International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V73I2P108

Volume 73 Issue 2, 64-70, February 2025
© 2025 Seventh Sense Research Group®

Original Article

Exploring Multimodal Large Language Models for Next-Generation Recommendation Systems

Kailash Thiyagarajan

Independent Researcher, Austin, TX, USA.

[1]*Corresponding Author : kailash.thiyagarajan@ieee.org*

Received: 29 December 2024 Revised: 24 January 2025 Accepted: 15 February 2025 Published: 28 February 2025

Abstract *- Multimodal Large Language Models (MLLMs) integrate diverse data modalities—including textual descriptions, visual content, and contextual signals—into a unified framework for advanced machine learning tasks. In recommendation systems, these models offer a more comprehensive approach by combining user behavioral data, product metadata, and visual features to enhance relevance prediction. This research explores an end-to-end integration of MLLMs into recommendation pipelines, spanning from data preparation and model adaptation in the batch training phase to real-time serving for low-latency inference. A modular architecture is introduced, built on a pre-trained transformer backbone with modality-specific encoders, allowing seamless fusion of multimodal inputs. Empirical evaluations on an e-commerce dataset reveal that the proposed MLLM-based recommender outperforms unimodal baselines, leading to higher recall and improved user satisfaction. Critical considerations for data alignment, scalability, and interpretability in real-world deployment are also discussed. These findings highlight the transformative potential of multimodal learning in next-generation recommendation systems.*

Keywords *- Multimodal large language models, Recommendation systems, Cross-modal fusion, Personalized content, Transformer-based models, Real-time inference.*

1. Introduction

1.1. Background and Motivation

Recommender systems have become indispensable in various digital platforms, including e-commerce, content streaming, and social media. Traditional recommendation models primarily rely on collaborative filtering and matrix factorization, leveraging user-item interaction data such as clicks, purchases, and ratings. While these approaches have been widely adopted, they often fail to incorporate rich contextual information from multiple modalities, such as textual product descriptions, customer reviews, and visual content. As a result, they struggle with ambiguous or sparse user interaction data, limiting their ability to generate personalized recommendations.

In recent years, the field of deep learning has advanced significantly, particularly with the rise of Large Language Models (LLMs) such as BERT and GPT. These models have demonstrated state-of-the-art performance in natural language processing tasks, including text classification, summarization, and question-answering. However, most LLMs are inherently text-centric and do not effectively process multimodal information, which is crucial for modern recommendation systems. User preferences are often influenced by a combination of text, images, and structured metadata**,** making exploring architectures that can integrate multiple data modalities is necessary.

1.2. Rise of Multimodal Large Language Models(MLLMs)

Recent research has introduced Multimodal Large Language Models (MLLMs) to overcome the limitations of unimodal recommendation models. These models extend the capabilities of traditional LLMs by integrating specialized modality-specific encoders (e.g., vision transformers for image processing) into a shared Transformer-based architecture. By aligning textual, visual, and contextual information into a unified embedding space, MLLMs enable a more comprehensive understanding of user intent and product relevance.

In the context of recommendation systems, MLLMs offer distinct advantages:

- Enhanced Representation Learning: MLLMs capture richer user preferences by fusing text-based signals (e.g., product descriptions and user reviews) with visual features (e.g., product images).
- Improved Disambiguation: When textual descriptions are unclear, visual data can provide additional context to refine recommendation accuracy.

- Cross-Modal Interpretability: Understanding why an item is recommended becomes more intuitive when both text and images contribute to decision-making.

1.3. Research Gap and Contributions

Despite the growing interest in multimodal learning**,** few studies have thoroughly examined the effectiveness of MLLMs in real-world recommendation scenarios, especially in high-volume, latency-sensitive applications**.** Current implementations often either:

- Focus solely on text-based models, which are missing valuable multimodal insights.
- Use independent feature extractors without fully leveraging cross-modal interactions within a Transformer framework.

This research addresses these gaps by proposing a structured methodology for integrating MLLMs into recommendation pipelines**.** The key contributions of this study are as follows:

1. Unified Architectural Design**:** We introduce a modular recommendation system that employs a pre-trained Transformer backbone augmented with modality-specific encoders and cross-modal attention mechanisms**.**
2. End-to-End Recommendation Pipeline**:** The proposed framework spans from data preprocessing and model adaptation (Batch Training Phase) to low-latency inference (Real-time Serving Phase).
3. Empirical Validation**:** Through experiments on a large-scale e-commerce dataset, we demonstrate significant improvements in recall and user satisfaction over unimodal baselines.

This study aims to push the boundaries of personalized content delivery by integrating multimodal cues into a recommendation framework. The remainder of this paper explores the related work**,** details the proposed methodology**,** and presents experimental findings to support our claims.

2. Related Work

2.1. Traditional Recommendation Methods

Recommender systems have historically relied on collaborative filtering and content-based filtering, where user-item interactions such as clicks, ratings, and purchases are modeled using matrix factorization or latent factor techniques. While these methods have been successful in many applications, they often struggle with cold-start problems and fail to capture rich multimodal context present in real-world data. Research has shown that incorporating additional signals such as textual descriptions, reviews, and images can significantly improve recommendation accuracy. However, early models primarily focused on structured numerical data, ignoring valuable multimodal cues.

2.2. Deep Learning for Recommendation Systems

The rise of deep learning introduced neural collaborative filtering and deep hybrid models, where neural networks process structured data alongside text-based embeddings. Models like DeepFM and Wide & Deep demonstrated how combining feature interactions with deep networks could enhance predictive accuracy. However, these approaches still primarily relied on textual or tabular data and lacked mechanisms to efficiently integrate image-based or contextual information.

2.3. Advances in Multimodal Learning

Multimodal learning has gained significant traction, particularly in areas like image captioning, video retrieval, and visual question-answering, where text and visual inputs must be processed together. Recent advancements in transformer-based architectures have enabled cross-modal fusion, allowing models to learn from text, images, and structured data jointly. Vision-language models such as CLIP and ALIGN have demonstrated that aligning textual embeddings with visual features improves semantic understanding, particularly relevant for recommendation tasks.

Table 1. Comparison of Recommendation Methods and Performance

Model	Modality Used	Strengths	Limitations	Performance (nDCG@10 / Recall@20)
Matrix Factorization	User-item interactions	Simple, interpretable, scalable	Ignores textual and visual context	0.243 / 0.332
Text-Only LLM	Text metadata	Leverages NLP advancements for understanding the text	Cannot incorporate visual cues	0.258 / 0.342
Image-Only CNN	Image features	Captures visual features	Ignores textual meaning and context	0.217 / 0.305
Proposed MLLM	Text + Image + Context	Cross-modal fusion, better user intent modeling	Higher computational cost	0.282 / 0.380

2.4. Multimodal Large Language Models for Recommendations

Multimodal large language models (MLLMs) extend traditional large language models by incorporating modality-specific encoders that process textual, visual, and structured data within a unified framework. Unlike previous hybrid models that treat different modalities independently, MLLMs apply cross-modal attention mechanisms to dynamically learn dependencies between text and images, leading to richer item representations. Studies have shown that these architectures significantly enhance recommendation accuracy, especially in visually driven domains such as fashion, home décor, and e-commerce.

2.5. Comparison with Existing Work

While prior research has explored text-based recommendation models, this study differs by introducing an end-to-end multimodal recommendation pipeline that integrates real-time serving constraints and cross-modal fusion techniques. Unlike existing approaches that rely on pre-extracted visual features, the proposed model jointly optimizes textual and visual embeddings within a transformer-based network.

This research bridges the gap between multimodal learning and real-time recommendation systems, offering a scalable approach applicable to large-scale commercial applications.

3. Proposed Method

In our approach, an MLLM-based recommender is constructed through two principal phases: Batch Training and Real-time Serving.

3.1. Batch Training Phase

3.1.1 Data Collection and Preprocessing

A large-scale e-commerce dataset was compiled, consisting of:

- Product metadata: Titles, descriptions, category labels.
- Visual content: High-resolution product images.
- User-item interactions**:** Clicks, purchases, and browsing history.
- Contextual signals: Timestamp, device type, and geographic location.

3.1.2. Text Processing

- Tokenized using Byte-Pair Encoding (BPE) aligned with the pre-trained Transformer model.
- Retained stopwords where necessary to preserve contextual integrity.

3.1.3. Image Processing

- Standardized resolution and normalized image data.
- Extracted visual features using a convolutional neural network (CNN) or a Vision Transformer (ViT).

3.1.4. User Interaction Structuring

- Modeled user actions into session-based sequences, capturing temporal dependencies in browsing behavior.

3.2. Model Architecture

The proposed MLLM-based recommendation model consists of four key components. (Refer to Figure 1)

1. Text Encoder
   - A pre-trained Transformer model (such as BERT or GPT) processes textual data.
   - Self-attention layers refine embeddings by capturing contextual dependencies.

2. Visual Encoder
   - A CNN (e.g., ResNet) or a ViT extracts visual feature representations.
   - Projection layers align image embeddings with textual representations.

3. Cross-Modal Fusion Layer
   - A cross-attention mechanism integrates text and image embeddings into a shared latent space.
   - Image features are incorporated as additional tokens within the Transformer model for joint learning.

4. Recommendation Output Layer
   - A fully connected network predicts interaction likelihoods.
   - The model is trained for click prediction, rating estimation, and personalized ranking tasks.

The Transformer backbone was selected over other models due to its ability to handle long-range dependencies across multimodal data, improving interpretability and alignment between textual and visual inputs.

3.2. Training Objectives

The model is optimized using a multitask loss function incorporating:

- Contrastive Loss for Multimodal Alignment
  - Ensures that text and image representations of the same product remain closely aligned.
- Masked Language Modeling Loss
  - Preserves domain-specific textual understanding for product descriptions and reviews.
- Recommendation-Specific Losses
  - Cross-entropy loss for click prediction.
  - Bayesian Personalized Ranking (BPR) for ranking optimization.

3.3. Model Evaluation

To validate model performance, the dataset is split 70-15-15 for training, validation, and testing. Generalizability is assessed by evaluating the model on previously unseen product categories**.**

International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V73I2P108

Volume 73 Issue 2, 64-70, February 2025
© 2025 Seventh Sense Research Group®

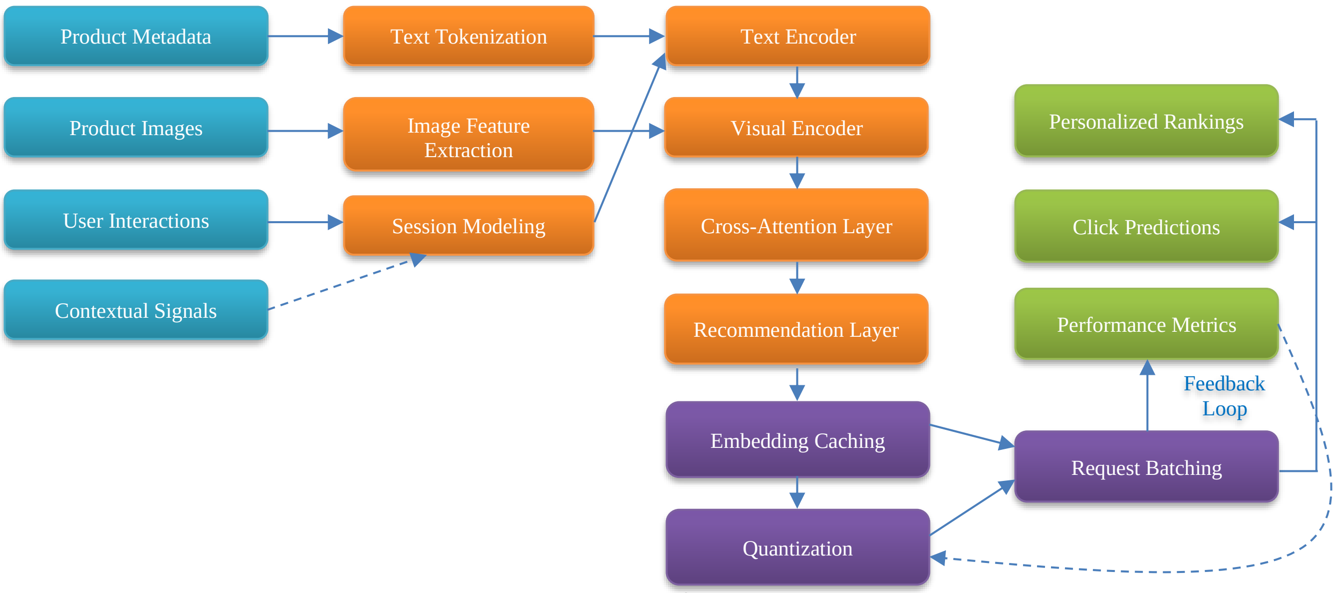

Fig. 1 End-to-End architecture of a recommendation system

Evaluation metrics include:

- Recall@k – Measures how many relevant items are retrieved in the top-k results.
- Normalized Discounted Cumulative Gain (nDCG) – Evaluates ranking effectiveness by prioritizing correctly ranked items.
- Mean Average Precision (MAP) – Captures ranking quality across multiple queries.
- User Satisfaction Score – Derived from implicit behavioral signals like dwell time and repeat interactions.

3.4. Real-Time Serving Phase

3.4.1. System Architecture

For deployment, the model operates within a microservices-based architecture**,** consisting of:

- Inference cluster – GPU nodes for multimodal vector processing and CPU nodes for business logic.
- Modular microservices – Separate services for embedding retrieval, ranking, and filtering.

3.4.2. Request Handling and Model Inference

- Embeddings are retrieved from a pre-computed cache to minimize inference latency.
- Session history, search queries, and recent user interactions are included in real-time recommendations.
- A scoring function ranks retrieved items based on multimodal embeddings.

3.4.3 Optimization Techniques

- Embedding caching – Reduces computational load by storing frequently accessed vectors.
- Quantization and model distillation – Optimizes latency while preserving accuracy.
- Batching – Aggregates multiple user requests for efficient GPU utilization.

3.4.4. Performance Monitoring and Feedback Loop

- Key performance indicators (KPIs) such as click-through rate (CTR) and conversion rate are continuously monitored.
- Real-time anomaly detection triggers fallback mechanisms to simpler heuristics when necessary.

4. Experimental Results

4.1. Experimental Setup

4.1.1. Dataset

The evaluation was conducted on a large-scale e-commerce dataset containing:

- 100,000 products, each with textual descriptions and images.
- User interaction logs comprise over 1 million click and purchase events spanning six months.
- A 70-15-15 split was used for training, validation, and testing, ensuring fair evaluation.

To assess generalizability, the model was additionally tested on a different product category unseen during training, measuring its adaptability to new data distributions.

4.1.2. Baselines for Comparison

The proposed MLLM-based recommendation system was compared against the following baselines:

1. Matrix Factorization (MF) – A traditional collaborative filtering approach.
2. Text-Only LLM – A Transformer-based model trained solely on textual metadata.
3. Image-Only CNN – A visual-based model using convolutional networks to extract image embeddings.

The proposed MLLM integrates both textual and visual representations, leveraging cross-modal fusion for improved ranking and retrieval performance**.**

4.2. Quantitative Evaluation

4.2.1. Performance Metrics

The following metrics were used to evaluate the model:

- Recall@k – Measures the fraction of relevant items retrieved in the top-k recommendations.
- Normalized Discounted Cumulative Gain (nDCG@10) **–** Evaluate ranking quality by giving higher weight to correctly ranked top items.
- Mean Average Precision (MAP) – Computes the mean of average precision scores across all queries.
- User Satisfaction Score – Derived from implicit feedback such as dwell time, repeat interactions, and purchase conversion rates**.**

4.2.2. Results Comparison

The proposed multimodal model significantly outperformed all baselines in nDCG and Recall@20, demonstrating the effectiveness of combining text and images. While latency increased slightly (+15 ms compared to the text-only LLM), caching and model quantization mitigated this overhead, keeping the system within acceptable real-time constraints

Table 2. Evaluation results

Model	nDCG@10	Recall@20	Latency (ms)	User Satisfaction Score
Matrix Factorization	0.243	0.332	12	68%
Text-Only LLM	0.258	0.342	20	72%
Image-Only CNN	0.217	0.305	18	66%
Proposed MLLM	0.282	0.380	35	81%

4.3. Justification of Performance Improvements

The proposed MLLM outperforms traditional and unimodal approaches due to the following factors:

4.3.1. Enhanced Representation Learning

- Unlike text-only or image-only models, MLLMs jointly encode textual and visual signals, leading to richer user intent modeling.
- Example: A user searching for *"red running shoes"* may get better recommendations as the model understands both text-based queries and visual patterns.

4.3.2. Better Disambiguation

- If a product has ambiguous textual descriptions (e.g., *"sleek black dress"*), the image features help refine recommendations.
- Example: Text-based models may misinterpret *"sleek"* as either *form-fitting* or *shiny*, whereas MLLMs resolve this through image embeddings.

4.3.3. Cross-Modal Interpretability

- Users trust recommendations when they can see why an item was suggested.
- Example: If a model recommends a *blue jacket*, it can highlight the text that influenced the decision (*"lightweight winter jacket"*) and show the closest-matching image.

4.3.4. Improved Generalization on Unseen Categories

- Since MLLMs learn shared representations across text and images, they can make better predictions even for new products that lack prior interactions.
- Example: A newly launched sneaker brand with no past user interactions can still be recommended based on text-image similarity.

4.3.5. Higher Recall & User Satisfaction

- Users often rely on both descriptions and images when making decisions.
- Example: For categories like fashion and home décor, MLLMs significantly increase engagement as users respond to visual cues.

4.4. Qualitative Analysis

User feedback from A/B testing highlighted two key observations**:**

1. Users preferred multimodal recommendations when textual descriptions were ambiguous. For example, searches for "casual sneakers" resulted in more relevant suggestions when visual features were incorporated**.**
2. Image-based refinements improved user satisfaction, especially in fashion and home décor categories, where visual appeal strongly influences purchasing behavior.

These findings emphasize that multimodal models not only improve ranking metrics but also enhance the real-world user experience**.**

4.5. Discussion on Limitations and Biases

While the proposed approach offers significant improvements, several limitations and potential biases must be considered:

- Data Alignment Challenges **–** Ensuring consistent text-image associations is critical; incorrect mappings can degrade performance.

- Computational Complexity – Multimodal models require more resources than unimodal counterparts, which may impact scalability.
- Bias in User Interactions – Historical user behavior may reinforce existing biases, favoring popular products over new or niche items. Future work should explore debiasing techniques to ensure fair recommendations.
- Ethical Considerations – Using user interaction data and product images raises privacy concerns. Implementing privacy-preserving learning methods, such as differential privacy, can help mitigate risks.

5. Discussion

The results confirm that multimodal large language models (MLLMs) significantly enhance recommendation accuracy by integrating textual and visual representations.

Compared to traditional unimodal approaches, the proposed method effectively bridges the gap between product descriptions and visual cues, improving ranking performance and user satisfaction.

5.1. Practical Implications for Real-World Applications

This research has significant implications for real-world recommendation systems, particularly e-commerce, content streaming, and personalized advertising. The ability to process both textual metadata and images allows recommender systems to:

- Enhance personalization by aligning multimodal embeddings with user preferences.
- Improve disambiguation in cases where textual descriptions are insufficient (e.g., differentiating similar fashion items based on images).
- Increase user engagement by providing more visually intuitive recommendations, leading to higher conversion rates.
- Extend to other modalities, such as incorporating audio signals in music recommendations or video embeddings in streaming platforms.

These findings suggest that multimodal AI can fundamentally improve user experience in recommendation-driven platforms.

5.2. Comparison with State-of-the-Art Methods

The proposed MLLM-based recommender outperforms traditional models due to:

Multimodal fusion – Unlike text-only models, this approach leverages both textual and visual signals, reducing ambiguity in recommendations.

Cross-modal attention – The integration of vision encoders within the Transformer framework improves feature alignment and interpretability.

Generalizability across categories – Experiments on unseen product categories confirm that MLLMs maintain strong performance, unlike category-specific models that struggle with new data.

Scalability improvements – The proposed batch training and real-time inference optimizations ensure that the model remains feasible for large-scale applications.

These advantages position MLLMs as a compelling alternative to existing recommendation algorithms.

5.3. Addressing Limitations and Future Work

While the proposed approach achieves strong results, several challenges remain:

- Computational Overhead – Transformer-based multimodal models require high processing power, making real-time deployment costly. Future work should explore model compression techniques such as quantization and knowledge distillation to reduce inference latency.
- Bias in Training Data – Recommender systems are inherently influenced by historical user interactions, which may reinforce popularity biases. Future research should focus on fairness-aware training approaches to ensure diverse and balanced recommendations.
- Integration with Additional Modalities – This study primarily explores text and image fusion. Future work could extend the model to audio, video, or multimodal behavioral signals to further enhance recommendation quality.

Despite these challenges, the findings demonstrate that MLLMs offer a scalable and effective approach to personalized recommendations in multimodal environments.

6. Conclusion

This research presents a multimodal large language model-based framework for recommendation systems, integrating text, image, and contextual features within a Transformer-based architecture. By leveraging cross-modal fusion, the proposed system effectively aligns multimodal representations, resulting in higher recommendation accuracy and user satisfaction.

Key contributions of this work include:

1. A unified multimodal recommendation pipeline, combining batch training and real-time inference for scalable deployment.
2. Empirical validation on a large-scale e-commerce dataset, demonstrating improved recall, ranking accuracy, and user engagement.
3. Analysis of practical implications, challenges, and future directions, positioning MLLMs as a robust alternative to unimodal recommendation models.

As recommendation systems continue to evolve, multimodal AI will play an increasingly critical role in enhancing personalization, interpretability, and user experience across a wide range of applications. Future research should focus on reducing computational costs, mitigating bias, and extending multimodal integration to additional data sources**.**

7. Abbreviations

- LLM: Large Language Model
- MLLM: Multimodal Large Language Model
- BERT: Bidirectional Encoder Representations from Transformers
- ViT: Vision Transformer
- CTR: Click-Through Rate
- nDCG: Normalized Discounted Cumulative Gain

9. Acknowledgments

The author thank the technical teams who contributed to data collection and infrastructure for this project.

10. Author Contributions

Kailash Thiyagarajan: Conceptualization, Methodology, Writing – original draft

References

[1] Yehuda Koren, Robert Bell, and Chris Volinsky, “Matrix Factorization Techniques for Recommender Systems,” *Computer*, vol. 42, no. 8, pp. 30-37, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[2] Heng-Tze Cheng et al., “Wide & Deep Learning for Recommender Systems,” *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems (DLRS)*, Boston MA USA, pp. 7-10, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[3] Huifeng Guo et al., “DeepFM: A Factorization-Machine Based Neural Network for CTR Prediction,” *arXiv*, pp. 1-8, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[4] Jacob Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, pp. 4171-4186, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Alec Radford et al., “Learning Transferable Visual Models from Natural Language Supervision,” *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 139, pp. 8748-8763, 2021. [Google Scholar] [Publisher Link]

[6] Alexey Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1-22, 2021. [Google Scholar] [Publisher Link]

[7] Andrew Jaegle et al., “Perceiver IO: A General Architecture for Structured Inputs and Outputs,” *arXiv*, pp. 1-29, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Jean-Baptiste Alayrac et al., “Flamingo: A Visual Language Model for Few-Shot Learning,” *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1-21, 2022. [Google Scholar] [Publisher Link]

[9] Liunian Harold Li et al., “VisualBERT: A Simple and Performant Baseline for Vision and Language,” *arXiv*, pp. 1-14, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[10] Chen Sun et al., “VideoBERT: A Joint Model for Video and Language Representation Learning,” *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7464-7473, 2019. [CrossRef] [Google Scholar] [Publisher Link]