A Comprehensive Survey on Self-Supervised Learning in Computer Vision

SurveyForge

Abstract— Self-supervised learning (SSL) has emerged as a pivotal paradigm in computer vision, facilitating robust representation learning from vast unlabeled datasets, thus addressing the challenges of data annotation in traditional supervised methods. Our comprehensive survey examines the scope and objectives of SSL, focusing on primary dimensions such as contrastive, non-contrastive, and generative frameworks. We explore architectural innovations, including Vision Transformers, and analyze their role in enhancing feature extraction and global context understanding. Key findings highlight SSL's capability in tasks like image classification and segmentation, as well as specialized applications in medical imaging and remote sensing. Current challenges such as computational scalability, dataset biases, and robustness limitations are addressed, emphasizing the need for lightweight, efficient solutions and improved evaluation protocols. We conclude by suggesting future directions, including hybrid methodologies, multimodal integration, and dynamic adaptation strategies, to advance SSL's utility across diverse domains. SSL's transformative potential in redefining representation learning underscores its significance in continuing advancements in visual intelligence, promising scalable and equitable solutions for global challenges.

Index Terms—Self-Supervised Learning, Contrastive Learning Frameworks, Vision Transformers Integration

1 Introduction

Self-supervised learning (SSL) has emerged as one of the most transformative paradigms in computer vision, aiming to address fundamental challenges in modern deep learning. Its rise can be attributed to the exponentially increasing demand for labeled data required by traditional supervised learning methods. Annotating large-scale datasets is expensive, time-consuming, and domain-specific, creating substantial barriers in fields such as medical imaging, remote sensing, and autonomous systems. By harnessing the vast reservoir of unlabeled data, SSL eliminates this dependency, substantially lowering the cost of training data while simultaneously improving scalability and adaptability. Importantly, SSL holds the promise of learning rich, generalizable representations without human intervention, aligning closely with the broader goal of unsupervised intelligence. Its applications extend far beyond mere annotation cost reduction, encompassing robustness to distributional shifts [1], systematic adaptation to new domains [2], and pretraining in data-scarce environments [3], [4].

The conceptual foundation of SSL lies in utilizing intrinsic structures and properties within data to generate supervisory signals, typically through pretext tasks that design surrogate objectives. Such tasks range from generative techniques like reconstruction [5], to predictive methods [5], and contrastive frameworks that exploit relationship modeling [6]. Although early heuristic approaches tackled simple surrogate targets, recent advancements in SSL leverage profound theoretical insights, such as mutual information maximization [7] and redundancy reduction [8]. Prominent frameworks such as SimCLR, MoCo, BYOL, and Masked Autoencoders (MAE) have exemplified state-of-the-art results across tasks ranging from classification and detection to dense prediction tasks like segmentation [9], [10].

One of the key milestones in SSL's historical progression is its extension from simple handcrafted feature extraction to complex, architecture-driven representation learning. Convolutional neural networks (CNNs) initially dominated SSL research, exhibiting high performance in localized feature extraction [11]. However, Vision Transformers (ViTs) have gained traction, demonstrating advantages in capturing rich global context due to their reliance on sequence modeling and self-attention mechanisms [12], [13]. Techniques such as DINO and SelfPatch illustrate how ViT-based SSL methods can even outperform supervised pretraining in downstream applications [14]. Furthermore, hybrid paradigms integrating contrastive and generative learning have demonstrated remarkable versatility [8], [15].

While self-supervised learning is distinguishable from supervised approaches in its independence from labeled data, its relationship with unsupervised and semi-supervised learning is more intertwined. Unlike unsupervised learning, which infers structure without explicit supervisory signals, SSL explicitly defines surrogate labels, pushing models toward task-relevant invariances. Meanwhile, semi-supervised learning frequently complements SSL by applying minimal label supervision to refine representations extracted from self-supervised objectives [16]. SSL's efficacy also extends to continual and domain-adaptive tasks, where dynamic distributions require robust and transferable features [17], [18].

Despite its rapid advancements, SSL faces open challenges. A key limitation lies in the design of pretext tasks that can avoid trivial solutions or overfitting to dataset-specific biases. Current frameworks often rely heavily on augmentations that are brittle or computationally intensive [19]. Furthermore, computational scalability presents another bottleneck, especially for large-scale foundational

models requiring significant energy resources [4]. Finally, evaluation protocols for SSL models often lag behind practical requirements, favoring benchmarks and metrics that fail to capture cross-domain transfer or multimodal applicability [20].

In summary, self-supervised learning represents a paradigm shift in computer vision research, fundamentally redefining how models derive supervision from data. By simultaneously addressing practical constraints and advancing theoretical understanding, SSL pushes the boundaries of what deep learning can achieve in resource-constrained, unlabeled, or noisy environments. The continued evolution of SSL—through the development of scalable architectures, robust task definitions, and thoughtful evaluation frameworks—offers a promising trajectory toward general-purpose, unsupervised intelligence, with profound implications across disciplines.

2 Theoretical Foundations and Pretext Tasks

2.1 Principles and Fundamentals of Self-Supervised Learning

Self-supervised learning (SSL) represents a transformative shift in machine learning by enabling models to learn useful representations without relying on costly labeled data. Instead, it leverages intrinsic structures and patterns within the data itself to generate pseudo-labels, creating supervisory signals derived entirely from the data's inherent properties. This paradigm is particularly influential in computer vision, where vast reservoirs of unlabeled visual data can be utilized to train deep models capable of performing diverse downstream tasks, such as classification, object detection, and segmentation. Here, we explore the theoretical underpinnings of SSL, its practical mechanisms, and emerging insights into its operation.

At its core, the principle of SSL lies in designing auxiliary tasks, termed pretext tasks, where ground truth is artificially constructed based on properties or transformations of the unlabeled data. The ultimate goal is to learn high-quality, transferable features that perform well when applied to downstream supervised tasks, even with minimal labeled data. For example, contrastive learning frameworks such as SimCLR and MoCo [21] define pretext tasks by treating separate augmentations of the same image as positive pairs, while viewing other samples as negative examples. The representations learned through these tasks are guided by objectives like InfoNCE loss, which maximizes agreement between positive pairs in latent space while enforcing separation from negative pairs. These objectives are theoretically linked to mutual information maximization, aiming to preserve shared information relevant for downstream tasks [7].

Another fundamental principle underpinning SSL is invariance, where models are trained to produce consistent embeddings, irrespective of variations in input such as augmentations, geometric transformations, or noise. This principle aligns learning with latent semantic structures that generalize across tasks. Approaches like BYOL and SimSiam [11] accomplish this by leveraging strategies such as asymmetric predictor networks or stop-gradient mechanisms. The avoidance of collapse into trivial solutions—where representations lack diversity—remains central to their success. Theoretical studies [22] further dissect these dynamics, showing how specific architectural and optimization choices prevent collapse and facilitate meaningful representation learning.

Complementary to invariance, SSL encourages decorrelation of redundant features to ensure compact and disentangled embeddings. Approaches like VICReg and Barlow Twins [12] achieve this through redundancy reduction principles, promoting representations with low dependence among embedding dimensions. This is particularly advantageous for creating versatile features transferable across tasks and domains.

One critical advantage of SSL is its cost-efficiency and scalability to large datasets that would be prohibitively expensive to annotate. For instance, scaling SSL to one billion uncurated images has produced state-of-the-art results, demonstrating that SSL methods like SEER and ReLICv2 [23] can outperform supervised approaches in representation quality when trained at scale.

However, these techniques are not without limitations. Pretext tasks may lead to representation biases, particularly if shortcuts are exploited—such as models leveraging low-level cues (e.g., color or texture) instead of more abstract structures—and thus failing to generalize to challenging downstream tasks. To address this, hybrid and multi-task frameworks combine complementary pretext objectives to balance trade-offs between generalization and task-specific learning [3], [24].

The theoretical foundation of SSL is deepened by connections to information theory, particularly the interpretation of SSL objectives as maximizing expected mutual information between positive pairs while filtering task-irrelevant details [7]. Recent insights propose that SSL representation effectiveness lies in aligning representations with downstream tasks, balancing invariance and selectivity. For generative SSL approaches, such as masked autoencoders (MAEs), pretext tasks like reconstructing masked image parts demonstrate the potential for deeper semantic understanding from holistic data [10].

Emerging challenges and opportunities include designing richer pretext tasks to learn hierarchical semantic abstractions dynamically adapted to the specifics of input data [25]. Advancements in modeling multimodality—integrating language, vision, and audio—also illustrate SSL's broadening applicability beyond pure vision tasks, opening new avenues for research in cross-modal coherence [26].

As SSL evolves, its potential to reshape machine learning pipelines endures, offering a scalable and resource-efficient alternative to supervised learning. By refining theoretical insights and practical implementations, future research is poised to unlock deeper connections between pretext objectives, representation quality, and downstream effectiveness.

2.2 Pretext Tasks and Their Design Principles

Pretext tasks serve as a cornerstone in self-supervised learning (SSL) by defining surrogate objectives that facilitate the extraction of meaningful representations from unlabeled

data. These tasks enable the generation of "pseudo-labels" derived intrinsically from the data through handcrafted transformations, perturbations, or relationships, providing training signals for the learning process. The design and selection of pretext tasks play a pivotal role in determining the quality, versatility, and applicability of the representations learned by SSL models. This subsection explores pretext task design principles and strategies, aligning them with downstream objectives and broader learning goals.

At their essence, pretext tasks are formulated to embed representations with semantic richness and transferability, ensuring that features learned during pretraining generalize well across a wide range of downstream tasks. Effective pretext task design strikes a balance between domain-specific constraints and universality, fostering invariance to irrelevant variations (e.g., rotation, scale) while preserving intrinsic factors crucial for semantic understanding. Notable examples like relative patch location prediction [27] and solving jigsaw puzzles [27] utilize spatial context to prompt locality-sensitive feature learning, producing representations highly effective for downstream tasks such as object detection and segmentation.

A unifying principle in pretext task design is invariance induction, which enforces robustness to specific data transformations. Tasks such as recognizing geometric transformations (e.g., rotation prediction) encourage models to capture shape and structural invariances [25]. Similarly, contrastive-based pretext tasks (e.g., instance discrimination, where augmented views of the same image are matched) enforce invariance across data augmentations while simultaneously maximizing mutual information between views [28]. These methods frequently optimize objectives like the InfoNCE loss, which is theoretically underpinned by mutual information estimates [29], with strategies such as multi-crop or color jittering augmentations employed to ensure diverse and challenging training pairs. However, meticulous augmentation design is essential to avoid the risk of enabling shortcut learning—where models rely on low-level cues rather than meaningful patterns.

Preventing shortcut learning is an enduring challenge in SSL. Poorly designed pretext tasks can lead models to exploit superficial features (e.g., image color statistics) that fail to generalize. Generative pretext tasks, such as inpainting or masked image modeling (e.g., masked autoencoders), mitigate this issue by tasking the model with reconstructing occluded image portions [10]. Such tasks inherently require the model to focus on global context and high-level semantics. The emergence of masked autoencoding paradigms has further demonstrated these tasks' scalability, particularly when applied to architectures such as vision transformers [30].

Evaluating the effectiveness of pretext tasks extends beyond accuracy on pretraining objectives. Metrics such as linear separability, transfer learning performance, and robustness under domain shifts provide a more comprehensive assessment [31]. For example, linear probing evaluates SSL-trained features by freezing the backbone and training a lightweight classifier, offering insights into feature generality. Additional robustness analyses, particularly in domain generalization contexts, highlight the adaptability of SSL representations to diverse, non-i.i.d. data [32].

Emerging trends in pretext task design emphasize hybridization of auxiliary objectives and dynamic adaptability. Combining complementary paradigms—such as contrastive learning with predictive modeling—enables representations that balance capturing global invariances with fine-grained discriminative details [33]. Context-specific developments, such as self-supervised co-training across modalities (e.g., integrating RGB and optical flow cues for videos), illustrate SSL's expanding application in multi-modal and temporal domains [34]. However, challenges remain, including reducing over-reliance on augmentations, mitigating pretext-specific biases, and ensuring computational efficiency at scale.

As SSL methodologies continue to evolve, enhancing the theoretical grounding of pretext tasks remains critical. Novel paradigms such as pretext-invariant representations [27] and maximum entropy coding mechanisms for bias mitigation [35] offer promising avenues for designing generalizable, robust, and scalable frameworks. By incorporating such advancements, future task designs will further strengthen SSL's ability to uncover rich, transferable representations across diverse datasets and application domains.

2.3 Context-based Self-Supervision Frameworks

Context-based self-supervision frameworks focus on leveraging the spatial and relational contextual information inherent in visual data to establish pretext tasks for self-supervised learning (SSL). These approaches are motivated by the premise that understanding the structural and semantic arrangement of visual elements is key to extracting transferable representations for downstream tasks. The pretext tasks in this category enforce the model to learn relationships and dependencies among spatially or temporally related components of an image or sequence, thereby fostering robust feature learning.

A prime example of such a pretext task is relative position prediction, wherein the model predicts the spatial arrangement of image patches. This task was popularized through methods that divide an image into patches and challenge the model to identify the relative position of one patch with respect to another, implicitly encoding spatial consistency and object-centric features. Pretext-Invariant Representation Learning (PIRL) integrates this concept by enforcing invariance under transformations for tasks like solving jigsaw puzzles, thereby improving representation semantics and robustness [27]. Jigsaw-based approaches, which rearrange image patches and task the model with reconstructing the original layout, are particularly effective in encouraging the understanding of spatial context, as demonstrated by their ability to outperform supervised methods in certain detection benchmarks [11], [36].

Another notable task under this framework is predicting patch order within a spatial sequence. Models trained with such objectives are compelled to exploit not just individual patch features but also their relational dependencies, which are crucial for semantic understanding. These dependencies are found to embed meaningful object-level information, which translates to superior performance on object detection tasks when compared to other generative pretext tasks like masked pixel reconstruction [37].

Beyond patch-based tasks, some frameworks adapt to object-centric spatial contexts by incorporating mechanisms to detect or prioritize objects within a scene. For example, in tasks like motion-based segment recovery, large-scale datasets of videos are exploited, where the motion cues provide segmentation masks for training models to understand object-level semantics and their relations [38]. These approaches improve on traditional context-based methods by embedding dynamic relationships that underpin temporal coherence, making them especially effective for video-based learning tasks.

Temporal extensions of context-based methods further exploit the continuity in video data. Tasks such as frame ordering or temporal alignment involve learning to predict the chronological order of frames or align frames across multiple views. These tasks capture temporal consistency and dynamic scene understanding, which have been shown to significantly improve downstream video-action recognition performance [24], [39].

While context-based self-supervision frameworks provide interpretable and empirically effective SSL representations, they face certain limitations. One challenge is their sensitivity to dataset-specific biases; for instance, models may overly rely on surface features like texture rather than learning deeper semantics, as discussed in works addressing shortcut removal strategies [40]. Similarly, these frameworks are often constrained by the design of their relational structure, such as fixed grid-based patch divisions, which may inadequately capture global contextual relationships in certain applications, such as those involving unstructured environments or 3D data [41].

Emerging trends aim to resolve these issues by integrating attention mechanisms or transformers, which inherently capture global feature contexts. For instance, self-supervised vision transformers (ViTs) combined with patch-based relational tasks align well with the goals of context-based learning by modeling interactions across broader spatial extents [14], [30]. Additionally, increasing emphasis is placed on multi-modal SSL tasks that enhance context understanding by fusing relational information across modalities, such as image-text alignment pretext tasks [42].

Future research in context-based self-supervision should focus on designing more adaptive relational structures for dynamic contexts, particularly for multimodal or 3D environments. Expanding frameworks to handle noisier, real-world data and minimizing dataset biases without compromising the semantic richness of learned representations are also promising directions. Lastly, hybridizing context-based tasks with predictive and contrastive methodologies could unlock further synergies for capturing diverse and robust features, advancing the generalizability of SSL approaches [39], [43].

2.4 Transformation-based Pretext Tasks for Invariance Learning

Transformation-based pretext tasks utilize semantic-preserving transformations to drive the learning of representations that are invariant to geometric, photometric, or structural alterations in visual data. Unlike the context-based pretext tasks discussed earlier, which emphasize spatial and relational dependencies, these approaches focus on transformation invariance, aligning model representations with the fundamental properties of visual data. By doing so, transformation-based methods enhance model robustness and transferability across diverse downstream applications. Additionally, they serve as a bridge to generative pretext tasks, contributing complementary invariance properties that can be exploited in reconstruction or hybrid SSL frameworks.

A foundational example in this category is geometric transformation prediction, particularly rotation angle classification. Here, models are trained to predict rotation angles (e.g., 0°, 90°, 180°, 270°) applied to an image—a task that exploits the object-level semantics and spatial awareness inherent in visual data. This approach, as demonstrated in [44], enables models to capture spatially aware feature representations essential for general-purpose feature learning. Moreover, local variants such as patch-wise rotation [27] compound these benefits by focusing on rotational transformations at a finer granularity, yielding context-rich and granular local representations. Despite their success, these methods are often limited in ambiguous scenarios where visual cues such as texture or repetitive patterns obscure orientation information.

Beyond rotation, scaling and affine transformation recognition tasks extend the scope of geometric transformation learning by enabling models to identify changes in scale, translation, or distortions within images. This approach fosters invariance to size and spatial deformations, encouraging models to capture multi-scale hierarchical structures and spatial relationships in visual data [45]. However, such pretext tasks can be heavily dependent on the design of objectives that sufficiently generalize across scaling or complex structural variations, making them susceptible to overfitting on specific datasets or transformations.

Transformation-based objectives also synergize with augmentation-driven contrastive learning, as seen in approaches such as SimCLR and PIRL. These methods represent different augmented views of the same image as "positives" in latent space, enforcing semantic consistency between transformed views. PIRL [27], in particular, incorporates structural perturbations like cropping, jittering, or flipping, encouraging models to maintain semantic alignment despite spatial or photometric distortions. This contrastive approach effectively expands representation capacity, countering shortcut learning tendencies where models might exploit trivial correlations within data. Nevertheless, such methods carry risks of overreliance on augmentation pipelines, which can bias the model towards specific transformations and limit representation versatility [6].

Advancements in Vision Transformers (ViTs) have unlocked new avenues for incorporating transformation contrast at a more granular level. For instance, SelfPatch [14] introduces patch-wise contrastive objectives, wherein individual image patches are perturbed or rearranged while ensuring embedding consistency. This advancement deviates from earlier geometric approaches, emphasizing local-global coherence by leveraging the patch tokenization architecture of ViTs. These methods are particularly suited for dense prediction tasks, such as segmentation and object localization, complementing the global structural aware-

ness achieved through generative and reconstruction-based tasks.

Despite their versatility, transformation-based tasks face intrinsic trade-offs. Predictive objectives, such as rotation or affine recognition, provide interpretability and simplicity but may struggle with representation collapse in cases lacking sufficient task complexity. On the other hand, contrastive objectives inherently mitigate this limitation by leveraging negative examples, albeit at the cost of computational overhead due to the need for large batch sizes or memory queues [4], [46]. Additionally, task-specific augmentation reliance remains a fundamental challenge, as transformations tailored for controlled datasets may fail to generalize in real-world or unstructured environments [47].

Future research in this domain should aim to address these limitations by integrating adaptive augmentation frameworks and hybrid multi-pretext systems. By combining global transformation prediction and patch-level objectives, it may be possible to design multi-scale invariance learning approaches. Moreover, incorporating weaker supervisory signals, such as temporal coherence in video datasets [48], could further enhance structural and temporal consistency learning. These innovations hold significant promise for transforming transformation-based pretext tasks into scalable, robust, and domain-agnostic tools for general-purpose visual representation learning, aligning seamlessly with the generative reconstruction frameworks outlined in the subsequent section.

2.5 Generative Pretext Tasks for Reconstruction-Based SSL

Generative pretext tasks in reconstruction-based self-supervised learning (SSL) leverage the inherent structure of visual data to guide representation learning by reconstructing missing or corrupted content. These methods prioritize capturing holistic data characteristics, thereby enabling models to uncover the underlying visual semantics and global structure of images. Unlike contrastive or predictive approaches, generative tasks inherently focus on the reconstruction of the input signal in either input or latent spaces—a process tightly tied to the underlying data distribution. This subsection examines key reconstruction-based techniques, analyzes their strengths and limitations, and highlights trends in this branch of SSL.

Reconstruction-based generative tasks often rely on reconstructing masked or missing parts of data while preserving its semantic coherence. Masked image modeling techniques, such as Masked Autoencoders (MAEs), have gained prominence for their ability to generate global contextual representations by input occlusion and subsequent prediction [37], [49]. MAEs split images into patches, mask a subset, and task the model to reconstruct missing patches, allowing the encoder to focus on global structures and contextual dependencies. These approaches excel at embedding high-level features that generalize well across dense prediction tasks, such as segmentation and object detection. The introduction of latent-space reconstruction objectives, exemplified in techniques such as Context Autoencoders, has further refined representation quality by promoting compact feature encoding, wherein only essential elements for reconstruction are retained [37].

Traditional pixel-based reconstruction tasks, including denoising autoencoders and inpainting, have also demonstrated significant success in self-supervised learning. These methods reconstruct visual data corrupted by noise or missing content, promoting robustness in learned embeddings. Denoising autoencoders reduce semantic noise sensitivity while retaining crucial features, aiding in improving resilience to adversarial perturbations or domain shifts. Meanwhile, image inpainting—tasked with filling missing pixels—is particularly effective at capturing spatial coherence and object-level context, as these tasks encourage feature extraction that respects both the local structure and global distributions [49].

The field has complemented traditional pixel-level objectives by incorporating latent-space reconstruction tasks, such as those involving Variational Autoencoders (VAEs) or hybrid frameworks like generative adversarial networks (GANs). While VAEs model the underlying data distribution via approximate posterior inference, they tend to prioritize smoothness over fine-grained detail. By contrast, GAN-based reconstruction frameworks employ adversarial objectives to refine reconstruction quality, which, despite yielding sharper outputs, often suffer from training instability or mode collapse [37]. A notable extension lies in reconstructing domain-invariant features, where specialized models like DiMAE (Domain-Invariant Masked AutoEncoders) promote cross-domain generalization by reconstructing style-noisy inputs across heterogeneous data distributions, a critical advancement for diverse real-world applications [50].

Challenges faced by generative pretext tasks revolve around potential redundancies in pixel-space objectives and inherent biases introduced by trivial solutions. Tasks like pixel-intensity reconstruction may orient models toward low-level details at the expense of abstract semantic features critical for downstream tasks. Methods such as pretext-invariant representation constraints (e.g., PIRL) mitigate these shortcomings by aligning generative objectives with downstream invariance requirements [27]. Moreover, efficient masking techniques, including semantic-aware masking or transformer-based random masking, have been proposed to strike a balance between information preservation and task complexity [14], [49].

Recent advancements have also shown generative tasks converging with vision transformers (ViTs), particularly through their patch-based processing pipelines, enabling fine-grain attention mechanisms to amplify the reconstruction process [14], [49]. Additionally, multi-modal extensions have emerged, such as cross-modal reconstruction (e.g., image-to-text mapping), showcasing how generative reconstruction aligns with multimodal pretraining goals [50].

Looking forward, the expansion of generative pretext tasks into hybrid frameworks, integrating contrastive and predictive paradigms, holds significant promise. These hybrid methods can harness the strengths of generative modeling—global structure comprehension—while addressing its weakness in task-specific invariance or localized learning objectives. Furthermore, efficient scalability mechanisms, such as progressive masking or the use of lightweight decoders, may alleviate computational burdens associated with generative tasks. Ultimately, advancements in reconstruction-based SSL are poised to underpin increas-

ingly flexible, robust, and general-purpose visual representation learning frameworks.

2.6 Advanced Hybrid Frameworks and Multi-modal Pretext Tasks

Advanced hybrid frameworks in self-supervised learning (SSL) represent an ambitious evolution, combining diverse pretext task strategies to leverage their complementary benefits and enhance representation learning. Traditional SSL methods typically revolve around singular paradigms, such as contrastive learning or reconstruction-based tasks, each excelling within specific contexts yet limited in their ability to holistically capture visual semantics. In contrast, hybrid approaches synergize orthogonal objectives, unifying discriminative, predictive, and generative strengths into more comprehensive pretraining pipelines. Additionally, these frameworks extend to multi-modal SSL, integrating diverse data modalities (e.g., vision and language or 3D spatial data) to harness cross-domain synergies, substantially expanding the applicability of SSL across various real-world scenarios.

Hybrid pretext strategies enable the fusion of contrastive learning's instance discrimination capabilities with the contextual and structural richness offered by predictive or reconstruction-based paradigms. For example, predictive tasks like geometric transformation prediction, which encode invariance to spatial variations, pair effectively with contrastive objectives, which focus on discriminative alignment at the instance level [51]. Similarly, masked image modeling methods like Masked Autoencoders (MAEs), known for their capacity to uncover pixel-level contextual correlations, have been incorporated into contrastive SSL frameworks to achieve complementary optimization [52]. These dual-objective pipelines promote a balance between preserving fine-grained details and disentangling high-level features, enabling generalized transferability across downstream vision tasks. However, successfully reconciling conflicting gradients between objectives remains a challenge, often complicating optimization. Iterative disentangling solutions, such as Iterative Partition-based Invariant Risk Minimization (IP-IRM), have emerged to address this issue, improving gradient alignment and task-specific representation quality [53].

The progression to multi-modal SSL builds on hybrid principles, amplifying model expressiveness by combining complementary modalities. State-of-the-art vision-language architectures such as CLIP integrate contrastive pretext tasks to unify representations across image and text domains, facilitating robust zero-shot generalization [54]. These frameworks employ modality-specific encoders while aligning their outputs into a shared embedding space, exploiting semantic correspondences between modalities. To address inherent biases that arise from modality discrepancies, models like DiMAE (Domain-Invariant Masked AutoEncoder) employ cross-domain augmentations and custom decoder configurations, achieving greater cross-modal consistency [50].

Similarly, 3D vision tasks have begun integrating spatial data with hybrid SSL workflows. Tasks like point cloud reconstruction, in combination with global contrastive pretraining, improve representations for applications requiring 3D spatial comprehension, such as autonomous navigation and object recognition [14]. Moreover, temporal coherence objectives using video data, including motion prediction or frame ordering, complement spatial understanding by embedding dynamics-sensitive features that transfer effectively across video-centric applications [55]. Such combinations represent a growing trend where hybrid SSL frameworks converge with domain-specific task requirements.

Despite demonstrated improvements, hybrid and multi-modal SSL still face critical challenges. Hybrid frameworks often struggle with high computational demands due to the complexity of integrating and optimizing multiple loss functions, while multi-modal SSL relies heavily on extensive pre-aligned datasets. Tackling these bottlenecks requires novel approaches, such as automated curriculum learning strategies that dynamically prioritize tasks based on evolving model competence [56]. Additionally, regularization techniques involving adversarially-driven masking policies or entropy-constrained embedding spaces offer promising avenues for disentangling biases while improving generalization [40].

Future directions for hybrid and multi-modal SSL emphasize adaptability and scalability. Dynamic adjustment of object-specific optimization weights—guided by feature redundancy or task-specific interactions—has shown potential to improve multi-task synergy without excessive computational trade-offs. Emerging techniques like latent masked modeling further blur the line between pixel-level precision and latent-space abstraction, promising efficient yet expressive representation learning [57]. Furthermore, domain-specific augmentations and meta-learning frameworks focused on out-of-distribution generalization are likely to play a key role in the evolution of both hybrid and multi-modal approaches. As these frameworks continue to mature, their ability to adapt to specialized tasks while maintaining versatility and robustness across domains underscores their foundational role in the next generation of self-supervised representation learning.

3 Contrastive Learning and Information Maximization Frameworks

3.1 Fundamentals of Contrastive Learning

Contrastive learning has emerged as one of the most influential paradigms in self-supervised learning, offering a robust framework for learning meaningful representations by contrasting positive and negative instances. At its core, contrastive learning leverages the principle of instance discrimination, where similar samples (positive pairs) are brought closer in the latent space, and dissimilar samples (negative pairs) are pushed apart. This subsection delves into the underlying mechanisms of contrastive loss, the roles of positive and negative pairs, and its connection to mutual information maximization.

The primary objective of contrastive learning is operationalized through the contrastive loss function, which actively encourages the alignment of positive samples and the repulsion of negatives. Among various formulations, the

InfoNCE loss is the most commonly used and defined as follows [6], [7]:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(z_i, z_k)/\tau)},$$

where z_i and z_j are the latent representations of the anchor sample and its positive counterpart, sim denotes a similarity measure (typically cosine similarity), τ is the temperature parameter, and k indexes the negative samples. This formulation encourages the embeddings of z_i and z_j to be maximally similar while ensuring z_i remains distinct from all negatives. The negative denominator term plays a critical role in preventing a trivial solution, though it introduces computational challenges when working with large datasets.

Central to the success of contrastive learning is the proper definition of positive and negative pairs. Positive pairs typically originate from different augmentations of the same sample, ensuring invariance to transformations such as cropping, flipping, or color jittering. Augmentation diversity is essential to imbue the model with desired invariances [6], [7]. On the other hand, negative pairs are drawn from different data samples, requiring strategies that ensure hard negatives (distinct but contextually similar samples) are selected to optimize the contrastive objective. Improper sampling of negatives can lead to issues such as false negatives, where semantically similar samples are incorrectly treated as dissimilar. Mitigation strategies, such as memory banks (e.g., MoCo [2]) or larger batch sizes (e.g., SimCLR [4]), effectively enhance negative instance diversity and overall performance, though at the cost of increased computational complexity.

Contrastive learning frameworks are fundamentally grounded in the theoretical underpinnings of mutual information maximization. The training objective implicitly maximizes a lower bound on the mutual information between positive pairs, thereby capturing shared features that retain semantic relevance [7], [58]. This connection to information theory further motivates the adoption of augmentations and sampling procedures that preserve informative content while discarding task-irrelevant variability. However, as recent works argue, existing contrastive methods have limitations in fully capturing higher-order dependencies or disentangling fine-grained semantics, often requiring large-scale datasets to approach supervised baselines [4].

Non-contrastive approaches, such as BYOL and SimSiam, have recently challenged the necessity of explicit negative samples by leveraging asymmetric architecture designs and stop-gradient operations to prevent representational collapse [22]. While these methods eschew traditional contrastive mechanisms, they achieve comparable or superior performance, exposing an intriguing duality between contrastive and non-contrastive paradigms [8].

Despite its empirical successes, contrastive learning faces notable challenges. The reliance on large batch sizes, high-capacity models, and extensive augmentations limits its accessibility to resource-constrained environments [4], [29]. Furthermore, contrastive frameworks are sensitive to hyperparameter choices, such as the temperature parameter τ, which drastically affects the tightness of cluster formations [59]. Future research could explore adaptive temperature scaling mechanisms or hybrid objectives that combine contrastive and redundancy-reduction principles [4].

Looking ahead, contrastive learning shows promise in applications beyond traditional vision tasks. Extensions into multimodal domains (e.g., CLIP [7]) and data-efficient frameworks for video and graph representations exemplify its versatility. Incorporating domain-specific augmentations and addressing sampling biases further stand as compelling research directions. By bridging contrastive principles with broader self-supervised objectives, contrastive learning reinforces its role as a cornerstone of representation learning in computer vision.

3.2 Key Architectures in Contrastive Learning

Contrastive learning has emerged as a cornerstone of self-supervised visual representation learning, with its foundations grounded in the principle of discriminating between positive and negative pairs to align related representations while diverging unrelated ones. This subsection critically explores prominent architectures in contrastive learning, highlighting their design principles, mechanisms, and contributions to advancing the field while building upon the discussion of augmentations, sampling strategies, and intervention mechanisms presented earlier.

SimCLR exemplifies the simplicity and effectiveness of contrastive frameworks, relying on strong data augmentations and the InfoNCE loss to learn meaningful representations [28]. Its augmentation pipeline—comprising random cropping, color distortion, and Gaussian blurring—generates positive pairs by applying different transformations to the same image. These representations are contrasted against negatives sampled from a mini-batch, following the foundational principles discussed previously. SimCLR uniquely introduces a multi-layer perceptron (MLP) projection head that projects features into a latent space optimized specifically for contrastive training. During downstream applications, the projection head is discarded, and intermediate backbone representations are used instead. While effective, SimCLR requires large batch sizes to ensure a sufficient number of negatives in each update, introducing significant computational demands [60].

To address these computational challenges, **MoCo (Momentum Contrast)** decouples the reliance on large batch sizes by introducing a momentum-based encoder and a dynamic memory queue [11]. The memory queue stores a bank of negative samples from previous mini-batches, enabling efficient training by maintaining a wide diversity of negatives. A momentum encoder updates slowly compared to the primary network, ensuring stable representations for stored negatives. This innovation facilitates resource-efficient contrastive learning, which proved particularly effective in downstream tasks like object detection on COCO, emphasizing its practical applicability. As highlighted in earlier discussions, the interplay between architectural mechanisms like MoCo's memory bank and augmentation strategies can significantly influence both the representation robustness and scalability of contrastive frameworks [11].

Departing from traditional paradigms, **Bootstrap Your Own Latent (BYOL)** eliminates the need for explicit negative pairs, defying the longstanding assumption that

negatives are essential to avoid representational collapse [22]. BYOL features two network branches—a student and a momentum-updated teacher—that interact via a stop-gradient operation applied to the teacher outputs. A predictor network on the student branch aligns the representations without resulting in trivial solutions, thanks to architectural innovations such as asymmetric parameter updates. BYOL's ability to achieve state-of-the-art results on ImageNet and other benchmarks underscores its effectiveness, particularly in scenarios where the computational overhead associated with negative sampling is prohibitive [22]. This non-contrastive paradigm connects to emerging trends that emphasize the synergy between methods that integrate or forgo traditional contrastive losses.

Another non-contrastive approach, **VICReg (Variance-Invariance-Covariance Regularization)**, emphasizes a systematic method to prevent collapse by combining invariance objectives, variance constraints, and redundancy reduction techniques [33]. The variance term ensures sufficient feature distribution across dimensions, while the covariance term minimizes redundancy. By sidestepping negative pairs altogether, VICReg simplifies training dynamics while retaining representation robustness. As discussed in prior subsections, these strategies align with broader efforts to balance robustness, scalability, and computational efficiency through redundancy reduction mechanisms.

Despite their varying implementations, these architectures commonly rely on data augmentations to generate consistent positive pairs, a topic covered in depth earlier [61]. However, challenges persist across the board, including the sensitivity of contrastive paradigms to augmentation parameters and computational bottlenecks arising from large-scale training requirements or architectural redundancy constraints.

Emerging research increasingly aims to unify the strengths of these diverse frameworks. Hybrid methods that integrate principles from contrastive models like SimCLR and MoCo with the simplicity of non-contrastive approaches such as BYOL and VICReg seek to optimize robustness without compromising computational efficiency [8]. Furthermore, recent efforts extend these principles to multimodal domains, as exemplified by frameworks like CLIP, which align visual and textual representations through cross-modal contrastive objectives [26].

As contrastive learning continues to evolve, its future lies in addressing resource efficiency, minimizing dependency on hand-designed augmentations, and designing architectures that generalize seamlessly across diverse applications. The collective progress of methods like SimCLR, MoCo, BYOL, and VICReg highlights a promising trajectory, anchored in the synergy between robust augmentation pipelines, innovative architectural mechanisms, and novel optimization strategies. Such advancements pave the way for contrastive self-supervision to tackle increasingly complex and diverse real-world challenges.

3.3 The Role of Data Augmentations and Sampling

Contrastive learning has emerged as a highly effective paradigm for self-supervised representation learning, centralizing the role of contrasting positive and negative sample pairs to learn robust semantic features. However, this framework fundamentally hinges on the quality of two key components: data augmentations for generating meaningful positive pairs and sampling strategies to select informative negative pairs. This subsection examines how the design of augmentations and sampling strategies influences representation quality, robustness to downstream tasks, and scalability across diverse settings.

Data augmentations play a pivotal role in generating strong positive pairs, whereby two augmented views of the same instance encode shared semantics while differing in appearance. Techniques such as random cropping, horizontal flipping, color jittering, resizing, and Gaussian blurring have become the cornerstone of augmentation pipelines, as exemplified in SimCLR [11]. These augmentations aim to capture invariances by training models to recognize that visually dissimilar variations of the same instance still represent the same underlying concept. For instance, random cropping enhances the model's spatial invariance, while blurring enforces robustness to low-frequency image distortions [62]. However, improper augmentation tuning can inadvertently lead to semantic collapse, as overly aggressive augmentations risk removing essential features required for pair consistency [62]. To address this, studies such as PIRL [27] advocate incorporating augmentation techniques tailored specifically to downstream tasks to maintain feature relevance for specific applications.

On the other hand, the efficacy of sampling strategies in constructing negatives is equally critical. Effective contrastive learning relies on ensuring that negative pairs (samples from different instances) are sufficiently diverse and meaningful. Without careful oversight, improperly selected negatives may introduce false negatives—pairs that mistakenly group semantically similar samples as negatives—leading to representation degradation [63]. MoCo, for example, mitigates this challenge by using a dynamic memory bank to sample negatives, decoupling batch size from negative pool size, thereby facilitating robust training on small batches [11]. Additionally, hard negative mining, which explicitly focuses on negatives that are challenging to distinguish from positives, has shown promise in improving both contrastive losses and the quality of learned embeddings [28]. However, hard negative sampling introduces computational overhead, as identifying such samples necessitates similarity-based ranking, and it risks amplifying semantic ambiguities if false negatives are inadvertently emphasized.

Recently, innovative perspectives on augmentation-sampling interplay have been gaining traction. Techniques such as multi-view agreements expand contrastive learning to capture not just local invariances but broader variations across spatial and temporal dimensions [64]. For instance, DINO-MC explores multi-sized crops to introduce hierarchical semantics in view alignment [64], while GenView employs generative models to achieve high-quality, semantically controlled augmentations, dynamically balancing between preserving shared content and introducing diversity [65].

Another promising direction lies in addressing sampling biases caused by dataset and domain-specific limitations. For example, false negatives arise when instances representing different views of the same class are incorrectly con-

trasted as negatives. Solutions such as Positive-Unlabeled Contrastive Learning mitigate these biases by modifying contrastive objectives to accommodate unlabeled positive pairs, thus reducing penalization of unintentional misclassification [63]. Parallelly, VICReg and Barlow Twins move beyond contrastive frameworks altogether by leveraging redundancy reduction strategies, obviating the need for negative samples while retaining strong representation learning capabilities [66].

While augmentation designs and sampling techniques have individually matured, their joint optimization remains an open research challenge. Striking the right balance between augmentation strength and sampling diversity requires transition-aware objectives that dynamically align augmentations to the current state of the learning process [67]. Additionally, domain-specific applications, such as medical imaging and remote sensing, emphasize the need for nuanced augmentation pipelines and task-adapted sampling methods that respect structural or semantic constraints [64].

In conclusion, data augmentation and sampling are indispensable components of contrastive learning architectures, offering both opportunities and challenges. Future directions should explore more adaptive, dynamic strategies that integrate augmentation trajectories with context-aware sampling protocols, yielding representations that are robust, semantically meaningful, and transferable across a broader range of downstream scenarios. As highlighted by frameworks like GenView and DINO-MC, leveraging generative augmentations and multi-view alignment mechanisms may allow contrastive learning to extend its utility into richer, more complex real-world applications.

3.4 Extensions Beyond Negative Samples

A significant advancement in self-supervised learning has been the emergence of methods that discard explicit reliance on negative samples, driving innovations in frameworks focused on maximizing representational quality through alternative mechanisms. Traditional contrastive learning approaches, such as InfoNCE-based objectives, have achieved remarkable success by distinguishing augmented positive pairs from a curated pool of negative samples. However, these methods face limitations, including potential sampling biases—such as false negatives in limited datasets—and computational overheads linked to maintaining large negative sets [6]. To address these challenges, a new class of approaches, including redundancy reduction methods and joint embedding architectures, has emerged to enable effective self-supervised representation learning without explicit negative samples.

Variance-Invariance-Covariance Regularization (VICReg) exemplifies this shift by introducing an objective that enforces three distinct constraints: variance, invariance, and covariance. The variance term ensures that embeddings maintain sufficient spread within the representational space to avoid degeneration, while the invariance term encourages consistency between embeddings of augmented views. Additionally, the covariance term reduces redundant correlations among embedding dimensions, fostering decorrelated and comprehensive representations. This design allows VICReg to bypass the need for negative samples, addressing the challenge of false negatives inherent in contrastive learning frameworks. However, VICReg's reliance on pre-defined regularization coefficients may introduce hyperparameter sensitivity, particularly when applied to diverse datasets at scale.

Building on the concept of redundancy reduction, Barlow Twins optimizes a cross-correlation matrix of embeddings from positive pairs, encouraging it to approach an identity matrix. This ensures that the learned representations are both invariant to augmentations and decorrelated across dimensions. Compared to VICReg, Barlow Twins directly integrates redundancy minimization into its objective, offering a computationally straightforward yet effective solution [11]. Nevertheless, its reliance on balanced datasets with well-calibrated augmentations may constrain its generality when applied to heterogeneous datasets or multimodal scenarios.

Joint embedding predictive architectures, such as BYOL (Bootstrap Your Own Latent) [27], represent a further advancement by entirely eliminating the need for negatives while avoiding collapsing solutions. BYOL achieves robust representation alignment by incorporating architectural asymmetry—using stop-gradient operations on one of its branches—while maximizing agreement between positive pairs. This approach promotes stable optimization and mitigates degeneracies, enabling effective feature extraction. Extending these principles, SimSiam further demonstrates that reduced architectural complexity can still yield competitive results [68]. Importantly, these methods forego the computational demands of memory banks or momentum queues that characterize frameworks like MoCo [6], offering a conspicuous advantage in resource-constrained environments.

The shared advantage of these non-contrastive approaches lies in their intuitive objectives, which replace contrastive loss with mechanisms designed to shape the embedding space for compactness and diversity. By eliminating explicit negatives, these methods are particularly robust to dataset noise and class imbalances, challenges that often undermine traditional contrastive learning frameworks [46]. Moreover, their computational efficiency and scalability make them well-suited for large-scale training scenarios, addressing practical concerns in resource-constrained settings [4].

Despite these strengths, non-contrastive methodologies present certain limitations. A key drawback is their lack of explicit enforcement of inter-class separation, a feature inherent to contrastive frameworks through the use of negative pairs. Consequently, embedding spaces produced by these methods may be less discriminative for downstream tasks requiring strict inter-class boundaries, such as open-set recognition or zero-shot classification [28]. Additionally, their reliance on carefully designed augmentations as implicit substitutes for contrast introduces sensitivity to augmentation quality, a recognized bottleneck across self-supervised frameworks [11].

Emerging trends offer promising directions for improving these non-contrastive paradigms. Incorporating domain-specific constraints, such as spatiotemporal structures in video data [69] or object-level consistency in scene

understanding [70], could enhance their task-specific utility. Similarly, integrating cross-modal signals, as in vision-language alignment tasks, may mitigate issues of weak inter-class discrimination [39]. Furthermore, adaptive objectives capable of dynamically balancing invariance and diversity requirements hold potential for extending these methods to more complex and heterogeneous data distributions.

By challenging the necessity of negative samples, these innovations signify a paradigm shift in self-supervised learning, fostering models that are computationally efficient, noise-robust, and scalable across diverse domains. Moving forward, synthesizing the strengths of non-contrastive methods with the discriminative advantages of traditional contrastive techniques presents a compelling opportunity to create more unified and versatile self-supervised frameworks.

3.5 Multi-modal Contrastive Learning

Multi-modal contrastive learning represents a transformative avenue in self-supervised learning by expanding contrastive methodologies to incorporate data from different modalities, such as vision, language, and audio. The overarching goal is to align embeddings across modalities while preserving intra-modal discriminative properties, thereby ensuring coherent and robust cross-modal representations. By leveraging complementary features from heterogeneous sources, multi-modal contrastive frameworks bolster downstream tasks ranging from vision-language modeling to video-audio understanding.

A seminal example of multi-modal contrastive learning is the CLIP framework, which employs contrastive objectives to match images with their associated text descriptions [71]. In such setups, the embeddings of paired modalities are brought closer in a shared latent space using a contrastive loss, such as InfoNCE, while unrelated pairs are pushed further apart. The efficacy of CLIP lies in its ability to distill semantic correspondences between modalities, resulting in representations that generalize exceptionally well to zero-shot tasks. However, CLIP's reliance on extensive curated datasets for pairwise alignment introduces scalability challenges, particularly in domains where labeled text-image pairs are scarce.

To address some of these limitations, frameworks like FactorCL introduce the notion of factorized representations [63]. These approaches emphasize disentangling shared and modality-unique features, ensuring that cross-modal alignment retains essential modality-specific information while focusing on shared semantics. For example, by capturing unique textual nuances (e.g., syntax or domain-specific jargon) alongside visual content (e.g., spatial relations or object properties), such factorized methods produce feature representations capable of handling diverse task-specific requirements. Nevertheless, the challenge here lies in tailoring representation disentanglement without inadvertently suppressing meaningful inter-modal interactions.

The robustness and versatility of multi-modal contrastive learning extend beyond image-text models into more complex multimodalities like video and audio. Recent works like Video-Audio Contrastive Learning demonstrate the utility of temporal coherence and audiovisual synchronization in learning spatiotemporal features [72]. By contrasting synchronized representations of video frames and audio sequences, these models foster robustness to modality-specific noise, such as background clutter in frames or irrelevant sounds. While effective, one limitation is the high computational overhead associated with maintaining and processing long-horizon temporal dependencies inherent in video-audio tasks.

From a theoretical standpoint, many multi-modal contrastive frameworks can be formalized as joint alignment problems across views, leading to loss formulations such as the generalized InfoNCE or mutual information objectives [28]. Formally, given multi-modal data $M = \{x_1, x_2, \ldots, x_k\}$, where x_i represents individual modalities, the objective is to maximize the mutual information $I(h(x_1), h(x_2), \ldots, h(x_k))$ while ensuring decorrelation with non-positive pairs. Alignment constraints across paired modalities guarantee robust cross-modal consistency, while intra-modal constraints mitigate embedding collapse.

Despite their successes, multi-modal contrastive learning still contends with several challenges. Dependence on paired datasets restricts scalability to uncurated or domain-specific modalities where paired annotations are sparse or unavailable [73]. Furthermore, optimizing joint embeddings often necessitates sophisticated augmentation pipelines or modality-specific losses to achieve desirable invariance across a range of transformations [74]. These pipelines can exacerbate computational inefficiency, limiting real-world applicability.

Emerging trends in multi-modal contrastive learning explore novel solutions to address these challenges. One promising approach involves the integration of generative models for data augmentation to create semantically consistent and diverse pseudo-pairs [65]. Additionally, frameworks like DINO-MC leverage multiple scales of views to handle varying object sizes and ensure cross-modal alignment in specialized datasets such as remote sensing imagery [64]. These innovations point to the potential of using domain-agnostic generative augmentation or scaling techniques to mitigate dataset bias while enhancing multi-modal feature representations.

Looking forward, future research can focus on minimizing dependence on explicit pairs by leveraging pseudo-labeling mechanisms to align modalities in unstructured datasets [73]. Automated approaches for optimizing task-specific invariances during multi-modal pretraining also represent a fertile ground for exploration [67]. Moreover, frameworks like DiMAE highlight the need for domain-invariant fusion techniques to accommodate both shared and divergent inter-domain features [50].

In conclusion, multi-modal contrastive learning offers a compelling framework for aligning heterogeneous data modalities into unified representations, with applications across domains such as vision-language models, video-audio tasks, and cross-domain generalization. However, unlocking its full potential requires improved scalability, reduced dependency on paired data, and mechanisms for balancing shared and modality-specific features systematically. Research in this space is likely to define the next generation of robust and high-performing multi-modal self-

supervised systems.

3.6 Evaluation Metrics and Open Challenges in Contrastive Frameworks

Evaluating the effectiveness of contrastive learning frameworks is critically tied to understanding how well representations capture meaningful semantic features and generalize to diverse downstream tasks. Traditional metrics have centered around linear probing, fine-tuning, and transfer learning, but with contrastive learning evolving toward multi-modal and increasingly complex data contexts, assessing representation quality has grown more nuanced. This subsection explores the current landscape of evaluation methods, highlighting their benefits, limitations, and the challenges that persist in achieving a comprehensive assessment framework.

Linear probing has long been a popular evaluation technique, relying on training a simple linear classifier over frozen representations to test their linear separability for specific downstream tasks. While this provides insight into the global structure of learned representations, it oversimplifies real-world scenarios where complex, non-linear decision boundaries are often needed. This has driven the adoption of fine-tuning, which evaluates model performance under complete weight adaptation. However, fine-tuning introduces dependencies on hyperparameters and downstream task-specific requirements, complicating the ability to make consistent comparisons across different methods [75], [76].

To provide a more theoretically grounded approach, metrics based on mutual information (MI) have emerged, derived from objectives like InfoNCE. Metrics such as normalized MI or entropy-based measures quantify the shared information preserved between augmented image pairs [77]. These metrics, however, face challenges with estimation biases in finite batch regimes that may lead to inaccurate MI values. Furthermore, MI alone does not capture the downstream efficacy of representations, necessitating its integration with explicit evaluation tasks, such as clustering separability or transfer robustness [76]. The tradeoff between maximizing abstract information and ensuring practical task-relevance therefore remains an open question.

Contrastive learning's reliance on negative samples introduces additional complexities in evaluation. Negative sampling strategies are critical for avoiding feature collapse, with metrics such as negative alignment scores and hard negative distributions providing some insight into their quality and diversity. However, challenges arise from false negatives, where samples belonging to the same class are mistakenly treated as negatives due to overlapping augmentations or data transformations. Queue-based sampling protocols, such as those in Momentum Contrast (MoCo) and Positive-Unlabeled Contrastive Learning (PUCL), attempt to mitigate this through mechanisms like intra-class exclusion, but no universally accepted metric has emerged for quantitatively evaluating negative sample design across datasets of varying complexity [60], [78]. This lack of standardization hinders the ability to systematically study sampling biases.

In multi-modal settings, where the alignment of representations across modalities such as vision, language, and audio is crucial, additional evaluation challenges arise. Metrics like cross-modal retrieval accuracy and latent space alignment have become standard practice, but they often fail to capture modality-specific characteristics that are essential for real-world applications [79]. Furthermore, the inherent modality imbalance in multi-modal datasets can distort optimization dynamics, necessitating balanced evaluation protocols capable of reflecting these nuances without introducing biases. Current metrics inadequately address these intricacies, limiting their applicability in evaluating the robustness and versatility of multi-modal methods.

The computational burden and resource requirements for existing evaluation methodologies further exacerbate these challenges. Large-scale datasets like ImageNet remain the gold standard for benchmarking representation quality, but their resource-intensive nature raises concerns about accessibility for resource-constrained institutions and the environmental cost of large-scale experiments [55]. Scaled-down datasets and low-resolution embeddings have emerged as alternatives to alleviate these concerns, yet they risk excluding critical dimensions of performance evaluation, such as generalization on diverse or large-scale data distributions.

Moving forward, there is a pressing need for standardized and comprehensive evaluation protocols that integrate diverse dimensions of representation quality. These should incorporate tests of linear separability, clustering performance, robustness to distribution shifts, and domain adaptation into cohesive frameworks capable of reflecting real-world deployment conditions. Special attention should also be given to cross-domain generalization, with emerging benchmarks for domain-invariant representations, such as those proposed in [50], offering promising directions. Concurrently, lightweight and memory-efficient evaluation protocols are required to democratize research, making advanced contrastive learning techniques accessible to a wider audience. With the increasing prevalence of hybrid self-supervised methods that blur the lines between contrastive and non-contrastive learning, existing metrics must also evolve to accommodate diverse optimization goals.

In summary, the evaluation of contrastive learning frameworks has not yet matched the rapid innovation seen in the methods themselves. Persistent challenges, from false negative biases to inadequate cross-modal and computational efficiency metrics, constrain our ability to holistically assess representation quality. Addressing these gaps will unlock the broader applicability of contrastive learning systems, enabling them to meet the demands of both practical and theoretical advancement in self-supervised learning research.

4 Architectures and Training Strategies for Self-Supervised Learning

4.1 Neural Network Architectures for Self-Supervised Learning

Self-supervised learning (SSL) has achieved remarkable success in leveraging unlabeled data to train neural network

architectures, enabling robust representation learning without manual annotations. This subsection examines the architectural advancements that have optimized convolutional neural networks (CNNs), vision transformers (ViTs), and hybrid designs for extracting rich and transferable visual features in SSL settings. These architectures play a crucial role in the efficacy of SSL methods, as their design directly impacts feature diversity, global context understanding, and computational efficiency.

Convolutional neural networks (CNNs) remain a foundational choice for SSL tasks due to their well-established ability to extract hierarchical, spatially localized features. Architectures like ResNet have been extensively employed in SSL frameworks, such as SimCLR and MoCo, to facilitate contrastive learning by mapping visual instances into discriminative embedding spaces [6], [11]. ResNet's modular design allows its intermediate layers to encode representations across multiple scales, which are pivotal for both semantic and fine-grained tasks like object detection and segmentation. However, recent investigations highlight that deeper networks (e.g., ResNet-101) tend to outperform shallower counterparts, particularly in multi-task SSL setups, where high-dimensional embeddings are better suited for diverse downstream evaluations [39].

Despite their success, CNNs face limitations in modeling long-range dependencies and global context, which are crucial for tasks requiring a holistic understanding of visual content. This has catalyzed a shift toward vision transformers (ViTs), which leverage self-attention mechanisms to encode global interdependencies between image regions. ViTs have demonstrated significant potential in SSL, as the patch-based tokenization paradigm aligns naturally with masked image modeling and dense predictive tasks [14], [30]. For example, the masked autoencoder (MAE) framework, which reconstructs masked patches of input images, has emerged as a state-of-the-art SSL technique, utilizing ViTs' ability to dynamically focus on relevant image regions to learn compact yet expressive representations [10]. Additionally, self-supervised ViTs display emergent properties, such as implicit semantic segmentation capabilities without relying on annotated datasets, as investigated through methods like DINO [12].

The adaptability of transformers to multi-modal SSL has further extended their dominance beyond unimodal settings. By processing diverse data formats (e.g., visual and textual embeddings) through unified self-attention mechanisms, ViTs have enabled SSL frameworks like CLIP to achieve both robust unimodal capability and cross-modal alignment [26]. Nonetheless, ViTs suffer from high computational overhead due to their quadratic complexity with respect to input size, which underscores the need for scalable designs or training accelerations such as multi-crop augmentation [4].

Hybrid architectures, integrating the localized inductive biases of CNNs with the global receptive fields of transformers, have emerged as a promising alternative. Recent works reveal that such architectures achieve complementary feature extraction, making them versatile across a broader range of SSL tasks. Cross-architectural strategies, such as convolutional tokenization layers feeding into transformer encoders, address the weaknesses of both CNNs (limited global understanding) and ViTs (inefficient local pattern recognition) [14], [59]. By unifying these paradigms, hybrid models achieve superior generalization compared to purely convolutional or transformer-based architectures.

Despite these advances, challenges persist in optimizing architectures for SSL. First, CNNs and ViTs each face trade-offs in terms of scalability, data efficiency, and compatibility with modern SSL losses like contrastive and predictive objectives [8]. Second, computational inefficiencies in training large architectures, especially in resource-constrained settings, highlight the demand for lightweight yet powerful designs [80]. Furthermore, robust cross-architecture interactions remain an underexplored area; emerging methods must balance architectural consistency with adaptability to diverse tasks.

Future directions in SSL architectures emphasize dynamic adaptability—optimizing networks to tailor their learned representations for specific data distributions or task domains. Innovations such as domain-dependent masking strategies in MAEs or task-aware attention heads in hybrid models offer a path to more efficient and resilient SSL frameworks [10], [17]. By synergizing architectural efficiency with SSL's pretext-task diversity, the field is poised to advance the scalability, versatility, and interpretability of self-supervised learning frameworks.

4.2 Training Efficiency and Scalability in Self-Supervised Learning

Scalability and training efficiency are critical considerations in self-supervised learning (SSL), particularly in the context of handling large-scale datasets and deploying energy-efficient models. This subsection delves into strategies to improve training efficiency and scalability, forming a natural progression from the architectural advancements discussed previously and laying foundational insights that resonate with paradigms such as masked image modeling (MIM) explored later. Key areas of focus include distributed training architectures, augmentation methodologies, and optimization practices tailored to SSL's unique challenges.

Efficient utilization of hardware and computational resources is essential for extending SSL to real-world applications. Distributed training emerges as a pivotal strategy to address computational bottlenecks, enabling SSL models to process large-scale datasets by distributing workloads across multiple GPUs or compute nodes. Techniques such as gradient synchronization and memory optimization algorithms ensure scalability without compromising convergence stability. For instance, memory-efficient queue-based mechanisms, as seen in frameworks like MoCo, are instrumental in managing massive datasets by reducing dependence on large batch sizes [12]. Momentum contrast methods leverage dynamic queues to synchronize representations from distributed workers, exemplifying how computational cost can be balanced without sacrificing the quality of learned features. Despite these advances, challenges like communication overhead and hardware constraints persist, particularly in resource-limited environments.

Augmentation strategies play a complementary role in boosting SSL scalability and training efficiency. Multi-crop augmentations, popularized by methods such as SwAV,

enable models to encode representations that encompass both global and localized visual features, enhancing the diversity of learned representations [7]. Thoughtful augmentation design is critical to avoid redundancy and ensure computational feasibility, especially when dealing with high-resolution inputs. Adaptive augmentation techniques further refine this process by dynamically tailoring transformations to dataset-specific characteristics, thereby maximizing semantic richness without introducing noise or artifacts [61]. These tailored techniques are especially relevant for domains like medical imaging or remote sensing, where preserving the integrity of critical information is paramount [81].

Optimization algorithms are equally central to enhancing scalability. A major focus has been on accelerating convergence and ensuring stability during large-scale SSL training. Cyclic learning rates, as evidenced in approaches like BYOL, foster dynamic learning trajectories that prevent optimization from stagnating, reducing the overall number of pretraining epochs required [22]. VICReg exemplifies how representation collapse can be alleviated through covariance regularization, enabling efficient learning even with smaller batch sizes [33]. Similarly, weight-averaging techniques, such as the use of exponential moving averages, have become integral to maintaining consistent optimization dynamics and preventing performance degradation during extended pretraining sessions [82].

Nevertheless, the heavy computational overhead inherent to many SSL techniques limits accessibility, restricting cutting-edge advancements to institutions equipped with state-of-the-art resources. To address this, lightweight methodologies, such as whitening-based approaches, have been introduced to reduce the reliance on negatives in contrastive learning, thereby simplifying training pipelines and improving scalability to resource-constrained settings [77]. Low-rank approximations and feature compression strategies provide further promise, maintaining representation quality while minimizing memory and computational requirements [83]. These techniques highlight the ongoing innovation aimed at democratizing SSL by making powerful frameworks viable on smaller datasets or less resource-intensive systems.

Emerging directions include dynamic task allocation strategies and modular optimization frameworks that are attuned to SSL's iterative training processes. For example, pretext tasks that adapt dynamically to convergence metrics can improve training efficiency by redistributing computational effort without diminishing the diversity of learned representations [7]. Additionally, the use of generative models for producing task-specific augmentations holds potential for introducing richer semantic variation into training datasets, complementing existing augmentation strategies while maintaining efficiency [15]. When integrated with distributed training mechanisms, these approaches could further bridge the divide between computational feasibility and high-quality representation learning, aligning with the scalability goals central to modern SSL.

In summary, training efficiency and scalability in SSL represent an intricate balance between computational innovation, augmentation design, and optimization techniques. These factors collectively build upon the architectural foundations of SSL, as highlighted in prior discussions, and set the stage for efficient frameworks like those in MIM. While significant strides have been made, particularly in reducing the reliance on labeled data and ensuring adaptability across tasks, challenges like resource demands and domain-specific adaptations temper broader adoption. By focusing on lightweight adaptations and dynamic methodologies, SSL can continue to evolve toward practical, scalable solutions that advance both research and real-world deployment, underpinning its growing influence in computer vision.

4.3 Masked Image Modeling (MIM) and Partial Input Reconstruction

Masked Image Modeling (MIM) has emerged as a prominent framework in self-supervised learning (SSL) that focuses on reconstructing corrupted or missing portions of input data, enabling models to learn spatial-semantic relationships and capture hierarchical representations of visual information. Rooted in strategies like masked language modeling in natural language processing (e.g., BERT), MIM leverages masking to introduce an information bottleneck, ensuring that the learning process is conditioned on the visible context. This subsection delves into the implementation, progress, and implications of MIM and partial input reconstruction in computer vision.

The core idea behind MIM is to mask a portion of the input image, requiring the model to reconstruct either the pixel-level details or latent feature representations of the masked regions. Recent studies have shown that operating in the latent space, rather than reconstructing pixel values directly, enhances the semantic quality of learned features by shifting the focus from low-level textures to global image understanding. For instance, He et al.'s Context Autoencoder (CAE) reformulates MIM by leveraging contextual information to regress representations of masked patches in the latent space while simultaneously reconstructing the pixel values through a decoder [37]. This separation between representation learning and pretext task completion allows CAE's learned features to transfer more effectively to downstream tasks.

Masking strategies are instrumental in determining reconstruction difficulty and model generalization, with recent works exploring diverse approaches. Randomized masking, as seen in Masked Autoencoders (MAEs), employs stochastic removal of patches, ensuring broad applicability by preventing task shortcuts [10]. Conversely, adaptive masking methods like semantic-aware masking select patches more strategically, often focusing on content richer in information, critical in domains where data distributions are imbalanced or diverse (e.g., remote sensing) [50]. Such strategies optimize feature generalization by balancing information redundancy and representational diversity.

A significant trade-off in MIM techniques lies between pixel-level reconstruction and prediction within the abstract latent space. Pixel-level methods target detailed reconstructions but risk overfitting the model to textures and local data biases. In contrast, latent-space approaches shift toward semantic abstractions, which enhance downstream generalization but may neglect fine-grained detail necessary

for tasks like segmentation. Hybrid frameworks attempt to bridge this divide, as seen in works like Cross-Attention Masked Autoencoders, which integrate multi-scale attention to retain patch-level interdependencies and reconcile contextual and detailed feature learning [14].

The recent adoption of vision transformers (ViTs) within MIM frameworks underscores their suitability for patch-based tokenization and global context integration. ViT-based models like MAEs have demonstrated exceptional pretraining efficiency and transferability, leveraging their natural capacity to process global feature relationships across patches [30]. Notably, their ability to scale seamlessly with larger datasets and architectures positions MIM methods at the forefront of scalable SSL for vision.

Current limitations of MIM pertain to computational overheads and the need for large-scale datasets. Transformer-based architectures are particularly resource-intensive during pretraining, with masking strategies introducing complexity by requiring iterative reconstructions. Simple decoder designs, as proposed in works like Hierarchical PreTraining (HPT), could offer a promising solution, enabling faster convergence and reduced training costs [84].

Emerging trends include extending MIM principles to multimodal domains (such as vision-language fusion), leveraging masking to enforce cross-modal consistency (e.g., masked vision-language models like SLIP) [42]. Additionally, constructing domain-invariant masking objectives, especially for tasks like medical imaging and remote sensing, could further enhance robustness across diverse datasets [50].

In conclusion, MIM continues to enable advancements in SSL by reinforcing generalizable feature learning rooted in partial input reconstruction. As research develops, integrating adaptive masking, efficient latent-space reconstruction, and cross-domain adaptability will define its trajectory, enabling broader applicability in both academic and industrial contexts. Ensuring scalability, computational efficiency, and task-specific tuning will remain critical for MIM to achieve its full potential across diverse vision applications.

4.4 Multimodal Learning Architectures for Self-Supervised Learning

Multimodal self-supervised learning architectures harness the complementary strengths of diverse data modalities, such as vision and language, to uncover intricate relationships and learn unified, enriched feature representations. These architectures have grown in prominence alongside the success of large-scale foundational models spanning multiple modalities. By aligning and integrating varied streams of information through shared self-supervised objectives, these models unlock sophisticated capabilities across numerous downstream tasks, including vision-language understanding, video-audio correlation, and 3D spatial reasoning.

A foundational approach in multimodal self-supervised learning is joint embedding architectures, which aim to establish shared latent spaces where inputs from different modalities are semantically aligned. For instance, methods like CLIP [28] leverage contrastive paradigms to bridge modalities by treating paired modality-specific inputs (e.g., an image and its caption) as positives, while contrasting them against unpaired negatives. This strategy enforces a high degree of semantic alignment between modalities, enabling robust multimodal retrieval and alignment. Similarly, the Image-based Joint-Embedding Predictive Architecture [85] extends this paradigm by incorporating predictive objectives, further refining representation quality through tasks such as reconstructing or predicting masked modality-specific features.

Complementarily, modality fusion architectures, often implemented through vision transformers with cross-attention mechanisms, provide a structured paradigm for dynamically aggregating and adapting information across modalities. For example, hierarchical attention models [30] process vision and language representations by leveraging transformers that attend to modality-specific features before integrating them into unified embeddings. These approaches frequently employ auxiliary tasks, such as masked token prediction or cross-modal matching, to enhance robustness in both unimodal and multimodal scenarios. Furthermore, hybrid models like those discussed in [86] combine the inductive biases of convolutional neural networks (CNNs) with the global attention capabilities of transformers, achieving state-of-the-art performance in visual and multimodal video tasks.

Multimodal learning often benefits from auxiliary supervision grounded in modality-spanning data characteristics. In video-audio tasks, for example, temporal synchronization acts as an inherent supervisory signal, allowing models to align and capture dependencies between modalities, as demonstrated in [34]. This method employs RGB and optical flow features in tandem, using cross-modal co-training to improve video understanding. However, the dependency on large-scale paired datasets presents challenges, especially for supervised contrastive objectives. To address this, techniques like those in [87] utilize clustering across visual and audio modalities to generate pseudo-labels, enabling training on unpaired multimodal datasets.

Despite these innovations, significant challenges remain for multimodal architectures, particularly regarding dependency on high-quality paired data for contrastive learning and sensitivity to domain shifts between modalities. For instance, modalities such as text and vision may encode information in fundamentally different structural formats, complicating alignment. Emerging paradigms, like dynamic task adaptation [48], mitigate these limitations by embedding constraints such as temporal or cyclic consistency across modalities, thereby reducing reliance on extensive paired data. Additionally, domain-invariant alignment techniques are being developed to disentangle shared and modality-specific information, further addressing discrepancies in multimodal datasets.

Collaborative learning frameworks are increasingly expanding the scope of multimodal self-supervised learning by incorporating additional modalities such as auditory, textual, and spatiotemporal data into unified systems. Frameworks like BraVe [88] showcase the advantages of augmenting unimodal tasks with auxiliary multimodal signals. Future avenues may investigate more granular alignment techniques, such as the dynamic fusion of multimodal streams during training, and underexplored applications, such as 3D spatial reasoning in multimodal environments.

In summary, multimodal self-supervised learning offers transformative opportunities for cross-modal intelligence by leveraging the diversity of information encoded in individual modalities as well as their synergistic interactions. Key progress in joint embedding, modality fusion, and auxiliary supervision frameworks underscores the field's potential. However, overcoming persistent hurdles, including computational scalability, domain alignment, and paired data dependency, will be critical to expanding the applicability of these models to real-world scenarios and broader tasks.

4.5 Specialized Training for Enhanced Feature Quality

Specialized training strategies for enhancing feature quality in self-supervised learning (SSL) are designed to prioritize robust, generalizable, and transferable representations, aligned with domain-specific challenges or constrained learning environments. These tailored approaches focus on learning representations that are resilient to noise, adaptable across tasks and domains, and reflective of the underlying semantic structures of the data. This section explores domain-specific pretraining methods, the incorporation of auxiliary and multi-task learning frameworks, and mechanisms to improve robustness and transferability in out-of-distribution conditions.

Domain-specific SSL pretraining involves adapting self-supervised strategies to unique features and constraints of specific application domains, such as medical imaging, autonomous navigation, or remote sensing. Unlike general-purpose image datasets like ImageNet, domain-specific data often exhibit unique challenges, such as limited diversity, high inter-class similarity, or varying spatial resolutions. For example, in medical imaging applications, feature extraction must account for modality-specific properties like high-resolution textures in CT or MRI images, often requiring dedicated pretext tasks like volume reconstruction or modality-specific masking strategies [27]. Similarly, recent work on remote sensing imagery leverages multi-scale data augmentations and global-local alignment to address relatively limited object diversity while capturing both broad land-use features and fine-grained textures [64].

Auxiliary and multi-task learning frameworks complement primary SSL objectives by introducing additional learning signals that regularize the resulting representations. Auxiliary tasks, such as predicting augmentation parameters [89] or enforcing consistency across augmented views with varying spatial or temporal resolutions [90], help reduce overfitting to spurious correlations and guide the model toward learning diverse and robust features. Multi-task setups, which combine multiple pretext objectives within a unified framework, improve representation quality by aggregating complementary perspectives. For instance, hybrid methods that integrate transformation-invariant contrastive losses with diversity-enforcing reconstruction tasks have been shown to achieve significant performance gains by exploiting distinct pretext task synergies [33], [91]. Furthermore, domain-adapted frameworks that dynamically adjust auxiliary objectives, such as learning representations aligned with domain-specific augmentation distributions, enable improved adaptability to downstream tasks without resorting to supervised fine-tuning [50].

Enhancing robustness and out-of-distribution generalization within SSL frameworks is a growing area of focus, particularly for applications in safety-critical fields like autonomous systems and healthcare. Robust feature extraction strategies leverage adversarial regularization techniques to prevent models from depending on shortcut correlations or dataset-specific artifacts. For instance, lenses trained to introduce adversarial perturbations to the pretext task overcome this shortcut exploitation, encouraging models to prioritize more salient, generalizable features [40]. Consistency regularization approaches, such as enforcing invariance over multiple domain-perturbed variations of data [92], also improve neural network resilience against environmental changes and domain shifts. Moreover, hybrid SSL methods that balance invariance and equivariance objectives (e.g., preserving color sensitivity for fine-grained classification tasks) enable robust yet domain-appropriate representations that accommodate diverse downstream task requirements [93].

An emerging challenge is the computational efficiency of specialized training strategies. Many domain-specific and auxiliary frameworks rely on advanced masking techniques, multi-head architectures, or iterative task adaptation, which substantially increase training complexity and resource requirements. Recent innovations, such as lightweight task-adaptive networks and localized feature pretraining approaches, show promise in minimizing such overhead without compromising representational power [14], [94]. Future research may focus on combining efficient task-adaptive pipelines with improved pretext task modularity for scalable SSL in resource-constrained environments [95].

Looking ahead, specialized training strategies will likely benefit from tighter integration of multimodal and multi-label self-supervised objectives, leveraging shared representations across data modalities or hierarchical labels. For example, tasks involving both spatial and temporal consistency, such as video representation learning, could inspire enhanced frameworks that seamlessly scale across spatial and semantic granularities [72], [96]. Another promising direction is dynamic task reorganization during training, wherein task priors are adjusted iteratively based on model performance or dataset-specific signals, to improve both convergence and feature robustness. Ultimately, addressing these challenges will unlock higher-quality representations and broader applicability for SSL in specialized domains.

4.6 Emerging Trends and Challenges in Architectures and Training

Self-supervised learning (SSL) has revolutionized computer vision by enabling the self-directed extraction of robust and domain-agnostic features from unlabeled data. However, as SSL architectures and training paradigms evolve, emerging trends illuminate both opportunities for advancement and persistent challenges that demand critical attention. This subsection delves into these developments, spanning computational efficiency, architectural synergies, multimodal integration, adaptive training strategies, and the broader obstacles in developing scalable and generalizable SSL models—providing a cohesive bridge between specialized strategies and future directions in SSL.

A noticeable shift in SSL architecture is the increasing preference for Vision Transformers (ViTs) over conventional convolutional neural networks (CNNs). ViTs, with their natural aptitude for capturing global relationships and contextual information, have demonstrated superior performance in several SSL frameworks, including those driven by contrastive and masked image modeling objectives [10], [97]. However, this transition is not without complications. The computational demands of ViTs—both in terms of memory usage and training time—create barriers to scalability, particularly for large datasets and organizations with limited computational resources [98]. Recent strategies like local masked reconstruction [99] and adaptive token selection have emerged to address these inefficiencies, offering potential solutions to balance representation quality with algorithmic practicality. Still, achieving a universally efficient and accessible SSL paradigm remains an active area of exploration.

Another promising development is the emergence of hybrid architectures that blend the local feature-capturing efficiency of CNNs with the global reasoning strengths of ViTs [14], [97]. These architectures have shown substantial improvements in downstream tasks requiring fine-grained detail or dense prediction, such as segmentation. Yet, challenges remain. Differences in optimization dynamics and computational patterns between CNNs and transformers can complicate the design of unified frameworks, necessitating further innovations in cross-architecture integration [10]. Specific optimization strategies capable of harmonizing these architectural components are critical to unlocking the potential of such hybrid systems.

In parallel, multimodal SSL frameworks are gaining traction, particularly in scenarios where vision converges with other data modalities, such as text and audio. Models like CLIP, which align image and text modalities during representation learning, have demonstrated the power of such approaches in tasks like visual question answering and cross-modal retrieval [100]. Nonetheless, the necessity of large paired datasets for pretraining remains a significant bottleneck, especially in underexplored domains lacking such extensive multimodal data. Innovations like pseudo-labeling and domain-specific alignment objectives are promising avenues to increase the applicability of multimodal SSL systems while reducing reliance on vast paired datasets [53].

Concurrently, adaptive training paradigms in SSL are undergoing significant evolution. Dynamic optimization techniques, which adapt loss functions and augmentations in response to model performance, represent a growing trend that seeks to create more diverse and high-quality representations. For instance, hard example mining [101] and adversarial contrastive learning [43] introduce task complexity dynamically, enhancing the robustness and generalization of representations. However, these methods often significantly escalate computational requirements due to iterative data sampling or adversarial perturbation steps. Research focused on formulating lightweight adaptive objectives could mitigate these challenges, promoting cost-effective scalability.

A critical and persistent challenge in SSL, however, lies in mitigating the effects of shortcut learning. Models frequently rely on spurious correlations in pretext tasks, which can produce features with limited transferability to downstream tasks. To counteract this, techniques involving adversarial feature perturbations have emerged to remove shortcut features and encourage the extraction of more meaningful representations [40]. That said, these approaches remain underexplored across diverse datasets and domains. A deeper theoretical understanding of how SSL models prioritize features during training could provide insights for more robust pretext task design [22].

Finally, sustainability emerges as an overarching concern. Many state-of-the-art SSL models require extensive computational resources, raising environmental and accessibility concerns, especially for under-resourced settings [55], [98]. Initiatives to democratize SSL, such as scaled-down architectures and minimal data augmentation pipelines, have underscored the potential of lightweight, efficient frameworks [98]. These approaches aim to make self-supervised techniques more accessible while minimizing their environmental impact.

Looking forward, future advancements in SSL must confront these challenges head-on, prioritizing computational efficiency, cross-domain scalability, and sustainability. The synergy between adaptive optimization, multimodal learning, and hybrid architectures holds significant promise while necessitating meticulous attention to data quality, fairness, and generalization capacity. By addressing these pressing concerns, the potential of SSL to transform computer vision and extend its impact across diverse domains can be fully realized.

5 Evaluation Protocols and Performance Benchmarks

5.1 Datasets and Benchmark Selection

The increasing prominence of self-supervised learning (SSL) in computer vision has necessitated diverse and robust evaluation protocols. Central to such protocols is the selection of datasets that not only reflect SSL models' pretraining potential but also comprehensively assess their downstream task performance. This subsection examines commonly used datasets in SSL research, comparing their utility, scalability, and domain-specific relevance. Additionally, it delves into dataset selection strategies, trade-offs, and emerging challenges, establishing a framework for robust benchmarking practices.

For SSL pretraining, large-scale, varied, and unlabeled datasets play a pivotal role. ImageNet, though originally designed for supervised vision tasks, remains one of the most commonly used datasets for SSL due to its rich semantic diversity and balanced class distribution. Many studies have demonstrated that SSL models pretrained on ImageNet exhibit competitive performance when fine-tuned on downstream tasks such as object detection and semantic segmentation [4], [11]. However, questions arise concerning its continued dominance, as its proposed closed-world assumption does not align with SSL's fundamental goal of leveraging vast, uncurated, and unlabeled data. Addressing this limitation, datasets such as SEER, consisting of a billion uncurated images [23], have shown promise in assessing SSL models' ability to generalize beyond curated settings.

Similarly, domain-specific datasets like those in remote sensing [102] and medical imaging [81] are underscoring niche applicability of SSL frameworks in specialized and real-world contexts where annotation scarcity is prevalent.

Downstream evaluation datasets are equally critical and play the role of assessing feature transferability across various tasks. Benchmarks such as PASCAL VOC and COCO have been widely adopted to evaluate SSL-pretrained models in object detection, segmentation, and other dense prediction tasks, where performance relies on robust semantic representations [6], [11]. Smaller datasets like CIFAR-10, CIFAR-100, and STL-10 often serve as proxies for low-resource tasks, enabling SSL approaches to be tested for efficiency in low-label regimes [11], [103]. Additionally, video-specific datasets such as Kinetics and AVE assess temporal dynamics modeled by SSL in video-based downstream tasks [34]. Emerging applications are also leveraging domain-adaptive datasets, such as SA-1B for object detection, to reveal SSL's capacity in complex real-world transfer settings [104].

Despite the utility of benchmark datasets, limitations persist. Pretraining datasets like ImageNet often contain cultural and geographic biases that may inadvertently skew downstream performance evaluations and hinder cross-domain generalization [4], [105]. Moreover, the standard practice of relying on curated datasets during pretraining (e.g., ImageNet) introduces implicit biases that may diminish SSL's capacity to leverage broader, noisier data distributions [4], [6]. To counter these limitations, researchers are exploring uncurated sources (e.g., SEER [23]) and dynamic selection of benchmark tasks to reflect practical, non-curated real-world distributions. Domain-specific benchmarks further provide tangible evidence of SSL's potential beyond conventional settings, such as in medical imaging with datasets designed for CT scans and pathology tissue [81] and remote sensing with domain-adapted multispectral benchmarks [59].

Moving forward, expanding the diversity of SSL benchmark datasets across tasks is necessary to establish the practical relevance of SSL systems. Incorporating multimodal datasets, such as those combining vision and language (e.g., vision-language models evaluated on datasets like VQA or COCO captions), might present more intricate benchmarks for evaluating SSL in multi-faceted contexts [7]. Additionally, rigorous cross-domain evaluations are necessary to reflect model robustness, such as the ability to generalize across unseen data distributions [18]. Benchmarks emphasizing fairness, robustness to adversarial perturbations, and common corruptions—such as the recent inclusion of out-of-domain datasets for downstream evaluation [1]—represent pressing areas of research interest.

Finally, the field would benefit from standardized evaluation frameworks, such as methods introduced in platforms like solo-learn, which streamline benchmarking pipelines across data domains [80]. A unified framework combining multimodal perspectives, domain-specific extensions, and dynamic pretraining and evaluation techniques would magnify SSL's relevance and enable direct, equitable comparisons of future approaches.

5.2 Evaluation Metrics for Representation Quality

Evaluating the quality of learned representations in self-supervised learning (SSL) is essential for understanding their utility across diverse visual tasks. This subsection explores key metrics and their relevance in quantifying pretraining success and downstream generalization capability, thereby bridging the gap between label-independent feature learning and task-specific performance.

A foundational metric in SSL evaluation is *linear probing*, whereby a simple linear classifier is trained on frozen feature representations to assess their utility for classification tasks. This technique provides insights into the linearly separable structure embedded within the representation, serving as a proxy for downstream performance [106]. Linear probing is computationally efficient, offering an early glimpse into representation quality, but it inherently overlooks higher-order semantic relationships that are inaccessible to linear decision boundaries. To address this limitation, *fine-tuning performance* serves as a complementary metric, allowing the pretrained model to adapt entirely to target tasks [11]. Fine-tuning uncovers the full potential of learned features under the lens of gradient updates, but its utility as a standalone measure is constrained due to entanglement with task-specific optimization factors.

In addition to classification accuracy, clustering-based metrics provide an alternative lens to evaluate the semantic structure of representations. Metrics like Adjusted Rand Index (ARI) and normalized Mutual Information (nMI) measure the alignment between learned clusters in embedding space and ground-truth labels [107]. These metrics offer insights into the coherence of groups formed within the representation space, revealing how well the features capture semantic distributions. However, their reliance on post hoc alignment with labels limits their applicability in purely unsupervised contexts, necessitating supplementary evaluation tools.

Robustness evaluations are increasingly central in assessing SSL representation quality. Metrics that measure model performance under adversarial perturbations, data corruptions, or random noise—such as accuracy under common corruptions or adversarial testing datasets like ImageNet-C—highlight the invariance and generalizability of learned features to real-world variabilities [1]. This robustness is particularly vital for applications in high-stakes domains like autonomous driving and medical imaging, where reliability under imperfect conditions is paramount [31].

Another critical dimension of evaluation is cross-domain generalization, which examines representation performance when transferred to datasets with significant domain shifts. These shifts may stem from variations in data distribution, texture, or style, as seen in benchmarks like PACS and VLCS [33], [82]. Models such as BYOL and VICReg have demonstrated robust transfer capabilities, excelling in domain adaptation scenarios. Metrics that quantify this generalization reveal the adaptability of features across diverse contexts while underscoring the versatility of SSL approaches.

Nearest-neighbor (kNN) evaluation offers a lightweight and training-free alternative to assess feature quality directly. By classifying data points based on proximity within

the learned feature space, kNN methods provide valuable insights into the neighborhood structure of embeddings [12]. While computationally efficient for small-scale settings, kNN struggles with scalability for large datasets due to its reliance on pairwise comparisons, limiting its applicability for extensive evaluations.

Emerging metrics are expanding the evaluation landscape to accommodate the evolving goals of SSL. For instance, *dimensionality metrics* examine how well features are distributed across embedding dimensions to avoid representational collapse, aligning with methods like Barlow Twins and VICReg that emphasize redundancy reduction [33], [108]. These metrics highlight the importance of balanced feature representations, ensuring that embeddings are informative and decorrelated. Additionally, cross-modal evaluation metrics, such as cross-modal alignment scores or retrieval metrics like Recall@K, are gaining prominence as SSL models increasingly incorporate multimodal data [26]. These metrics enable assessments of shared embeddings for tasks like image captioning and text-image retrieval [102].

Despite these advancements, challenges in standardizing SSL evaluation persist. Many current metrics operate under idealized settings, often overlooking dynamic factors in real-world applications, such as dataset biases or the effects of data augmentation [31]. There is a growing need for unified benchmarks that balance robustness, fairness, and interpretability, facilitating unbiased comparisons across SSL methods. Developing task-agnostic evaluation techniques rooted in theoretical concepts, such as mutual information maximization [28], could further illuminate the intrinsic properties of learned representations.

In conclusion, the evaluation of SSL representation quality spans a diverse spectrum, encompassing task performance, robustness, and theoretical guarantees. As SSL moves toward broader multimodal and out-of-distribution applications, integrating these metrics into cohesive, generalized benchmarking frameworks will be critical for driving fair and comprehensive comparisons. Such frameworks will ensure that SSL models are equipped not only to succeed in idealized benchmarks but also to thrive in real-world, dynamic environments.

5.3 Evaluation Protocols and Transferability Assessment

Evaluation of self-supervised learning (SSL) models has become increasingly nuanced, mirroring the complexity and diversity of tasks for which these representations are employed. The primary focus of evaluation lies in assessing the transferability of learned representations across diverse downstream tasks, ensuring the utility of SSL pretraining beyond the original domain. This subsection explores prevalent evaluation protocols, discusses their trade-offs, and highlights challenges and emerging trends in measuring the transferability and reliability of SSL features.

A cornerstone of SSL evaluation is linear probing, wherein a linear classifier is trained atop frozen SSL-learned representations to evaluate how effectively these features encode information relevant to downstream tasks. Linear probing is lightweight and interpretable, offering a quick approximation of representation quality across domains. However, it may obscure deeper insights into the adaptability of representations under task-specific fine-tuning. Recent studies [11], [109] demonstrate the utility of this approach but argue for complementary evaluations that probe multi-level feature utility.

Fine-tuning, on the other hand, involves adapting the entire SSL-pretrained model to a downstream task, thus evaluating the full capacity of the representation and its compatibility with gradient-based optimization. This method often reveals discrepancies in SSL methods that show strong linear evaluation scores but underperform when adapted to tasks requiring intricate feature refinements, such as semantic segmentation and object detection [11], [56]. Fine-tuning provides a more complete understanding of representational transferability but is computationally expensive and sensitive to hyperparameter choices, underscoring the need for robust evaluation frameworks.

Another evaluative tool is k-nearest neighbors (kNN) classification, which offers a non-parametric assessment of feature quality without additional training. Recent work [109] highlights its utility in scenarios where interpretability and computational efficiency are prioritized. However, kNN can underrepresent deeper semantic structures in the data compared to more sophisticated evaluation protocols.

Cross-domain transfer evaluation has gained traction as a critical benchmark for generalization. This involves testing SSL-powered representations on datasets distinct from those used during pretraining. Several studies [110], [111] reinforce the importance of evaluating SSL under domain-shift scenarios, revealing that robustness to shifts often correlates with higher-quality general-purpose features. Domain-adaptation tasks, such as on VLCS and PACS benchmarks [32], provide complementary insights into domain-agnostic representation learning. However, challenges remain, particularly in ensuring that cross-domain robustness does not inadvertently degrade task-specific performance within the original domain [32].

To unify and standardize evaluation practices, frameworks such as VISSL and solo-learn have emerged as vital tools for consolidating diverse SSL benchmarks. Such platforms facilitate comparisons across methods, streamline experimental reproducibility, and reduce biases arising from inconsistent implementation details [112]. Yet, hyperparameter sensitivity during evaluations, such as the impact of learning rate schedules or batch sizes, remains a non-trivial challenge, often complicating the fair comparison of SSL methods [11].

Emerging methodologies also focus on task-specific assessment techniques. For instance, studies employing feature disentanglement [63] or probing task-relevant invariances [67] aim to refine understanding of the alignment between SSL-learned features and downstream task requirements. This perspective shifts the evaluation discourse toward identifying situations where specific SSL paradigms excel or break down. For instance, contrastive approaches often learn robust global features for coarse-grained tasks but may falter on fine-grained or localized applications without adaptation [14], [111].

Looking ahead, more holistic evaluations are imperative—benchmarks should seamlessly incorporate measures for fairness, robustness to perturbations, efficiency, and

interpretability. Incorporating metrics such as bias amplification, behavioral performance under domain-shifts, and energy efficiency in large-scale self-supervised tasks [62], [113] will ensure representation quality aligns with practical deployment demands. Additionally, frameworks for multimodal evaluation, wherein task objectives span multiple data domains like vision, language, or even 3D spatial representation [42], [50], represent another promising research frontier.

In sum, while existing protocols provide valuable insights into representational quality, the SSL community would benefit from the establishment of comprehensive and predictive metrics that holistically evaluate transferability and representation competency across both traditional and emerging tasks. Such advancements will catalyze the deployment of SSL in real-world, high-stakes scenarios.

5.4 Benchmarking Frameworks and Practical Considerations

Benchmarking frameworks serve as critical infrastructure for evaluating and comparing Self-Supervised Learning (SSL) models, ensuring reproducibility, fairness, and meaningful insights into their performance. In computer vision, benchmarking SSL models entails unique challenges due to the diversity of pretraining objectives, architectures, and downstream evaluation tasks. This subsection explores existing benchmarking frameworks, their strengths and limitations, and key considerations essential for designing robust, scalable, and fair evaluations.

To standardize SSL evaluation, unified frameworks such as VISSL and solo-learn have emerged, addressing the need for consistent experimentation. VISSL [4] supports large-scale SSL pretraining and downstream evaluations, incorporating a modular design for implementing state-of-the-art algorithms like MoCo, SimCLR, and BYOL. Its comprehensive suite of transfer learning metrics—ranging from linear probing and fine-tuning to clustering—enables structured comparisons across diverse tasks. Similarly, solo-learn consolidates SSL tools into a unified interface, streamlining experimentation with contrastive, generative, and hybrid SSL paradigms [4]. By enforcing consistent hyperparameters, augmentation pipelines, and evaluation protocols, these platforms reduce experimental variability and promote fairness in comparative studies.

However, the effectiveness of these frameworks hinges on the benchmarks and datasets they employ. ImageNet has served as a cornerstone benchmark [4], [9], achieving widespread adoption for pretraining and downstream evaluations. Nonetheless, its object-centric nature does not capture the heterogeneous, unstructured environments encountered in real-world applications, which limits generalization and risks creating biased evaluations [70]. As a response, emerging efforts now employ datasets like MS COCO and PACS for more diverse scene-level tasks [114]. In addition, video datasets such as UCF101, HMDB51, and Kinetics enable the evaluation of SSL models in spatiotemporal domains, highlighting the growing significance of multi-modal and temporal coherence in benchmarks [26], [115].

Another pivotal consideration is hyperparameter sensitivity, as the performance of SSL models can depend heavily on optimization schedules, augmentations, and evaluation configurations [6]. Addressing this issue calls for adaptive frameworks and rigorous sensitivity analyses to disentangle genuine model improvements from those arising due to hyperparameter tuning. VISSL partially addresses this challenge by incorporating configuration sweeps to improve the reliability of comparisons [4]. Furthermore, resource efficiency is an ongoing concern, as SSL methods often rely on computationally intensive pretraining over millions of images, making it prohibitive for smaller research groups or institutions [4]. Lightweight pretraining solutions [116], including multi-scale evaluation strategies that adapt data and computational requirements dynamically, represent promising directions for improving accessibility and scalability.

Beyond computational concerns, benchmarking must address emerging dimensions of SSL effectiveness, such as fairness, robustness, and applicability in unstructured environments. Fairness-oriented evaluation frameworks evaluate model resilience to demographic or dataset biases, as pretrained SSL models risk amplifying biases present in source datasets [46]. Unsupervised object discovery benchmarks, independent of curated datasets, are also gaining prominence as SSL is increasingly applied beyond object-centric tasks [117]. Furthermore, robustness metrics, including sensitivity to adversarial perturbations and domain shifts, are underexplored but crucial for assessing SSL's reliability in diverse real-world applications [1].

In summary, benchmarking frameworks like VISSL and solo-learn have made significant progress in standardizing SSL evaluations, fostering fairer comparisons through modular pipelines and consistent benchmarks. Nonetheless, to meet the growing need for broader applicability, future efforts must prioritize diversity in datasets, improved hyperparameter tuning protocols, and fairness-oriented approaches. Additionally, heightened focus on computational efficiency, robustness, and the evaluation of multi-modal capabilities will be instrumental in building benchmarking frameworks that reflect real-world complexities. These advancements will ultimately enable SSL to transition smoothly from controlled research environments to practical applications across diverse domains.

5.5 Emerging Trends in SSL Evaluation

A recent shift in evaluating self-supervised learning (SSL) frameworks has focused on broadening the scope of assessment, ensuring that evaluation mechanisms align with the increasingly complex demands of real-world applications. Traditional metrics such as linear probing accuracy, fine-tuning evaluations, and clustering efficiency have dominated SSL benchmarks, but emerging trends are introducing deeper dimensions such as fairness, interpretability, and multimodality.

Fairness in SSL Evaluation: A growing recognition of model biases has spotlighted the need for evaluating SSL methods on fairness criteria, particularly when pretrained models are applied across diverse datasets. Biases introduced during pretraining on curated datasets, such as ImageNet, often amplify demographic inequities across socially sensitive attributes, reflecting systemic gaps in SSL evaluation approaches [6], [118]. Techniques that

measure disparate impacts and error rates across demographic groups are being explored, though their application remains nascent. For example, methodologies leveraging self-supervision to disentangle features aligned with sensitive attributes from those aligned with downstream tasks have been proposed to mitigate bias amplification [119]. The inclusion of fairness-aware augmentation strategies or domain-specific benchmarks may address these biases, particularly when dealing with datasets outside the scale of mainstream pretrained corpora.

Interpretability and Explainability: Another emerging priority is the interpretability of SSL-pretrained models. Recent work emphasizes quantifying the contribution of learned features towards downstream tasks, employing interpretability metrics such as saliency maps and attribution-based methods to explain model predictions [49], [120]. For example, spatially consistent learning frameworks have demonstrated that augmenting SSL training with geometric consistency constraints can improve interpretability by aligning feature maps to meaningful object locations, addressing the black-box nature of SSL systems [121]. Advances in attention mechanisms, such as those embedded in vision transformers, are also explored to illustrate hierarchical feature attribution across input samples [95]. Yet, challenges persist in explaining contrasts between representations and their semantic alignment across varied input distributions.

Multimodal Embedding and Alignment: As SSL moves beyond single-modality data, its evaluation must include benchmarks that measure cross-modal alignment and representation. Multi-modal SSL frameworks, including those integrating text, image, and audio data, require embedding spaces that encode shared semantics across modalities, a capability central to tasks like vision-language navigation or cross-view geo-localization [41], [73]. Multimodal contrastive approaches, such as leveraging shared augmentations between disparate modalities, have been shown to enhance transferability [35], [122]. Evaluation protocols for such models must extend beyond unimodal tasks to jointly measure task-specific and cross-modal performance under adversarial conditions, where inter-modal dependencies are tested.

Broader Implications and Challenges: As the field embraces these emerging evaluation paradigms, several challenges arise. For fairness, the definition of universal metrics applicable across unimodal and multimodal tasks remains elusive, particularly under distributional shifts [67]. While interpretability methods enrich our understanding of SSL models, they often trade off with computational complexity. Multimodal systems, similarly, raise questions about how well adaptation protocols generalize to dramatically under-explored domains, such as remote sensing or creative AI [61], [64].

Future directions must focus on synthesizing these goals through unified benchmarks tailored for fairness, multimodality, and task-centric evaluations across diverse domains. For instance, the development of standardized pretraining datasets that ensure demographic and domain diversity could mitigate unintended biases. Additionally, augmenting SSL interpretability methods with causal reasoning promises more human-aligned models that are not just effective but also trustworthy. Furthermore, multimodal evaluation frameworks should account for dynamic task definitions, evolving inputs, and interpretability constraints to continuously measure robustness and usability. These trends underscore the evolution of SSL evaluation into a more nuanced, application-relevant discipline, reflecting the expanding horizon of its utility across artificial intelligence applications.

6 Applications of Self-Supervised Learning in Computer Vision

6.1 Image-Level Tasks

Self-supervised learning (SSL) has emerged as a powerful paradigm for addressing image-level computer vision tasks by leveraging the vast amount of unlabeled image data to learn meaningful visual representations. This capability has significantly mitigated the dependency on extensive labeled datasets, enabling improved performance across various fundamental tasks such as image classification, object detection, and semantic segmentation.

In image classification, SSL methods like SimCLR and MoCo have demonstrated state-of-the-art results by learning discriminative features through contrastive objectives that maximize similarity between augmented views of the same image while distinguishing them from other images in the dataset. SimCLR, for instance, relies on aggressive augmentation strategies and large batch sizes to generate diverse positive and negative pairs, enabling the model to learn invariant and robust representations [4]. MoCo, on the other hand, introduces a memory bank mechanism for dynamic negative sample generation, making it computationally efficient in smaller batch settings [11]. Despite their successes, these approaches face challenges such as reliance on meticulous data augmentations, high memory costs, and the need for large-scale datasets to prevent representation collapse—a limitation partially addressed by non-contrastive methods like BYOL and SimSiam [11], [22].

Object detection, which requires localized representations of objects within an image, also benefits significantly from SSL. Pretraining via self-supervised methods like SwAV and DINO has shown remarkable improvements over conventional supervised pretraining on benchmarks like COCO, achieving comparable or even superior precision for bounding box regression and object classification [9], [12]. These methods leverage joint embedding architectures to enhance spatial feature capture, with Vision Transformers (ViTs) such as DINO showcasing emergent localization properties within self-supervised features [12]. Additionally, hybrid methods like HASSOD have introduced innovations such as hierarchical clustering of region proposals, improving self-supervised detection performances in challenging datasets like LVIS [104].

For semantic segmentation, SSL methods exploit pixel-wise feature consistency through dense prediction tasks. For instance, masked autoencoders (MAEs) reconstruct partially visible images, compelling models to learn complex contextual relationships across pixels. These approaches, well-suited to dense prediction tasks, outperform supervised baselines on semantic segmentation benchmarks like ADE20K [10]. Furthermore, patch-level self-supervised

methods, such as SelfPatch, leverage the architecture of Vision Transformers to model relationships across disjoint image regions, thereby enhancing granularity in predictions for segmentation tasks like COCO instance segmentation and Cityscapes [14].

An emerging trend across these tasks is the integration of multi-task SSL frameworks. Multi-task learning approaches, such as those combining generative and predictive objectives, foster richer feature embeddings by sharing representations across related pretext tasks, thus better aligning with downstream requirements [5], [39]. However, challenges remain in mitigating task interference and optimizing the balance between pretext and downstream task relevance.

Despite SSL's progress, obstacles persist. One key issue is the scalability of SSL methods to uncurated, imbalanced image datasets, which often contain noisy or irrelevant samples. Methods like SEER have shown promising results by extending SSL to uncurated internet-scale datasets, but they demand significant computational resources, raising concerns about energy efficiency and accessibility [23]. Additionally, biases introduced during pretraining (e.g., dataset-specific augmentations) may limit generalizability to diverse real-world domains, an issue pointed out by works such as "Demystifying Contrastive Self-Supervised Learning: Invariances, Augmentations and Dataset Biases" [6].

Looking forward, the integration of SSL with other paradigms like weak supervision and domain generalization offers exciting potential to improve transferability and robustness for image-level tasks. For instance, combining pseudo-label generation with SSL has demonstrated gains in adapting to domain shifts, particularly for applications in medical imaging and remote sensing, where annotations are sparse [81], [106]. Future research must also address interpretability concerns by unraveling the high-dimensional feature space learned by SSL and aligning representations with human-recognizable concepts, as investigated in works on theoretical understandings of SSL [56].

In sum, SSL has revolutionized image-level computer vision tasks, offering scalable and versatile representation learning frameworks. Innovations in pretext task design, architectural advancements, and multi-task objectives will likely drive the next wave of SSL applications, underscoring the paradigm's relevance across both academic research and industry applications.

6.2 Video Analysis and Temporal Modeling

Self-supervised learning (SSL) has made significant advances in video analysis and temporal modeling by addressing the unique challenges posed by video data, such as high dimensionality, temporal dynamics, and sequential dependencies. These advancements have facilitated the development of robust spatiotemporal representations from unlabeled video datasets, enabling diverse applications such as action recognition, event detection, and video summarization.

A key enabler of SSL's success in video analysis lies in the design of pretext tasks that emphasize temporal coherence and motion dynamics. For instance, temporal context prediction has proven effective, with tasks like frame order permutation and video pace prediction compelling models to capture temporal dependencies. Studies demonstrate that predicting the correct chronological order of video frames allows models to better understand complex sequential patterns, thereby improving the recognition of actions unfolding over time [11]. Similarly, pace prediction tasks encourage models to encode fine-grained motion-sensitive features by predicting the playback speed of frame sequences, enabling enhanced temporal modeling [34]. These approaches vividly illustrate the capacity of SSL methods to discover latent temporal structures in video data without relying on manual annotations.

Contrastive learning—a cornerstone of SSL—has also been adapted for spatiotemporal representation learning in videos. Temporal contrastive learning frameworks maximize similarity between temporally aligned positive samples while contrasting these with temporally misaligned negatives. Cross-view strategies, which utilize complementary input streams such as RGB frames and optical flow, have proven especially effective in jointly capturing appearance and motion features [34]. However, this adaptation presents challenges, particularly when identifying hard negative pairs in lengthy video streams. Semantically similar segments can be misclassified as negatives, leading to suboptimal contrastive learning objectives and necessitating more sophisticated negative sampling strategies.

Generative approaches, such as reconstruction-based methods, have also played a critical role in SSL for video data. Masked autoencoders, for example, have been extended to video analysis by reconstructing occluded or missing temporal segments, forcing models to infer holistic and long-range spatiotemporal dependencies. Such techniques excel in dense prediction tasks like video segmentation and temporal action localization, where understanding the continuity and context of motion is paramount. Additionally, predictive modeling—where models anticipate future frames from past sequences—has emerged as a complementary strategy, embedding anticipatory motion cues into learned representations [10]. Nonetheless, these generative frameworks often come with high computational demands due to the inherent complexity of modeling dynamic video sequences, limiting their scalability for large-scale applications.

An exciting development in video SSL is the adoption of multimodal learning paradigms. These approaches leverage visual, auditory, and textual information to enhance temporal modeling, enabling better handling of ambiguities in one modality by complementing it with cues from another. For example, aligning video frames with audio tracks, such as dialogue or environmental sounds, has been shown to yield robust and noise-resistant representations [32]. Such multimodal frameworks present promising opportunities for downstream applications, including audiovisual scene understanding and video-audio synchronization.

Despite the progress, challenges persist in scaling SSL frameworks to handle the diverse and often unpredictable nature of video data. Variability in temporal resolution, abrupt scene transitions, and occlusions can disrupt the learning of consistent spatiotemporal representations. Additionally, simplistic pretext tasks may lead to shortcut exploitation, where models learn irrelevant biases instead

of meaningful temporal patterns [31]. Future research could address these issues by introducing adaptive temporal masking strategies or dynamically optimized learning objectives that account for complex temporal variations across videos. Moreover, the integration of richer contextual supervision—such as leveraging textual metadata or external knowledge graphs—may significantly advance the generalization and domain adaptability of SSL models for videos.

In summary, SSL has transformed video analysis and temporal modeling by offering scalable approaches to learn spatiotemporal representations from unlabeled data. Through innovative pretext tasks, generative objectives, and multimodal integration, these frameworks have expanded the potential for solving challenges in video-related applications. Addressing computational overheads, ensuring robustness to temporal variability, and promoting cross-domain generalization remain crucial directions for the continued advancement of SSL in video analysis.

6.3 Medical Imaging Applications

The application of self-supervised learning (SSL) to medical imaging has shown transformative potential in addressing the acute scarcity of labeled datasets and dependency on expert annotations prevalent in this domain. Medical imaging data, often characterized by multimodal formats such as X-rays, computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound, possesses rich structural and semantic complexities that SSL approaches are increasingly effective at leveraging. These methods enable the extraction of robust and generalizable visual representations by exploiting intrinsic properties of the data, fostering advancements in diagnostic accuracy, segmentation of anatomical regions, and disease characterization.

SSL finds particular utility in classification tasks across medical imaging modalities. For example, models pretrained using SSL have demonstrated significant improvements in disease detection on both 2D (e.g., chest X-rays) and 3D imaging modalities (e.g., CT scans, MRIs). Techniques like contrastive learning and masked autoencoding are commonly employed, facilitating the generation of invariant, semantically meaningful representations without reliance on human-annotated labels. Prior studies employing methods akin to those in [27] have shown that encoding high-level semantic structures early on allows for improved generalization to downstream tasks, such as detecting pleural effusion or classifying intracranial hemorrhages. This aligns with findings in broader SSL research [11], reinforcing the value of pretext-invariant objectives in achieving domain adaptability.

In segmentation, SSL has enabled high levels of performance in delineating anatomical structures and pathological regions, particularly in modalities like 3D MRI and CT imaging. Methods such as masked autoencoders and self-distillation have been adapted to 3D spatial data, with innovations emphasizing volumetric data consistency. For instance, [37] provides a compelling framework wherein patch-level reconstruction drives an encoder-decoder architecture to learn spatially aware representations, crucial for tasks such as tumor delineation or cardiac ventricle segmentation. Notably, dynamic masking mechanisms tailored to medical data characteristics, such as intensity thresholds for specific tissue types, have extended these methods' efficacy.

The design of domain-specific pretext tasks is another critical enabler for medical SSL. For instance, reconstructive tasks modeled after frameworks like [10] have been adapted by masking domain-relevant features (e.g., organs or anomalies) and requiring their reconstruction. Other approaches, inspired by advances in multi-modal SSL [42], integrate radiological images with textual diagnostic reports, enabling the alignment of visual representations with semantic embeddings derived from clinical descriptions. This fusion of vision and language modalities offers a unique pathway to enhancing decision-support systems in medical imaging.

Despite these advances, challenges remain. Medical images often exhibit domain-specific artifacts such as noise, low contrast, and scanner-specific variations, complicating pretraining and feature extraction processes. For models to generalize effectively, these variations necessitate pretext tasks aligned to the domain, such as contrastive frameworks tailored to texture-based or spectral nuances ([64]). Furthermore, the potential for shortcut learning—where models leverage irrelevant biases—emphasizes the need for robust adversarial strategies, as evidenced in [40]. Moving forward, future work could explore hybrid architectures, as in [42], for integrating cross-modal and hierarchical features into a unified SSL framework.

Emerging trends suggest a growing role for SSL in longitudinal disease monitoring and temporal modeling, areas currently underexplored in medical contexts. Techniques leveraging time-series imaging data, building upon approaches such as [123], may yield significant diagnostic improvements in progressive conditions like Alzheimer's disease or cancer metastasis. Additionally, efforts to standardize evaluation protocols in SSL for clinical-grade modeling will underpin its successful deployment. Unified benchmarks, as recommended broadly in [112], are an urgent necessity in medical imaging to ensure interoperability and reproducibility across datasets, modalities, and diagnostic tasks.

In conclusion, SSL in medical imaging demonstrates unparalleled promise in reducing the reliance on annotated data and expert supervision, unlocking new possibilities for precision diagnostics. Future advancements will likely emerge from domain-specific pretext tasks, robust modeling against diverse data artifacts, and innovations in multimodal representation learning. By addressing current challenges, these methods hold the potential to redefine automated medical image analysis while ensuring scalability and accessibility across healthcare systems.

6.4 Remote Sensing and Environmental Applications

Self-supervised learning (SSL) has positioned itself as a transformative methodology in remote sensing, addressing the vast and complex landscape of satellite imagery where high-quality labels are frequently sparse or unavailable. Leveraging the intrinsic structures and patterns inherent in geospatial data, SSL minimizes the dependence on annotated datasets while enhancing the generalizability of learned representations across diverse tasks and domains.

This progress has significantly advanced various remote sensing applications, including land use classification, environmental monitoring, and disaster assessment, making SSL an increasingly vital tool within geospatial AI.

Land use and land cover classification stand out as foundational applications of SSL in remote sensing. In this context, SSL methods have demonstrated the ability to effectively harness the spectral-spatial characteristics of multispectral and hyperspectral datasets, deriving robust features without requiring detailed pixel-level annotations. Contrastive learning approaches, for instance, empower models to align similar patches sampled from temporally or spatially consistent locations, as shown by work that leverages geo-tagged data to define positive pairs amid spatiotemporal alignments [46]. Additionally, pretext tasks explicitly designed for geospatial domains, such as spectral band reconstruction and spatiotemporal pixel correlation, further ensure that the learned representations capture both spatial textures and spectral uniqueness [46].

Environmental monitoring exemplifies another domain where SSL's promise continues to grow, addressing critical tasks like deforestation detection, climate change assessment, and natural hazard identification. The learning of temporal dynamics in remote sensing time-series data is particularly advantageous here. Temporal contrastive objectives, which create temporal positive pairs by splitting input sequences into subwindows, enable the detection of gradual land changes as well as abrupt anomalies such as landslides or floods [124]. Furthermore, masked modeling techniques, including masked autoencoders, excel at reconstructing missing spectral or spatial information, effectively capturing contextual relationships at both local and global scales. These capabilities are vital for anomaly detection in ecosystems, where slight deviations from baseline conditions can provide critical early warnings [37].

A prevailing challenge in remote sensing remains the domain adaptation problem, caused by significant variations in satellite imagery due to differences in sensors, resolutions, and lighting conditions. SSL frameworks offer a compelling approach to this issue by enabling visual representations to align across datasets with divergent distributions. Methods such as contrastive pretraining, combined with spatial and spectral augmentations, enhance interoperability across multi-source Earth observation datasets [124]. Moreover, dynamic augmentation pipelines that manipulate spectral bands or simulate atmospheric effects contribute to the robustness and adaptability of these frameworks.

Despite these advancements, challenges persist. Models pretrained on biased datasets, such as those dominated by urban imagery like ImageNet, may exhibit limited efficacy over rural or ecological regions where training data is sparse. Innovative approaches, such as utilizing uncurated satellite imagery from diverse geographies and adopting adaptive sampling strategies, are critical to addressing these disparities [4]. Additionally, the multiscale dependencies unique to geospatial data often remain underutilized in current SSL approaches. This limitation underscores the necessity for hybrid frameworks that integrate fine-grained, pixel-level objectives with broader scene-level contextual modeling.

Emerging trends further highlight the potential of combining SSL with multimodal learning for geospatial AI. Integrating satellite imagery with auxiliary data modalities—like climate variables or textual geographic reports—has the potential to generate more nuanced and actionable representations [124]. Similarly, advancements in graph-based SSL, wherein spatial entities are represented as nodes, hold promise for capturing the intricate non-Euclidean geometries and environmental interdependencies inherent in these datasets.

To fully realize the potential of SSL in remote sensing, future research should concentrate on adapting frameworks to the geospatial context, developing representative benchmarks that encompass real-world challenges, and embedding SSL-driven models into operational workflows for time-critical applications such as disaster response and sustainable resource management. By fostering scalability, adaptability, and cross-domain generalization, SSL can position remote sensing as a powerful tool for addressing pressing environmental challenges and promoting global sustainability, harmonizing seamlessly with broader efforts in multimodal and cross-domain learning.

6.5 Multimodal and Vision-Language Applications

Multimodal learning, particularly the integration of vision and language, has emerged as a pivotal area within self-supervised learning (SSL), enabling advancements in tasks like visual question answering (VQA), cross-modal retrieval, and image captioning. Within this domain, SSL offers a framework to leverage the vast amounts of unlabeled multimodal data, effectively aligning visual and textual representations without requiring explicit annotations. By designing pretext tasks that bridge these modalities, SSL promises versatile representations capable of generalization across downstream tasks.

Key advancements in multimodal SSL hinge on aligning visual and language embeddings through contrastive objectives. For example, methodologies such as CLIP and its derivatives utilize a large corpus of paired image-text data where a contrastive loss aligns the vision and language representations in a joint embedding space. The primary advantage of this approach lies in its ability to derive semantic correspondences between images and text without labeled fine-tuning, enabling zero-shot generalization capabilities. This property has shown state-of-the-art performance on tasks like text-to-image retrieval and open-world object recognition. Meanwhile, similar concepts extend beyond text-image alignment to involve spatiotemporal modalities like video and audio, as in frameworks for self-supervised video transformers, which exploit alignment cues across varying frame rates and spatial resolutions [90].

Generative-based multimodal SSL methods also demonstrate significant potential by predicting or reconstructing one modality based on another. Masked autoencoders, for instance, have been extended to multimodal setups where textual prompts guide the reconstruction of masked regions in images [37]. Such strategies enforce a dense understanding of cross-modal correspondences, thus fostering representations advantageous to VQA and multimodal summarization tasks. Unlike contrastive approaches that focus on global correspondence at the instance level, generative

approaches often focus on capturing the fine-grained relationships between modalities.

Vision-language learning has also benefited from advanced augmentation strategies. Techniques like positive temporal contrast, introduced for adapting SSL to remote sensing and video datasets [64], demonstrate how domain-specific augmentations can enhance robustness. By leveraging diverse pretext tasks, these methods improve alignment in datasets featuring subtle inter-modality inconsistencies, such as those caused by perspective or temporal variations.

A prominent challenge remains in addressing the inherent ambiguities and biases in multimodal datasets. For instance, SSL models may inadvertently overfit to dominant spurious correlations, requiring carefully designed pretext losses to mitigate such biases. GenView leverages generative models to synthesize diverse, semantically meaningful views of an image while maintaining consistency with its textual embedding, thereby enhancing robustness in multimodal SSL [65]. Similarly, pseudo-labeling techniques that group weakly aligned visual and language representations iteratively refine cross-modal embeddings by rectifying noisy pairings [73].

Despite the progress, significant hurdles remain in scaling multimodal SSL models to specialized domains such as remote sensing, medical imaging, and 3D vision. In fields like remote sensing, the constrained availability of paired multimodal data exacerbates the reliance on pretraining routines or handcrafted pretext tasks tailored to domain-specific nuances [64]. Meanwhile, adapting vision-language transformers to efficiently handle volumetric data or sparse cues imposes considerable computational and modeling challenges.

Emerging trends suggest that compositionality and disentanglement may be instrumental for the next generation of multimodal SSL frameworks. By disentangling content- and domain-specific (e.g., style) features, models stand to gain enhanced generalizability, as evidenced by domain-invariant autoencoders that preserve content across stylistic variations [50]. Furthermore, the integration of causal representation learning into SSL opens new avenues for aligning modalities based on underlying semantic structures over transient correlations.

In summary, multimodal SSL exemplifies a transformative approach to bridging the semantic gap between vision and language. While contrastive, generative, and augmentation-based methods excel in large-scale paired data scenarios, challenges remain in addressing spurious multimodal correlations and scalability to under-explored domains. Future developments are likely to focus on improving pretext task granularity, embracing causality-driven objectives, and devising domain-adaptable architectures to unlock the full potential of vision-language SSL methods.

6.6 Automotive and Navigation Systems

The integration of self-supervised learning (SSL) into automotive and navigation systems has significantly transformed perception mechanisms for autonomous vehicles (AVs) and robotics, aligning with its broader impact across computer vision and multimodal domains. By leveraging vast amounts of unlabeled sensory data, SSL provides robust feature representations that enable real-time scene understanding, efficient object detection, and dynamic motion prediction in increasingly complex driving environments. This subsection examines advancements, challenges, and future directions for SSL in autonomous systems, building on the foundational principles discussed in vision-language learning and extending them into spatiotemporal contexts.

Scene segmentation and understanding serve as critical components for AV navigation, requiring precise semantic comprehension of roads, lanes, obstacles, and traffic participants to ensure safe and efficient operation. SSL advancements have propelled this task through innovative architectures like masked autoencoders (MAEs), which excel in learning spatially explicit visual representations by reconstructing masked content. For instance, approaches such as LoMaR demonstrate the feasibility of locally reconstructing masked patches, achieving efficiency and accuracy in dense segmentation tasks [99]. These methods are particularly advantageous in automotive scenarios, where high-resolution scene representations are essential for lane detection, road surface analysis, and obstacle recognition. Further extending robustness, Domain-Invariant Masked Autoencoders utilize style-mix augmentation strategies to enable generalization across diverse driving conditions, accommodating environmental variations such as weather, road textures, and lighting [50].

Dynamic object and motion detection constitutes another essential capability enabled by SSL in AVs, addressing the temporal dimension critical for tracking and trajectory prediction. Predictive and contrastive SSL frameworks facilitate the learning of temporally coherent representations, enabling improved performance in motion forecasting. For example, SSL-Lanes introduces specialized pretext tasks like velocity consistency prediction, outperforming traditional supervised models even in challenging datasets like Argoverse, enhancing motion prediction accuracy under uncertain traffic scenarios [125]. Emerging techniques incorporating contrastive mechanisms, such as core-tuning, refine forecasting models by aligning inter-frame representations while ensuring critical separation between distinct object trajectories [78].

Domain generalization remains a persistent challenge for SSL-powered AV perception systems, as vehicles encounter geographically and environmentally diverse contexts. SSL offers compelling alternatives to conventional supervised transfer learning via domain-agnostic frameworks such as DiMAE, which employs cross-domain reconstruction tasks to achieve invariant representations that generalize effectively across urban and rural settings [50]. Complementary innovations like adaptive masking techniques emphasize context-aware patches, further improving cross-domain transferability [126]. Such strategies enable SSL-driven AV systems to adapt to varying traffic rules and environmental dynamics, ensuring consistent and robust performance across regions with distinct operational requirements.

Despite these advancements, SSL faces scalability and deployment challenges in real-world navigation scenarios. A primary concern is computational efficiency, as AV systems necessitate models capable of balancing accuracy and latency under stringent resource constraints. Lightweight MAE variants offer promising solutions, reducing energy

consumption while retaining competitive accuracy in reconstruction tasks [99]. Moreover, ensuring robustness under adversarial conditions such as sensor corruption or extreme weather remains critical. Techniques like denoising autoencoders exhibit early success in extracting clean representations from noisy sensor inputs, offering tangible pathways to enhance AV reliability in adverse environments [127].

Looking ahead, the fusion of SSL with multi-modal capabilities promises to further advance AV perception systems. Integration of modalities such as LiDAR, radar, and audio into shared SSL-driven representation spaces can provide a comprehensive understanding of spatial and semantic contexts. Additionally, hybrid SSL frameworks incorporating limited annotated data for guiding representation learning could better address domain-specific challenges [128]. Novel pretext tasks emphasizing dynamic sequence alignment and time-series consistency offer exciting directions for improving spatiotemporal reasoning critical for navigating dynamic, real-world environments.

In conclusion, SSL is redefining the landscape of autonomous driving and navigation. By leveraging vast unlabeled datasets, SSL fosters transferable, robust, and efficient representations that align with advancements across vision-language and emerging specialized domains. Continuous innovation in multi-modal integration, domain adaptation, and computational efficiency will further democratize SSL applications, facilitating its role as a cornerstone of safe, scalable, and intelligent automotive systems.

6.7 Emerging and Specialized Applications

The adaptability of self-supervised learning (SSL) in computer vision has enabled its expansion into emerging and specialized domains, illustrating its versatility and potential to drive innovation in fields with unique challenges. SSL's ability to learn from unlabeled data is particularly valuable in areas where manual annotation is expensive, infeasible, or requires domain-specific expertise. This subsection explores cutting-edge applications of SSL in 3D computer vision, robotics, and creative AI, highlighting novel methodologies, challenges, and opportunities for further research.

SSL has increasingly been adopted for 3D computer vision tasks such as point cloud segmentation, 3D object detection, and scene reconstruction, where data inherently exhibits rich spatial and geometric properties. Traditional 3D data, often presented as point clouds or volumetric grids, is irregular and sparsely structured, imposing challenges that differ from 2D image data. Methods like CrossPoint [129] leverage self-supervised contrastive learning to align 3D point cloud representations with 2D image data, enabling a richer understanding of spatial relationships. These techniques often involve multi-modal alignment, allowing cross-referencing between modalities to enhance representation quality. CrossPoint achieves significant performance gains in 3D segmentation and classification tasks, demonstrating the potential of SSL to bridge gaps between 2D and 3D domains. However, challenges such as scale invariance, occlusion handling, and computational overhead in large-scale 3D datasets remain open research directions.

Robotics is another domain where SSL is making profound impacts, particularly in tasks requiring interaction with dynamic and unstructured environments. Here, SSL frameworks enable robotic agents to learn effective control policies and object manipulation strategies without labeled supervision. Approaches like MViTac [130] combine visual and tactile modalities to develop robust multi-sensory representations. By employing intra- and inter-modality losses, MViTac facilitates robotic systems in adapting to alternating textures, materials, and object geometries, which are crucial for tasks like grasp prediction and material classification. Building on multi-modal SSL, further research could explore the inclusion of other sensory data, such as auditory signals, providing avenues for developing truly holistic robotic systems. However, limitations persist in scaling these systems to real-world datasets due to uneven access to large multi-modal datasets and the computational complexity of fusing diverse signals.

In parallel, creative domains have embraced SSL to push the boundaries of generative and artistic AI applications. For example, techniques such as Representation-Conditioned Generation (RCG) [131] leverage self-supervised representations in latent space to improve semantic coherence in generated outputs. RCG has demonstrated an unprecedented ability to produce high-fidelity images by addressing long-standing challenges in unconditional generation. Additionally, frameworks such as Ge^2-AE [132] have introduced innovative paradigms by reconstructing data in both pixel and frequency domains. These methods enhance artistic outputs in computational design, style transfer, and generative art, where capturing global structures and fine-grained details simultaneously is essential. Despite these advancements, open challenges exist in scaling SSL generative models to multimodal creative tasks, such as combining textual prompts with visual representations for complex artistic objectives.

Looking forward, SSL's potential in these emerging applications identifies several key future directions. First, multi-modal integration—whether across visual, tactile, auditory, or textual domains—remains a promising yet underexplored dimension. Techniques must address the challenges of cross-modal alignment, robust redundancy reduction, and efficient handling of noise, as discussed in [28]. Second, improving computational efficiency and scalability will be critical for deploying SSL models in resource-constrained domains. Innovative sparsity mechanisms like LoMaR [99] provide promising solutions to mitigate the computational bottlenecks of traditional methods. Finally, ethical considerations and bias mitigation are paramount for SSL's deployment in specialized domains, particularly in robotics and creative applications where societal impact is direct. Integrating fairness criteria, as highlighted in [133], could guide SSL's ethical adoption.

In summary, the expansion of SSL into specialized domains underscores its transformative potential. By enabling autonomous learning in 3D vision, robotics, and creative AI, SSL not only addresses long-standing challenges but also spawns novel research opportunities. However, achieving generalization, efficiency, and ethical consistency across specialized applications requires sustained interdisciplinary innovation and rigorous evaluation frameworks.

7 Challenges, Limitations, and Opportunities in Self-Supervised Learning

7.1 Computational and Scalability Challenges

Self-supervised learning (SSL) in computer vision has achieved remarkable advancements by leveraging unlabeled data to pretrain models that yield strong generalization in downstream tasks. However, its scalability is constrained by significant computational and energy demands, especially when applied to training large neural networks on vast uncurated datasets. These challenges hinder practical adoption, particularly for resource-limited institutions, necessitating an exploration of computational and scalability barriers, along with emerging solutions.

The training of large-scale SSL models, such as Vision Transformers (ViTs), requires substantial computational resources, often leveraging distributed systems with hundreds or even thousands of GPUs. Techniques like masked autoencoding and contrastive learning involve complex optimization over high-dimensional feature spaces and necessitate large batch sizes for stable training. For example, SimCLR relies heavily on large batch sizes to ensure meaningful negative sample diversity, which imposes significant memory and hardware constraints [4]. Similarly, scaling pretraining frameworks like SEER to 1 billion uncurated images has demonstrated superior performance but at the cost of unsustainable energy consumption and hardware requirements, making such approaches inaccessible to smaller research labs or enterprises [23].

Additionally, the computational efficiency of many SSL methods is hindered by the costly augmentations and pairwise comparisons necessary for contrastive frameworks. Hard negative mining and augmentation-based sample generation in contrastive learning pipelines exacerbate the problem, leading to inefficient exploration of the data distribution [6]. Although non-contrastive methods like BYOL and SimSiam alleviate the dependency on negative pairs, their reliance on synchronized training across momentum encoders still demands considerable computational resources [22]. Moreover, recent studies have underscored training instability in SSL models, particularly for newer architectures like ViTs. The high dimensionality of transformer models introduces additional challenges in achieving convergence without significant tuning and computational overhead [13].

Data quality and scale further exacerbate computational inefficiencies. Training SSL models on uncurated, large-scale datasets introduces noise, redundancy, and class imbalance—a problem noted in endeavors to scale SSL to naturalistic, unfiltered datasets [6], [23]. Addressing these issues requires sophisticated strategies for filtering and curating datasets or designing loss functions robust to noisy and imbalanced data distributions. For example, hierarchical adaptive clustering and multi-modal learning paradigms offer promising pathways to reconcile efficiency with scalability [7], [104].

To mitigate resource requirements, several innovative approaches have emerged. Masked autoencoders (MAEs) have gained traction as highly efficient SSL strategies that reconstruct masked regions, enabling architectures like ViTs to focus only on informative patches, thereby reducing the computational burden compared to contrastive counterparts [10]. Similarly, model compression techniques, such as distillation-based methods, have enabled the transfer of knowledge from large pretrained SSL models to smaller and more practical student models without significant downstream performance degradation [134]. In addition, multi-crop training, where fewer unique samples are used per batch but with higher augmentation diversity, reduces overall memory requirements without hindering convergence [12].

Future research must prioritize the development of lightweight algorithms capable of achieving state-of-the-art performance with constrained resources. Strategies such as dynamic loss adjustment, gradient sparsification, and adaptive masking have potential to optimize training pipelines and computational efficiency [4]. Furthermore, addressing energy consumption concerns is imperative for sustainable SSL deployment, with techniques like energy-efficient network designs and on-device learning frameworks representing promising opportunities. By aligning scalability innovations with environmental sustainability and equity in access to computational resources, SSL can move closer to democratizing its transformative potential.

7.2 Dataset Biases and Robustness Limitations

Dataset biases and robustness limitations present persistent obstacles for self-supervised learning (SSL) in computer vision, restricting the ability of models to generalize across diverse and unseen scenarios. While SSL has unlocked the potential of utilizing large-scale unlabeled datasets, such datasets are inherently shaped by various biases, including imbalanced representations, sampling artifacts, and over-representation of dominant data distributions. These biases not only constrain the diversity of learned features but also compromise model performance in out-of-distribution (OOD) settings, where the data diverge significantly from the pretraining distribution.

A prominent source of dataset bias stems from the reliance on curated benchmark datasets, such as ImageNet, which often fail to capture the full spectrum of real-world variability. This curation process injects systemic biases, as the datasets lack adequate diversity along demographic, geographical, and contextual axes [31]. Consequently, SSL models pretrained on such datasets may overfit to dominant classes or feature distributions, leaving them vulnerable to underperformance in tasks requiring nuanced recognition or minority-class identification. Imbalanced data distributions further exacerbate the issue, leading SSL representations to prioritize dominant features at the expense of underrepresented classes or inter-class variation, thereby limiting fairness and adaptability [31].

Augmentation pipelines, which are integral to SSL frameworks, play a dual role in mitigating and perpetuating dataset biases. On one hand, augmentations such as cropping, flipping, and color jittering are essential for inducing invariances and enriching feature representations; on the other hand, they may inadvertently amplify dataset-specific artifacts or encourage spurious correlations [61]. Insufficient diversity in augmentations can restrict a model's ability to generalize, while over-aggressive augmentations risk corrupting task-relevant information. Designing augmentation

strategies that introduce sufficient variability while preserving semantic structure is crucial for simulating diverse real-world conditions [61].

A related challenge arises from the robustness of SSL-trained models to domain and distribution shifts. Empirical findings suggest that models pretrained on curated datasets like ImageNet struggle to adapt to datasets with distinct statistical properties, such as remote sensing imagery, medical scans, or geographically diverse data [102]. This gap is rooted in SSL objectives optimized heavily on source dataset distributions, often failing to yield representations transferable to unfamiliar domains. Preliminary efforts, such as domain-agnostic pretext tasks or normalization strategies, aim to address this limitation, though these methods are often constrained by computational demands and domain-specific design [32].

Compounding these issues, SSL models are prone to memorization, wherein dataset-specific noise or artifacts are inadvertently preserved in learned representations, adversely affecting generalization to downstream tasks. While memorization may enhance within-domain results, it is detrimental to robustness in OOD settings [135]. Techniques such as redundancy reduction and feature decorrelation have emerged as effective regularization strategies, explicitly enforcing diversity and mitigating memorization pitfalls [33], [108].

To combat dataset biases and improve robustness, future work must prioritize automated data curation approaches that leverage clustering and similar techniques to create datasets with enhanced balance, diversity, and domain coverage—reducing the manual biases inherent in traditional dataset collection [136]. Additionally, causally grounded augmentation strategies could help disentangle spurious correlations introduced by conventional methods, ensuring that learned representations remain semantically valid and transferable [61].

The evaluation of SSL models must also evolve to move beyond traditional benchmarks, incorporating metrics like fairness, OOD robustness, and cross-domain generalization as core performance criteria. Establishing unified evaluation frameworks alongside robust, well-balanced pretraining datasets will be pivotal in positioning SSL as a universally adaptable learning paradigm. Achieving this will not only improve SSL's effectiveness in diverse real-world applications but also promote equitable and unbiased deployment of machine learning systems across various domains and populations.

7.3 Interpretability and Explainability of SSL Models

The interpretability and explainability of self-supervised learning (SSL) models remain significant challenges, especially given their increasing adoption in critical applications like healthcare, autonomous driving, and security. While SSL has shown remarkable progress in learning transferable representations from unlabeled data, understanding the internal mechanisms and what features these models prioritize remains largely opaque. This lack of clarity raises concerns about trust, reliability, and usability, particularly in high-stakes domains.

One of the primary barriers to interpretability in SSL arises from the nature of pretext tasks. Unlike supervised models with direct mappings between inputs and labeled outputs, SSL representations are optimized to solve auxiliary tasks, such as contrastive discrimination or masked reconstruction, which are only loosely connected to downstream objectives. For example, contrastive methods like SimCLR and MoCo prioritize invariances to specific transformations through contrastive loss, but the representations learned may obscure semantically meaningful distinctions that could be essential for downstream tasks [11], [63]. The complexity of these learned invariances makes it difficult to decipher whether the representations align with human-understandable concepts or if they capture spurious correlations inherent in pretraining datasets [62].

Existing interpretability methods like saliency maps, which highlight critical input regions for predictions, are not directly applicable to SSL given the absence of labeled outputs during training. Some efforts attempt to dissect the representations post hoc. Techniques such as feature visualization or clustering representations into interpretable units have been employed to identify groupings aligned to semantic classes or tasks. For instance, PIRL [27] uses pretext-invariance properties to assess the semantic consistency of learned features, achieving improved transferability alongside limited interpretability. However, these techniques often fall short of producing human-aligned explanations, as SSL-generated features are densely distributed and lack clear semantic disentanglement.

A promising avenue for improving interpretability in SSL is the use of concept-based methods, where representations are explicitly disentangled into components representing semantically meaningful features, such as shapes, textures, or object parts. Multi-task systems like MuST [137] partially achieve this by training SSL models on complementary pretext tasks involving human-recognizable constructs. Similarly, emerging frameworks like Geminated Gestalt Autoencoders [138] leverage both pixel-space and frequency-space reconstructions to generate features with distinct semantic meanings, facilitating interpretability.

Another critical aspect is uncertainty quantification, which is currently underexplored in SSL. Incorporating mechanisms to assess and explain uncertainty in prediction pipelines could improve trust in SSL models, particularly in critical applications like diagnostics. For example, probabilistic extensions to SSL algorithms that model uncertainty alongside feature learning could provide nuanced explanations, alerting users to cases where the learned representations or downstream classifications may be unreliable [68].

Additionally, SSL frameworks leveraging attention mechanisms, such as Vision Transformers (ViTs) and their self-supervised variants like SiT [30], could inherently aid interpretability by visualizing attention weights across patches. By aligning these weights with human-understood concepts, attention-based SSL models offer a natural mechanism for explaining predictions. However, while initial results show promise, the interpretability of transformer-based SSL models remains an open area with much work needed to better associate attention mechanisms with semantic concepts.

Practical challenges for interpretability include dataset biases that propagate to SSL representations, potentially leading to poor generalization or discriminatory behavior in

downstream tasks [139]. For instance, SSL representations generated from biased datasets may learn features irrelevant or even harmful for their intended application. Techniques like adversarial training to remove shortcuts [40] or masking strategies targeted at dataset-specific augmentation biases [50] represent early steps toward equipping SSL models with fairer and more interpretable features.

Future directions in interpretability research for SSL must focus on integrating interpretability objectives directly into SSL pretraining. This may involve enforcing constraints, such as disentanglement and semantic alignment, during the pretraining process itself rather than retrofitting interpretability post hoc. Additionally, developing unified evaluation protocols to measure interpretability alongside representation quality in various downstream tasks will be critical to incentivizing progress. Innovations in explainable SSL could also be driven by integrating explainability techniques from multimodal learning paradigms [42], enabling richer explanations grounded in both visual and textual contexts.

In conclusion, while SSL has demonstrated powerful performance and versatility, the interpretability of its learned representations remains a frontier rife with challenges and opportunities. Addressing these issues will require interdisciplinary innovation at the intersection of machine learning, cognitive science, and human-computer interaction, ensuring that SSL systems can be trusted and meaningfully integrated into real-world, user-driven applications.

7.4 Ethical and Societal Considerations in SSL

The rise of self-supervised learning (SSL) has revolutionized the extraction of representations from large-scale unlabeled datasets, yet its application brings significant ethical and societal challenges, spanning privacy, bias amplification, and misuse in sensitive contexts. These issues arise from SSL's reliance on uncurated datasets and the opacity of its learned representations, necessitating dedicated efforts to identify risks and implement effective mitigation strategies to ensure responsible and equitable deployment.

One prominent ethical concern is the inadvertent incorporation of private or sensitive information embedded in datasets. SSL methods, by design, exploit intrinsic patterns in data without manual oversight, risking privacy violations when datasets contain personally identifiable information (PII) or sensitive content. For areas like healthcare, SSL-based models trained on medical imaging risk encoding private patient details if datasets are insufficiently anonymized [45]. Furthermore, multimodal SSL systems enhance this risk by fusing data modalities, such as vision and language, potentially uncovering and correlating sensitive attributes in unintended ways [116]. Standard anonymization protocols might fall short in mitigating these risks, emphasizing the urgent need for privacy-preserving techniques like federated learning and differential privacy embedded directly into SSL frameworks.

The challenge of bias amplification represents another substantial obstacle in SSL applications. By learning statistical patterns from uncurated, inherently biased datasets, SSL models risk perpetuating and even reinforcing societal inequities. Biases in demographic representation, for instance, lead to performance disparities in downstream applications such as facial recognition and object detection, which can have discriminatory implications in high-stakes domains like law enforcement [6]. A multifaceted approach is required to mitigate these effects, including bias-aware algorithms that adaptively counter dataset imbalances and post-hoc fairness adjustments to SSL-derived features [46]. Such solutions must be coupled with principled dataset curation practices to reduce systemic inequities embedded in the pretraining stages.

Compounding these concerns is the opaque nature of SSL-learned representations, which hinders interpretability and accountability in critical applications. Unlike supervised learning frameworks, where outputs are aligned with labeled targets, SSL relies on pretext tasks that often lack direct semantic coherence with downstream objectives. This opacity can conceal harmful patterns or shortcuts learned during pretraining, leading to downstream misalignments that escape detection. For instance, SSL models may exploit trivial correlations in data, bypassing semantically meaningful concepts crucial for ethical decision-making, especially in domains like autonomous driving [47]. To address this limitation, research into disentangling SSL representations and aligning features with human-understandable semantics is essential [140]. Advancements in uncertainty quantification are also necessary to flag unreliable or ethically questionable predictions in safety-critical domains.

Another pressing concern is the misuse of SSL models, especially in surveillance and security applications, where these frameworks could enable large-scale privacy violations or exacerbate societal inequalities. For instance, SSL-powered unsupervised object localization can facilitate mass surveillance systems, while multimodal SSL approaches incorporating geospatial data can heighten tracking and profiling capabilities [117], [124]. Beyond surveillance, SSL may also perpetuate misinformation, with models potentially generating misleading synthetic data or amplifying harmful content. These risks underline the urgency of establishing ethical guidelines and stakeholder oversight to govern SSL's deployment in sensitive domains.

Addressing these challenges requires extending beyond technical innovations to include robust societal interventions. Ethical deployment demands fairness assessments, transparent consent protocols for dataset use, and active engagement with stakeholders to incorporate societal values into SSL system design [141]. Benchmarks that evaluate societal impacts—spanning fairness, privacy adherence, and bias amplification—must complement performance-oriented metrics to foster responsible technological progress [9]. Interdisciplinary collaborations among researchers, ethicists, and policymakers are indispensable to ensuring SSL technologies advance responsibly, respecting both societal and ethical imperatives.

In conclusion, while the transformative potential of SSL in representation learning is undeniable, its deployment must navigate significant ethical complexities. By proactively addressing challenges tied to privacy, bias, and misuse, alongside fostering interpretability and fairness, the SSL community can ensure its innovations lend themselves to equitable and socially responsible applications.

7.5 Integration with Multimodal and Weakly Supervised Architectures

The integration of self-supervised learning (SSL) with multimodal and weakly supervised architectures presents transformative opportunities to enhance SSL's capability in tackling resource-scarce and complex multimodal domains. By combining complementary strengths across modalities and leveraging minimal supervision, these integrations address the limitations of single-modality SSL while unlocking enriched, transferable representation learning for underexplored applications.

Multimodal SSL leverages the mutual enrichment between modalities such as vision, text, audio, and spatiotemporal data to generate robust representations. Multimodal contrastive methods like CLIP [71] demonstrate the feasibility of aligning heterogeneous data modalities (e.g., vision and language) using shared self-supervision. By pairing image-text embeddings through alignment objectives, CLIP has enabled breakthroughs in few-shot and zero-shot learning tasks. Building on this, variations such as DINO-MC [64] extend alignment strategies to integrate high-resolution spatial features tailored for remote sensing data. These multimodal approaches enhance generalizability by exploiting semantic consistency across modalities. However, challenges persist, notably the reliance on large-scale paired data, as seen in CLIP's dependence on curated vision-language datasets, and the lack of effective integration techniques for modalities with asynchronous resolutions (e.g., video-audio alignment).

Weakly supervised architectures offer complementary avenues by incorporating small quantities of labeled data alongside unlabeled sets to enrich SSL signal quality and accelerate convergence, particularly in data-scarce applications like medical imaging and remote sensing. For instance, strategies like learning cross-modal pseudo-labels have demonstrated promise in domains requiring alignment without full annotations [73]. Moreover, methods such as SEAM [92] leverage self-supervised attention mechanisms to bridge the gap between weak supervision and pixel-level tasks like segmentation. Here, consistency in equivariance-driven representations allows models to leverage coarse labels while progressively refining granular predictions. While effective, these approaches often require careful tuning of weak supervision signals to avoid amplifying dataset biases.

Emerging architectures integrate SSL innovations with multimodal and weakly supervised learning objectives through novel mechanisms like pretext task decomposition and domain-invariant representations. For example, VICRegL [122] balances local and global representation learning across spatial contexts in dense downstream tasks, making it compatible with both image-level multimodal applications and fine-grained tasks dependent on minimal supervision. Similarly, the fusion of structured masking techniques, such as DiMAE [50], not only reinforces domain generalization but also facilitates cross-domain reconstruction in multimodal scenarios, addressing shifts between training and unseen domains. These advancements underscore the potential of unified frameworks capable of collapsing the divide between SSL's paradigms.

Nonetheless, integrating multimodal and weakly supervised methods introduces computational and structural trade-offs. For multimodal SSL, the primary challenges stem from designing scalable training strategies that align modalities with differing temporal and spatial resolutions. Techniques like multi-modal-specific masking [50] show promise but increase computational complexity. In weakly supervised settings, scalability remains contingent on curating representative datasets that ensure semantically meaningful alignments while minimizing bias amplification [92]. Furthermore, both paradigms require novel evaluation strategies to benchmark performance comprehensively on multimodal and weakly supervised benchmarks, as tasks like multimodal video-text alignment or object-centric segmentation often lack diverse, standardized testbeds.

Looking forward, integrating SSL with emerging paradigms like generative pretraining and cross-modal knowledge distillation [65] provides a new frontier for multimodal learning. Techniques such as controllable view generation for synthetic data amplifications hold the potential to address gaps in divergent modalities or missing supervision. Moreover, harmonizing multimodal SSL with dynamic task adaptation frameworks [67] could significantly boost low-resource training by learning modality-specific invariances flexibly. Exploring task-agnostic pretraining pipelines that integrate weak supervision seamlessly into multimodal SSL pipelines offers exciting possibilities for advancing theories of unified representation learning.

In summary, the blend of SSL with multimodal and weakly supervised architectures shows immense promise for resource-scarce and complex domains. While challenges in scale and alignment persist, innovations in alignment frameworks, adaptive masking, and domain-agnostic representation learning provide a robust foundation for future exploration in this space. These paradigms, when effectively integrated, promise to redefine state-of-the-art capabilities across disciplines like healthcare, autonomous navigation, and multimodal understanding.

7.6 Advancing SSL Through Novel Pretext Tasks and Evaluation Protocols

The design of pretext tasks plays an essential role in the progress of self-supervised learning (SSL), as it directly impacts the quality, robustness, and transferability of the learned representations. Beyond traditional paradigms such as jigsaw puzzles and rotation prediction, recent innovations in pretext task design have prioritized deeper alignment with high-level semantics and downstream relevance. Simultaneously, evaluation protocols for SSL models have come under increasing scrutiny, as they often fail to rigorously assess robustness, adaptability, and fairness across diverse tasks and domains. This subsection delves into the interconnected advancements in pretext task design and evaluation frameworks, examining their potential to overcome existing limitations and propel SSL forward.

The shift from heuristic, low-level tasks to semantically meaningful pretext objectives has fundamentally enhanced SSL's capacity for robust representation learning. For instance, Pretext-Invariant Representation Learning (PIRL) [27] demonstrated the efficacy of combining jigsaw puzzles

with invariance to input transformations in capturing richer semantic embeddings. Masked image modeling (MIM) has emerged as another transformative class of pretext objectives. Frameworks like Masked Autoencoders (MAEs) [37] focus on reconstructing masked regions, promoting holistic understanding of global structures. When coupled with transformer architectures, extensions such as SiameseIM [97] leverage spatial sensitivity to augment semantic alignment. Nonetheless, these methods often contend with low-level biases: by prioritizing pixel reconstruction, they may inadvertently hinder the generality of learned representations.

To address these shortcomings, hybrid approaches to pretext task design have gained prominence. Core-tuning [78] combines contrastive learning with hard example mining, enabling finer disambiguation of positive pairs and enhancing representation specificity. Similarly, adversarial augmentation strategies introduced in HEXA [101] employ challenging conditions to mitigate shortcut learning and improve adaptability to downstream tasks. Unified frameworks, such as DiRA [142], integrate generative, discriminative, and adversarial pretext tasks, fostering domain scalability and achieving fine-grained semantic representation learning in fields like medical imaging. These approaches demonstrate the versatility of hybrid designs in navigating complex data structures, extending SSL applicability to domain-specific challenges.

Despite advancements in task design, the evaluation of SSL methods lags behind, raising concerns over the reliability and universality of existing protocols. Traditional approaches like linear probing and fine-tuning often fail to comprehensively evaluate models' adaptability, particularly in capturing spatial generalization or cross-modal alignment. For instance, while frameworks such as VICReg improve robustness by reducing redundancy during pretraining [60], they rely heavily on narrow benchmarks that may not generalize to broader contexts. Furthermore, as SSL encompasses more complex datasets and multimodal scenarios, the absence of standardized evaluation protocols across diverse modalities, such as vision and language [143], exacerbates challenges in fair and meaningful assessment.

Emerging trends in SSL emphasize the necessity for unified benchmarks addressing practical constraints like low-resource settings [98] and efficient cross-domain adaptability [50]. For example, Latent Masked Image Modeling [57], which focuses on reconstructing latent feature spaces rather than pixel values, offers a promising solution toward computationally efficient and semantically robust pretext tasks and evaluation strategies. Such developments underscore the growing demand for scalable and generalizable protocols for evaluating SSL models across domains and modalities.

Future research in SSL must focus on refining pretext task designs to integrate hierarchical objectives that balance low-level feature constraints with high-level semantic abstractions. Concurrently, evaluation frameworks need to evolve to measure SSL models' performance on fairness, robustness to noise, and domain generalization. Approaches such as semi-supervised refinement [144] could enable balanced SSL performance while adhering to increasingly complex benchmarks. By fostering innovation in pretext task design and evaluation, SSL can achieve more cohesive and resilient representations, paving the way for advancements across diverse and resource-constrained real-world scenarios.

8 Conclusion

Self-supervised learning (SSL) in computer vision has emerged as a transformative paradigm, leveraging the intrinsic structure of unlabeled data to train models without manual annotation. This shift away from labor-intensive labeling processes has substantially reduced barriers to deploying deep learning systems at scale and across diverse domains, including those with limited annotated datasets. The foundational promise of SSL lies in its ability to produce generalized and robust representations that rival or even surpass those obtained via supervised learning paradigms in specific downstream tasks [16], [106]. Over the past decade, the field has witnessed rapid advancements in methods, architectures, and applications, reshaping the foundations of visual representation learning.

Central to SSL's progress is the diversity of methodologies tailored to distinct objectives, ranging from contrastive learning approaches, such as SimCLR and MoCo, to generative frameworks like masked autoencoders (MAEs). Contrastive learning, particularly, introduced robust mechanisms to maximize agreement between augmented views of the same data while contrasting them against negative samples. These methods have demonstrated remarkable scalability and have defined benchmarks in representation transferability and robustness [4], [6]. Non-contrastive and redundancy-reduction approaches, such as BYOL and VICReg, further advanced the field by eliminating the reliance on explicit negative pairs, achieving stability and improved performance through innovative loss functions [8], [22]. Generative methods, epitomized by masked prediction models, have renewed interest in pretext tasks that refine both global and local feature representations, mirroring the success of masked language models in natural language processing [10], [30].

The strength of SSL lies in its versatility, enabling applications across diverse vision tasks. Image-level tasks, such as classification and segmentation, have benefited from pretrained SSL features, which exhibit superior transferability and reduced annotation dependencies compared to traditional supervised methods [106]. Domain-specific applications, including medical imaging and remote sensing, have similarly highlighted SSL's ability to adapt to specialized data distributions while maintaining robust performance [81], [102]. Furthermore, SSL's integration with multi-modal learning frameworks has opened new frontiers in vision-language tasks, such as visual question answering (VQA) and image captioning, where cross-modal consistency plays a pivotal role [7], [26].

Despite these successes, SSL faces several persistent challenges. Computational scalability remains a critical limitation, as state-of-the-art methods often require substantial resources and hardware infrastructure for training on large-scale datasets such as ImageNet or uncurated datasets [4], [23]. Robustness to out-of-distribution data and sensitivity

to dataset biases also remain unresolved issues, particularly in scenarios where augmentations or dataset curation introduce unwanted artifacts [6], [18]. Additionally, the interpretability of representations learned through SSL continues to lag behind supervised approaches, creating obstacles for deploying these models in critical applications such as healthcare and autonomous driving [17], [45].

Looking ahead, multiple avenues offer promising directions for the evolution of SSL. The dynamic adaptation of SSL strategies during training, such as task switching or augmentation refinement, could mitigate dataset biases and improve generalization [19], [103]. The integration of generative and contrastive principles into hybrid frameworks may further optimize data utilization while improving robustness across downstream tasks [10], [56]. Moreover, SSL's expansion into multi-modal setups and time-series domains will likely yield impactful developments in areas ranging from video analysis to real-time sensor fusion [34], [79].

In conclusion, self-supervised learning has redefined how we approach representation learning in computer vision by minimizing reliance on labeled data while achieving unprecedented performance levels across diverse tasks and settings. While significant challenges remain, particularly concerning computational efficiency, robustness, and interpretability, the field's foundational breakthroughs and emerging trends point toward a future where SSL models dominate both academic research and practical applications. By addressing these challenges and capitalizing on its current momentum, SSL has the potential to unlock more scalable, equitable, and innovative solutions for visual understanding.

REFERENCES

[1] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," *ArXiv*, vol. abs/1906.12340, 2019. 1, 17, 19

[2] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," *arXiv: Computer Vision and Pattern Recognition*, 2017. 1, 7

[3] S. Gidaris, A. Bursuc, N. Komodakis, P. Pérez, and M. Cord, "Boosting few-shot visual learning with self-supervision," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8058–8067, 2019. 1, 2

[4] P. Goyal, D. Mahajan, A. Gupta, and I. Misra, "Scaling and benchmarking self-supervised visual representation learning," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6390–6399, 2019. 1, 2, 5, 7, 9, 12, 16, 17, 19, 20, 23, 26, 30

[5] J. Sauder and B. Sievers, "Self-supervised deep learning on point clouds by reconstructing space," in *Neural Information Processing Systems*, 2019, pp. 12 942–12 952. 1, 21

[6] S. Purushwalkam and A. Gupta, "Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases," *ArXiv*, vol. abs/2007.13916, 2020. 1, 4, 7, 9, 12, 17, 19, 21, 26, 28, 30, 31

[7] Y.-H. H. Tsai, Y. Wu, R. Salakhutdinov, and L.-P. Morency, "Self-supervised learning from a multi-view perspective," *arXiv: Learning*, 2020. 1, 2, 7, 13, 17, 26, 30

[8] Q. Garrido, Y. Chen, A. Bardes, L. Najman, and Y. LeCun, "On the duality between contrastive and non-contrastive self-supervised learning," *ArXiv*, vol. abs/2206.02574, 2022. 1, 7, 8, 12, 30

[9] N. Tomašev, I. Bica, B. McWilliams, L. Buesing, R. Pascanu, C. Blundell, and J. Mitrovic, "Pushing the limits of self-supervised resnets: Can we outperform supervised learning without labels on imagenet?" *ArXiv*, vol. abs/2201.05119, 2022. 1, 19, 20, 28

[10] C. Zhang, C. Zhang, J. Song, J. S. K. Yi, K. Zhang, and I.-S. Kweon, "A survey on masked autoencoder for self-supervised learning in vision and beyond," *ArXiv*, vol. abs/2208.00173, 2022. 1, 2, 3, 12, 13, 16, 20, 21, 22, 26, 30, 31

[11] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1920–1929, 2019. 1, 2, 3, 7, 8, 9, 12, 16, 17, 18, 20, 21, 22, 27

[12] M. Caron, H. Touvron, I. Misra, H. J'egou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. 1, 2, 12, 18, 20, 26

[13] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9620–9629, 2021. 1, 26

[14] S. Yun, H. Lee, J. Kim, and J. Shin, "Patch-level representation learning for self-supervised vision transformers," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8344–8353, 2022. 1, 4, 5, 6, 12, 14, 15, 16, 18, 21

[15] A. Afkanpour, V. R. Khazaie, S. Ayromlou, and F. Forghani, "Can generative models improve self-supervised representation learning?" *ArXiv*, vol. abs/2403.05966, 2024. 1, 13

[16] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1476–1485, 2019. 1, 30

[17] E. Fini, V. Costa, X. Alameda-Pineda, E. Ricci, A. Karteek, and J. Mairal, "Self-supervised models are continual learners," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9611–9620, 2021. 1, 12, 31

[18] J. Liang, R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Generalized semi-supervised learning via self-supervised feature adaptation," *ArXiv*, vol. abs/2405.20596, 2024. 1, 17, 31

[19] J. Ye, Q. Xiao, J. Wang, H. Zhang, J. Deng, and Y. Lin, "Cosleep: A multi-view representation learning framework for self-supervised learning of sleep stage classification," *IEEE Signal Processing Letters*, vol. 29, pp. 189–193, 2022. 1, 31

[20] K. Zhang, Q. Wen, C. Zhang, R. Cai, M. Jin, Y. Liu, J. Zhang, Y. Liang, G. Pang, D. Song, and S. Pan, "Self-supervised learning for time series analysis: Taxonomy, progress, and prospects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 6775–6794, 2023. 2

[21] J. Yu, H. Yin, X. Xia, T. Chen, J. Li, and Z.-L. Huang, "Self-supervised learning for recommender systems: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, pp. 335–355, 2022. 2

[22] Y. Tian, X. Chen, and S. Ganguli, "Understanding self-supervised learning dynamics without contrastive pairs," in *International Conference on Machine Learning*, 2021, pp. 10 268–10 278. 2, 7, 8, 13, 16, 20, 26, 30

[23] P. Goyal, M. Caron, B. Lefaudeux, M. Xu, P. Wang, V. Pai, M. Singh, V. Liptchinsky, I. Misra, A. Joulin, and P. Bojanowski, "Self-supervised pretraining of visual features in the wild," *ArXiv*, vol. abs/2103.01988, 2021. 2, 16, 17, 21, 26, 30

[24] M. C. Schiappa, Y. Rawat, and M. Shah, "Self-supervised learning for videos: A survey," *ACM Computing Surveys*, vol. 55, pp. 1 – 37, 2022. 2, 4

[25] X. Wang, K. He, and A. Gupta, "Transitive invariance for self-supervised visual representation learning," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1338–1347, 2017. 2, 3

[26] S. Deldari, H. Xue, A. Saeed, J. He, D. V. Smith, and F. D. Salim, "Beyond just vision: A review on self-supervised representation learning on multimodal and temporal data," *ArXiv*, vol. abs/2206.02353, 2022. 2, 8, 12, 18, 19, 30

[27] I. Misra and L. Maaten, "Self-supervised learning of pretext-invariant representations," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6706–6716, 2019. 3, 4, 5, 8, 9, 15, 22, 27, 29

[28] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Neural Information Processing Systems*, 2019, pp. 15 509–15 519. 3, 7, 8, 9, 10, 14, 18, 25

[29] C. Tosh, A. Krishnamurthy, and D. J. Hsu, "Contrastive learning, multi-view redundancy, and linear models," *ArXiv*, vol. abs/2008.10150, 2020. 3, 7

[30] S. A. A. Ahmed, M. Awais, and J. Kittler, "Sit: Self-supervised vision transformer," *ArXiv*, vol. abs/2104.03602, 2021. 3, 4, 12, 14, 27, 30

[31] H. Liu, J. Z. HaoChen, A. Gaidon, and T. Ma, "Self-supervised learning is more robust to dataset imbalance," *ArXiv*, vol. abs/2110.05025, 2021. 3, 17, 18, 22, 26

[32] S. Wang, L. Yu, C. Li, C.-W. Fu, and P. Heng, "Learning from extrinsic and intrinsic supervisions for domain generalization," *ArXiv*, vol. abs/2007.09316, 2020. 3, 18, 21, 27

[33] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," *ArXiv*, vol. abs/2105.04906, 2021. 3, 8, 13, 15, 17, 18, 27

[34] T. Han, W. Xie, and A. Zisserman, "Self-supervised co-training for video representation learning," *ArXiv*, vol. abs/2010.09709, 2020. 3, 14, 17, 21, 31

[35] X. Liu, Z. Wang, Y. Li, and S. Wang, "Self-supervised learning via maximum entropy coding," *ArXiv*, vol. abs/2210.11464, 2022. 3, 20

[36] S. Gidaris, A. Bursuc, N. Komodakis, P. P'erez, and M. Cord, "Learning representations by predicting bags of visual words," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6926–6936, 2020. 3

[37] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context autoencoder for self-supervised representation learning," *ArXiv*, vol. abs/2202.03026, 2022. 3, 5, 13, 22, 23, 30

[38] D. Pathak, R. B. Girshick, P. Dollár, T. Darrell, and B. Hariharan, "Learning features by watching objects move," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6024–6033, 2016. 4

[39] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2070–2079, 2017. 4, 10, 12, 21

[40] M. Minderer, O. Bachem, N. Houlsby, and M. Tschannen, "Automatic shortcut removal for self-supervised representation learning," in *International Conference on Machine Learning*, 2020, pp. 6927–6937. 4, 6, 15, 16, 22, 28

[41] A. Bhunia, P. N. Chowdhury, Y. Yang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Vectorization and rasterization: Self-supervised learning for sketch and handwriting," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5668–5677, 2021. 4, 20

[42] N. Mu, A. Kirillov, D. A. Wagner, and S. Xie, "Slip: Self-supervision meets language-image pre-training," *ArXiv*, vol. abs/2112.12750, 2021. 4, 14, 19, 22, 28

[43] C. Zhang, K. Zhang, C. Zhang, A. Niu, J. Feng, C. D. Yoo, and I.-S. Kweon, "Decoupled adversarial contrastive learning for self-supervised adversarial robustness," in *European Conference on Computer Vision*, 2022, pp. 725–742. 4, 16

[44] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," *ArXiv*, vol. abs/1603.09246, 2016. 4

[45] N. A. Koohbanani, B. Unnikrishnan, S. Khurram, P. Krishnaswamy, and N. Rajpoot, "Self-path: Self-supervision for classification of pathology images with limited annotations," *IEEE Transactions on Medical Imaging*, vol. 40, pp. 2845–2856, 2020. 4, 28, 31

[46] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 161–10 170, 2020. 5, 9, 19, 23, 28

[47] J. Wang, Y. Gao, K. Li, Y. Lin, A. J. Ma, and X. Sun, "Removing the background by adding the background: Towards background robust self-supervised video representation learning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 799–11 808, 2020. 5, 28

[48] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2561–2571, 2019. 5, 14

[49] X. Kong and X. Zhang, "Understanding masked image modeling via learning occlusion invariant feature," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6241–6251, 2022. 5, 20

[50] H. Yang, M. Chen, Y. Wang, S. Tang, F. Zhu, L. Bai, R. Zhao, and W. Ouyang, "Domain invariant masked autoencoders for self-supervised learning from multi-domains," *ArXiv*, vol. abs/2205.04771, 2022. 5, 6, 10, 11, 13, 14, 15, 19, 24, 28, 29, 30

[51] S. Jenni, H. Jin, and P. Favaro, "Steering self-supervised feature learning beyond local pixel statistics," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6407–6416, 2020. 6

[52] L. Wang, F. Liang, Y. Li, H. Zhang, W. Ouyang, and J. Shao, "Repre: Improving self-supervised vision transformer with reconstructive pre-training," *ArXiv*, vol. abs/2201.06857, 2022. 6

[53] T. Wang, Z. Yue, J. Huang, Q. Sun, and H. Zhang, "Self-supervised learning disentangled group representation as feature," *ArXiv*, vol. abs/2110.15255, 2021. 6, 16

[54] L. Wu, H. Lin, Z. Gao, C. Tan, and Stan.Z.Li, "Self-supervised learning on graphs: Contrastive, generative, or predictive," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 4216–4235, 2021. 6

[55] F. Bordes, R. Balestriero, and P. Vincent, "Towards democratizing joint-embedding self-supervised learning," *ArXiv*, vol. abs/2303.01986, 2023. 6, 11, 16

[56] Y. Dubois, T. Hashimoto, S. Ermon, and P. Liang, "Improving self-supervised learning by characterizing idealized representations," *ArXiv*, vol. abs/2209.06235, 2022. 6, 18, 21, 31

[57] Y. Wei, A. Gupta, and P. Morgado, "Towards latent masked image modeling for self-supervised visual representation learning," in *European Conference on Computer Vision*, 2024, pp. 1–17. 6, 30

[58] J. Z. HaoChen and T. Ma, "A theoretical study of inductive biases in contrastive learning," *ArXiv*, vol. abs/2211.14699, 2022. 7

[59] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1182–1191, 2021. 7, 12, 17

[60] K. Gupta, T. Ajanthan, A. Hengel, and S. Gould, "Understanding and improving the role of projection head in self-supervised learning," *ArXiv*, vol. abs/2212.11491, 2022. 7, 11, 30

[61] J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Scholkopf, M. Besserve, and F. Locatello, "Self-supervised learning with data augmentations provably isolates content from style," in *Neural Information Processing Systems*, 2021, pp. 16 451–16 467. 8, 13, 20, 26, 27

[62] I. Bendidi, A. Bardes, E. Cohen, A. Lamiable, G. Bollot, and A. Genovesio, "No free lunch in self supervised representation learning," *ArXiv*, vol. abs/2304.11718, 2023. 8, 19, 27

[63] T. Xiao, X. Wang, A. A. Efros, and T. Darrell, "What should not be contrastive in contrastive learning," *ArXiv*, vol. abs/2008.05659, 2020. 8, 9, 10, 18, 27

[64] X. Wanyan, S. Seneviratne, S. Shen, and M. Kirley, "Extending global-local view alignment for self-supervised learning with remote sensing imagery," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2443–2453, 2023. 8, 9, 10, 15, 20, 22, 24, 29

[65] X. Li, Y. Yang, X. Li, J. Wu, Y. Yu, B. Ghanem, and M. Zhang, "Genview: Enhancing view quality with pretrained generative model for self-supervised learning," *ArXiv*, vol. abs/2403.12003, 2024. 8, 10, 24, 29

[66] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," *ArXiv*, vol. abs/2103.03230, 2021. 9

[67] R. Chavhan, H. Gouk, J. Stuehmer, C. Heggan, M. Yaghoobi, and T. M. Hospedales, "Amortised invariance learning for contrastive self-supervision," *ArXiv*, vol. abs/2302.12712, 2023. 9, 10, 18, 20, 29

[68] J. Lee, Q. Lei, N. Saunshi, and J. Zhuo, "Predicting what you already know helps: Provable self-supervised learning," *ArXiv*, vol. abs/2008.01064, 2020. 9, 27

[69] J. Wang, J. Jiao, and Y. Liu, "Self-supervised video representation learning by pace prediction," *ArXiv*, vol. abs/2008.05861, 2020. 9

[70] J. Xie, X. Zhan, Z. Liu, Y. Ong, and C. C. Loy, "Unsupervised object-level representation learning from scene images," in *Neural Information Processing Systems*, 2021, pp. 28 864–28 876. 10, 19

[71] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," *ArXiv*, vol. abs/2002.05709, 2020. 10, 29

[72] L. Jing, X. Yang, J. Liu, and Y. Tian, "Self-supervised spatiotemporal feature learning via video rotation prediction." *arXiv: Computer Vision and Pattern Recognition*, 2018. 10, 15

[73] H. Li, C. Xu, W. Yang, H. Yu, and G.-S. Xia, "Learning cross-view visual geo-localization without ground truth," *IEEE Transactions*

on Geoscience and Remote Sensing, vol. 62, pp. 1–17, 2024. 10, 20, 24, 29

[74] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning," *ArXiv*, vol. abs/2005.10243, 2020. 10

[75] D. Kim, D. Cho, D. Yoo, and I.-S. Kweon, "Learning image representations by completing damaged jigsaw puzzles," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 793–802, 2018. 11

[76] A. Tendle and M. R. Hasan, "A study of the generalizability of self-supervised representations," *ArXiv*, vol. abs/2109.09150, 2021. 11

[77] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, "Whitening for self-supervised representation learning," *ArXiv*, vol. abs/2007.06346, 2020. 11, 13

[78] Y. Zhang, B. Hooi, D. Hu, J. Liang, and J. Feng, "Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning," *ArXiv*, vol. abs/2102.06605, 2021. 11, 24, 30

[79] Z. Liu, A. Alavi, M. Li, and X. Zhang, "Self-supervised learning for time series: Contrastive or generative?" *ArXiv*, vol. abs/2403.09809, 2024. 11, 31

[80] V. G. T. D. Costa, E. Fini, M. Nabi, N. Sebe, and E. Ricci, "Solo-learn: A library of self-supervised methods for visual representation learning," *J. Mach. Learn. Res.*, vol. 23, pp. 56:1–56:6, 2021. 12, 17

[81] S. Shurrab and R. Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: a survey," *PeerJ Computer Science*, vol. 8, 2021. 13, 17, 21, 30

[82] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *ArXiv*, vol. abs/2006.10029, 2020. 13, 17

[83] J. Pan, P. Zhu, K. Zhang, B. Cao, Y. Wang, D. Zhang, J. Han, and Q. Hu, "Learning self-supervised low-rank network for single-stage weakly and semi-supervised semantic segmentation," *International Journal of Computer Vision*, vol. 130, pp. 1181 – 1195, 2022. 13

[84] C. Reed, X. Yue, A. Nrusimha, S. Ebrahimi, V. Vijaykumar, R. Mao, B. Li, S. Zhang, D. Guillory, S. L. Metzger, K. Keutzer, and T. Darrell, "Self-supervised pretraining improves self-supervised pretraining," *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1050–1060, 2021. 14

[85] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. G. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15 619–15 629, 2023. 14

[86] S. Guo, Z. Xiong, Y. Zhong, L. Wang, X. Guo, B. Han, and W. Huang, "Cross-architecture self-supervised video representation learning," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19 248–19 257, 2022. 14

[87] Y. M. Asano, M. Patrick, C. Rupprecht, and A. Vedaldi, "Labelling unlabelled videos from scratch with multi-modal self-supervision," *ArXiv*, vol. abs/2006.13662, 2020. 14

[88] A. Recasens, P. Luc, J.-B. Alayrac, L. Wang, F. Strub, C. Tallec, M. Malinowski, V. Patraucean, F. Altch'e, M. Valko, J.-B. Grill, A. van den Oord, and A. Zisserman, "Broaden your views for self-supervised video learning," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1235–1245, 2021. 14

[89] W. Moon, J. Kim, and J.-P. Heo, "Tailoring self-supervision for supervised learning," in *European Conference on Computer Vision*, 2022, pp. 346–364. 15

[90] K. Ranasinghe, M. Naseer, S. H. Khan, F. Khan, and M. Ryoo, "Self-supervised video transformer," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2864–2874, 2021. 15, 23

[91] J. Zhu, R. M. Moraes, S. Karakulak, V. Sobol, A. Canziani, and Y. LeCun, "Tico: Transformation invariance and covariance contrast for self-supervised visual representation learning," *ArXiv*, vol. abs/2206.10698, 2022. 15

[92] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 272–12 281, 2020. 15, 29

[93] A. Devillers and M. Lefort, "Equimod: An equivariance module to improve self-supervised learning," *ArXiv*, vol. abs/2211.01244, 2022. 15

[94] O. J. H'enaff, S. Koppula, J.-B. Alayrac, A. van den Oord, O. Vinyals, and J. Carreira, "Efficient visual pretraining with contrastive detection," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 066–10 076, 2021. 15

[95] Y. Liu, E. Sangineto, W. Bi, N. Sebe, B. Lepri, and M. D. Nadai, "Efficient training of visual transformers with small datasets," in *Neural Information Processing Systems*, 2021, pp. 23 818–23 830. 15, 20

[96] P. Zhou, Y. Zhou, C. Si, W. Yu, T. K. Ng, and S. Yan, "Mugs: A multi-granular self-supervised learning framework," *ArXiv*, vol. abs/2203.14415, 2022. 15

[97] C. Tao, X. Zhu, G. Huang, Y. Qiao, X. Wang, and J. Dai, "Siamese image modeling for self-supervised vision representation learning," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2132–2141, 2022. 16, 30

[98] Y. Cao and J. Wu, "Rethinking self-supervised learning: Small is beautiful," *ArXiv*, vol. abs/2103.13559, 2021. 16, 30

[99] J. Chen, M.-J. Hu, B. A. Li, and M. Elhoseiny, "Efficient self-supervised vision pretraining with local masked reconstruction," *ArXiv*, vol. abs/2206.00790, 2022. 16, 24, 25

[100] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, and D. Tao, "A survey on self-supervised learning: Algorithms, applications, and future trends." *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2023. 16

[101] C. Li, X. Li, L. Zhang, B. Peng, M. Zhou, and J. Gao, "Self-supervised pre-training with hard examples improves visual representations," *ArXiv*, vol. abs/2012.13493, 2020. 16, 30

[102] Y. Wang, C. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, pp. 213–247, 2022. 17, 18, 27, 30

[103] W. V. Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Gool, "Learning to classify images without labels," *ArXiv*, vol. abs/2005.12320, 2020. 17, 31

[104] S. Cao, D. Joshi, L. Gui, and Y.-X. Wang, "Hassod: Hierarchical adaptive self-supervised object detection," *ArXiv*, vol. abs/2402.03311, 2024. 17, 20, 26

[105] L. Schmarje, M. Santarossa, S.-M. Schroder, and R. Koch, "A survey on semi-, self- and unsupervised learning for image classification," *IEEE Access*, vol. 9, pp. 82 146–82 168, 2020. 17

[106] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 4037–4058, 2019. 17, 21, 30

[107] Y. M. Asano, C. Rupprecht, and A. Vedaldi, "Self-labelling via simultaneous clustering and representation learning," *ArXiv*, vol. abs/1911.05371, 2019. 17

[108] T. Hua, W. Wang, Z. Xue, Y. Wang, S. Ren, and H. Zhao, "On feature decorrelation in self-supervised learning," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9578–9588, 2021. 18, 27

[109] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Processing Magazine*, vol. 39, pp. 42–62, 2021. 18

[110] A. Islam, C.-F. Chen, R. Panda, L. Karlinsky, R. Radke, and R. Feris, "A broad study on the transferability of visual representations with contrastive learning," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8825–8835, 2021. 18

[111] E. Cole, X. S. Yang, K. Wilber, O. M. Aodha, and S. J. Belongie, "When does contrastive visual representation learning work?" *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 01–10, 2021. 18

[112] M. Marks, M. Knott, N. Kondapaneni, E. Cole, T. Defraeye, F. Pérez-Cruz, and P. Perona, "A closer look at benchmarking self-supervised pre-training with image classification," *ArXiv*, vol. abs/2407.12210, 2024. 18, 22

[113] T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, M. Carbin, and Z. Wang, "The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 301–16 311, 2020. 19

[114] O. J. H'enaff, S. Koppula, E. Shelhamer, D. Zoran, A. Jaegle, A. Zisserman, J. Carreira, and R. Arandjelovi'c, "Object discovery and representation networks," *ArXiv*, vol. abs/2203.08777, 2022. 19

[115] L. Tao, X. Wang, and T. Yamasaki, "Self-supervised video representation learning using inter-intra contrastive framework," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 19

[116] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, "Efficient self-supervised learning with contextualized target representations for vision, speech and language," in *International Conference on Machine Learning*, 2022, pp. 1416–1429. 19, 28

[117] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. P'erez, R. Marlet, and J. Ponce, "Localizing objects with self-supervised transformers and no labels," *ArXiv*, vol. abs/2109.14279, 2021. 19, 28

[118] W. V. Gansbeke, S. Vandenhende, S. Georgoulis, and L. Gool, "Revisiting contrastive methods for unsupervised learning of visual representations," in *Neural Information Processing Systems*, 2021, pp. 16 238–16 250. 19

[119] T. Wang, Q. Sun, S. Pranata, J. Karlekar, and H. Zhang, "Equivariance and invariance inductive bias for learning from insufficient data," *ArXiv*, vol. abs/2207.12258, 2022. 20

[120] N. Park, W. Kim, B. Heo, T. Kim, and S. Yun, "What do self-supervised vision transformers learn?" *ArXiv*, vol. abs/2305.00729, 2023. 20

[121] B. Roh, W. Shin, I. Kim, and S. Kim, "Spatially consistent representation learning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1144–1153, 2021. 20

[122] A. Bardes, J. Ponce, and Y. LeCun, "Vicregl: Self-supervised learning of local visual features," *ArXiv*, vol. abs/2210.01571, 2022. 20, 29

[123] F. Xiao, K. Kundu, J. Tighe, and D. Modolo, "Hierarchical self-supervised representation learning for movie understanding," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9717–9726, 2022. 22

[124] Y. Chen, W. Huang, K. Zhao, Y. Jiang, and G. Cong, "Self-supervised learning for geospatial ai: A survey," *ArXiv*, vol. abs/2408.12133, 2024. 23, 28

[125] P. Bhattacharyya, C. Huang, and K. Czarnecki, "Ssl-lanes: Self-supervised learning for motion forecasting in autonomous driving," *ArXiv*, vol. abs/2206.14116, 2022. 24

[126] V. Hondru, F.-A. Croitoru, S. Minaee, R. T. Ionescu, and N. Sebe, "Masked image modeling: A survey," *ArXiv*, vol. abs/2408.06687, 2024. 24

[127] Q. Wu, H. Ye, Y. Gu, H. Zhang, L. Wang, and D. He, "Denoising masked autoencoders help robust classification," in *International Conference on Learning Representations*, 2022. 25

[128] Y. Chen, M. Mancini, X. Zhu, and Z. Akata, "Semi-supervised and unsupervised deep visual learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 1327–1347, 2022. 25

[129] M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, and R. Rodrigo, "Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9892–9902, 2022. 25

[130] V. Dave, F. Lygerakis, and E. Rueckert, "Multimodal visual-tactile representation learning through self-supervised contrastive pre-training," *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8013–8020, 2024. 25

[131] T. Li, D. Katabi, and K. He, "Return of unconditional generation: A self-supervised representation generation method," 2023. 25

[132] H. Liu, X. Jiang, X. Li, A. Guo, D. Jiang, and B. Ren, "The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training," *ArXiv*, vol. abs/2204.08227, 2022. 25

[133] E. Amrani, R. Ben-Ari, D. Rotman, and A. Bronstein, "Noise estimation using density estimation for self-supervised multimodal learning," *ArXiv*, vol. abs/2003.03186, 2020. 25

[134] S. A. Koohpayegani, A. Tejankar, and H. Pirsiavash, "Compress: Self-supervised learning by compressing representations," *ArXiv*, vol. abs/2010.14713, 2020. 26

[135] W. Wang, M. A. Kaleem, A. Dziedzic, M. Backes, N. Papernot, and F. Boenisch, "Memorization in self-supervised learning improves downstream generalization," *ArXiv*, vol. abs/2401.12233, 2024. 27

[136] H. V. Vo, V. Khalidov, T. Darcet, T. Moutakanni, N. Smetanin, M. Szafraniec, H. Touvron, C. Couprie, M. Oquab, A. Joulin, H. J'egou, P. Labatut, and P. Bojanowski, "Automatic data curation for self-supervised learning: A clustering-based approach," *ArXiv*, vol. abs/2405.15613, 2024. 27

[137] G. Ghiasi, B. Zoph, E. D. Cubuk, Q. V. Le, and T.-Y. Lin, "Multi-task self-training for learning general representations," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8836–8845, 2021. 27

[138] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, and D. Larlus, "Hard negative mixing for contrastive learning," *ArXiv*, vol. abs/2010.01028, 2020. 27

[139] J. Jia, Y. Liu, and N. Gong, "Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning," *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2043–2059, 2021. 28

[140] A. Ziegler and Y. M. Asano, "Self-supervised learning of object parts for semantic segmentation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14 482–14 491, 2022. 28

[141] M. Seitzer, M. Horn, A. Zadaianchuk, D. Zietlow, T. Xiao, C.-J. Simon-Gabriel, T. He, Z. Zhang, B. Scholkopf, T. Brox, and F. Locatello, "Bridging the gap to real-world object-centric learning," *ArXiv*, vol. abs/2209.14860, 2022. 28

[142] F. Haghighi, M. Taher, M. Gotway, and J. Liang, "Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20 792–20 802, 2022. 30

[143] U. Ozbulak, H. J. Lee, B. Boga, E. T. Anzaku, H. min Park, A. V. Messem, W. D. Neve, and J. Vankerschaver, "Know your self-supervised learning: A survey on image-based generative and discriminative training," *ArXiv*, vol. abs/2305.13689, 2023. 30

[144] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *ArXiv*, vol. abs/2001.07685, 2020. 30