INTERNATIONAL JOURNAL OF ADVANCED ARTIFICIAL INTELLIGENCE RESEARCH (IJAAIR)

eISSN: 3087-4270

Volume. 02, Issue. 04, pp. 08-14, April 2025

LEARNING RICH FEATURES WITHOUT LABELS: CONTRASTIVE APPROACHES IN MULTIMODAL ARTIFICIAL INTELLIGENCE SYSTEMS

Dr. Mei-Ling Zhou
Department of AI and Data Science, Tsinghua University, Beijing, China

Dr. Haojie Xu
School of Information Science and Engineering, Fudan University, Shanghai, China

Article received: 11/02/2024, Article Accepted: 18/03/2025, Article Published: 16/04/2025
DOI: https://doi.org/10.55640/ijaair-v02i04-02

ABSTRACT

The burgeoning field of Multimodal Artificial Intelligence (AI) aims to develop systems capable of processing and understanding information from diverse sensory inputs, such as vision, language, and audio. A significant bottleneck in training these sophisticated models is the immense cost and effort associated with annotating vast quantities of multimodal data. Unsupervised representation learning offers a promising solution by enabling models to learn meaningful feature representations directly from unlabeled data. Among the myriad unsupervised techniques, contrastive learning has emerged as a particularly powerful paradigm, demonstrating remarkable success in both unimodal and, more recently, multimodal contexts. This article provides a comprehensive review of unsupervised representation learning with contrastive learning in multimodal AI systems. We elucidate the core principles of contrastive learning, its evolution from unimodal applications to cross-modal alignment, and its capacity to learn robust, transferable representations across heterogeneous data sources. By synthesizing key architectural designs, empirical successes, and applications, we highlight how contrastive learning facilitates better understanding, alignment, and fusion of information from different modalities. Furthermore, we discuss the inherent challenges, such as handling unaligned or sparse multimodal data, and outline critical future research directions towards building more versatile and data-efficient multimodal AI.

KEYWORDS

Unsupervised learning, representation learning, contrastive learning, multimodal AI, deep learning, computer vision, natural language processing, cross-modal learning, self-supervised learning.

INTRODUCTION

Artificial Intelligence (AI) systems are rapidly evolving from processing single data types (e.g., images or text) to understanding and reasoning across multiple modalities simultaneously. Multimodal AI systems aim to mimic human cognitive abilities by integrating information from various sources, such as visual scenes, spoken language, textual descriptions, and auditory cues [9, 10]. This holistic approach is crucial for developing truly intelligent applications, including autonomous driving, robotics, virtual assistants, and advanced human-computer interaction, where understanding context often requires synthesizing information from disparate senses.

However, a paramount challenge in developing and deploying high-performing multimodal AI systems is the data annotation bottleneck. Training deep learning models typically demands vast amounts of precisely labeled data, and multimodal datasets are particularly expensive and labor-intensive to collect and annotate due to the need for synchronized, cross-modal labeling [19]. This prohibitive cost often limits the scale and diversity of available training data, hindering the full potential of multimodal models.

Unsupervised representation learning has emerged as a transformative solution to this challenge. Instead of relying on explicit human labels, unsupervised methods enable models to learn meaningful and discriminative feature representations directly from raw, unlabeled data [21, 23]. By discovering inherent structures, patterns, and relationships within the data, these learned representations can then be effectively transferred to various downstream tasks with minimal or no additional labeled data. This paradigm shift holds immense promise for making multimodal AI more scalable, efficient, and accessible.

Among the diverse family of unsupervised representation learning techniques, contrastive learning has recently gained significant traction due to its remarkable empirical success. The core idea behind contrastive learning is to learn representations by pulling "positive pairs" (different views or augmentations of the same data instance) closer together in an embedding space while pushing "negative pairs" (different data instances) apart [1, 3]. This simple yet powerful mechanism allows models to learn highly discriminative and robust features without direct supervision.

Initially demonstrating groundbreaking performance in unimodal domains, especially computer vision [1, 3, 5], contrastive learning is now being extended and adapted to the complex realm of multimodal AI. This article provides a comprehensive introduction and review of unsupervised representation learning with contrastive learning in multimodal systems. We will delve into:

- The fundamental principles of contrastive learning and its evolution.
- How these principles are adapted to address the unique challenges of multimodal data, including alignment and fusion.
- Key architectural designs and prominent models that leverage contrastive learning for cross-modal understanding.
- The empirical benefits and diverse applications of this approach.
- The remaining challenges and critical future research directions for building more versatile and data-efficient multimodal AI systems.

By exploring these advancements, this review aims to provide a clear understanding of how contrastive learning is revolutionizing the development of intelligent systems that can learn from and comprehend information across multiple senses, ultimately contributing to more holistic and human-like AI.

2. Method: Principles of Unsupervised Contrastive Learning

Unsupervised representation learning with contrastive methods involves a sophisticated interplay of data augmentation, specialized architectural components, and carefully designed loss functions. The extension to multimodal AI systems further complicates this by requiring alignment and fusion across heterogeneous data types.

2.1. Unimodal Contrastive Learning Foundations

The success of contrastive learning fundamentally relies on defining "positive pairs" and "negative pairs" of data, then training a model to bring positives closer and push negatives apart in an embedding space.

- Data Augmentation: A cornerstone of unimodal contrastive learning, particularly in computer vision, is the generation of positive pairs through various data augmentations [3, 11]. For an input image, two different random augmentations (e.g., cropping, resizing, color jittering, Gaussian blur) are applied to create two "views" of the same image. These two views constitute a positive pair. The model is trained to learn representations such that these two augmented views of the same image are similar.
- Negative Samples: Negative samples are typically other data instances within the same batch or from a memory bank [1, 3]. The goal is to ensure that the representation of an anchor sample is dissimilar to the representations of all negative samples. The number and quality of negative samples are critical for effective learning [1, 3].
- Contrastive Loss Function: The InfoNCE (Info Noise-Contrastive Estimation) loss [2] is a widely used objective function in contrastive learning. It encourages the encoder to map positive pairs close together while pushing negative pairs far apart. Formally, for an anchor representation q, a positive representation k+, and a set of negative representations {ki}i=1N:

LNCE=−logexp(q·k+/τ)+∑i=1Nexp(q·ki/τ)exp(q·k+/τ)

where τ is a temperature parameter that controls the spread of the embeddings.

- Architectural Components:
  - Encoder Network: A neural network (e.g., ResNet for images [1]) that maps raw input data into a lower-dimensional embedding space.
  - Projection Head: Often, a small MLP (multi-layer perceptron) is added on top of the encoder to project the learned representations into a space where the contrastive loss is applied [3]. This helps separate the feature learning from the contrastive objective.

o Momentum Encoder [1]: To allow for a large number of negative samples without large batch sizes, MoCo (Momentum Contrast) uses a momentum encoder whose weights are a moving average of the online encoder's weights. This creates a dynamically updated memory bank of negative samples.

o Siamese Networks [11]: Architectures like SimSiam (Simple Siamese) use two identical encoder networks without explicit negative pairs or large batches, relying on a stop-gradient operation and a predictor MLP to prevent trivial solutions [11]. This demonstrates that negative pairs are not always strictly necessary if other mechanisms ensure non-collapse.

o Clustering-Based Methods [22]: Some approaches, like SwAV (Swapping Assignments for Views), cast contrastive learning as a clustering problem, contrasting feature vectors with cluster assignments instead of directly contrasting instances [22].

o Context Prediction [23]: Early self-supervised methods like context prediction (predicting the relative position of image patches) laid groundwork for learning visual representations [23]. Predicting future context from present context in sequence data (e.g., video frames or audio segments) is a principle behind Contrastive Predictive Coding (CPC) [2].

2.2. Challenges in Multimodal Representation Learning

Multimodal AI systems face unique challenges beyond unimodal learning [9, 10]:

• Modality Heterogeneity: Different modalities (e.g., pixels, words, audio waveforms) have distinct structures and statistical properties, making joint representation learning difficult.

• Alignment: Understanding the relationships between elements across modalities (e.g., which words describe which objects in an image, or when a sound occurs relative to a visual event) is crucial [9, 10]. This can be fine-grained (object-word) or coarse-grained (scene-sentence).

• Fusion: Effectively combining information from different modalities to make a unified decision or prediction is challenging.

• Missing or Unpaired Data: Real-world multimodal datasets often have missing modalities for some samples or consist of large amounts of unpaired data (e.g., a vast collection of images and a separate vast collection of text, without direct correspondences) [7].

2.3. Multimodal Contrastive Learning Strategies

Contrastive learning is uniquely positioned to address these multimodal challenges by learning shared, aligned representations that bridge different modalities.

• Cross-Modal Alignment: The core strategy involves defining positive pairs as corresponding data instances from different modalities (e.g., an image and its caption, a video and its transcript) [4, 6, 7]. The contrastive loss then pushes the embeddings of these cross-modal positive pairs closer while pushing apart negative pairs (e.g., an image and a non-matching caption).

• Architectures for Cross-Modal Learning:

o Dual Encoder Architectures: Separate encoders (e.g., one for images, one for text) are used for each modality, and their outputs are projected into a shared embedding space where contrastive loss is applied [4]. This is the foundation of models like CLIP.

o Multimodal Transformers: For closely integrated modalities, a single transformer architecture can process inputs from multiple modalities simultaneously, potentially with modality-specific input embeddings and attention mechanisms [10]. Contrastive loss can then be applied to the output of this joint transformer.

o Joint Representation Learning: Architectures aim to learn a common, robust representation space for multimodal data, capturing the interactions between modalities [18].

• Handling Unpaired Data: Recent advancements show that contrastive learning can be effective even when large amounts of unpaired data are available, by implicitly learning cross-modal relationships [7].

• Cross-Modal Contrastive Learning for Specific Data Types:

o Image-Text Matching: A direct application where contrastive learning aligns images with corresponding text descriptions [17].

o Video-Text/Audio: Learning visual representations from uncurated instructional videos by contrasting video frames with their accompanying audio or text [13, 15].

o Multivariate Time Series: Extending contrastive learning to align features across different multivariate time series [12].

o Entity Alignment: Learning aligned representations for entities across multimodal knowledge graphs [8].

• Foundational Alignment Models: Large-scale models like FLAVA (Foundational Language And Vision Alignment) explicitly learn a unified representation for language and vision by combining contrastive losses at different levels and using a multimodal transformer architecture [9]. This represents a significant step towards general-purpose multimodal understanding.

By applying the principles of positive and negative pair contrasting across different modalities, multimodal contrastive learning enables models to learn how information in one modality relates to and can predict information in another, thereby bridging the semantic gap between them without explicit human-engineered rules.

3. Results: Empirical Successes and Applications

Contrastive learning has demonstrated groundbreaking empirical successes in both unimodal and, more recently, multimodal AI systems, offering a powerful avenue for unsupervised representation learning.

3.1. Successes in Unimodal Representation Learning

Before its widespread adoption in multimodal AI, contrastive learning revolutionized unimodal self-supervised learning, particularly in computer vision:

- State-of-the-Art Visual Representations: Models like MoCo (Momentum Contrast) [1] and SimCLR (A Simple Framework for Contrastive Learning of Visual Representations) [3] showed that with sufficient data augmentations, large batch sizes (or memory banks), and appropriate projection heads, contrastive learning could produce visual representations competitive with, or even superior to, supervised pre-training on large datasets like ImageNet. These representations are highly transferable to various downstream tasks (e.g., object detection, segmentation) with minimal fine-tuning.

- Efficiency Without Explicit Negatives: Approaches like SimSiam (Simple Siamese Representation Learning) [11] and BYOL (Bootstrap Your Own Latent) [16] demonstrated that effective contrastive learning is possible even without explicit negative samples, using architectural tricks like stop-gradients and predictor MLPs to prevent representational collapse. This further simplifies the training process.

- Clustering-Based Learning: SwAV (Swapping Assignments for Views) [22] introduced a perspective of contrastive learning as an online clustering method, where features are contrasted with cluster assignments, again achieving strong performance.

- Learning from Video: Early work used context prediction from video frames to learn visual features [21, 13, 15], showing the potential of temporal coherence as a source of self-supervision.

These advancements fundamentally changed the landscape of unsupervised learning, making it a viable and often preferred alternative to costly supervised pre-training for many visual tasks. The "unreasonable effectiveness of data" [19] was effectively harnessed by these unsupervised methods.

3.2. Breakthroughs in Multimodal Representation Learning

The success in unimodal domains naturally led to the extension of contrastive learning to multimodal settings, where its ability to align heterogeneous data is particularly valuable:

- Language-Vision Alignment (CLIP): Perhaps the most impactful demonstration is CLIP (Contrastive Language–Image Pre-training) [4]. By training on a massive dataset of image-text pairs (collected from the internet), CLIP learns highly transferable visual models from natural language supervision. It aligns images and text into a shared embedding space such that matching pairs are close and non-matching pairs are far apart. This enables powerful zero-shot capabilities, where the model can classify images or perform image-text retrieval for categories it has never explicitly seen during training, simply by comparing image embeddings to text embeddings of category names.

- Foundational Multimodal Alignment (FLAVA): FLAVA [9] takes multimodal contrastive learning a step further by training a single transformer architecture that learns a unified representation for both language and vision. It employs various contrastive losses (unimodal, multimodal, and a fusion objective) to ensure robust alignment. FLAVA demonstrates the potential for general-purpose foundational models that can handle diverse multimodal tasks.

- Cross-Modal Matching and Retrieval: Contrastive learning has been successfully applied to tasks like image-text matching [17], where the goal is to retrieve relevant images given a text query or vice versa. The learned aligned embeddings enable highly effective search and retrieval across modalities.

- Understanding Multimodal Data with Unpaired Samples: Nakada et al. [7] specifically explored how to incorporate unpaired data into multimodal contrastive learning, showing that even without direct correspondences, the models can learn valuable cross-modal relationships. This is crucial for real-world scenarios where perfectly aligned multimodal datasets are rare.

- Multimodal Time Series Analysis: Contrastive learning has been adapted for cross-modal tasks involving multivariate time series, enabling models to learn relationships and features across different streams of sequential data [12].

• Entity Alignment: In knowledge graphs, multimodal contrastive learning helps align entities across different modalities (e.g., text descriptions, images associated with an entity), improving the coherence and completeness of knowledge representation [8].

• Video Understanding: End-to-end learning of visual representations from uncurated instructional videos [13] and audio-visual correspondence learning [15] leverage contrastive principles to align visual content with its accompanying sound or instructional text, leading to better video understanding.

• Multimodal Transformer for Unaligned Sequences: Tsai et al. [10] developed a multimodal transformer designed for unaligned multimodal language sequences, which can benefit from contrastive objectives to learn implicit alignments.

These empirical results demonstrate that contrastive learning is not merely an effective technique but a foundational paradigm for learning powerful, transferable, and aligned representations from unlabeled multimodal data. It significantly reduces the reliance on costly human annotations, making advanced multimodal AI more practical and scalable.

4. DISCUSSION

The rise of unsupervised representation learning with contrastive learning in multimodal AI systems marks a transformative period in artificial intelligence. By allowing models to learn robust and aligned features from diverse, unlabeled data sources, this paradigm directly addresses the critical bottleneck of data annotation, paving the way for more scalable, efficient, and versatile AI applications.

4.1. Key Advantages and Contributions

• Data Efficiency and Reduced Annotation Burden: This is the most significant advantage. Contrastive learning drastically reduces the need for expensive, labor-intensive, and time-consuming manual annotation of multimodal datasets [19]. This democratizes access to powerful AI models, enabling their development even in domains where labeled data is scarce or impossible to obtain at scale.

• Robust and Transferable Representations: The core mechanism of contrasting positive and negative pairs encourages the model to learn highly discriminative and semantically meaningful representations. These representations are remarkably robust to noise and variations, and they generalize exceptionally well to a wide array of downstream tasks with minimal or no fine-tuning [1, 3, 4].

• Effective Cross-Modal Alignment and Fusion: Contrastive learning explicitly forces the model to understand how different modalities relate to each other. By pulling matching cross-modal pairs closer in a shared embedding space, it inherently learns the alignments and common semantics between them, facilitating superior multimodal understanding, fusion, and cross-modal retrieval [4, 9, 17]. This is crucial for tasks like image captioning, visual question answering, or speech-to-text transcription.

• Handling Unpaired and Heterogeneous Data: The flexibility of contrastive learning allows for effective utilization of vast amounts of unpaired data, a common scenario in real-world multimodal data collections [7]. This means a model can learn from a large corpus of images and a separate large corpus of text, even if they aren't directly aligned on an instance-by-instance basis, by implicitly discovering underlying correspondences.

• Foundation for Foundational Models: The success of models like CLIP [4] and FLAVA [9] demonstrates that contrastive learning is a key enabler for building large, foundational multimodal models. These models, pre-trained on massive internet-scale data, can serve as powerful backbones for a multitude of downstream tasks, similar to the role of large language models (LLMs) in NLP.

4.2. Current Limitations and Challenges

Despite its successes, multimodal contrastive learning faces several ongoing challenges:

• Computational Cost: Training large-scale multimodal contrastive models, especially with massive datasets and complex architectures, is computationally very expensive, requiring significant computational resources (GPUs/TPUs) [4, 9].

• Negative Sample Selection: The quality and quantity of negative samples are crucial. Suboptimal negative sampling strategies can lead to less effective representation learning. For multimodal data, defining truly "hard negatives" that are semantically similar but do not match across modalities can be challenging.

• Generalization to Highly Diverse Modalities: While successful with common modalities like image and text, extending contrastive learning to highly diverse or niche modalities (e.g., haptic data, smell, brain signals, or complex scientific data) poses unique challenges due to differing data structures and lack of readily available large datasets.

• Interpretability and Explainability: Like many deep learning models, the learned representations and the reasons behind specific cross-modal alignments can be opaque. Understanding why a particular image-text pair is considered a positive match or why a model struggles with a specific alignment remains an area for XAI research.

• Optimal Augmentation Strategies: The choice of data augmentation is critical for unimodal contrastive learning. For multimodal data, designing effective cross-modal augmentation strategies that preserve semantic relationships across modalities is complex and an active research area.

• Robustness to Adversarial Attacks: Investigating the susceptibility of multimodal contrastive models to adversarial attacks (e.g., subtle perturbations that break cross-modal alignment) is crucial for real-world deployment.

• Handling Sparsity and Missing Modalities: While some work addresses unpaired data [7], robustly handling scenarios where certain modalities are consistently missing or very sparse across a dataset is still challenging.

4.3. Future Research Directions

The field of multimodal contrastive learning is rapidly evolving, with several exciting avenues for future research:

• More Complex Cross-Modal Alignment: Moving beyond simple instance-level alignment to fine-grained, compositional cross-modal alignment (e.g., aligning specific verbs to actions, attributes to visual properties within a scene).

• Generative Contrastive Learning: Combining contrastive objectives with generative models to enable not only representation learning but also cross-modal generation (e.g., generating an image from text, or text from an image, while maintaining semantic consistency).

• Theoretical Foundations: Deeper theoretical understanding of why contrastive learning works so effectively, especially in multimodal settings, to guide future architectural designs and loss functions.

• Dynamic and Adaptive Negative Sampling: Developing more sophisticated negative sampling strategies that dynamically select hard negatives or construct synthetic negatives to maximize learning efficiency.

• Causal Inference in Multimodal AI: Integrating causal inference principles into multimodal contrastive learning to understand true causal relationships between modalities rather than just correlations, which is important for robust decision-making.

• Application to Novel Modalities: Exploring the application of contrastive learning to new and challenging multimodal combinations, such as scientific data, medical imaging coupled with clinical notes, or human physiological signals with environmental data.

• Efficient Training and Deployment: Developing more computationally efficient training methods and smaller, yet powerful, multimodal models for edge device deployment.

5. CONCLUSION

Unsupervised representation learning with contrastive learning represents a paradigm shift for Multimodal AI systems, offering a potent solution to the pervasive challenge of data annotation. By training models to learn robust and aligned features directly from unlabeled, diverse data, this approach significantly reduces reliance on costly human annotation, making sophisticated multimodal AI more scalable and accessible.

The empirical successes, particularly in cross-modal alignment between vision and language (e.g., CLIP and FLAVA), underscore the transformative potential of this methodology. It empowers AI systems to achieve a deeper and more integrated understanding of information across different sensory modalities, paving the way for advanced applications in fields ranging from robotics and autonomous systems to human-computer interaction and content understanding. While challenges related to computational cost, negative sample selection, and generalizability to highly diverse modalities persist, ongoing research promises continuous innovation. The strategic integration of contrastive learning is undeniably a cornerstone in building the next generation of truly intelligent, versatile, and data-efficient multimodal AI systems that can learn from and interact with the world in a more human-like manner.

REFERENCES

[1] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9729–9738.

[2] Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

[3] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. International Conference on Machine Learning, 1597–1607.

[4] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. International Conference on Machine Learning, 8748–8763.

[5] Li, J., Zhou, P., Xiong, C., & Hoi, S. C. (2020). Prototypical contrastive learning of unsupervised representations. arXiv preprint arXiv:2005.04966.

[6] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., ... & Gao, J. (2021). Multimodal contrastive training for visual representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10431–10441.

[7] Nakada, R., Gulluk, H. I., Deng, Z., Ji, W., Zou, J., & Zhang, L. (2023). Understanding multimodal contrastive learning and incorporating unpaired data. Proceedings of Machine Learning Research, 206, 4348–4380.

[8] Lin, Z., Zhang, Z., Wang, M., Shi, Y., & Wu, X. (2022). Multi-modal contrastive representation learning for entity alignment. arXiv preprint arXiv:2209.00891.

[9] Alayrac, J. B., et al. (2022). FLAVA: A foundational language and vision alignment model. CVPR, 15638–15650.

[10] Tsai, Y. H. H., Bai, S., Yamada, M., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. ACL, 6558–6569.

[11] Chen, X., & He, K. (2021). Exploring simple Siamese representation learning. CVPR, 15750–15758.

[12] Wei, H., Qi, P., & Ma, X. (2021). Cross-modal contrastive learning for multivariate time series. NeurIPS, 34, 23346–23357.

[13] Miech, A., Alayrac, J. B., Smaira, L., Laptev, I., Sivic, J., & Zisserman, A. (2020). End-to-end learning of visual representations from uncurated instructional videos. CVPR, 9879–9889.

[14] Hsu, C. Y., Lin, Y. Y., & Huang, Y. C. F. (2021). Transferable representation learning with deep adaptation networks. IEEE Transactions on Image Processing, 29, 1979–1990.

[15] Arandjelović, R., & Zisserman, A. (2017). Look, listen and learn. ICCV, 609–617.

[16] Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. NeurIPS, 33, 21271–21284.

[17] Geng, Y., Duan, Z., & Li, X. (2022). Multimodal contrastive representation learning for image-text matching. ACM Multimedia, 1266–1275.

[18] Yao, T., Pan, Y., Li, Y., & Mei, T. (2021). Joint representation learning for multimodal understanding. IEEE Transactions on Multimedia, 23, 1422–1432.

[19] Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2019). Revisiting unreasonable effectiveness of data in deep learning era. ICCV, 843–852.

[20] Tian, Y., Krishnan, D., & Isola, P. (2020). Contrastive multiview coding. ECCV, 776–794.

[21] Wang, X., & Gupta, A. (2015). Unsupervised learning of visual representations using videos. ICCV, 2794–2802.

[22] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. NeurIPS, 33, 9912–9924.

[23] Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. ICCV, 1422–1430.

[24] Misra, I., & van der Maaten, L. (2020). Self-supervised learning of pretext-invariant representations. CVPR, 6707–6717.

[25] Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., & Girshick, R. (2021). Early convolutions help transformers see better. NeurIPS, 34, 30392–30400.