Hystar: Hypernetwork-driven Style-adaptive Retrieval via Dynamic SVD Modulation

Yujia Cai[1] Boxuan Li[1] Chenghao Xu[2] Jiexi Yan[1]

[1] School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China

[2] School of Electronic Engineering, Xidian University, Xi'an, Shaanxi, China

{yujiacai, boxuanli ,chx}@stu.xidian.edu.cn, jxyan1995@gmail.com

Introduction & Motivation

Query-based image retrieval must operate under highly diverse query styles, including sketches, artworks, low-resolution images, and textual descriptions. While large vision–language representation models exhibit strong generalization ability, their retrieval performance often deteriorates when query distributions deviate from those observed during pretraining, primarily due to style-induced discrepancies in visual statistics and semantic abstraction. Although parameter-efficient fine-tuning methods provide a computationally attractive adaptation strategy, most existing approaches rely on static parameter mappings shared across inputs, which inherently limits their ability to capture instance-specific style variations. Consequently, retrieval systems remain vulnerable to unseen styles and cross-style semantic confusion, motivating the need for a framework that simultaneously achieves adaptability, stability, and parameter efficiency.

Methods

To address these challenges, we introduce Hystar, a dynamic multi-style retrieval framework that reformulates model adaptation as spectral modulation in the singular value space. Rather than predicting full weight updates or low-rank matrices, Hystar preserves the pretrained semantic subspaces while introducing structured singular value perturbations, thereby enabling geometry-consistent adaptation with reduced optimization instability. Specifically, dynamic singular value increments are generated through a lightweight hypernetwork conditioned on style-aware embeddings, which allows the model to adjust its behavior in an input-dependent manner. In parallel, static learnable singular value offsets provide global cross-style calibration, ensuring that style flexibility does not compromise training stability. Through this dynamic–static decomposition, Hystar achieves a principled balance between per-query adaptability and shared representational robustness.

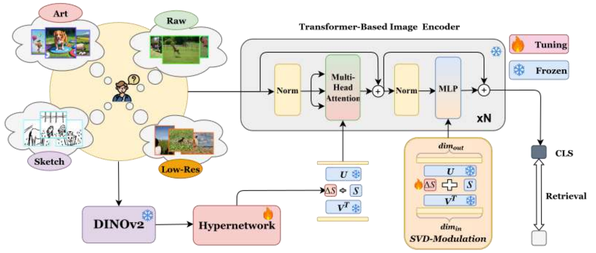

Figure 1: **Overview of the Hystar framework.** For multi-style queries, style features are first extracted using DINOv2. These features are fed into a hypernetwork to produce dynamic singular-value increments for attention layers, enabling style-conditioned modulation of the feature encoder. Additionally, static singular-value increments are applied to the MLP layers, serving as a fixed parameter modulation. Together, these mechanisms guide the encoder to produce style-diverse retrieval predictions.

Beyond architectural adaptation, we further observe that conventional contrastive learning objectives treat all negative samples uniformly, which is suboptimal in cross-style retrieval scenarios where semantic confusion is dominated by a small subset of hard negatives. To mitigate this limitation, we propose StyleNCE, an optimal-transport-weighted contrastive loss that reweights negative samples according to difficulty, thus emphasizing informative cross-style discrepancies while preventing gradients from being dominated by trivial negatives. By integrating difficulty-aware weighting with spectral modulation, the learning process more effectively captures style-dependent semantic relations.

StyleNCE extends the conventional InfoNCE objective by introducing difficulty-aware negative weighting. Instead of treating all negatives uniformly, the loss assigns adaptive importance coefficients ω_{ij} ,which emphasize hard cross-style negatives and mitigate the dominance of visually trivial samples. The balancing factor γ controls the relative contribution between positives and negatives.

$$\mathcal{L}_{\text{StyleNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(q_{i'}p_i)/\tau)}{\exp(\text{sim}(q_{i'}p_i)/\tau) + \gamma\sum_{j\neq i}\omega_{ij}\exp(\text{sim}(q_{i'}p_j)/\tau)}$$

To estimate the difficulty weights, we adopt an optimal transport formulation, where pairwise similarities define the transport cost and Sinkhorn iterations efficiently compute the assignment matrix. The resulting weights naturally amplify semantically confusing negatives while maintaining globally balanced contributions across the batch. This design improves cross-style discrimination and stabilizes retrieval learning under heterogeneous query distributions.

Quantitative Results

Method \ Query Style	Art		Sketch		Low-Res		Text	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CLIP	58.5	93.7	47.5	77.3	45.0	75.7	66.1	94.7
CLIP*	58.2	90.4	63.6	93.6	78.8	97.1	72.2	96.4
BLIP*	51.1	85.3	67.1	90.9	77.2	95.8	74.3	95.3
LoRA	63.8	96.5	72.8	96.5	79.7	95.1	70.4	97.1
VPT	66.7	96.5	73.3	97.0	81.4	96.0	69.9	96.1
$(IA)^3$	64.3	96.8	71.8	95.7	80.9	96.1	70.1	96.6
AdaptFormer	65.1	97.0	73.5	96.4	81.1	96.3	69.7	95.8
SSF	64.7	96.4	73.0	97.0	79.9	95.8	70.1	96.3
ImageBind	58.2	86.3	50.8	79.4	79.0	96.7	71.0	95.5
LanguageBind	67.5	92.9	63.6	89.1	78.6	94.5	79.7	98.1
FreestyleRet-CLIP	71.4	97.8	80.6	97.4	86.4	97.9	69.9	97.0
FreestyleRet-BLIP	74.5	97.4	81.2	97.1	90.5	98.5	81.6	99.2
Hystar-CLIP(Ours)	75.2	97.9	90.2	99.3	98.0	99.4	70.9	97.5
Hystar-BLIP(Ours)	**75.6**	**98.1**	**91.0**	**99.8**	**98.8**	**99.9**	**82.0**	**99.6**

Table 1: **Retrieval performance on the style-diverse QBIR task.** We evaluate Top-1 and Top-5 accuracy(%) on the DSR fine-grained benchmark. The two forms of our Hystar framework, Hystar-CLIP and Hystar-BLIP, outperform in multiple scenarios with different query styles compared with other baselines. Best results are highlighted in **bold**.

Method \ Query Style	Clipart		Sketch		Painting		Quickdraw		Infograph	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CLIP	60.9	77.0	49.1	67.0	59.2	75.2	9.1	15.8	41.2	**60.3**
LoRA	63.0	74.8	54.6	66.8	54.8	67.2	13.1	21.9	28.3	40.4
VPT	71.7	81.6	62.5	73.8	61.7	73.3	14.5	22.6	40.6	54.9
FreestyleRet	69.5	80.3	60.5	73.5	63.7	75.7	12.2	18.7	43.1	58.7
Hystar(Ours)	**75.7**	**86.4**	**65.8**	**78.1**	**65.5**	**78.3**	**19.0**	**29.9**	**43.7**	59.3

Table 2: **Zero-shot retrieval performance on unseen styles.** We evaluate Top-1 and Top-5 accuracy(%) on the DomainNet coarse-grained benchmark. Our proposed Hystar framework demonstrates strong performance in zero-shot category-level retrieval under unseen style conditions. Best results are highlighted in **bold**.

Method	Real	Clipart	Sketch	Painting	Quickdraw	Infograph
CLIP	82.4	72.8	64.9	68.1	14.6	53.0
LoRA	62.4	51.7	43.1	42.8	12.8	26.1
VPT	79.9	69.8	61.9	59.2	15.7	46.0
FreestyleRet	84.9	74.1	67.3	68.8	15.9	54.1
Hystar(Ours)	**85.7**	**79.5**	**71.2**	**71.0**	**22.9**	**54.6**

Table 3: **Zero-shot classification performance of different methods on DomainNet across six styles.** Accuracy (%) for each style is reported, with the best results highlighted in **bold**.

Method	Art	Sketch	Low-Res	Top-1 Avg
Triplet loss	65.6	71.9	89.5	75.7
InfoNCE loss	70.2	85.3	94.2	83.2
Circle loss	70.4	88.8	96.1	85.1
Triplet loss + Hard Negative Sampling	69.3	80.2	93.0	80.8
InfoNCE loss + Hard Negative Sampling	72.6	88.4	96.7	85.9
StyleNCE loss(Ours)	**75.2**	**90.2**	**98.0**	**87.8**

Table 4: Ablation study of different loss functions on the CLIP backbone.

Method	Parameters(M)	Additional Params (%)	Speed(ms)	Inference Time Increase (%)
CLIP	427	–	68	–
VPT	428	0.2	73	7.4
$(IA)^3$	427	0.1	71	2.9
AdaptFormer	429	0.5	74	8.8
FreestyleRet	476	11.5	96	41.2
Hystar(Ours)	442	3.5	108	58.8

Table 5: Computation comparison between our Hystar and representative baselines. For fairness, we only compare CLIP-based models; The percentages of additional parameters and inference-time increase are reported with respect to the CLIP baseline.

Interpretability & Visualization

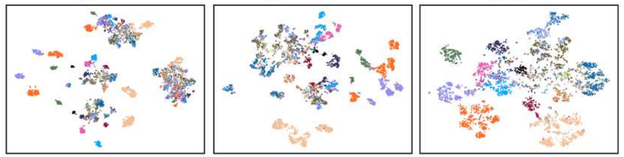

Figure 2: t-SNE visualization of feature embeddings derived by different methods on the DSR dataset. (a)**CLIP**: scattered, overlapping clusters. (b)**FreestyleRet**: more compact but some interclass entanglement. (c)**Hystar**: clearly separable, compact clusters, showing strong cross-style alignment.

Qualitative Results

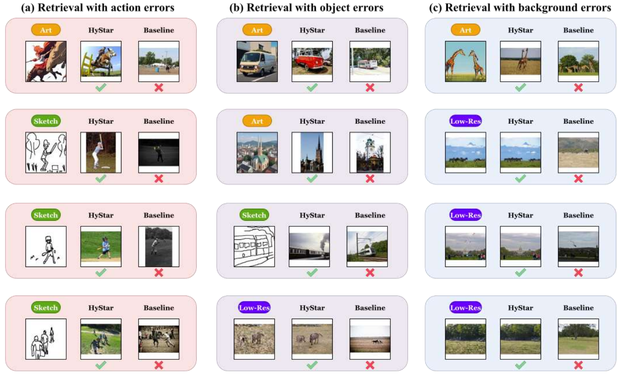

Figure 3: Qualitative retrieval examples on the DSR dataset. We illustrate three common error types made by baseline methods: (a) action errors, (b) object errors, and (c) background errors. Baselines often retrieve visually similar but semantically incorrect results, while our Hystar consistently retrieves the correct matches, highlighting its superior fine-grained alignment across multiple styles.

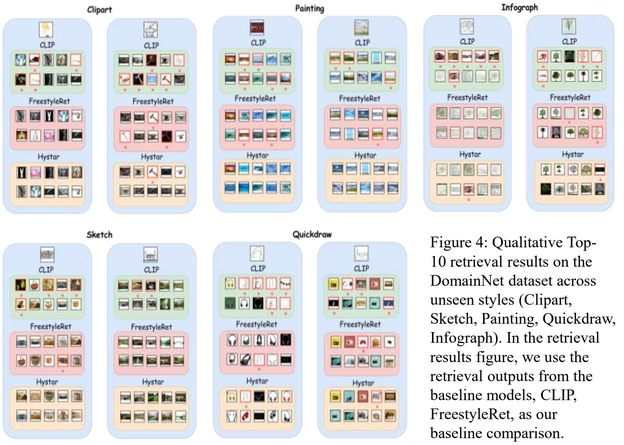

Figure 4: Qualitative Top-10 retrieval results on the DomainNet dataset across unseen styles (Clipart, Sketch, Painting, Quickdraw, Infograph). In the retrieval results figure, we use the retrieval outputs from the baseline models, CLIP, FreestyleRet, as our baseline comparison.