose other kinds of augmentations. MA-GCL [38] and SimGRACE [166] propose that the heterogeneous views can also be generated from the neural architecture instead of the graph instances.

4 CONTRASTIVE LEARNING FOR TASK HETEROGENEITY

An ultimate goal of CL is to train foundation models to extract useful representations without human annotations. The foundation models are usually trained through a pre-training and fine-tuning paradigm. During pre-training, various characteristics of the data are injected into the models by pre-training tasks. After pre-training, the models are fine-tuned to learn task-specific patterns of the downstream tasks. In this section, we discuss the heterogeneous pre-training tasks and downstream tasks of CL.

4.1 Pre-training Tasks

Different pre-training tasks can guide models to capture different aspects of the data. In general, there are four types of pre-training tasks, including *pretext tasks* [34, 51, 211], *supervised tasks* [73, 119, 180], *preference tasks* [22, 50] and *auxiliary tasks* [3, 117, 182].

4.1.1 Pretext Tasks. The pretext tasks are the pre-training tasks without expensive and time-consuming human labels, and their objective is to discriminate positive and negative instance pairs, which are determined either by *heuristics* [12, 34, 51, 121, 124] or *extra models* [62, 82, 143, 195].

Heuristics Based Pair Construction. The heuristics-based methods construct contrastive pairs based on either simple relationships between instances or heuristically designed data augmentations. For example, DIM [51] treats a pair of local and global embeddings from the same image as positive and treats local and global embeddings from different images as negative. SimCLR [12] treats two different augmented views of an image as a positive pair and treats two randomly sampled images as a negative pair. DeepWalk [121] leverages random walks to determine positive node pairs in a graph. CLIP [124] treats ground-truth image and text pairs as positive and other image and text pairs as negative. SRL [34] regards two adjacent sub-sequences in the same time series as a positive pair and two sub-sequences from different time series as a negative pair.

Model Based Pair Construction. Model-based methods leverage extra models, e.g., *clustering*, *view generation* and *image editing* models, to generate contrastive pairs. For *clustering*, ProtoNCE [82] leverages external clustering methods, e.g., K-Means, to obtain semantic clusters, and uses the cluster centers to reduce the semantic errors of random negative sampling. X-GOAL [62] extends ProtoNCE to graphs. For *view generation*, InfoMin [143] leverages flow-based models [23] to generate augmented views for an input image, and treats these generated views as positive pairs. AutoGCL [187] and InfoTS [101] extend InfoMin to graphs and time series. For *image editing*, FairCL [195] trains an image editor [49] to generate images with different sensitive labels, e.g., gender. The images generated from the same input image but having different sensitive labels are regarded as positive pairs.

4.1.2 Supervised Tasks. The data for supervised pre-training tasks is manually labeled before pre-training the models, which incorporates human knowledge. SupCon [73] proposes to maximize the similarity of a pair of instances that share the same label. Sel-CL [83] proposes to filter out noisy labels by selecting confident examples based on their representation similarity with their labels. FSCL [119] introduces a Fair Supervised Contrastive Loss (FSCL) for visual representation learning based on SupCon, which defines the positive and negative pairs based on both class labels, e.g., attractiveness, and sensitive attribute labels, e.g., gender. UniCL [180] unifies (image, label) and (image, text) pairs by expanding the label into a textual description, and then leverages image-to-text and text-to-image contrastive losses to pre-train the model. HeroCon [209] proposes the weighted supervised contrastive loss to weight the importance of positive and negative pairs based on the similarity of different label vectors in the multi-label setting.

4.1.3 Preference Tasks. In recent years, Human-In-The-Loop (HITL) machine learning has become popular, which induces human prior knowledge into models by including humans in the training process. Different from supervised tasks, where humans first label the data and then the data is used to train the model, humans iteratively evaluate the quality of the prediction made by the model and provide feedback to the model to adjust its learned knowledge in HITL machine learning. [50] derives contrastive preference loss for learning optimal behavior from human feedback using the regret-based model of human preferences. [22] proposes to combine CL loss to model exploratory actions and learn user preferences utilizing the data collected from an interactive signal design process, where the data collection process can be regarded as the functionality of HITL. [131] introduces the contrastive rewards to penalize uncertainty and improve robustness based on human feedback.

4.1.4 Auxiliary Tasks. The auxiliary tasks leverage external or metadata information to improve CL. For example, Knowledge-CLIP

[117] uses knowledge graph to guide CLIP [124] to encode more precise semantics by tasks related to knowledge graphs e.g., link prediction. KGCL [182] introduces a Knowledge Graph CL framework (KGCL) for the recommendation, which leverages knowledge graphs to provide side information of items via a knowledge-aware co-CL task. GeoCL [3] induces geo-location information into image embeddings by classifying geo-labels. MUSER [13] uses text metadata, e.g., lyrics and album description, to learn better music sequence representations by aligning text tokens with music tokens as CLIP [124]. TimesURL [92] uses a reconstruction error to preserve important temporal variation information. Additionally, other methods directly use downstream tasks as auxiliary tasks [67, 116, 135, 191, 196], which can also be regarded as *multi-task learning* methods (see Sec. 4.2.2).

4.2 Downstream Tasks

The effectiveness of the CL methods is usually measured by their performance on a variety of downstream tasks. In this subsection, we first briefly review representative downstream tasks and then discuss how to connect downstream tasks with CL strategies.

4.2.1 Typical Downstream Tasks. We briefly review the typical downstream tasks for different fields.

Computer Vision. Typical tasks include image classification [12, 14, 46], image clustering [82, 85, 147], objective detection [4, 156, 202], image generation [72, 196], style transfer [8, 44], etc.

Natural Language Processing. Typical tasks include machine translation [52, 87, 115], text classification [10, 151, 181, 193], topic modeling [63, 112, 132, 148], text summarization [45, 69, 98, 150, 173], and information extraction [41, 81, 183].

Graph Learning. Typical tasks include node classification [64, 120, 188, 213], node clustering [62, 79, 158, 200, 203], graph classification [102, 137, 149, 170], link prediction [47, 68, 76, 133], recommendation [11, 59, 67, 161], knowledge graph reasoning [138, 174, 178] and anomaly detection [7, 97, 105].

Time Series Analysis. Typical tasks include classification [27, 34, 53, 144], forecasting [65, 70, 101, 163], anomaly detection [5, 21, 127, 153] and imputation [16, 71, 92, 141].

4.2.2 Connecting Downstream Tasks with CL strategies. Different CL strategies, e.g., different views and pre-training tasks, usually have disparate impact on downstream tasks [101, 143, 186, 188]. Therefore, a fundamental challenge to train foundation models lies in how to construct suitable CL strategies for the desired downstream tasks. Given a downstream task and a set of available CL strategies, *Automated Machine Learning (AutoML)* [19, 101, 140, 143, 187, 214] and *prompt learning* [1, 157, 175, 215] methods could be used to discover the optimal CL strategies. Given the optimal CL strategies, one could either use these strategies to pre-train the model and then fine-tune on the downstream tasks, or train the model via *multi-task learning* [67, 135, 191, 192] by combining the CL strategies with downstream tasks. Additionally, some works also try to *reformulate* [24, 47, 73, 85, 138, 169] the downstream tasks as CL tasks since they are inherently related.

Automated Machine Learning. Being geared towards automating the procedure of machine learning, AutoML has gained a lot of attention in recent years [48]. AutoML methods formulate the problem of searching for the optimal CL strategies as a bi-level optimization problem [19, 61, 188]:

$$s^* = \arg\max \mathcal{R}(f_{\theta^*}, s) \; s.t. \; \theta^* = \arg\min \mathcal{L}(f_\theta, s) \tag{2}$$

where the lower-level problem is to minimize the loss $\mathcal{L}$ (e.g., cross-entropy loss) of the downstream task (e.g., classification) or a surrogate task (e.g., minimization of mutual information [143]) for the given model f_θ and the CL strategy s; the upper-level problem is to maximize the validation reward $\mathcal{R}$ (e.g., accuracy) for the pair of the trained model and CL strategy (f_{θ^*}, s). Existing AutoML methods can be categorized from two perspectives: *search space* and *search algorithms*. In terms of the search space, existing methods mainly focus on *data augmentations* [19, 125, 139, 188], *view constructions* [32, 101, 140, 143, 187], *pretext tasks* [61] and *overall CL strategies* [66]. For *data augmentations*, JOAO [188] could automatically select the most challenging data augmentation pairs for graph data based on the current contrastive loss. For *view constructions*, InfoMin [143] argues that good contrastive views should retain information relevant to downstream tasks while minimizing irrelevant nuisances, which constructs the optimal views by minimizing the mutual information between different views. For *pretext tasks*, AutoSSL [61] automatically searches for the optimal combination of pretext tasks for node clustering and node classification for graphs. For *overall CL strategies*, AutoCL [66] searches for all aspects of CL, including data augmentations, embedding augmentations, contrastive pair construction and loss functions for time series. In terms of the search algorithms, existing methods are mainly based on *reinforcement learning* [19, 66, 214], *adversarial learning* [33, 101, 139, 140, 143, 187, 188], *evolution strategy* [61] and *Bayesian optimization* [125]. For *reinforcement learning*, AutoCL [66] uses a controller network to sample CL strategies and uses the model's performance on the validation set to design reward, where the controller is optimized by maximizing the reward $\mathcal{R}$ via reinforcement learning. For *adversarial learning*, AD-GCL [139] measures the similarity of the node embeddings of two different graph views via mutual information, which formulates the lower-level and upper-level problems in Equation (2) as maximizing and minimizing the mutual information respectively. ARIEL [32] uses the same contrastive loss for both lower-level loss $\mathcal{L}$ and the upper-level reward $\mathcal{R}$, and uses the adversarial attack to maximize $\mathcal{R}$. For *evolution strategy*, For *Bayesian optimization*, SelfAugment [125] leverages image rotation prediction as the lower-level task, and uses Bayesian optimization [90] as the search algorithm to obtain the optimal data augmentations.

Prompt Learning. The contrastive-based prompt learning methods aim to combine CL with prompt learning for various purposes, such as maximizing the consistency of different representations [215], enabling fine-tuning in few-shot or zero-shot setting [1, 157, 204], and commonsense reasoning [118]. Specifically, [215] designs a multimodal prompt transformer to perform cross-modal information fusion and apply CL to maximize the consistency among the fused representation and the representation for each modality for the emotion recognition task. [1, 157, 204] combine CL with prompt learning to fine-tune the model in the few-shot or zero-shot setting. [17] devises visual prompt-based CL and guided-attention-based prompt ensemble algorithms to task-learn specific state representations from multiple prompted embeddings.

Multi-Task Learning. When we have prior knowledge about what characteristics of the data can be brought by CL strategies, in addition to the pre-training and fine-tuning paradigm, another simple yet effective way to connect downstream tasks and CL strategies is to combine the pre-training tasks and downstream tasks as a multi-task learning task. For example, CoCa [191] combines a contrastive loss with a caption generation loss to train image-text foundation models, where contrastive loss is used to learn global representations and captioning is used to learn fine-grained region-level features. XMC-GAN [196] leverages contrastive losses for various pairs, such as (image, sentence) and (generated image, real image), to improve the alignment between them for the text-to-image generation task. mRASP2 [116] combines a contrastive loss with cross-entropy for multilingual machine translation, where the contrastive loss is adopted to minimize the representation gap of similar sentences and maximize that of irrelevant sentences. SimCTG [135] leverages a contrastive loss to encourage language models to learn discriminative and isotropic token representations for neural text generation. SimGCL [192] discovers that InfoNCE loss helps models learn more evenly distributed user and item embeddings, which could mitigate the popularity bias.

Task Reformulation. Certain downstream tasks are inherently related to CL, such as *classification*, *clustering*, *link prediction*, *recommendation*, *anomaly detection* and *reinforcement learning*. Therefore, the loss functions of these downstream tasks can be reformulated as a contrastive loss. For example, in terms of *classification*, the previously mentioned SupCon [73] integrates image class labels into self-supervised contrastive losses and proposes a Supervised Contrastive (SupCon) loss. CLIP [124] reformulates the image classification task as an image-text alignment contrastive task. For *clustering*, CC [85] introduces a CL-based clustering objective function, called contrastive clustering, by regarding the embedding vector of an instance as the soft cluster labels. CRLC [24] reformulates the objective of clustering as a probability contrastive loss, which trains the parametric clustering classifier by contrasting positive and negative cluster probability pairs. For *link prediction*, since it is inherently a contrastive task: determining whether a pair of nodes is positive or not, most of the existing methods directly adopt CL losses as the objective functions to train the models [57]. For example, RotatE [138] trains knowledge graph link prediction models by the negative sampling loss [110]. For *recommendation*, it can be regarded as a link prediction task with ranking, and thus the training objectives are usually variants of contrastive losses [47, 177]. For example, CL4SRec [169] formulates the objective function of recommendation as a variant of InfoNCE [114]. For *anomaly detection*, NeuTraL [122] directly adopts a contrastive loss as the loss function as well as the anomaly score. For *reinforcement learning*, Contrastive RL [29] uses CL to directly perform goal-conditioned reinforcement learning by leveraging CL to estimate the Q-function for a certain policy and reward functions.

5 FUTURE DIRECTIONS

The past years have witnessed the rapid development of heterogeneous CL on foundation models. Building upon such progress, it opens the door to many exciting future opportunities to explore in this emerging area. Here, we summarize five promising future directions, focusing on contrastive foundation models.

Representation Redundancy and Uniqueness for CL Model. The current CL models mainly extract the shared representation by maximizing the similarity of two views for the same sample. However, some recent works [89, 206] have suggested the potential of extracting uniqueness via CL to improve the performance of the downstream tasks. However, the initial methods are only used to deal with the small model, and how to naturally combine it with foundation models remains a great challenge due to the extra computational cost and limited performance improvement.

Efficiency of CL Foundation Models. One major issue of training the foundation models with CL loss is the high GPU memory requirement as discussed in Section 3.2.1. Recently, Zeroth order optimization methods [106] have shown great potential to alleviate the computational cost by replacing the traditional forward-passing and backward-passing optimization scheme with a forward-passing-only optimization scheme. However, the zeroth order optimizer usually sacrifices the optimization efficiency for lower GPU requirements, as it requires significantly more steps than standard fine-tuning [106]. Efficiently training or fine-tuning a CL-based foundation model remains a great challenge.

Better Multi-view Benchmark Datasets for CL Models. Currently, high-quality multi-view benchmark datasets are urgently needed for constructing multi-modal foundation models. While many large-scale text-attributed graphs are collected from social media, e-commerce platforms, and academic domains [60], it's essential to acknowledge that real-world graphs span various domains, including finance, healthcare, transportation networks, and local infrastructure networks. Similarly, there is a high demand for large-scale text-image datasets to support the training of vision-language foundational models [160]. Moreover, concerns have been raised regarding potential biases present in many benchmark datasets. Studies [75, 113] have highlighted the existence of contextual, demographic, and stereotypical biases within benchmark datasets used for large language models

Trustworthy CL. Trustworthy machine learning refers to the development and deployment of machine learning models with a strong emphasis on interpretability, fairness, transparency, privacy, and robustness. While significant strides have been made in enhancing these aspects by CL-based regularization, inlcuding interpretability [35, 54], fairness considerations [195, 210], and out-of-distribution robustness [104, 206], these efforts are still in their nascent stages, e.g., training models on small datasets or failing to consider the view or task heterogeneity. Despite these early efforts, heterogeneous contrastive foundational models still encounter challenges related to interpretability, fairness, transparency, privacy, and robustness, which persist across multi-modal foundational models as well.

Understanding Mechanisms Between CL Strategies and Downstream Tasks. As introduced in Section 4.2, we present various CL strategies for downstream tasks. However, it remains unclear which CL strategies are good for specific downstream tasks and how can we evaluate the quality of CL strategies. In addition, how different CL strategies compete and cooperate in downstream tasks is expected to be better evaluated and understood by the researchers. Moreover, how to combine CL with other self-supervised methods to further improve the performance of the foundation models deserves great attention.

6 CONCLUSION

This paper provides a thorough exploration of heterogeneous CL for foundation models. We first delve into the traditional CL methods, particularly in addressing view heterogeneity, and elucidate the application of CL techniques in training and fine-tuning multi-view foundation models. Subsequently, we discuss CL methods tailored to tackle task heterogeneity, including pretraining and downstream tasks, and illustrate how CL combines different tasks for various objectives. Finally, we outline potential future research directions in heterogeneous CL for foundation models.

ACKNOWLEDGMENTS

This work is supported by National Science Foundation (IIS-2002540 and 2134079), the C3.ai Digital Transformation Institute, and IBM-Illinois Discovery Accelerator Institute - a new model of an academic-industry partnership designed to increase access to technology education and skill development to spur breakthroughs in emerging areas of technology. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

REFERENCES

[1] Amirhossein Abaskohi, Sascha Rothe, and Yadollah Yaghoobzadeh. 2023. LM-CPPF: Paraphrasing-Guided Data Augmentation for Contrastive Prompt-Based Few-Shot Fine-Tuning. In *ACL*. ACL, 670–681.

[2] Saleh Albelwi. 2022. Survey on Self-Supervised Learning: Auxiliary Pretext Tasks and Contrastive Learning Methods in Imaging. *Entropy* 24, 4 (2022), 551.

[3] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. 2021. Geography-aware self-supervised learning. In *ICCV*.

[4] Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. 2022. Contrastive Learning for Unsupervised Video Highlight Detection. In *CVPR*. IEEE, 14022–14032.

[5] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. 2021. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)* 54, 3 (2021), 1–33.

[6] William Brannon, Suyash Fulay, Hang Jiang, Wonjune Kang, Brandon Roy, Jad Kabbara, and Deb Roy. 2023. ConGraT: Self-Supervised Contrastive Pretraining for Joint Graph and Text Embeddings. *CoRR* abs/2305.14321 (2023).

[7] Bo Chen, Jing Zhang, Xiaokang Zhang, Yuxiao Dong, Jian Song, Peng Zhang, Kaibo Xu, Evgeny Kharlamov, and Jie Tang. 2022. Gccad: Graph contrastive learning for anomaly detection. *TKDE* (2022).

[8] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021. Artistic Style Transfer with Internal-external Learning and Contrastive Learning. In *NeurIPS*. 26561–26573.

[9] Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. In *ACL*. ACL, 2147–2157.

[10] Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. 2022. Contrastnet: A contrastive learning framework for few-shot text classification. In *AAAI*, Vol. 36. 10492–10500.

[11] Mengru Chen, Chao Huang, Lianghao Xia, Wei Wei, Yong Xu, and Ronghua Luo. 2023. Heterogeneous graph contrastive learning for recommendation. In *WSDM*. 544–552.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*. PMLR, 1597–1607.

[13] Tianyu Chen, Yuan Xie, Shuai Zhang, Shaohan Huang, Haoyi Zhou, and Jianxin Li. 2022. Learning music sequence representation from text supervision. In *ICASSP*. IEEE, 4583–4587.

[14] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020. Improved Baselines with Momentum Contrastive Learning. *CoRR* abs/2003.04297 (2020).

[15] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. 2023. AltCLIP: Altering the Language Encoder in CLIP for Extended Language Capabilities. In *ACL*. ACL, 8666–8682.

[16] MinGyu Choi and Changhee Lee. 2023. Conditional Information Bottleneck Approach for Time Series Imputation. In *ICLR*.

[17] Wonje Choi, Woo Kyung Kim, Seunghyun Kim, and Honguk Woo. 2023. Efficient Policy Adaptation with Contrastive Prompt Ensemble for Embodied Agents. In *NeurIPS*.

[18] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge J. Belongie. 2022. When Does Contrastive Visual Representation Learning Work?. In *CVPR*. IEEE, 1–10.

[19] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2018. Autoaugment: Learning augmentation policies from data. *ArXiv:1805.09501* (2018).

[20] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. 2021. Parametric Contrastive Learning. In *ICCV*. IEEE, 695–704.

[21] Shohreh Deldari, Daniel V Smith, Hao Xue, and Flora D Salim. 2021. Time series change point detection with self-supervised contrastive predictive coding. In *WWW*. 3124–3135.

[22] Nathaniel Steele Dennler, Stefanos Nikolaidis, and Maja Mataric. 2024. Using Exploratory Search to Learn Representations for Human Preferences. In *HRI*. 392–396.

[23] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2016. Density estimation using Real NVP. In *ICLR*.

[24] Kien Do, Truyen Tran, and Svetha Venkatesh. 2021. Clustering by maximizing mutual information across views. In *ICCV*. 9928–9938.

[25] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A Survey of Vision-Language Pre-Trained Models. In *IJCAI*.

[26] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2021. With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations. In *ICCV*. IEEE, 9568–9577.

[27] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2021. Time-Series Representation Learning via Temporal and Contextual Contrasting. In *IJCAI*. ijcai.org, 2352–2359.

[28] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP Learning Audio Concepts from Natural Language Supervision. In *ICASSP*. IEEE, 1–5.

[29] Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. 2022. Contrastive learning as goal-conditioned reinforcement learning. *NeurIPS* 35 (2022), 35603–35620.

[30] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving CLIP Training with Language Rewrites. In *NeurIPS*.

[31] Hongchao Fang and Pengtao Xie. 2020. CERT: Contrastive Self-supervised Learning for Language Understanding. *CoRR* abs/2005.12766 (2020).

[32] Shengyu Feng, Baoyu Jing, Yada Zhu, and Hanghang Tong. 2022. Adversarial graph contrastive learning with information regularization. In *WWW*. 1362–1371.

[33] Shengyu Feng, Baoyu Jing, Yada Zhu, and Hanghang Tong. 2024. Ariel: Adversarial graph contrastive learning. *TKDD* (2024).

[34] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. 2019. Unsupervised scalable representation learning for multivariate time series. *NeurIPS* 32 (2019).

[35] Takanori Fujiwara, Jian Zhao, Francine Chen, Yaoliang Yu, and Kwan-Liu Ma. 2020. Interpretable Contrastive Learning for Networks. *CoRR* abs/2005.12419 (2020).

[36] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*. 6894–6910.

[37] John M. Giorgi, Osvald Nitski, Bo Wang, and Gary D. Bader. 2021. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. In *ACL*. ACL, 879–895.

[38] Xumeng Gong, Cheng Yang, and Chuan Shi. 2023. MA-GCL: Model Augmentation Tricks for Graph Contrastive Learning. In *AAAI*. AAAI Press, 4284–4292.

[39] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *NeurIPS*.

[40] Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. In *ICLR*.

[41] Xinnan Guo, Wentao Deng, Yongrui Chen, Yang Li, Mengdi Zhou, Guilin Qi, Tianxing Wu, Dong Yang, Liubin Wang, and Yong Pan. 2023. CoMave: Contrastive Pre-training with Multi-scale Masking for Attribute Value Extraction. In *ACL*. ACL, 6007–6018.

[42] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*.

[43] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending Clip to Image, Text and Audio. In *ICASSP*. IEEE, 976–980.

[44] Yucheng Hang, Bin Xia, Wenming Yang, and Qingmin Liao. 2022. SCS-Co: Self-Consistent Style Contrastive Learning for Image Harmonization. In *CVPR*. IEEE, 19678–19687.

[45] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and Attend: Multimodal Summarization with Dual Contrastive Losses. In *CVPR*.

[46] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*. 9729–9738.

[47] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. 173–182.
[48] Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. AutoML: A survey of the state-of-the-art. *Knowledge-based systems* 212 (2021), 106622.
[49] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2019. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing* 28, 11 (2019), 5464–5478.
[50] Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. 2023. Contrastive Preference Learning: Learning from Human Feedback without RL. *CoRR* abs/2310.13639 (2023).
[51] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *ICLR*.
[52] Sathish Indurthi, Shamil Chollampatt, Ravi Agrawal, and Marco Turchi. 2023. CLAD-ST: Contrastive Learning with Adversarial Data for Robust Speech Translation. In *EMNLP*. ACL, 9049–9056.
[53] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *DMKD* 33, 4 (2019), 917–963.
[54] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive Explanations for Model Interpretability. In *EMNLP*. ACL, 1597–1611.
[55] Nihal Jain, Dejiao Zhang, Wasi Uddin Ahmad, Zijian Wang, Feng Nan, Xiaopeng Li, Ming Tan, Ramesh Nallapati, Baishakhi Ray, Parminder Bhatia, Xiaofei Ma, and Bing Xiang. 2023. ContraCLM: Contrastive Learning For Causal Language Model. In *ACL*. ACL, 6436–6459.
[56] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2020), 2.
[57] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems* (2021).
[58] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML*.
[59] Yangqin Jiang, Chao Huang, and Lianghao Huang. 2023. Adaptive graph contrastive learning for recommendation. In *SIGKDD*. 4252–4261.
[60] Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2023. Large Language Models on Graphs: A Comprehensive Survey. *CoRR* abs/2312.02783 (2023).
[61] Wei Jin, Xiaorui Liu, Xiangyu Zhao, Yao Ma, Neil Shah, and Jiliang Tang. 2021. Automated Self-Supervised Learning for Graphs. In *ICLR*.
[62] Baoyu Jing, Shengyu Feng, Yuejia Xiang, Xi Chen, Yu Chen, and Hanghang Tong. 2022. X-GOAL: Multiplex heterogeneous graph prototypical contrastive learning. In *CIKM*.
[63] Baoyu Jing, Chenwei Lu, Deqing Wang, Fuzhen Zhuang, and Cheng Niu. 2018. Cross-domain labeled LDA for cross-domain text classification. In *ICDM*. IEEE, 187–196.
[64] Baoyu Jing, Chanyoung Park, and Hanghang Tong. 2021. Hdmi: High-order deep multiplex infomax. In *WWW 2021*.
[65] Baoyu Jing, Hanghang Tong, and Yada Zhu. 2021. Network of tensor time series. In *WWW*. 2425–2437.
[66] Baoyu Jing, Yansen Wang, Guoxin Sui, Jing Hong, Jingrui He, Yuqing Yang, Dongsheng Li, and Kan Ren. 2024. Automated Contrastive Learning Strategy Search for Time Series. arXiv:2403.12641
[67] Baoyu Jing, Yuchen Yan, Kaize Ding, Chanyoung Park, Yada Zhu, Huan Liu, and Hanghang Tong. 2024. Sterling: Synergistic representation learning on bipartite graphs. *AAAI*.
[68] Baoyu Jing, Yuchen Yan, Yada Zhu, and Hanghang Tong. 2022. Coin: Co-cluster infomax for bipartite graphs. In *NeurIPS 2022 Workshop*.
[69] Baoyu Jing, Zeyu You, Tao Yang, Wei Fan, and Hanghang Tong. 2021. Multiplex Graph Neural Network for Extractive Text Summarization. In *EMNLP*. 133–139.
[70] Baoyu Jing, Si Zhang, Yada Zhu, Bin Peng, Kaiyu Guan, Andrew Margenot, and Hanghang Tong. 2022. Retrieval based time series forecasting. *ArXiv:2209.13525*.
[71] Baoyu Jing, Dawei Zhou, Kan Ren, and Carl Yang. 2024. CASPER: Causality-Aware Spatiotemporal Graph Neural Networks for Spatiotemporal Time Series Imputation. *ArXiv:2403.11960* (2024).
[72] Minguk Kang and Jaesik Park. 2020. Contragan: Contrastive learning for conditional image generation. *NeurIPS* 33 (2020), 21357–21369.
[73] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. *NeurIPS*.
[74] ByungSoo Ko and Geonmo Gu. 2022. Large-scale Bilingual Language-Image Contrastive Learning. *CoRR* abs/2203.14463 (2022).
[75] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*. 12–24.
[76] Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, and Bhaskar Biswas. 2020. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications* 553 (2020), 124289.
[77] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access* 8 (2020), 193907–193934.
[78] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. 2022. UniCLIP: Unified Framework for Contrastive Language-Image Pre-training. In *NeurIPS*.
[79] Bolian Li, Baoyu Jing, and Hanghang Tong. 2022. Graph communal contrastive learning. In *WWW 2022*.
[80] Feng Li, Hao Zhang, Yi-Fan Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, PengChuan Zhang, and Lei Zhang. 2022. Vision-Language Intelligence: Tasks, Representation Learning, and Large Models. *CoRR* abs/2203.01922 (2022).
[81] Jiaqi Li, Chuanyi Zhang, Miaozeng Du, Dehai Min, Yongrui Chen, and Guilin Qi. 2023. Three Stream Based Multi-level Event Contrastive Learning for Text-Video Event Extraction. In *EMNLP*. ACL, 1666–1676.
[82] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. 2020. Prototypical Contrastive Learning of Unsupervised Representations. In *ICLR*.
[83] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. 2022. Selective-supervised contrastive learning with noisy labels. In *CVPR*. 316–325.
[84] Yichuan Li, Kaize Ding, and Kyumin Lee. 2023. GRENADE: Graph-Centric Language Model for Self-Supervised Representation Learning on Text-Attributed Graphs. In *EMNLP*.
[85] Yunfan Li, Peng Hu, Jerry Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021. Contrastive Clustering. In *AAAI*. AAAI Press, 8547–8555.
[86] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In *ICLR*.
[87] Yaoyiran Li, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulic. 2022. Improving Word Translation via Two-Stage Contrastive Learning. In *ACL*. ACL, 4353–4374.
[88] Yingming Li, Ming Yang, and Zhongfei Zhang. 2018. A survey of multi-view representation learning. *IEEE TKDE*.
[89] Paul Pu Liang, Zihao Deng, Martin Q. Ma, James Y. Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2023. Factorized Contrastive Learning: Going Beyond Multi-view Redundancy. In *NeurIPS*.
[90] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. 2019. Fast autoaugment. *NeurIPS* 32 (2019).
[91] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. 2021. COMPLETER: Incomplete Multi-View Clustering via Contrastive Prediction. In *CVPR*.
[92] Jiexi Liu and Songcan Chen. 2023. TimesURL: Self-supervised Contrastive Learning for Universal Time Series Representation Learning. *ArXiv:2312.15709* (2023).
[93] Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S. Yu, and Chuan Shi. 2023. Towards Graph Foundation Models: A Survey and Beyond. *CoRR* abs/2310.11829 (2023).
[94] Pengfei Liu, Yiming Ren, and Zhixiang Ren. 2023. GIT-Mol: A Multi-modal Large Language Model for Molecular Science with Graph, Image, and Text. *CoRR* abs/2308.06911 (2023).
[95] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nat. Mac. Intell.* 5, 12 (2023), 1447–1457.
[96] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2023. Self-Supervised Learning: Generative or Contrastive. *TKDE* 35, 1 (2023), 857–876.
[97] Yixin Liu, Zhao Li, Shirui Pan, Chen Gong, Chuan Zhou, and George Karypis. 2021. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE transactions on neural networks and learning systems* 33, 6 (2021), 2378–2392.
[98] Yixin Liu and Pengfei Liu. 2021. SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization. In *ACL*. 1065–1072.
[99] Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023. MolCA: Molecular Graph-Language Modeling with Cross-Modal Projector and Uni-Modal Adapter. In *EMNLP*. ACL, 15623–15638.
[100] Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *ICLR*.
[101] Dongsheng Luo, Wei Cheng, Yingheng Wang, Dongkuan Xu, Jingchao Ni, Wenchao Yu, Xuchao Zhang, Yanchi Liu, Yuncong Chen, Haifeng Chen, et al. 2023. Time series contrastive learning with information-aware augmentations. In *AAAI*, Vol. 37. 4534–4542.
[102] Xiao Luo, Wei Ju, Meng Qu, Chong Chen, Minghua Deng, Xian-Sheng Hua, and Ming Zhang. 2022. Dualgraph: Improving semi-supervised graph classification via dual contrastive learning. In *ICDE*. IEEE, 699–712.
[103] Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023. MolFM: A Multimodal Molecular Foundation Model. *CoRR* abs/2307.09484 (2023).

[104] Xiaofei Ma, Cícero Nogueira dos Santos, and Andrew O. Arnold. 2021. Contrastive Fine-tuning Improves Robustness for Neural Rankers. In *ACL*. ACL, 570–582.

[105] Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z Sheng, Hui Xiong, and Leman Akoglu. 2021. A comprehensive survey on graph anomaly detection with deep learning. *TKDE* 35, 12 (2021), 12012–12038.

[106] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. 2024. Fine-tuning language models with just forward passes. *NeurIPS* 36 (2024).

[107] Haitao Mao, Zhikai Chen, Wenzhuo Tang, Jianan Zhao, Yao Ma, Tong Zhao, Neil Shah, Mikhail Galkin, and Jiliang Tang. 2024. Graph Foundation Models. *CoRR* abs/2402.02216 (2024).

[108] Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining. In *NeurIPS*. 23102–23114.

[109] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

[110] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *NeurIPS* 26 (2013).

[111] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2021. Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods. *J. Artif. Intell. Res.* 71 (2021), 1183–1317.

[112] Thong Nguyen and Anh Tuan Luu. 2021. Contrastive learning for neural topic model. *NeurIPS* 34 (2021), 11974–11986.

[113] Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. 2023. In-contextual bias suppression for large language models. *ArXiv:2309.07251* (2023).

[114] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv:1807.03748* (2018).

[115] Siqi Ouyang, Rong Ye, and Lei Li. 2023. WACO: Word-Aligned Contrastive Learning for Speech Translation. In *ACL*. ACL, 3891–3907.

[116] Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive Learning for Many-to-many Multilingual Neural Machine Translation. In *ACL*. ACL, 244–258.

[117] Xuran Pan, Tianzhu Ye, Dongchen Han, Shiji Song, and Gao Huang. 2022. Contrastive language-image pre-training with knowledge graphs. *NeurIPS* 35 (2022), 22895–22910.

[118] Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting Contrastive Explanations for Commonsense Reasoning Tasks. In *ACL*. ACL, 4179–4192.

[119] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. 2022. Fair contrastive learning for facial attribute classification. In *CVPR*. 10389–10398.

[120] Zhen Peng, Wenbing Huang, Minnan Luo, Qinghua Zheng, Yu Rong, Tingyang Xu, and Junzhou Huang. 2020. Graph representation learning via graphical mutual information maximization. In *WWW 2020*. 259–270.

[121] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *SIGKDD*. 701–710.

[122] Chen Qiu, Timo Pfrommer, Marius Kloft, Stephan Mandt, and Maja Rudolph. 2021. Neural transformation learning for deep anomaly detection beyond images. In *ICML*. PMLR, 8703–8714.

[123] Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Weizhu Chen, and Jiawei Han. 2021. CoDA: Contrast-enhanced and Diversity-promoting Data Augmentation for Natural Language Understanding. In *ICLR*.

[124] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, Vol. 139. PMLR, 8748–8763.

[125] Colorado J Reed, Sean Metzger, Aravind Srinivas, Trevor Darrell, and Kurt Keutzer. 2021. Selfaugment: Automatic augmentation policies for self-supervised learning. In *CVPR*. 2674–2683.

[126] Tal Reiss and Yedid Hoshen. 2023. Mean-shifted contrastive loss for anomaly detection. In *AAAI*, Vol. 37. 2155–2162.

[127] Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. 2019. Time-series anomaly detection service at microsoft. In *SIGKDD*. 3009–3017.

[128] Vin Sachidananda, Shao-Yen Tseng, Erik Marchi, Sachin Kajarekar, and Panayiotis G. Georgiou. 2022. CALM: Contrastive Aligned Audio-Language Multirate and Multimodal Representations. *CoRR* abs/2202.03587 (2022).

[129] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*. IEEE Computer Society, 815–823.

[130] Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A Simple but Tough-to-Beat Data Augmentation Approach for Natural Language Understanding and Generation. *CoRR* abs/2009.13818 (2020).

[131] Wei Shen, Xiaoying Zhang, Yuanshun Yao, Rui Zheng, Hongyi Guo, and Yang Liu. 2024. Improving Reinforcement Learning from Human Feedback Using Contrastive Rewards. *ArXiv:2403.07708* (2024).

[132] Tian Shi, Liuqing Li, Ping Wang, and Chandan K Reddy. 2021. A simple and effective self-supervised contrastive learning framework for aspect detection. In *AAAI*, Vol. 35. 13815–13824.

[133] William Shiao, Zhichun Guo, Tong Zhao, Evangelos E Papalexakis, Yozen Liu, and Neil Shah. 2022. Link Prediction with Non-Contrastive Learning. In *ICLR*.

[134] Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A Molecular Multimodal Foundation Model Associating Molecule Graphs with Natural Language. *CoRR* abs/2209.05481 (2022).

[135] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *NeurIPS* 35 (2022), 21548–21561.

[136] Dengdi Sun, Mingxin Cao, Zhuanlian Ding, and Bin Luo. 2022. Graph Contrastive Learning with Intrinsic Augmentations. In *BIC-TA*, Vol. 1801. Springer, 343–357.

[137] Fan-Yun Sun, Jordan Hoffman, Vikas Verma, and Jian Tang. 2019. InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization. In *ICLR*.

[138] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2018. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *ICLR*.

[139] Susheel Suresh, Pan Li, Cong Hao, and Jennifer Neville. 2021. Adversarial graph augmentation to improve graph contrastive learning. *NeurIPS* 34 (2021), 15920–15933.

[140] Jiabin Tang, Lianghao Xia, Jie Hu, and Chao Huang. 2023. Spatio-temporal meta contrastive learning. In *CIKM*. 2412–2421.

[141] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *NeurIPS* 34 (2021), 24804–24816.

[142] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *ECCV*. Springer, 776–794.

[143] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What Makes for Good Views for Contrastive Learning?. In *NeurIPS*.

[144] Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. 2020. Unsupervised Representation Learning for Time Series with Temporal Neighborhood Coding. In *ICLR*.

[145] Daniel J. Trosten, Sigurd Løkse, Robert Jenssen, and Michael Kampffmeyer. 2021. Reconsidering Representation Alignment for Multi-View Clustering. In *CVPR*. Computer Vision Foundation / IEEE, 1255–1265.

[146] Daniel J. Trosten, Sigurd Løkse, Robert Jenssen, and Michael C. Kampffmeyer. 2023. On the Effects of Self-supervision and Contrastive Alignment in Deep Multi-view Clustering. In *CVPR*. IEEE, 23976–23985.

[147] Tsung Wei Tsai, Chongxuan Li, and Jun Zhu. 2020. Mice: Mixture of contrastive experts for unsupervised image clustering. In *ICLR*.

[148] Ike Vayansky and Sathish AP Kumar. 2020. A review of topic modeling methods. *Information Systems* 94 (2020), 101582.

[149] Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2019. Deep graph infomax. *ICLR* (2019).

[150] Danqing Wang, Jiaze Chen, Hao Zhou, Xipeng Qiu, and Lei Li. 2021. Contrastive Aligned Joint Learning for Multilingual Summarization. In *ACL*. ACL, 2739–2750.

[151] Deqing Wang, Junjie Wu, Jingyuan Yang, Baoyu Jing, Wenjie Zhang, Xiaonan He, and Hui Zhang. 2021. Cross-lingual knowledge transferring by structural correspondence and space transfer. *IEEE Transactions on Cybernetics* 52, 7 (2021), 6555–6566.

[152] Lidong Wang. 2017. Heterogeneous data and big data analytics. *Automatic Control and Information Sciences* 3, 1 (2017), 8–15.

[153] Rui Wang, Chongwei Liu, Xudong Mou, Kai Gao, Xiaohui Guo, Pin Liu, Tianyu Wo, and Xudong Liu. 2023. Deep contrastive one-class time series anomaly detection. In *SDM*. SIAM, 694–702.

[154] Tianduo Wang and Wei Lu. 2022. Differentiable Data Augmentation for Contrastive Sentence Representation Learning. In *EMNLP*. ACL, 7640–7653.

[155] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. 2023. Large-scale Multi-modal Pre-trained Models: A Comprehensive Survey. *Mach. Intell. Res.* 20, 4 (2023), 447–482.

[156] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. 2021. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*. 3024–3033.

[157] Yau-Shian Wang, Ta-Chung Chi, Ruohong Zhang, and Yiming Yang. 2023. PESCO: Prompt-enhanced Self Contrastive Learning for Zero-shot Text Classification. In *ACL*. ACL, 14897–14911.

[158] Yanling Wang, Jing Zhang, Haoyang Li, Yuxiao Dong, Hongzhi Yin, Cuiping Li, and Hong Chen. 2022. Clusterscl: Cluster-aware supervised contrastive learning on graphs. In *WWW*. 1611–1621.

[159] Zehan Wang, Yang Zhao, Xize Cheng, Haifeng Huang, Jiageng Liu, Aoxiong Yin, Li Tang, Linjun Li, Yongqi Wang, Ziang Zhang, and Zhou Zhao. 2023. Connecting Multi-modal Contrastive Representations. In *NeurIPS*.

[160] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2022. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *ArXiv:2210.14896* (2022).

[161] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive learning for cold-start recommendation. In *MM*. 5382–5390.

[162] Zhihao Wen and Yuan Fang. 2023. Augmenting Low-Resource Text Classification with Graph-Grounded Pre-training and Prompting. In *SIGIR*. ACM, 506–516.

[163] Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. 2021. CoST: Contrastive Learning of Disentangled Seasonal-Trend Representations for Time Series Forecasting. In *ICLR*.

[164] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2CLIP: Learning Robust Audio Representations from Clip. In *ICASSP*. IEEE, 4563–4567.

[165] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: Contrastive Learning for Sentence Representation. *CoRR* abs/2012.15466 (2020).

[166] Jun Xia, Lirong Wu, Jintao Chen, Bozhen Hu, and Stan Z. Li. 2022. SimGRACE: A Simple Framework for Graph Contrastive Learning without Data Augmentation. In *WWW*. ACM, 1070–1079.

[167] Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. 2021. What Should Not Be Contrastive in Contrastive Learning. In *ICLR*.

[168] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. 2023. RA-CLIP: Retrieval Augmented Contrastive Language-Image Pre-Training. In *CVPR*.

[169] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *ICDE*. IEEE, 1259–1273.

[170] Dongkuan Xu, Wei Cheng, Dongsheng Luo, Haifeng Chen, and Xiang Zhang. 2021. Infogcl: Information-aware graph contrastive learning. *NeurIPS* (2021).

[171] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. 2022. Multi-level Feature Learning for Contrastive Multi-view Clustering. In *CVPR*. IEEE, 16030–16039.

[172] Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, et al. 2024. A survey of resource-efficient llm and multimodal foundation models. *ArXiv:2401.08092* (2024).

[173] Shusheng Xu, Xingxing Zhang, Yi Wu, and Furu Wei. 2022. Sequence level contrastive learning for text summarization. In *AAAI*, Vol. 36. 11556–11565.

[174] Yi Xu, Junjie Ou, Hui Xu, and Luoyi Fu. 2023. Temporal knowledge graph reasoning with historical contrastive learning. In *AAAI*.

[175] Ziyun Xu, Chengyu Wang, Minghui Qiu, Fuli Luo, Runxin Xu, Songfang Huang, and Jun Huang. 2023. Making Pre-trained Language Models End-to-end Few-shot Learners with Contrastive Prompt Tuning. In *WSDM*. ACM, 438–446.

[176] Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. 2021. Deep multi-view learning methods: A review. *Neurocomputing* 448 (2021), 106–129.

[177] Yuchen Yan, Baoyu Jing, Lihui Liu, Ruijie Wang, Jinning Li, Tarek Abdelzaher, and Hanghang Tong. 2024. Reconciling Competing Sampling Strategies of Network Embedding. *NeurIPS* 36 (2024).

[178] Yuchen Yan, Lihui Liu, Yikun Ban, Baoyu Jing, and Hanghang Tong. 2021. Dynamic knowledge graph alignment. In *AAAI*, Vol. 35. 4564–4572.

[179] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. *CoRR* abs/2211.01335 (2022).

[180] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. 2022. Unified contrastive learning in image-text-label space. In *CVPR*. 19163–19173.

[181] Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In *ACL*. ACL, 7617–7630.

[182] Yuhao Yang, Chao Huang, Lianghao Xia, and Chenliang Li. 2022. Knowledge graph contrastive learning for recommendation. In *SIGIR*. 1434–1443.

[183] Hongbin Ye, Ningyu Zhang, Shumin Deng, Mosha Chen, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Contrastive triple extraction with generative transformer. In *AAAI*, Vol. 35. 14257–14265.

[184] Seonghyeon Ye, Jiseon Kim, and Alice Oh. 2021. Efficient Contrastive Learning via Novel Data Augmentation and Curriculum Learning. In *EMNLP*. ACL, 1832–1838.

[185] Zhenhui Ye, Rongjie Huang, Yi Ren, Ziyue Jiang, Jinglin Liu, Jinzheng He, Xiang Yin, and Zhou Zhao. 2023. CLAPSpeech: Learning Prosody from Text Context with Contrastive Language-Audio Pre-Training. In *ACL*. ACL, 9317–9331.

[186] Chin-Chia Michael Yeh, Xin Dai, Huiyuan Chen, Yan Zheng, Yujie Fan, Audrey Der, Vivian Lai, Zhongfang Zhuang, Junpeng Wang, Liang Wang, et al. 2023. Toward a foundation model for time series data. In *CIKM*. 4400–4404.

[187] Yihang Yin, Qingzhong Wang, Siyu Huang, Haoyi Xiong, and Xiang Zhang. 2022. Autogcl: Automated graph contrastive learning via learnable view generators. In *AAAI*, Vol. 36. 8892–8900.

[188] Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. 2021. Graph Contrastive Learning Automated. In *ICML*, Vol. 139. PMLR, 12121–12132.

[189] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph Contrastive Learning with Augmentations. In *NeurIPS*.

[190] Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. 2022. Bringing Your Own View: Graph Contrastive Learning without Prefabricated Data Augmentations. In *WSDM*. ACM, 1300–1309.

[191] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *ArXiv:2205.01917* (2022).

[192] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. 2022. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *SIGIR*. 1294–1303.

[193] Yakun Yu, Mingjun Zhao, Shiang Qi, Feiran Sun, Baoxun Wang, Weidong Guo, Xiaoli Wang, Lei Yang, and Di Niu. 2023. ConKI: Contrastive Knowledge Injection for Multimodal Sentiment Analysis. In *ACL*. ACL, 13610–13624.

[194] Dejiao Zhang, Wei Xiao, Henghui Zhu, Xiaofei Ma, and Andrew O. Arnold. 2022. Virtual Augmentation Supported Contrastive Learning of Sentence Representations. In *ACL*. ACL, 864–876.

[195] Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, and Jun Xiao. 2022. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *ICLR*.

[196] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-Modal Contrastive Learning for Text-to-Image Generation. In *CVPR*. Computer Vision Foundation / IEEE, 833–842.

[197] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023. Vision-Language Models for Vision Tasks: A Survey. *CoRR* abs/2304.00685 (2023).

[198] Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. 2018. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science* 12, 2 (2018), 191–202.

[199] Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J. Passonneau. 2022. Contrastive Data and Learning for Natural Language Processing. In *NAACL*. ACL, Seattle, United States.

[200] Tianqi Zhang, Yun Xiong, Jiawei Zhang, Yao Zhang, Yizhu Jiao, and Yangyong Zhu. 2020. CommDGI: community detection oriented deep graph infomax. In *CIKM*.

[201] Wenqi Zhang, Yongliang Shen, Yanna Ma, Xiaoxia Cheng, Zeqi Tan, Qingpeng Nong, and Weiming Lu. 2022. Multi-View Reasoning: Consistent Contrastive Learning for Math Word Problem. In *EMNLP*. ACL, 1103–1116.

[202] Yanan Zhang, Jiaxin Chen, and Di Huang. 2022. CAT-Det: Contrastively Augmented Transformer for Multimodal 3D Object Detection. In *CVPR*. IEEE, 898–907.

[203] Han Zhao, Xu Yang, Zhenru Wang, Erkun Yang, and Cheng Deng. 2021. Graph Debiased Contrastive Learning with Joint Representation Clustering.. In *IJCAI*. 3434–3440.

[204] Kai Zheng, Qingfeng Sun, Yaming Yang, Tengchao Lv, Yeyong Pi, Changlin Zhao, Fei Xu, and Qi Zhang. 2023. Adversarial Knowledge Stimulated Contrastive Prompting for Few-shot Language Learners. In *ACL*. ACL, 13495–13507.

[205] Lecheng Zheng, Zhengzhang Chen, Jingrui He, and Haifeng Chen. 2024. MULAN: Multi-modal Causal Structure Learning and Root Cause Analysis for Microservice Systems. In *WWW*. ACM, 4107–4116.

[206] Lecheng Zheng, Zhengzhang Chen, Jingrui He, and Haifeng Chen. 2024. Multimodal Causal Structure Learning and Root Cause Analysis. *ArXiv:2402.02357* (2024).

[207] Lecheng Zheng, Yu Cheng, and Jingrui He. 2019. Deep Multimodality Model for Multi-task Multi-view Learning. In *SDM*.

[208] Lecheng Zheng, Yu Cheng, Hongxia Yang, Nan Cao, and Jingrui He. 2021. Deep Co-Attention Network for Multi-View Subspace Learning. In *WWW*. ACM / IW3C2, 1528–1539.

[209] Lecheng Zheng, Jinjun Xiong, Yada Zhu, and Jingrui He. 2022. Contrastive Learning with Complex Heterogeneity. In *SIGKDD*. ACM, 2594–2604.

[210] Lecheng Zheng, Yada Zhu, and Jingrui He. 2023. Fairness-aware multi-view clustering. In *SDM*. SIAM, 856–864.

[211] Dawei Zhou, Lecheng Zheng, Dongqi Fu, Jiawei Han, and Jingrui He. 2022. MentorGNN: Deriving Curriculum for Pre-Training GNNs. In *CIKM*. ACM, 2721–2731.

[212] Dawei Zhou, Lecheng Zheng, Yada Zhu, Jianbo Li, and Jingrui He. 2020. Domain Adaptive Multi-Modality Neural Attention Network for Financial Forecasting. In *WWW*. ACM / IW3C2, 2230–2240.

[213] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2021. Graph Contrastive Learning with Adaptive Augmentation. In *WWW*. ACM / IW3C2, 2069–2080.

[214] Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. 2020. Learning data augmentation strategies for object detection. In *ECCV*. Springer, 566–583.

[215] Shihao Zou, Xianying Huang, and Xudong Shen. 2023. Multimodal Prompt Transformer with Hybrid Contrastive Learning for Emotion Recognition in Conversation. In *MM*. ACM, 5994–6003.