Graduation Thesis:
Exploring the Boundaries of Unsupervised Learning for Fashion-MNIST via Momentum Constrast and Cluster-centric Finetuning

Zhiqi (ZKade) Liang

Aberdeen Institute of Data Science & Artificial Intelligence,

South China Normal University

2025.4

1 Introduction

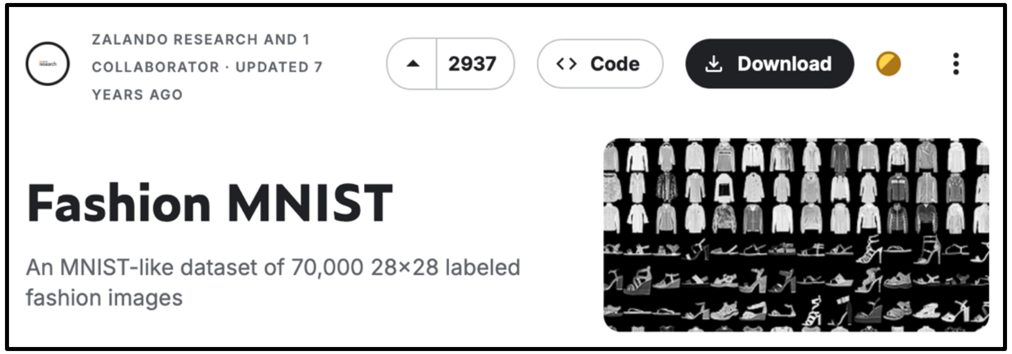
✓ **Fashion-MNIST**:

- Challenging benchmark in machine learning, e.g., diverse clothing categories (10), intricate textural patterns, and subtle inter-class variations.
- Supervised learning approaches have shown success, e.g., image classification.

✓ **Unsupervised Learning (UL) *vs.* Supervised Learning (SL)**:

- Weakness of SL: high annotation costs, inability to scale to novel classes, and poor generalization to complex tasks beyond classification.
- Advantages of UL: discover latent patterns and structures without annotation, better generalization to multiple classes and tasks.

How do UL approaches perform in Fashion-MNIST, especially clustering and data projection?



Rank	Model	Percentage error	Accuracy	Trainable Parameters	NMI	Power consumption	Paper	Code	Result	Year	Tags
1	pFedBreD_ns_mg		99.06				Personalized Federated Learning with Hidden Information on Personalized Prior			2022	
2	LR-Net		95.03				LR-Net: A Block-based Convolutional Neural Network for Low-Resolution Image Classification			2022	
3	Inception v3	5.56	94.44				CNN Filter DB: An Empirical Investigation of Trained Convolutional Filters			2022	

2020 2021 2022 2023 2024 2025

Other models · Models with highest Accuracy

2 Existing Work

✓ Naïve Solution: PCA + K-Means

- Principal Component Analysis (PCA) [1]: reduce image dimension, while capturing directions of greatest variance.
- K-Means [2]: cluster in the lower-dimensional space without supervised signal.
- **Problem**: *curse of dimensionality* (distance-based measurement, high computational cost), non-learnability (linearity assumptions, pre-defined rather than data driven optimzation, ill-suited).

✓ Deep Learning Era:

- E.g., Self Organizing Maps (SOM) [3], Deep Embedded Clustering (DEC) [4], Variational Deep Embedding (VaDE) [5], etc.
- **Problem**: intractable to train due to complexity of loss design, optimization stability issues; lack generalization.

✓ Self-Supervised Learning (SSL) Era:

- Train a *Generalist model* to learn robust representations of multi-modal data, e.g., GPT [6], MoCo [7], CLIP [8], etc.
- Improved trial: use **SSL**-based model to generate visual embeddings **instead of PCA** then cluster via K-Means.
- **Problem**: *static* representations are difficult to adapt to specific clustering tasks.

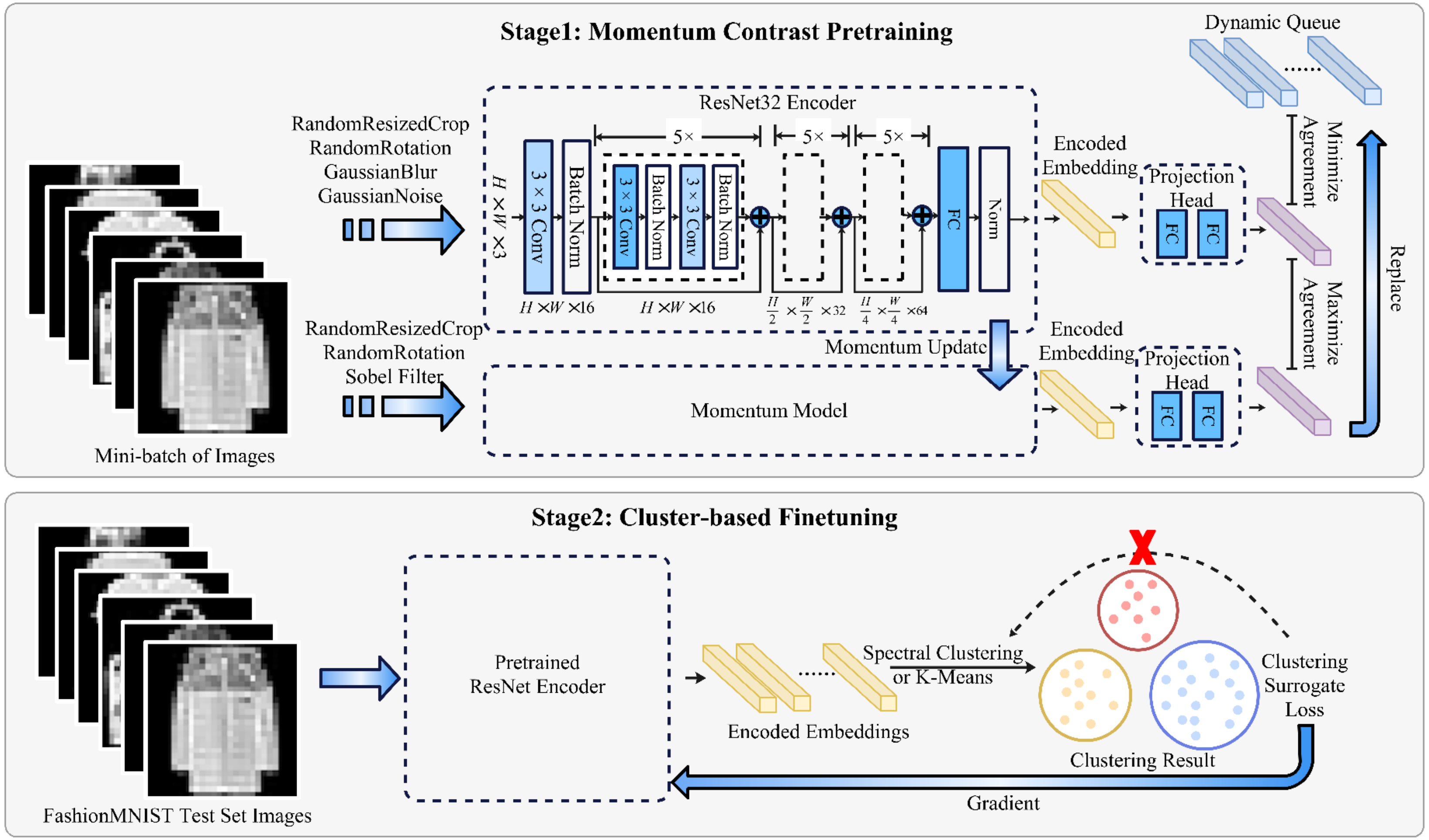
How to incorporate clustering into the optimization process while maintaining its generalization capability?

[1] H. Abdi and L. J. Williams, "Principal component analysis,", 2010.
[2] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding,", 2006.
[3] T. Kohonen, "The self-organizing map,", 1990.
[4] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis,", 2016.
[5] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: An unsu- pervised and generative approach to clustering,", 2016.
[6] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I, "Improving language understanding by generative pre-training, ", 2018.
[7] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning,", 2020.
[8] Radford, Alec, et al, "Learning transferable visual models from natural language supervision,", 2021.

3 Methodology - FashionMoCluster

✓ **A fully unsupervised two-stage framework**:

- ➢ Seamlessly integrates momentum contrast **pre-training** with clustering-based **fine-tuning**.
- ➢ Computationally efficient produces generalized, semantically coherent, and geometrically compact embeddings.



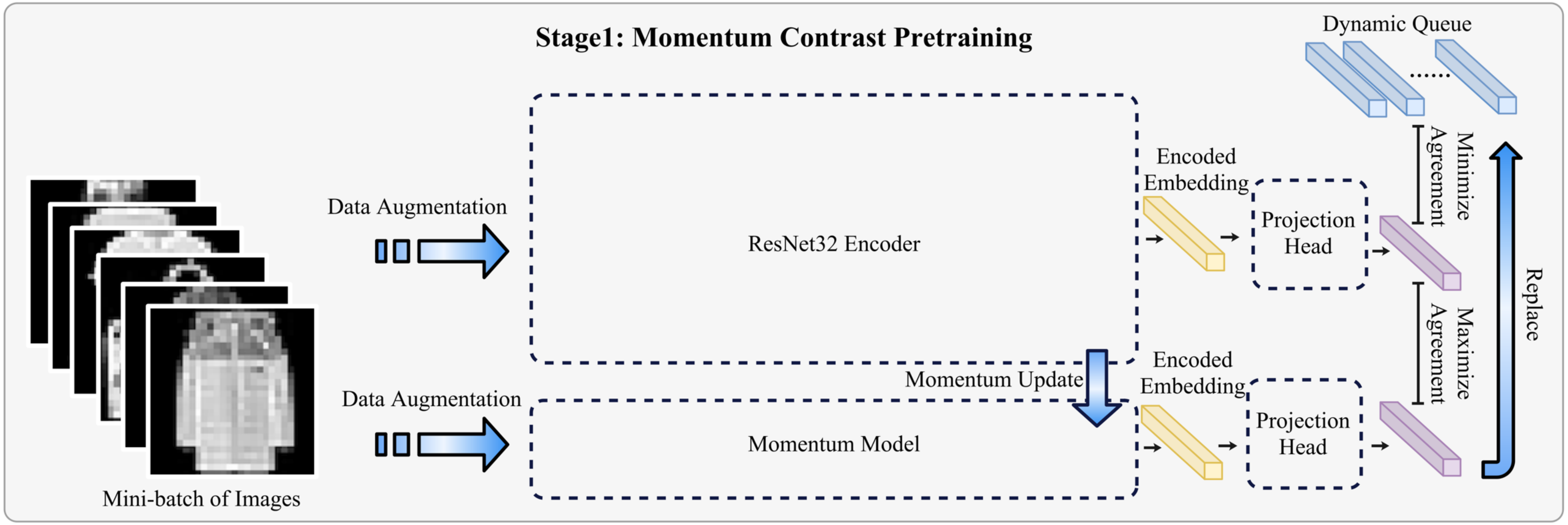
3 Methodology - Stage1: Momentum Contrast Pre-training

✓ Contrastive Learning (CL):

- For each image batch x, generate two **distinct augmentations** x^q (query) and x^k (key) and produce embeddings via query encoder f_q and key encoder f_k. Regard (x^q, x^{k_+}) as a positive pair and all other image x^{k_i} as negatives.
- Optimize with InfoNCE loss: $L_q = -\log \frac{\exp(f_q(x^q)\cdot f_k(x^{k^+})/\tau)}{\sum_{i=0}^{K}\exp(f_q(x^q)\cdot f_k(x^{k^i})/\tau)}$, *Maximize the agreement between positive sample, Minimize the agreement between negative samples*

✓ Instance Discrimination *vs.* Lable-instruct Contrastive Learning:

- "Label-instructed" contrastive approach forms positive pairs only **among samples sharing the same class label** and negatives from different classes, which **introduces supervision indirectly** and violates our goal of exploring purely UL.



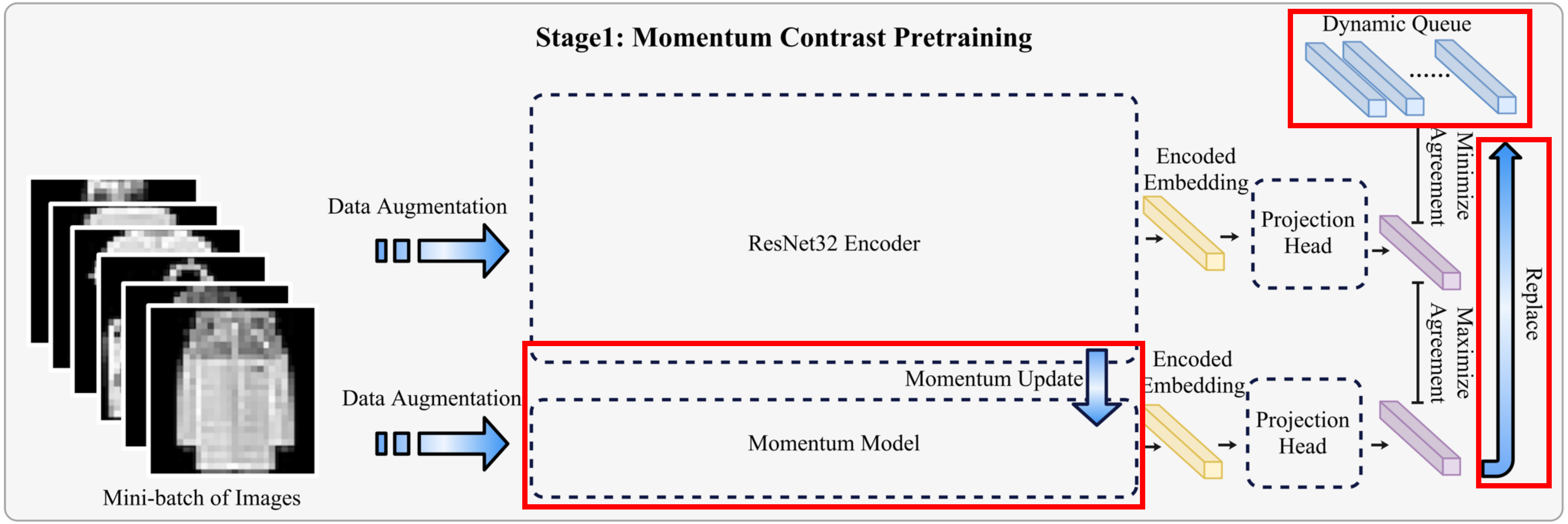
3 Methodology - Stage1: Momentum Contrast Pre-training

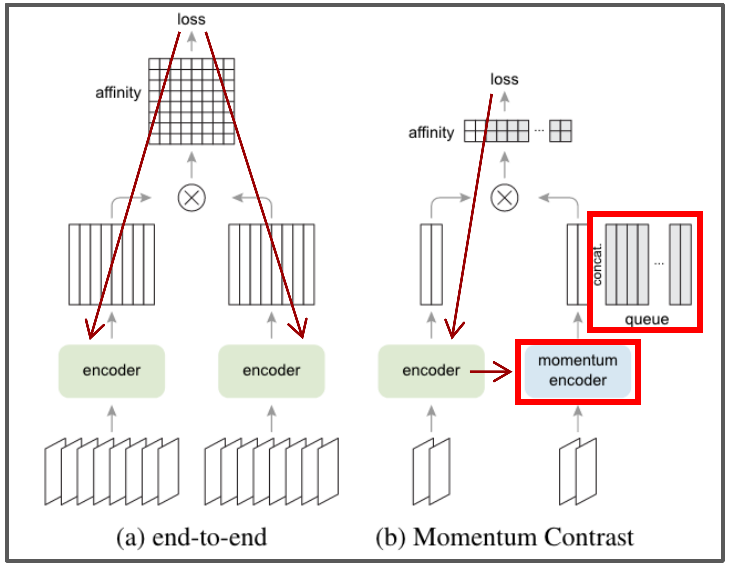
✓ Momentum Contrast:

- Dynamic queue → **Large Dictionary**:
  - Maintain a fixed-size queue (**8192**) as negative samples during contrastive learning.
  - Each iter, enqueue the encoded key embeddings drawn from recent mini-batches, dequeue the oldest.
- Momentum update → **Consistent Dictionary**: $\theta_k \leftarrow m\,\theta_k + (1-m)\,\theta_q$
  - Query encoder f_q is updated by backpropagation, while key encoder f_k is updated by exponential moving average (**0.99**) of f_q.

✓ MoCo *vs.* End-to-End Contrastive Learning:

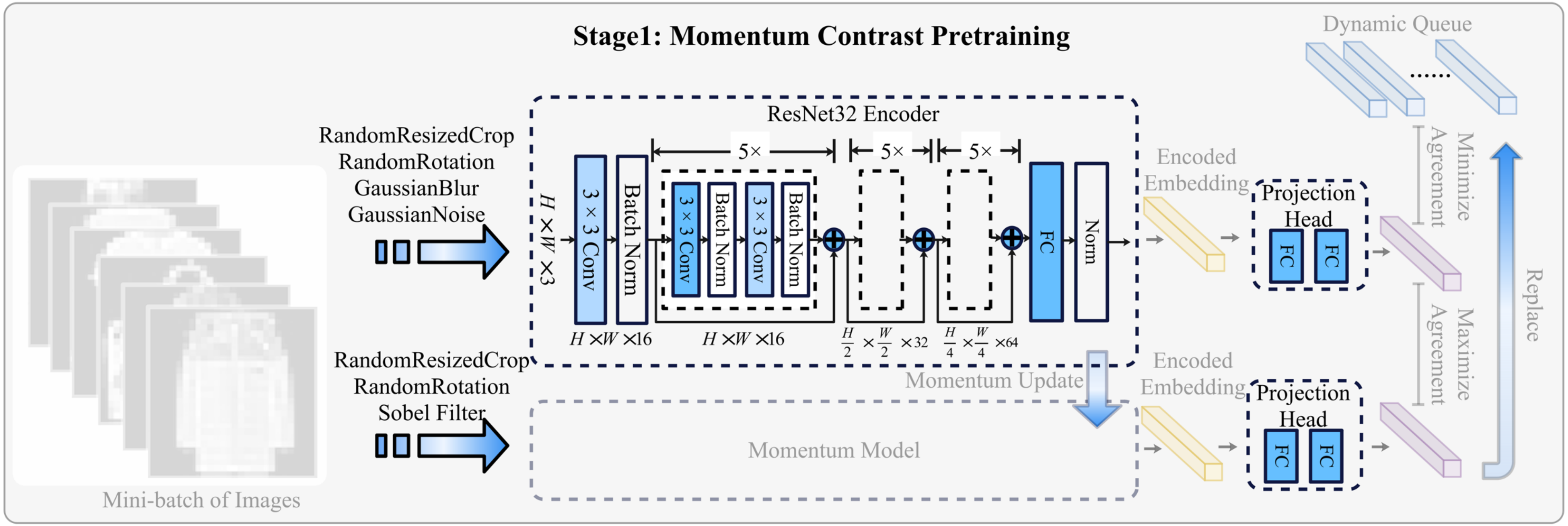
- Dictionary size: end-to-end is limited by batch size (constrained by GPU memory), while MoCo decouples them via queue.
- Train efficiency: end-to-end need to train both encoder, while only gradient of postive sample need to backward for MoCo.





3 Methodology - Stage1: Momentum Contrast Pre-training

- ✓ **ResNet32 Encoder**: Encode the original image (28×28×3) to embedding (128).
  - ➢ Residual connections combate the *vanishing gradient* problem.
- ✓ **Projection Head**: Implement with two-layer MLP head w/ ReLU.
  - ➢ Project output embedding to more focused feature space for CL tasks.
- ✓ **Dual-channel Augmentation**: Apply two different augmentations to query and key sample.
  - ➢ Introduce more randomness while forcing model learn underlying, invarient properties to help distinguish images, instead of irrelevant details (e.g., added noise) if apply same augmentation.



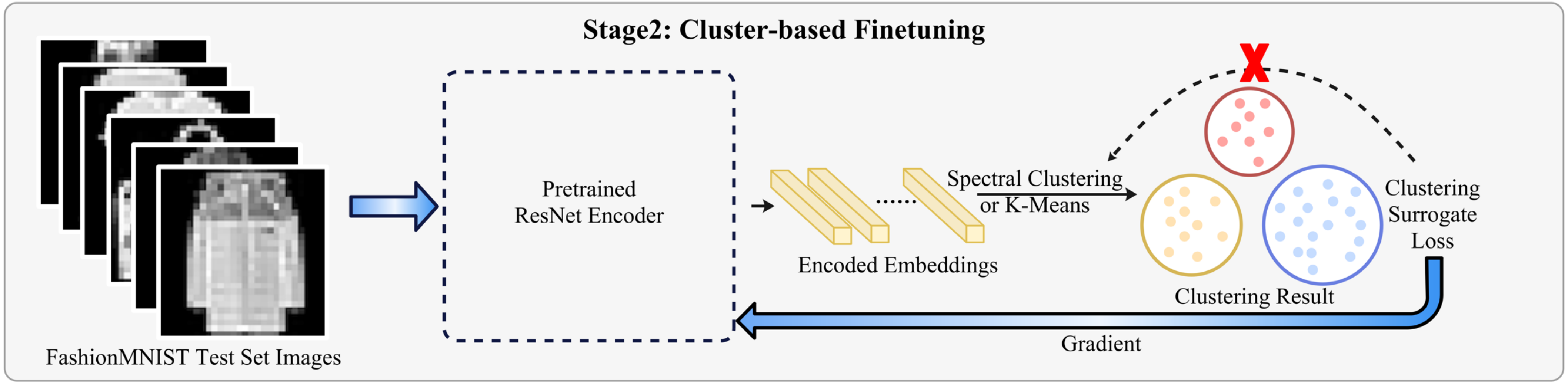
3 Methodology - Stage2: Cluster-based Fine-tuning

✓ **Unsupervised Clustering Surrogate Loss**:

- Silhouette Coefficient (Sil), Davies–Bouldin Index (DB), Calinski–Harabasz Index (CH): encourage higher within-cluster compactness and lower between-cluster seperation.
  - E.g., Sil: for each embedding z_i, compute the average intra-cluster distance (b_i) and minimum average distance (a_i) to any other cluster, then obtain the coefficient $\mathrm{sil}_i = \frac{b_i - a_i}{\max(a_i, b_i)}$, which ranges from -1 (clusters overlap) to +1 (well-separated clusters).
- Covert the metrics (e.g., neg, reciprocal, log) into loss terms and weighted sum: $L_{\text{cluster}} = \lambda_1 L_{\text{sil}} + \lambda_2 L_{\text{DB}} + \lambda_3 L_{\text{CH}}$

✓ **End-to-End Gradient Flow Management**:

- Generate embeddings by through pretrained ResNet encoder, retaining gradient on the encoder parameters.
- Apply K-Means or Spectral Clustering on detached embeddings to obtain cluster assignments, then compute the surrogate loss depending on the embeddings → gradients naturally flow back through and update the encoder.



4 Experiment

- **Experiment Design**: assesses FMC by answering
  - Q1: What is the effect of FMC compared to the standard machine learning models, the recent proposed self-supervised learning models and different composition paradigms? ← **Comparative Exp**
  - Q2: What is the contribution of each designed component to the overall performance of FMC? ← **Ablation Study**
  - Q3: Whether power-law scaling laws for model size and computation happens in our scenarios? ← **Ablation Study**
  - Q4: How to geometrically understand the impact of different models and training strategies on the embedding space? ← **Vis**
  - Q5: How inherent is FMC's actual groupings results of similar objects from the Fashion MNIST dataset? ← **Vis**
- **Evaluation Setup**:
  - Dataset: Fashion-MNIST test (10k images, 10 classes)
  - Hardware: NVIDIA RTX 4090, Intel Xeon CPUs
  - Metrics: NMI, ARI (supervised); silhouette, DB, CH (unsupervised)
  - Optimization: SGD, AMP, batch size 256, early stopping
  - FMC Training: Pretraining (lr=0.03, 300 epochs), Fine-tuning (lr=0.001, 100 epochs)

4 Experiment - Comparative Experiment (Q1)

- ✓ PCA: Marginal improvements, but reduces inference cost.
- ✓ Spectral Clustering (SC) after PCA: Boosts supervised metrics, weakens unsupervised.
- ✓ MoCo embeddings: Outperform end-to-end contrastive learning (performance, cost), but still lag behind.
  - ➢ Prioritizes local invariances (e.g., texture, grayscale, style invariances) among augmented views over global class separability.
- ✓ FMC (Pretraining-Finetuning): Outperforms all other approaches, achieving **best performance/cost trade-off** and maintaining training and inference efficiency. ←
  - ➢ Cluster-based fine-tuning refines pretrained manifold to align tightly with cluster structure.
  - ➢ Learnable centers fails due to **cluster collapse** ← substantial shifts in the encoder's output between epochs produce large, erratic gradients on the center embeddings, causing them to jump unpredictably.

Case	Method	NMI	ARI	Silhouette Coefficient	DBI	CHI
(a)	K-Means	0.4996	0.3414	0.1288	2.0691	0.6790
(b)	PCA+K-Means	0.5008	0.3429	0.1326	2.0242	0.6827
(c)	PCA+Spectral Clustering	**0.5956**	**0.4181**	0.0699	2.1916	0.6611
(d)	End-to-End+K-Means	0.4693	0.3076	0.0863	2.8426	0.6152
(e)	MoCo+K-Means	0.4997	0.3245	0.1021	2.5426	0.6461
(f)	MoCo+Spectral Clustering	0.5736	0.3592	0.0809	2.7012	0.6258
(g)	MoCo K-Means Jointtuning	0.5033	0.3296	0.3063	1.4118	0.8136
(h)	MoCo with K-Means Finetuning	0.5240	0.3582	**0.3619**	**0.9705**	**1.0304**
(i)	MoCo with Spectral Finetuning	0.5744	0.3810	0.3218	1.0568	0.7975
(j)	MoCo with Learnable Clustering Centers Finetuning	0.4276	0.2797	0.0724	2.9893	0.5972

Time and memory cost for a single training epoch

Train (per epoch)	End-to-End	MoCo	MoCo K-Means Joint Tuning	MoCo with Spectral Clustering Finetuning	MoCo with K-Means Finetuning
Time/s	78.33	13.66	18.86	12.94	6.39
Memory/MB	279.85	279.59	19138.59	18619.73	18601.32

Time and memory cost for Inference

Inference	K-Means	Spectral Clustering	PCA+K-Means	PCA+Spectral Clustering	MoCo K-Means
Time/s	15.10	9.79	5.57	7.93	10.96
Memory/MB	33.09	49.28	19.97	26.25	251.96

4 Experiment - Ablasion Study (Q2, Q3)

- ✓ Each component of MoCo **contributes meaningfully** to clustering accuracy, especially:
  - ➢ ResConn: Stabilize optimization, preserve spatial resolution for local features， e.g., neckline shape or sleeve length.
  - ➢ MLP projection head: Creates focused representation space with disentangled features.
  - ➢ Dual-channels augmentation: Forces model to learn underlying invariance through diverse perturbations.
- ✓ **Power scaling law emerges, but diminish returns** with larger model size and extended training for small-scale datasets, e.g., Fashion-MNIST test set.
  - ➢ Increasing model size (ResNet32): Slight performance decrease.
  - ➢ Extending training epochs: Best performance, but limited improvement ← ResNet20 may already saturate available information in Fashion-MNIST.

Table 4.4: Ablation of MoCo components with K-Means clustering performance. "ResConn" with a residual connection across two convolution blocks; "MLP": with an MLP projection head; "aug+": with dual-channels data augmentation; "cos": cosine learning rate schedule.

Case	Conv	ResConn	MLP	aug+	cos	size	epochs	NMI	ARI	Silhouette
(a)	✓	-	-	-	-	20	50	0.3900	0.2648	0.2266
(b)	✓	✓	-	-	-	20	50	0.4322	0.2992	0.2096
(c)	✓	✓	✓	-	-	20	50	0.4839	0.3582	0.0519
(d)	✓	✓	✓	✓	-	20	50	0.4905	0.3758	0.0326
(e)	✓	✓	✓	✓	✓	20	50	0.4969	0.3777	0.0381
(f)	✓	✓	✓	✓	✓	34	50	0.4782	0.3773	0.0709
(g)	✓	✓	✓	✓	✓	34	300	0.5075	0.3851	0.0921

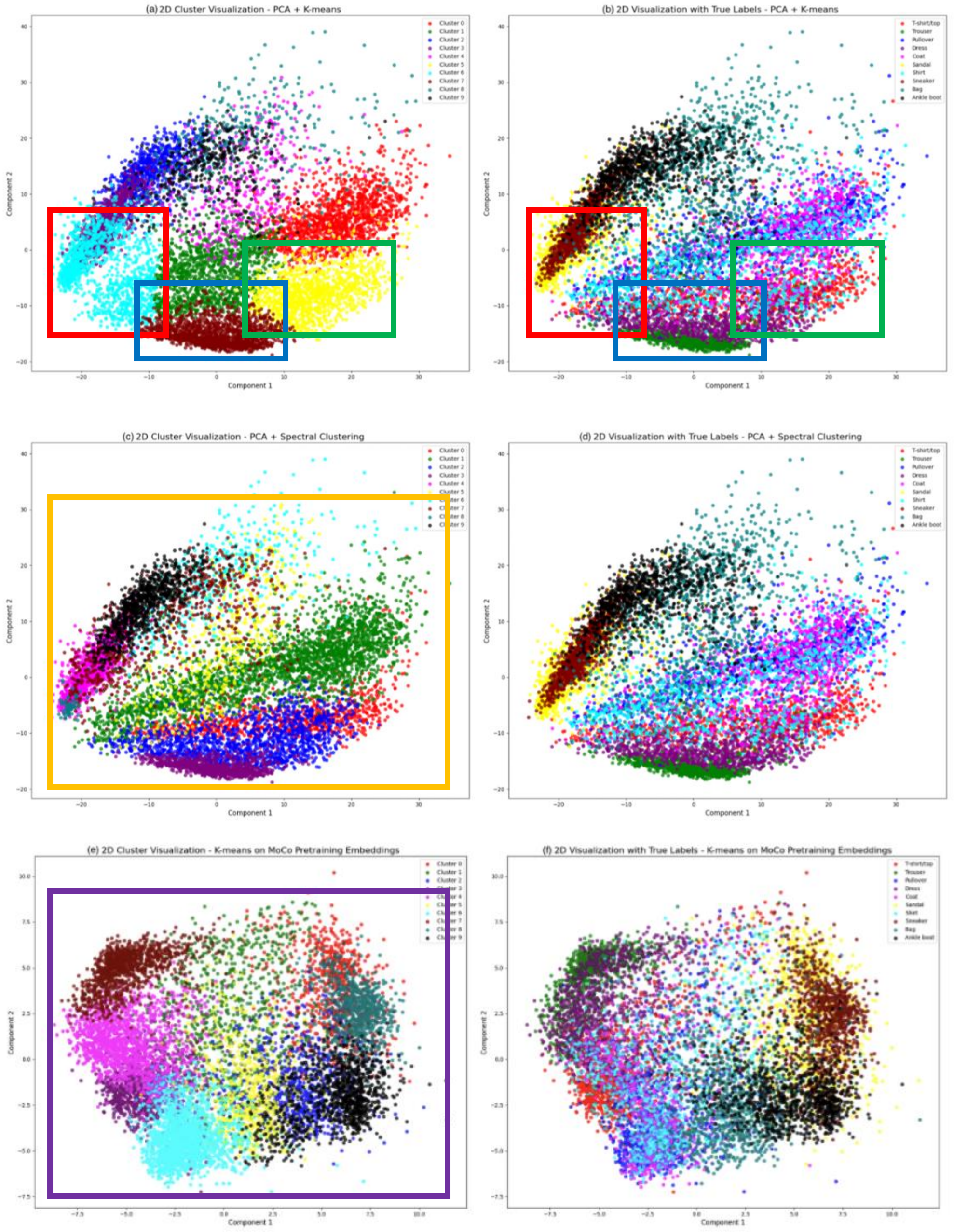
4 Experiment - 2D Embeddings Visualization (Q4)

✓ Distinctive clustering morphologies:

- **K-Means: Compact, convex, globular clusters** ← Objective function minimizes squared Euclidean distances to centroids.
- **Spectral Clustering: Non-convex, manifold-aligned clusters** ← Eigendecomposition of laplacian matrix preserves connectivity relationships rather than Euclidean proximity.
- Intutively observe the clustering performance (SC outperforms K-Means).

✓ Distinctive data projection morphologies:

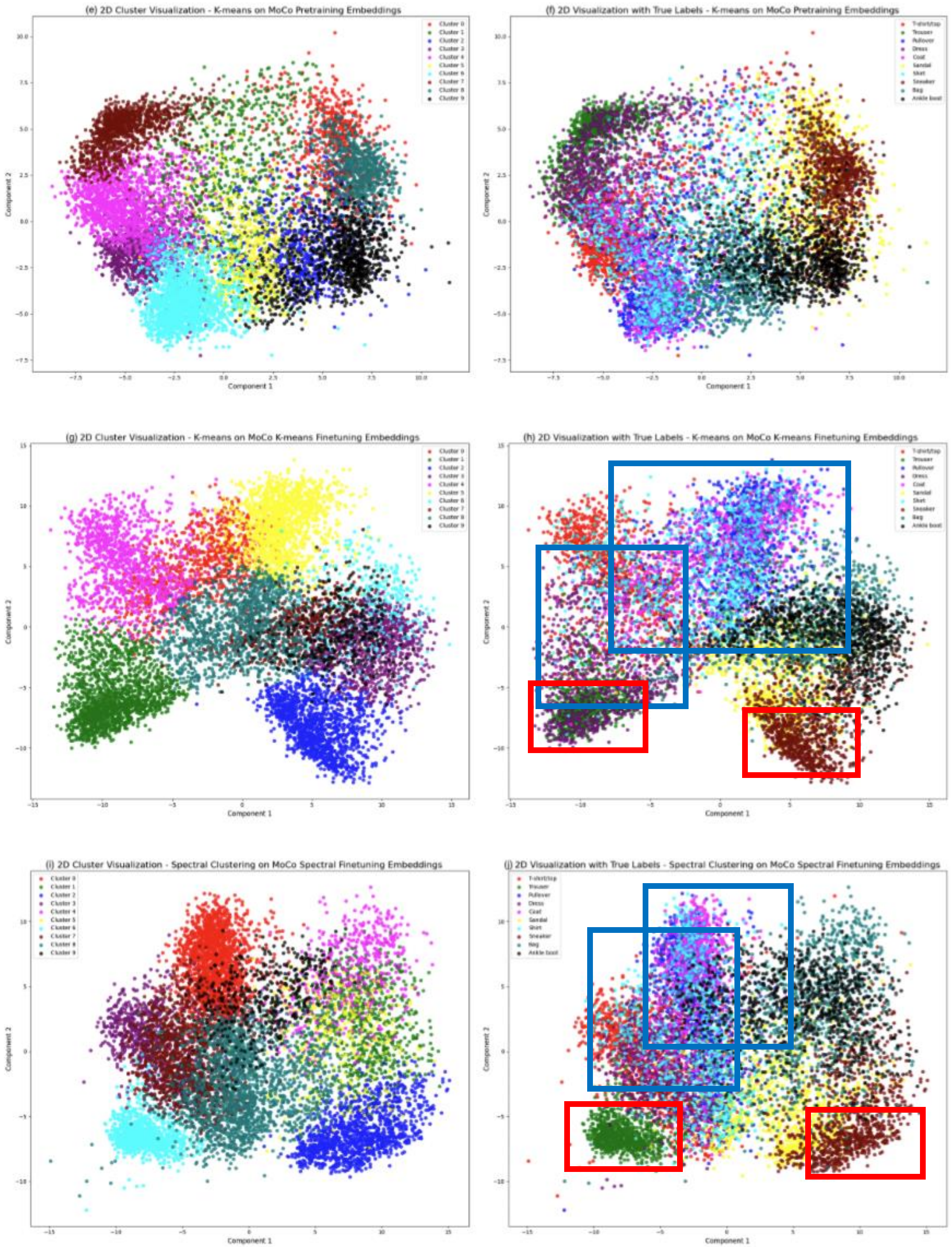
- **PCA embeddings**: Ground truth distributed **along continuous manifolds** with substantial overlap → PCA projects onto maximum variance directions without class separability.
- **MoCo embeddings**: **More distinct clusters** with improved semantic alignment → Loss function focuses low-level feature (e.g., texture details, material properties, local shape) instead of high-level feature for class discrimination.



4 Experiment - 2D Embeddings Visualization (Q4)

✓ **FMC embeddings**:

- ➢ **Visually distinct and semantically coherent clusters** ← Surrogate loss further improves intra-cluster compactness and inter-cluster separation (visually feel functioning process).
- ➢ Aligns with **unsupervised** metric improvements (higher within-cluster compactness and lower between-cluster seperation) seen in Q1**.**
- ➢ Remaining **Challenges**:
  - Boundary ambiguity between visually similar categories, e.g., sneakers *vs.* ankle boots, pullovers *vs.* coats.
  - Variable cluster density across classes:
    - ◆ Dense clusters: Categories with consistent visual patterns, e.g., trouser, sneaker.
    - ◆ Diffuse clusters: Categories with greater variability, e.g., dresses, coats.
  - Underutilized embedding regions suggests unbalanced exploitation of full embedding space.



4 Experiment - Class Distribution in Spectral Clusters Visualization (Q5)

- ✓ **PCA & FMC show moderate effectiveness (8/10 categories), aligning with supervised metrics in Q1.**
  - ➢ FMC achieves 92.2% purity for Ankle boots (+25% vs. PCA).
- ✓ **Consistently difficult in separating visually similar categories, e.g., Pullover, Coat, Shirt, Sandal ← Root causes:**
  - ➢ Silhouette Similarity:
    - ● e.g., Pullover & shirts are *"broader at the top and narrowing toward the bottom"*, sharing similar structural coontours. Some Sandal resemble those of upper-body garments after low-resolution grayscale compression.
  - ➢ Homogeneous Grayscale Intensity:
    - ● Most samples in confused cluster display medium gray levels with little black-white contrast, making it hard to classify via intensity-based discrimination.
  - ➢ Loss of Texture Information:
    - ● 28×28 resolution severely degrades fine texture details, which are useful to classify upper-body garments like Pullovers, Coats and Shirts.
  - ➢ Edge Cases Aggregation: tends to group borderline cases together.

Table 4.5: Result of Spectral-Clustering on PCA embeddings. Boldfaced numbers indicate the dominant classes (>50%) in each group. Both boldfaced and italic numbers indicate the confused classes (no dominant class & top2 or 3) in some groups.

Cluster_ID	Top (T-shirt/top)	Trsr (Trouser)	Pull (Pullover)	Dress	Coat	Sand (Sandal)	Shirt	Snkr (Sneaker)	Bag	Ankle (Ankle boot)
0	0.099	0.042	0.010	**0.638**	0.125	0.001	0.079	0.000	0.005	0.000
1	0.039	0.013	***0.360***	0.019	***0.302***	0.001	***0.244***	0.000	0.022	0.000
2	**0.743**	0.002	0.009	0.022	0.001	0.000	0.218	0.000	0.005	0.000
3	0.000	0.000	0.000	0.000	0.000	0.079	0.000	0.249	0.005	**0.667**
4	0.000	**0.997**	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	**0.984**	0.000	0.016	0.000	0.000
6	0.016	0.000	0.013	0.000	0.004	0.002	0.027	0.000	**0.938**	0.000
7	0.002	0.000	0.001	0.000	0.000	0.376	0.000	**0.593**	0.013	0.015
8	0.002	0.000	0.000	0.000	0.006	0.004	0.002	0.000	**0.986**	0.000
9	0.011	0.000	0.000	0.000	0.000	***0.479***	0.000	0.010	0.003	***0.498***

Table 4.6: Result of Spectral-Clustering on MoCo Pretraining Embeddings. Boldfaced numbers indicate the dominant classes (>50%) in each group. Both boldfaced and italic numbers indicate the confused classes (no dominant class & top2 or 3) in some groups.

Cluster_ID	Top (T-shirt/top)	Trsr (Trouser)	Pull (Pullover)	Dress	Coat	Sand (Sandal)	Shirt	Snkr (Sneaker)	Bag	Ankle (Ankle boot)
0	0.000	0.000	0.000	0.000	0.000	***0.284***	0.000	***0.265***	0.009	***0.442***
1	0.037	0.007	***0.366***	0.009	0.289	0.004	***0.257***	0.000	0.029	0.000
2	**0.765**	0.000	0.008	0.010	0.001	0.000	0.217	0.000	0.000	0.000
3	0.003	0.000	0.001	0.000	0.000	0.351	0.002	**0.610**	0.028	0.006
4	0.012	0.000	0.007	0.000	0.005	0.016	0.009	0.000	**0.950**	0.000
5	0.001	0.000	0.000	0.000	0.000	0.335	0.001	0.009	0.003	**0.651**
6	0.007	0.000	0.007	0.000	0.000	0.007	0.010	0.000	**0.967**	0.003
7	0.000	**0.997**	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	***0.114***	0.065	0.062	***0.439***	***0.189***	0.003	0.104	0.000	0.025	0.000
9	0.023	**0.640**	0.010	0.301	0.006	0.003	0.012	0.000	0.006	0.000

Table 4.7: Result of Spectral-Clustering on MoCo Spectral Finetuning Embeddings. Boldfaced numbers indicate the dominant classes (>50%) in each group. Both boldfaced and italic numbers indicate the confused classes (no dominant class & top 2 or 3) in some groups.

Cluster_ID	Top (T-shirt/top)	Trsr (Trouser)	Pull (Pullover)	Dress	Coat	Sand (Sandal)	Shirt	Snkr (Sneaker)	Bag	Ankle (Ankle boot)
0	0.017	0.002	***0.351***	0.014	***0.407***	0.000	0.187	0.000	0.022	0.000
1	0.000	0.000	0.000	0.000	0.000	0.067	0.000	0.278	0.004	**0.652**
2	0.002	0.000	0.001	0.000	0.000	0.367	0.000	**0.604**	0.015	0.011
3	**0.778**	0.000	0.009	0.018	0.000	0.000	0.194	0.000	0.000	0.000
4	0.010	0.000	0.002	0.004	0.008	0.004	0.002	0.000	**0.970**	0.000
5	0.008	0.000	0.008	0.000	0.000	0.003	0.025	0.000	**0.956**	0.000
6	0.001	**0.982**	0.001	0.013	0.002	0.000	0.000	0.000	0.001	0.000
7	0.105	0.032	0.008	**0.666**	0.112	0.000	0.075	0.000	0.002	0.000
8	0.118	0.025	***0.187***	0.058	0.073	***0.235***	***0.216***	0.014	0.055	0.019
9	0.000	0.000	0.000	0.000	0.000	0.071	0.000	0.007	0.000	***0.992***

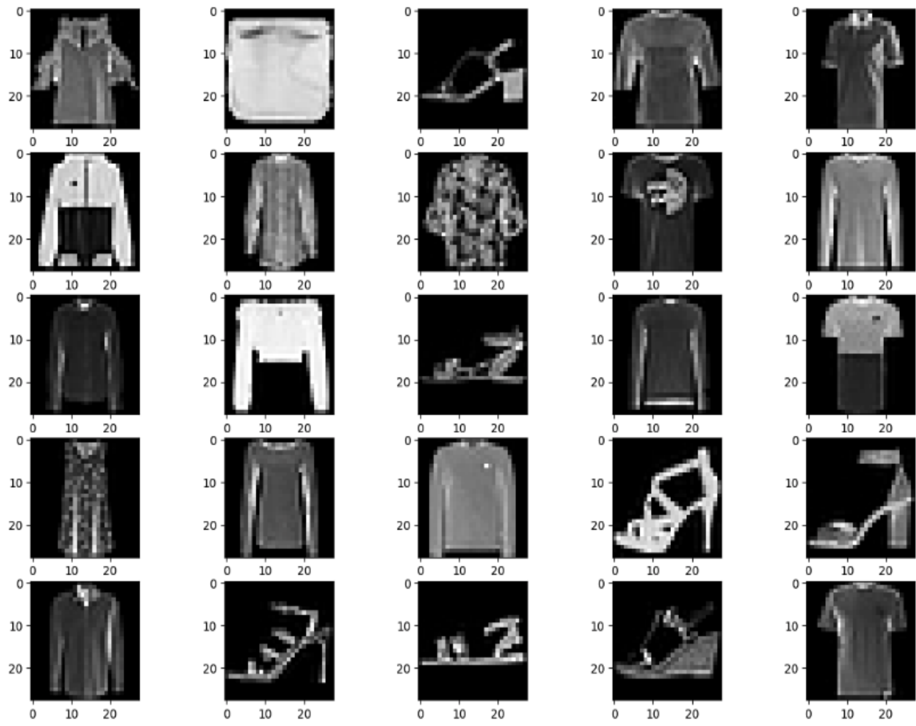
4 Experiment - Class Distribution in Spectral Clusters Visualization (Q5)

✓ **PCA & FMC show moderate effectiveness (8/10 categories), aligning with supervised metrics in Q1.**

- FMC achieves 92.2% purity for Ankle boots (+25% vs. PCA).

✓ **Consistently difficult in separating visually similar categories, e.g., Pullover, Coat, Shirt, Sandal ← Root causes:**

- Silhouette Similarity:
  - e.g., Pullover & shirts are *"broader at the top and narrowing toward the bottom"*, sharing similar structural coontours. Some Sandal resemble those of upper-body garments after low-resolution grayscale compression.
- Homogeneous Grayscale Intensity:
  - Most samples in confused cluster display medium gray levels with little black-white contrast, making it hard to classify via intensity-based discrimination.
- Loss of Texture Information:
  - 28×28 resolution severely degrades fine texture details, which are useful to classify upper-body garments like Pullovers, Coats and Shirts.
- Edge Cases Aggregation: tends to group borderline cases together.



5 Discussion & Future Direction

✓ Supervised Learning (SL) vs. Unsupervised Learning (UL) for Clustering:

- Problem Summary: UL exhibit consistent difficulty in separating **visually similar** categories (lack of semantic awareness).
- Fundamentally, UL models data distributions or underlying structural properties which generally and stably exists in the open physical world →
  - Advantages: uncover novel, non-obvious visual patterns beyond conventional semantic categories, which has better transferablility for multiple prediction tasks and datasets.
  - Disadvantages: results in diverse, semantically misaligned gradient flows, making task/metric-oriented optimization harder.
- In contrast, SL provides clear gradient directions driven by the objective of minimizing prediction–label differences, which stabilize the training process toward the objective.

✓ Introducing (semi-) supervised signal and advanced components / datasets:

- Integrate **semantic priors we discovered or textual priors** to the training process to further reduce atypical misgroupings.
- Introduce **weak supervision** (small number of labeled examples, hierarchical class information).
  - I do believe the introduction of supervised signal will provide more direct and stable gradient to those learnable centers, and push them to more discriminative positions in space, enabling fully regions utilization.
  - More regularization techniques (e.g., entropy penalities, repulsion losses) can be used to prevent cluster collapse.
- Replace the components with more advanced SSL (DINOv2, VICReg) and Clustering methods (DeepCluster, SwAV).
- Extending FMC to larger dataset (Fashion-MNIST train, mixed), or more realistic, high-resolution datasets (e.g., DeepFashion, ModaNet, Street2Shop) ← Data Hungry Properties.

6 Highlight & Conclusion

✓ We propose FMC, a purely unsupervised two-stage vision learning framework that efficiently learns robust and transferable fashion representations by combining Momentum Contrast with cluster-centric fine-tuning, empowering a wide range of downstream fashion applications and supporting future extension with more advanced mechanisms.

✓ We conduct a comprehensive evaluation of classical and state-of-the-art self-supervised clustering methods under fair experimental settings, revealing the performance boundaries of unsupervised learning on Fashion-MNIST, observing power-law scaling behavior, and offering intuitive insights into the functioning of both traditional algorithms and the proposed components in FMC, thereby establishing a strong foundation for the future advancement of unsupervised fashion representation learning.

Thanks for Listening!

Any comments or critiques are appreciated.

Zhiqi (ZKade) Liang

2025.4

Paper & Code Available: https://github.com/ZhiqiLiang/FashionMoCluster

Free to Connect: zkadelive1101@gmail.com