DIAGNOSING AND MITIGATING LATE-STAGE SHORTCUT EMERGENCE IN GENERATIVE SELF-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative self-supervised learning is commonly improved by extending pretraining, yet better in-distribution representations do not necessarily translate into uniformly better robustness. We study this discrepancy by tracking frozen representations throughout prolonged masked-image pretraining. Beyond controlled stylized shifts, we find that performance on several style-sensitive natural distribution shifts, including ImageNet-Sketch and a label-aligned subset of ImageNet-R, can peak substantially earlier than in-distribution accuracy and subsequently decline while in-distribution performance continues to improve. This behavior recurs across multiple settings but is shift-dependent, revealing a late-stage divergence between in-distribution fitting and transferable representation quality. Motivated by this observation, we formulate an operational invariance-aligned pretraining design principle: nuisance variation should not only be introduced at the input, but should also be paired with supervision that makes nuisance-dependent features unreliable for minimizing the pretraining objective. We instantiate this idea with HyGDL, which combines cross-perturbation self-distillation, low-rank geometric projection, and residual-conditioned reconstruction to encourage perturbation-stable representations. Across unsupervised domain generalization benchmarks, HyGDL improves robustness over generative SSL baselines and mitigates the late-stage degradation observed in standard pretraining. Representation-level probing further shows that the learned stable projection preserves substantially more semantic predictability than its orthogonal residual, while the residual remains non-trivially domain-predictive. Together, these results characterize a previously underexplored failure mode of prolonged generative pretraining and provide an operational approach for reducing shortcut-sensitive representation drift.

1 INTRODUCTION

Self-supervised learning (SSL) has become a central paradigm for learning transferable visual representations without human annotations. In particular, generative SSL based on masked image modeling has demonstrated strong representation learning capability and often continues to benefit from prolonged pre-training. This naturally encourages longer optimization whenever additional computation is available. However, the representation that performs best on the pre-training distribution need not be the representation that transfers best under distribution shift. This raises a basic but relatively underexplored question: *does representation robustness improve monotonically throughout generative self-supervised pre-training?*

We study this question from a learning-dynamics perspective by evaluating frozen representations throughout pre-training rather than only at the final checkpoint. Under controlled style shifts, we observe a clear divergence between in-distribution (ID) and out-of-distribution (OOD) behavior: while ID accuracy continues to improve or saturate, OOD performance reaches its optimum substantially earlier and subsequently declines. More importantly, a similar pattern also emerges on natural distribution shifts. As shown in Fig. 1a, we refer to this mismatch between the evolution of ID and transferable performance as **late-stage robustness divergence**.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

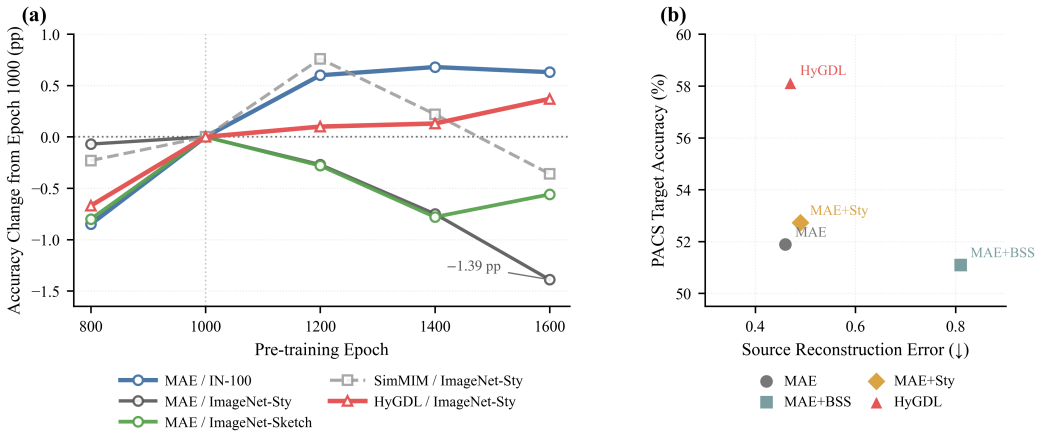

Figure 1: **Late-stage robustness divergence in generative self-supervised learning and its mitigation.** **(a)** Pre-training trajectories under controlled and natural distribution shifts. Across MAE and SimMIM, transfer performance on several shifts peaks before in-distribution representation quality, revealing a shift-dependent late-stage robustness divergence. HyGDL substantially changes this late-stage behavior and improves the stability of OOD transfer. **(b)** Relationship between source-domain reconstruction error and downstream OOD generalization on PACS. All methods are evaluated using the same masked pixel-reconstruction metric on the source training splits, and target accuracy is averaged over the four leave-one-domain-out evaluations. Lower reconstruction error does not consistently correspond to stronger target-domain transfer.

The key observation is therefore not simply that OOD accuracy is lower than ID accuracy, which is expected under distribution shift. Rather, ID and OOD performance can follow different *training trajectories* and reach their optimal checkpoints at different stages of pre-training. This rise-and-fall behavior recurs across multiple random seeds and with another generative SSL objective, SimMIM (Fig. 1a), although its magnitude varies across target distributions, as illustrated by the relatively stable ImageNet-R trajectory. Complementarily, Fig. 1b shows that, under the same PACS source splits and reconstruction metric, lower source reconstruction error across MAE-based variants does not consistently yield higher average target-domain accuracy. Thus, prolonged optimization can improve source-domain fitting while transferability under some distribution shifts stagnates or declines.

What may drive such divergence? A plausible explanation is that later-stage optimization increasingly exploits nuisance-sensitive visual statistics that remain useful for minimizing the pre-training objective but transfer poorly across appearance changes. This interpretation is closely related to shortcut learning, where models exploit statistically convenient but less transferable cues instead of relying exclusively on more stable structure. Importantly, the trajectory itself does not constitute a causal proof that a particular semantic or nuisance factor has been isolated. Rather, it provides a behavioral diagnostic: later representations can become increasingly specialized to the source distribution despite continued improvement under conventional ID evaluation. We use the term *late-stage shortcut emergence* when referring specifically to this nuisance-sensitive interpretation of the observed robustness divergence.

A seemingly direct remedy is to introduce stronger perturbations during generative pre-training. If appearance-related shortcuts contribute to the observed robustness divergence, exposing the model to larger style or frequency variations might be expected to improve invariance. However, modifying the training data or reconstruction target does not necessarily change which information remains useful for minimizing the objective. Indeed, in our downstream unsupervised domain generalization experiments, style- and Fourier-based perturbation variants improve robustness to different extents, but remain consistently weaker than explicitly aligning nuisance variation with representation-level invariance. This suggests that *exposure to nuisance variation and invariance to that variation are distinct learning objectives*.

This observation motivates an **invariance-aligned pre-training design principle**: nuisance variation should not only be introduced through the training data, but should also be paired with super-

vision that makes nuisance-dependent features unreliable for minimizing the pre-training objective. The goal is not to recover a universally correct semantic factorization. Instead, we seek an operational inductive bias that encourages representations to remain stable under a specified family of perturbations. This distinction is important because real-world nuisance factors can be entangled with semantically useful information, and no fixed perturbation distribution can perfectly separate "content" from "style" across all visual domains.

Guided by this principle, we introduce **HyGDL**, a hybrid generative–discriminative learning framework that explicitly couples nuisance perturbation with invariance-oriented supervision. Given an image and an online stylized counterpart, HyGDL first uses student–teacher self-distillation to identify low-rank representation directions that remain stable across the perturbation. The representation is then analytically decomposed into a *perturbation-stable projection* and an orthogonal residual. Rather than assuming that these components form a perfect semantic–style decomposition, we interpret them operationally: the stable projection captures variation consistently preserved across perturbations, whereas the residual contains information not explained by the stable subspace. A conditioned reconstruction objective then combines a masked student representation with residual-derived embeddings, thereby reducing the utility of nuisance-sensitive information within the shared representation pathway. This intervention changes the late-stage learning behavior observed with standard generative pre-training. On the controlled style-shift trajectory, MAE reaches its OOD peak around 1000 epochs and subsequently declines, whereas HyGDL maintains its OOD representation quality throughout continued training (Fig. 1a). HyGDL also consistently improves unsupervised domain generalization on PACS and DomainNet.

Taken together, these results suggest that evaluating only the final checkpoint or ID representation quality can obscure important changes in transferability during pre-training. More broadly, they motivate treating robustness not only as a property of a learned representation, but also as a quantity that evolves throughout optimization.

Our contributions are threefold:

- **We identify late-stage robustness divergence in generative SSL.** Transfer performance can peak earlier than ID performance under controlled and natural shifts, revealing a recurrent but distribution-dependent training dynamic.
- **We propose an invariance-aligned pre-training principle and instantiate it with HyGDL.** HyGDL combines cross-view alignment, low-rank geometric decomposition, and residual-conditioned reconstruction to reduce reliance on perturbation-sensitive information in the transfer pathway.
- **We validate the intervention behaviorally and at the representation level.** HyGDL mitigates late-stage OOD degradation, improves domain generalization, and yields a stable projection that preserves more semantic predictability than the orthogonal residual.

2 Related Work

Self-Supervised Learning and Shortcut Bias. Self-supervised learning (SSL) learns transferable representations through pretext objectives such as contrastive learning and masked image modeling (Grill et al., 2020; Caron et al., 2020; Chen et al., 2020a; He et al., 2022). Their learned invariances are largely shaped by the augmentations or reconstruction targets used during pre-training, and generative objectives may retain appearance-sensitive cues when they remain useful for reconstruction. Shortcut learning, including the well-known texture bias (Geirhos et al., 2019; 2020), has been extensively studied in supervised learning, but how robustness evolves throughout prolonged generative self-supervised pre-training remains less understood. Our work focuses on this training dynamic rather than only the final representation.

Invariant and Domain-Robust Representation Learning. Prior work improves robustness under domain shift through domain-aware architectures, reconstruction objectives, or transformations that reduce sensitivity to appearance variation (Yang et al., 2022; 2023; Scalbert et al., 2024). Rather than assuming an exact semantic–style factorization, we define invariance relative to a specified perturbation family and incorporate it directly into generative pre-training. HyGDL combines cross-

Table 1: **Late-stage representation dynamics and mitigation.** **(a)** Summary of the diagnosed dynamics across objectives and distribution shifts, where Δ_{post} denotes the change from the best checkpoint to epoch 1600. **(b)** ImageNet-Sty trajectories of MAE and HyGDL, reported as mean±standard deviation over three runs.

(a) Diagnostic summary

Method	Evaluation	Peak	Peak Acc.	Δ_{post}
MAE	ImageNet-100	1400	82.23±0.36	−0.05
MAE	ImageNet-Sty	1000	30.25±0.18	−1.39
SimMIM	ImageNet-Sty	1200	12.56±0.04	−1.12
MAE	ImageNet-Sketch	1000	28.19±0.13	−0.56
MAE	ImageNet-R_{overlap}	1000	36.73±0.22	−1.27
MAE	ImageNet-R	1400	25.03±0.14	−0.23

(b) ImageNet-Sty mitigation

Epoch	MAE	HyGDL
1000	30.25±0.18	30.14±0.17
1200	29.98±0.29	30.24±0.20
1400	29.50±0.54	30.27±0.37
1600	28.86±0.12	**30.51±0.17**

view alignment, low-rank geometric decomposition, and residual-conditioned reconstruction to reduce the utility of perturbation-sensitive information in the representation pathway used for transfer.

3 Late-Stage Robustness Divergence in Generative Self-Supervised Learning

Standard self-supervised evaluation typically focuses on the final checkpoint, implicitly assuming that representation quality improves monotonically with prolonged pre-training. We instead examine how ID and OOD representation quality evolve throughout optimization. To keep the main text focused, we report only the key trajectory statistics here; complete checkpoint-wise results, multi-seed statistics, dataset construction, and probing protocols are provided in the Appendix.

3.1 Diagnostic Protocol and Observation

We pre-train masked-image models on ImageNet-100 for 1600 epochs and evaluate frozen representations every 200 epochs using a fixed linear probing protocol. We first consider ImageNet-Sty, a controlled appearance shift that approximately preserves semantic content while altering style and texture statistics.

A clear mismatch emerges during MAE pre-training. ImageNet-Sty accuracy reaches its maximum of $30.25 \pm 0.18\%$ at 1000 epochs and subsequently decreases to $28.86 \pm 0.12\%$ at 1600 epochs. In contrast, ImageNet-100 accuracy continues to improve from 81.55% at 1000 epochs to 82.23% at 1400 epochs and remains essentially saturated thereafter. Thus, the checkpoint that maximizes transfer performance can occur substantially earlier than the one that maximizes ID representation quality. We refer to this mismatch as **late-stage robustness divergence**.

3.2 Reproducibility and Scope

The phenomenon is not specific to a single MAE run or objective. Across three runs, MAE consistently exhibits post-peak degradation on ImageNet-Sty. SimMIM shows a similar pattern on both ImageNet-Sty and ImageNet-Sketch, with the latter declining by 2.75 points after its epoch-1200 peak.

We further evaluate MAE on natural distribution shifts. As summarized in Table 1(a), ImageNet-Sketch and ImageNet-R_{overlap} both peak at 1000 epochs, earlier than the ID optimum at 1400 epochs, with post-peak decreases of 0.56 and 1.27 points, respectively. The full ImageNet-R evaluation, however, remains comparatively stable, with only a 0.23-point post-peak decrease. The effect is therefore recurrent across controlled and natural shifts, but its magnitude depends on the target distribution.

For reference, we also track HyGDL under the same controlled ImageNet-Sty protocol. Its accuracy evolves from $29.47 \pm 0.41\%$ at 800 epochs to $30.14 \pm 0.17\%$ at 1000, $30.24 \pm 0.20\%$ at 1200, $30.27 \pm 0.37\%$ at 1400, and $30.51 \pm 0.17\%$ at 1600 epochs. Unlike standard MAE, HyGDL does

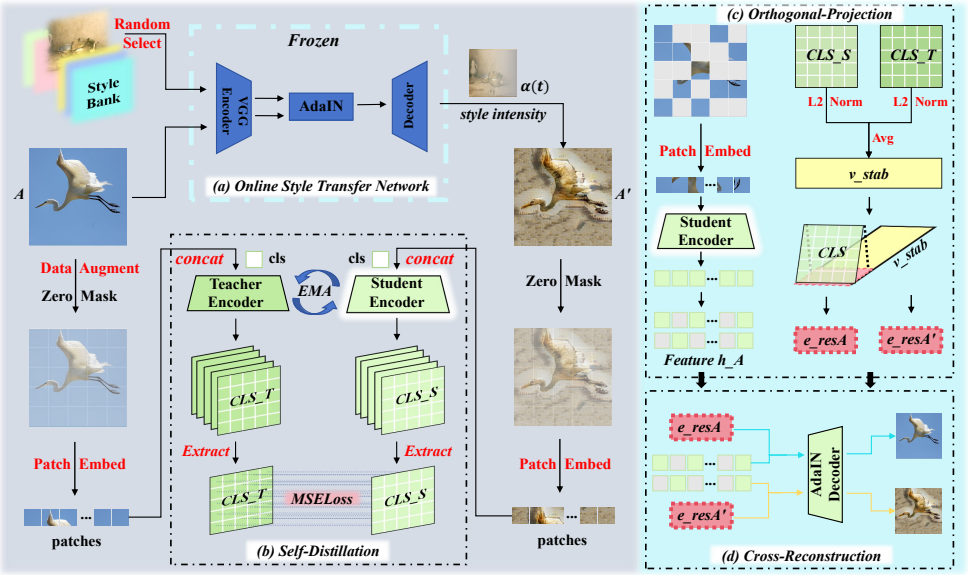

Figure 2: **Overview of HyGDL.** Cross-view self-distillation and low-rank geometric decomposition identify a perturbation-stable projection and its orthogonal residual from an image A and its perturbed view A'. A masked representation of A is then combined with residual information from either view to reconstruct the corresponding target.

not exhibit the same late-stage post-peak decline in this setting. We defer a detailed evaluation of this intervention to Sec. 5.

Importantly, these trajectories establish a behavioral phenomenon rather than a unique causal mechanism. We use *late-stage shortcut emergence* to describe the hypothesis that continued optimization increasingly exploits nuisance-sensitive statistics that remain useful on the source distribution but transfer poorly. The observation itself is the measurable divergence between ID and transfer trajectories; the shortcut interpretation motivates the invariance-aligned intervention introduced next.

4 Method: Invariance-Aligned Generative Pre-training

The analysis in Sec. 3 suggests that exposure to nuisance variation alone does not guarantee invariance: a generative objective may still benefit from encoding perturbation-dependent appearance cues for reconstruction. This motivates an **invariance-aligned pre-training design principle**: for a specified perturbation family, the representation used for transfer should remain stable across the perturbation, while perturbation-dependent information is represented separately when needed by the generative objective.

We instantiate this principle with **HyGDL**, a hybrid generative–discriminative learning framework comprising three coupled components: (i) online nuisance perturbation with cross-view self-distillation to promote perturbation-stable representations; (ii) low-rank geometric projection that decomposes the representation into a stable component and an orthogonal residual; and (iii) residual-conditioned reconstruction that encourages perturbation-dependent information to be carried outside the shared representation pathway. Figure 2 provides an overview.

4.1 Overall Framework

Given an image A, we generate an online perturbed view A' using a randomly sampled reference style and AdaIN. The framework only requires the perturbation to introduce diverse nuisance variation while approximately preserving semantic identity. Unmasked views of (A, A') are processed by the teacher and student encoders, f_{θ_t} and f_{θ_s}, respectively, where the teacher is updated by an exponential moving average (EMA) of the student. Their global representations provide the cross-view signal used to identify perturbation-stable representation directions.

In parallel, a masked version of A is encoded by the student for generative reconstruction. The resulting masked representation is combined with residual information derived from either A or A' to reconstruct the corresponding target. This yields two coupled pathways: a cross-view pathway for learning stable representation structure and a reconstruction pathway that retains perturbation-dependent information when required for generation. Figure 2 illustrates the overall framework.

4.2 Learning Perturbation-Stable Representations

Online nuisance randomization. For each image A, we sample a reference image online and transfer its appearance statistics to obtain a perturbed view A'. We use AdaIN as a lightweight perturbation operator (Fig. 2a), so the induced invariance should be understood relative to this perturbation distribution rather than to arbitrary real-world variation.

Cross-perturbation self-distillation. The student processes the perturbed view A', while the EMA teacher processes an independently augmented view of the original image A (Fig. 2b). Both branches are unmasked in this pathway, yielding global `[CLS]` representations

$$z_s = f_{\theta_s}(A'), \qquad z_t = f_{\theta_t}(A). \tag{1}$$

We align their normalized representations using

$$\mathcal{L}_{\text{distill}} = \left\| \frac{z_s}{\|z_s\|_2} - \text{sg}\left(\frac{z_t}{\|z_t\|_2} \right) \right\|_2^2, \tag{2}$$

which is equivalent, up to a constant factor, to maximizing cosine similarity. The teacher is updated by EMA, $\theta_t \leftarrow m\theta_t + (1-m)\theta_s$, and gradients are propagated only through the student. This objective operationally favors representation directions that remain predictable across the specified perturbation, without assuming that the aligned representation is purely semantic.

4.3 Low-Rank Geometric Decomposition

Cross-view alignment promotes stability but does not explicitly separate stable from perturbation-sensitive variation. We therefore introduce a low-rank geometric decomposition as an operational inductive bias, without assuming that the representation manifold is globally linear or that semantic and nuisance factors admit an exact linear factorization.

Let $\bar{Z}_s, \bar{Z}_t \in \mathbb{R}^{B\times d}$ denote the normalized student and teacher representations in a mini-batch. We compute their cross-view correlation matrix and its singular value decomposition,

$$C = \bar{Z}_t^\top \bar{Z}_s = U\Sigma V^\top, \tag{3}$$

and use the top-k left singular vectors to define the stable subspace, $V_{\text{stab}} = U_{[:,1:k]}$. We use $k = 2$ in the main experiments, and Fig. 2c illustrates the $k = 1$ case for clarity, while our formulation operates on a general k-dimensional stable subspace. For a teacher representation z_t, we then obtain

$$\begin{aligned} z_{\text{stab}} &= V_{\text{stab}} V_{\text{stab}}^\top z_t, \\ z_{\text{res}} &= z_t - z_{\text{stab}}. \end{aligned} \tag{4}$$

Thus, z_{stab} captures variation explained by the dominant cross-view-consistent directions, while z_{res} contains the orthogonal remainder. We do not interpret the latter as pure "style"; it is only expected to be relatively enriched in perturbation-sensitive information. This distinction is examined empirically in Sec. 5.5.

The subspace is estimated from each mini-batch and therefore evolves with the encoder during training. For reconstruction, the residual is mapped through a lightweight MLP,

$$e_{\text{res}} = g_{\text{res}}(z_{\text{res}}), \tag{5}$$

which supplies perturbation-dependent information to the decoder through a separate conditioning channel.

4.4 Residual-Conditioned Reconstruction

The geometric decomposition becomes useful only if the generative objective assigns different roles to its two components. Let A_{mask} be the masked input and $h_A = f_{\theta_s}(A_{\text{mask}})$ its student representation. The decoder conditions h_A on residual embeddings derived from either the original or perturbed view, yielding

$$\mathcal{L}_{\text{self}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left\| D(h_A, e^A_{\text{res}})_i - A_i \right\|_2^2, \qquad \mathcal{L}_{\text{cross}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left\| D(h_A, e^{A'}_{\text{res}})_i - A'_i \right\|_2^2. \quad (6)$$

where $\mathcal{M}$ denotes the masked patches. The self-reconstruction term preserves the standard generative learning signal, while cross-reconstruction keeps the shared representation h_A fixed and changes only the residual-conditioned appearance signal (Fig. 2d).

This conditioning provides the key intervention beyond perturbation-only training. Because the same h_A must support reconstruction of different appearance targets when paired with different residual embeddings, perturbation-dependent information that conflicts with the supplied residual becomes less useful in the shared representation pathway. HyGDL therefore encourages stable and perturbation-dependent information to play different roles, without assuming or guaranteeing complete disentanglement.

4.5 Unified Optimization with Curriculum Scheduling

HyGDL is trained end-to-end, with the student encoder, residual projector, and decoder optimized by gradient descent and the teacher updated by EMA. The overall objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{self}} + \lambda_{\text{align}}(t)\mathcal{L}_{\text{distill}} + \lambda_{\text{cross}}(t)\mathcal{L}_{\text{cross}}, \quad (7)$$

where $\alpha_{\text{style}}(t) \in [0, 0.5]$ controls stylization strength. For stable optimization, training first emphasizes masked reconstruction, then ramps up $\alpha_{\text{style}}(t)$ and $\lambda_{\text{align}}(t)$, and finally activates cross-reconstruction. This curriculum only schedules perturbation strength and loss weights within a single training process; its sensitivity is examined in Sec. 5.4.

5 Experiments

We evaluate HyGDL from four perspectives: (i) unsupervised domain generalization against perturbation-only alternatives; (ii) mitigation of the late-stage robustness divergence identified in Sec. 3; (iii) representation-level analysis of the stable projection and orthogonal residual; and (iv) ablations of the key design components.

5.1 Experimental Setup

Datasets and protocols. We evaluate unsupervised domain generalization on PACS and DomainNet. PACS follows the standard leave-one-domain-out protocol. For DomainNet, we use an asymmetric multi-source setting with Real, Painting, and Sketch as sources and Clipart, Infograph, and Quickdraw as targets. Main HyGDL results are averaged over three random seeds. Complete implementation and evaluation details are provided in the Appendix.

Comparisons. We compare HyGDL with representative source-only SSL/UDG methods, including MoCo v2, SimCLR v2, BYOL, AdCo, DARLING, DisMAE, and MS-UDG (Chen et al., 2020b; Hu et al., 2021; Zhang et al., 2022; Harary et al., 2022; Zhang et al., 2024; Pan et al., 2025). DisMAE and MS-UDG are reproduced under the same source-only/from-scratch protocol as HyGDL. We also include perturbation-only MAE baselines: MAE+BSS uses Fourier-domain style perturbation, while MAE+Sty uses online AdaIN. External ImageNet-pretrained methods are reported separately.

Computational cost. HyGDL increases pre-training cost from 8.34 to 11.6 GPU-hours in our PACS setting, while downstream inference uses only the student encoder and therefore retains the same architecture as MAE. Detailed efficiency results are deferred to the Appendix.

Table 2: **Unsupervised domain generalization on PACS and DomainNet.** PACS follows leave-one-domain-out evaluation. For DomainNet, Real/Painting/Sketch are sources and Clipart/Infograph/Quickdraw are unseen targets. HyGDL reports mean±standard deviation over three runs. External ImageNet-pretrained methods are reported separately.

	PACS					DomainNet			
Method	Photo	Art	Cartoon	Sketch	Avg.	Clipart	Infograph	Quickdraw	Avg.
Source-only / from-scratch									
ERM	43.29	24.27	32.62	20.84	30.26	15.10	9.39	7.11	10.53
MoCo v2	59.86	28.58	48.89	34.79	43.03	32.46	18.54	8.05	19.68
SimCLR v2	67.45	43.60	54.48	34.73	50.06	37.11	19.87	12.33	23.10
BYOL	41.42	23.73	30.02	18.78	28.49	14.55	8.71	5.95	9.74
AdCo	58.59	29.81	50.19	30.45	42.26	32.25	17.96	11.56	20.59
DARLING	68.86	41.53	56.89	37.51	51.20	35.15	20.88	**15.69**	23.91
MAE+BSS	66.40	47.11	54.94	35.58	51.01	46.32	20.91	9.42	25.55
MAE+Sty	70.59	46.58	53.84	39.91	52.73	41.92	17.53	8.41	22.62
DisMAE	68.93	50.64	55.16	40.23	53.74	44.13	24.56	13.74	27.48
MS-UDG	76.23	48.97	51.39	**43.21**	54.96	47.23	**26.12**	14.13	29.16
HyGDL	**76.75±0.58**	**55.70±0.27**	**58.72±0.47**	41.41±1.41	**58.15±0.33**	**51.49±0.56**	23.86±0.13	12.76±0.24	**29.37±0.29**
External ImageNet pre-training									
DIMAE	78.99	63.23	59.44	55.89	64.39	70.78	38.06	27.39	45.41
BrAD	–	–	–	–	–	68.27	26.60	34.03	42.97
CycleMAE	90.72	75.34	69.33	50.24	71.41	74.87	38.42	28.32	47.20

5.2 Unsupervised Domain Generalization

Table 2 compares HyGDL with source-only SSL/UDG methods, closely related disentanglement-based approaches, and perturbation-only MAE variants.

On PACS, HyGDL achieves $58.15 \pm 0.33\%$, compared with 53.74% for DisMAE and 54.96% for MS-UDG under the matched source-only/from-scratch protocol, and improves over MAE+Sty and MAE+BSS by 5.42 and 7.14 points, respectively.

On DomainNet, HyGDL reaches $29.37 \pm 0.29\%$, while matched DisMAE and MS-UDG obtain 27.48% and 29.16%, respectively, indicating comparable performance to MS-UDG in this setting. Overall, the results show that nuisance perturbation alone does not consistently match HyGDL, while its perturbation-relative formulation remains competitive with recent representation-factorization methods.

5.3 Does HyGDL Mitigate Late-Stage Robustness Divergence?

Improved downstream UDG accuracy does not by itself establish that HyGDL addresses the training dynamics diagnosed in Sec. 3. We therefore return to the controlled ImageNet-Sty experiment and track HyGDL over the same prolonged pre-training horizon.

The contrast is most apparent after 1000 epochs. MAE decreases from 30.25% to 28.86% by epoch 1600, corresponding to a 1.39-point decline. HyGDL instead increases from 30.14% to 30.51% over the same interval (Table 1(b)). Thus, HyGDL does not merely improve the final checkpoint; it changes the late-stage trajectory associated with the motivating failure mode.

This experiment provides an intervention-based complement to the diagnostic evidence in Sec. 3. The rise-and-fall behavior is observed with standard generative objectives, whereas explicitly coupling perturbation with invariance-aligned supervision substantially suppresses the post-peak degradation. We interpret this result as evidence that the proposed intervention reduces sensitivity to the nuisance variation used in the controlled diagnostic, rather than as proof that all forms of OOD degradation share the same mechanism.

5.4 Ablation Study

We ablate the main design choices of HyGDL on PACS. Table 3(a) shows that removing any core objective degrades performance, with the largest drop occurring without self-reconstruction. Removing cross-reconstruction also reduces accuracy from 58.13% to 56.33%, supporting the role of residual-conditioned reconstruction beyond standard generative training.

Table 3: **Ablation and representation analysis on PACS. (a)** Average target-domain accuracy (%) for key HyGDL variants. **(b)** Linear probing of the native HyGDL representation geometry. Full per-domain ablations and additional probing analyses are provided in the Appendix.

(a) Ablation study

Core Objectives		*Alignment*		*Conditioning*	
Variant	Avg.	Variant	Avg.	Variant	Avg.
Full	**58.13**	**Full**	**58.13**	**Full**	**58.13**
w/o Distill	51.76	Contrastive	52.45	FiLM	49.04
w/o Self-Recon	37.69	DINO-style	43.40	LayerNorm	52.46
w/o Cross-Recon	56.33				

(b) Representation probing

Feature	Semantic	Domain
Original	27.90±1.15	52.80±1.65
Stable	26.43±1.30	48.75±1.50
Residual	17.31±0.43	42.84±0.25

Alternative alignment and conditioning strategies are consistently weaker. Replacing direct cross-view alignment with contrastive or DINO-style distillation reduces average accuracy to 52.45% and 43.40%, respectively, while FiLM and LayerNorm conditioning achieve 49.04% and 52.46%. Overall, the gains arise from the coordinated use of cross-view alignment, geometric decomposition, and residual-conditioned reconstruction rather than from a single component. Additional sensitivity analyses, including the subspace rank and curriculum schedule, are provided in the Appendix.

5.5 Representation-Level Analysis

To examine whether the geometric decomposition behaves as intended, we probe three frozen representations from HyGDL: the original teacher feature z, the stable projection z_{stab}, and the orthogonal residual z_{res}. Linear probes are trained for semantic class and domain prediction on PACS, with results averaged over three seeds.

As shown in Table 3(b), the stable projection preserves substantially more semantic predictability than the residual (26.43% vs. 17.31%), while the residual still retains non-trivial domain information, exceeding the domain majority baseline of 39.26%. This supports an asymmetric, but not complete, separation: the stable component preferentially preserves class-predictive information, whereas the residual is relatively depleted of semantic information. We emphasize that this does not imply semantic–style disentanglement: the stable projection remains domain-predictive, and the separation weakens after per-sample ℓ_2 normalization. Additional analyses of normalization sensitivity, orthogonality, feature norms, and reconstruction exactness are provided in the Appendix.

6 Conclusion

We studied how representation robustness evolves during prolonged generative self-supervised pre-training and identified *late-stage robustness divergence*: on several appearance-sensitive distribution shifts, transfer performance peaks earlier than in-distribution representation quality and subsequently deteriorates. This behavior recurs across MAE and SimMIM and extends from controlled style perturbations to natural shifts, although its magnitude is distribution-dependent.

Motivated by this observation, we introduced an *invariance-aligned pre-training design principle* and instantiated it with HyGDL. HyGDL combines cross-perturbation self-distillation, low-rank geometric decomposition, and residual-conditioned reconstruction to reduce the utility of perturbation-sensitive information in the representation pathway used for transfer. Across PACS and DomainNet, HyGDL consistently improves over perturbation-only MAE variants and mitigates the late-stage degradation observed in our diagnostic setting. Representation probing further shows that the stable projection preserves substantially more semantic predictability than the orthogonal residual.

Our findings should be interpreted within the scope of the studied perturbations and training regimes. The observed divergence is shift-dependent rather than universal, and its interpretation as shortcut emergence does not establish a unique causal mechanism. Likewise, HyGDL induces invariance relative to the chosen perturbation distribution, and its geometric decomposition is an operational, scale-sensitive separation rather than a semantic–style factorization. Extending trajectory-level analysis and invariance-aligned training to larger models, broader distribution shifts, and more diverse self-supervised objectives remains an important direction.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature search and analysis, code review and debugging, and language editing of the manuscript. Generative AI tools were not used to generate experimental data, formulate mathematical claims or proofs, or independently design the proposed methodology or experiments. All literature identified or summarized with AI assistance was manually verified against the original sources. AI-assisted code suggestions were reviewed, tested, and validated by the authors before being incorporated into the experimental pipeline. AI-assisted manuscript editing was limited to improving clarity, grammar, and readability, without altering the scientific claims or conclusions. We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including all text, claims, code, experiments, and results.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility. Detailed descriptions of the proposed method, training procedure, model configurations, and implementation choices are provided in the main text and Appendix. The Appendix further documents the datasets, data preprocessing, evaluation protocols, hyperparameters, and additional experimental settings used throughout our experiments. We also provide an anonymous code repository containing the source code, configuration files, training and evaluation scripts, and instructions for reproducing the main results presented in this paper. The repository is available at `https://anonymous.4open.science/r/HyGDL-31FD`.

REFERENCES

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022. URL https://arxiv.org/abs/2106.08254.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/70feb62b69f16e0238f741fab228fec2-Paper.pdf.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020a. URL https://proceedings.mlr.press/v119/chen20j.html.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020b. URL https://arxiv.org/abs/2003.04297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024. doi: 10.1109/TPAMI.2024.3415112.

Sivan Harary, Eli Schwartz, Assaf Arbelle, Peter Staar, Shady Abu-Hussein, Elad Amrani, Roei Herzig, Amit Alfassy, Raja Giryes, Hilde Kuehne, Dina Katabi, Kate Saenko, Rogerio S. Feris, and Leonid Karlinsky. Unsupervised domain generalization by learning a bridge across domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5280–5290, June 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1074–1083, June 2021.

Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2506–2517, 2024. doi: 10.1109/TPAMI.2023.3336525.

Juwon Kang, Sohyun Lee, Namyup Kim, and Suha Kwak. Style neophile: Constantly seeking novel styles for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7130–7140, June 2022.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2023a. doi: 10.1109/TKDE.2021.3090866.

Yuan Liu, Songyang Zhang, Jiacheng Chen, Kai Chen, and Dahua Lin. Pixmim: Rethinking pixel reconstruction in masked image modeling, 2023b. URL https://arxiv.org/abs/2303.02416.

Pingchuan Ma, Xiaopei Yang, Yusong Li, Ming Gui, Felix Krause, Johannes Schusterbauer, and Björn Ommer. Scflow: Implicitly learning style and content disentanglement with flow models, 2025. URL https://arxiv.org/abs/2508.03402.

Lilian Ngweta, Subha Maity, Alex Gittens, Yuekai Sun, and Mikhail Yurochkin. Simple disentanglement of style and content in visual representations, 2023. URL https://arxiv.org/abs/2302.09795.

Tan Pan, Kaiyu Guo, Dongli Xu, Zhaorui Tan, Chen Jiang, Deshu Chen, Xin Guo, Brian Lovell, LIMEI HAN, Yuan Cheng, and Mahsa Baktashmotlagh. Minimal semantic sufficiency meets unsupervised domain generalization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 23011–23038. Curran Associates, Inc., 2025. doi: 10.52202/085713-0777. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2170283104cd365e28e7b1136c3d3ccf-Paper-Conference.pdf.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415, 2019. doi: 10.1109/ICCV.2019.00149.

Marin Scalbert, Maria Vakalopoulou, and Florent Couzinié-Devy. Multi-source domain adaptation via supervised contrastive learning and confident consistency regularization, 2021. URL https://arxiv.org/abs/2106.16093.

Marin Scalbert, Maria Vakalopoulou, and Florent Couzinié-Devy. Towards domain-invariant self-supervised learning with batch styles standardization, 2024. URL https://arxiv.org/abs/2303.06088.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2132–2141, June 2023.

Luya Wang, Feng Liang, Yangguang Li, Honggang Zhang, Wanli Ouyang, and Jing Shao. Repre: Improving self-supervised vision transformer with reconstructive pre-training, 2022. URL https://arxiv.org/abs/2201.06857.

Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Disentangling content and style via unsupervised geometry distillation, 2019. URL https://arxiv.org/abs/1905.04538.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16684–16693, June 2021.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9653–9663, June 2022.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14383–14392, June 2021.

Haiyang Yang, Shixiang Tang, Meilin Chen, Yizhou Wang, Feng Zhu, Lei Bai, Rui Zhao, and Wanli Ouyang. Domain invariant masked autoencoders for self-supervised learning from multi-domains. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 151–168, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19821-2.

Haiyang Yang, Xiaotong Li, SHIXIANG TANG, Feng Zhu, Yizhou Wang, Meilin Chen, LEI BAI, Rui Zhao, and Wanli Ouyang. Cycle-consistent masked autoencoder for unsupervised domain generalization. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=wC98X1qpDBA.

An Zhang, Han Wang, Xiang Wang, and Tat-Seng Chua. Disentangling masked autoencoders for unsupervised domain generalization. *arXiv preprint arXiv:2407.07544*, 2024.

Xingxuan Zhang, Linjun Zhou, Renzhe Xu, Peng Cui, Zheyan Shen, and Haoxin Liu. Towards unsupervised domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4910–4920, June 2022.

Nanxuan Zhao, Zhirong Wu, Rynson W.H. Lau, and Stephen Lin. Distilling localization for self-supervised representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10990–10998, May 2021. doi: 10.1609/aaai.v35i12.17312. URL https://ojs.aaai.org/index.php/AAAI/article/view/17312.

Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2543–2555. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/14c4f36143b4b09cbc320d7c95a50ee7-Paper.pdf.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2022. URL https://arxiv.org/abs/2111.07832.

A Appendix

A.1 Extended Related Work

This section provides additional context for the concise related-work discussion in the main paper.

A.1.1 Self-Supervised Learning Paradigms

Visual self-supervised learning (SSL) broadly follows discriminative and generative paradigms (Gui et al., 2024; Liu et al., 2023a). Discriminative methods learn invariance across augmented views, ranging from contrastive approaches such as SimCLR (Zhao et al., 2021; Zheng et al., 2021; Xie et al., 2021; Chen et al., 2020a) to non-contrastive student–teacher methods such as BYOL and DINO (Siméoni et al., 2025; Grill et al., 2020; Caron et al., 2020). Their invariances are largely determined by the augmentations used to construct positive views.

Generative SSL, particularly masked image modeling (MIM), instead learns by recovering missing image content (Zhou et al., 2022; Xie et al., 2022; Devlin et al., 2019; Bao et al., 2022). Methods such as MAE encourage representations that preserve information useful for reconstruction, which may include both semantic structure and appearance-specific cues. Recent hybrid approaches combine reconstruction and discriminative objectives, for example RePre (Wang et al., 2022), CMAE (Huang et al., 2024), and SiameseIM (Tao et al., 2023). These methods primarily exploit the complementary strengths of generative and discriminative learning. Our work instead focuses on how robustness evolves during prolonged generative pre-training and explicitly aligns the generative objective with a specified invariance.

A.1.2 Shortcut Learning and Robustness in SSL

Shortcut learning describes the tendency of neural networks to exploit predictive but non-generalizable cues. A canonical example is texture bias: ImageNet-trained models can rely heavily on texture rather than shape, while stylized training can reduce this dependence (Geirhos et al., 2019; 2020). Related observations also arise in SSL. For example, PixMIM modifies reconstruction targets to reduce reliance on low-level image statistics (Liu et al., 2023b), while strong augmentations in contrastive learning encourage invariance to selected photometric and geometric variations (Scalbert et al., 2021; Kang et al., 2022; Xu et al., 2021).

Most prior work studies shortcut mitigation through objective or augmentation design. In contrast, we examine its *training dynamics*: whether transferability continues to improve throughout generative pre-training. Our experiments reveal that OOD performance can peak before in-distribution representation quality, motivating the notion of late-stage robustness divergence and the hypothesis of late-stage shortcut emergence.

Table 4: **Conceptual comparison with closely related UDG methods.** All three methods separate transferable and variation-related information, but differ in how the decomposition is constructed and how reconstruction is used.

Method	Representation mechanism	Reconstruction and primary distinction
DisMAE	Separate semantic and variation encoders; domain supervision and adaptive contrastive learning encourage domain-invariant semantic features.	Recombines or swaps semantic and variation representations for reconstruction; explicitly targets semantic–variation disentanglement.
MS-UDG	Learns semantic and variation components under a minimal-sufficiency formulation using contrastive and disentanglement objectives.	Uses variation-conditioned / swapped reconstruction to support minimal sufficient semantic representations.
HyGDL	Estimates a low-rank perturbation-stable subspace via mini-batch cross-view SVD; the orthogonal residual is not assumed to be pure nuisance.	Keeps the masked shared representation fixed while varying residual-derived conditioning; targets perturbation-relative invariance and late-stage robustness dynamics rather than exact semantic–variation factorization.

A.1.3 Invariant and Factorized Representation Learning

A related line of work aims to separate transferable information from domain- or appearance-specific variation. Methods such as DIMAE and CycleMAE (Yang et al., 2022; 2023) employ domain-aware generative structures, while BSS (Scalbert et al., 2024) manipulates Fourier amplitude statistics to reduce sensitivity to style variation. Other approaches impose architectural, transformation-based, or representation-level priors to encourage factorized features (Wu et al., 2019; Ma et al., 2025; Ngweta et al., 2023).

Relation to Recent Disentanglement-based UDG Methods. HyGDL is closely related to recent UDG approaches that separate transferable information from domain- or appearance-specific variation. As summarized in Table 4, DisMAE (Zhang et al., 2024) learns separate semantic and variation representations and combines domain-aware contrastive learning with reconstruction, while MS-UDG (Pan et al., 2025) formulates UDG through minimal sufficient semantic representations and jointly employs contrastive, disentanglement, and reconstruction objectives. We therefore do not claim that combining representation separation with reconstruction is unique to HyGDL.

HyGDL differs in both its starting point and its mechanism. Our motivation is the late-stage evolution of robustness during generative SSL: we first diagnose a divergence between in-distribution fitting and shifted-domain transfer, and then design the training objective to modify this trajectory. Rather than learning an explicit semantic–variation factorization through separate encoders or prediction heads, HyGDL defines invariance relative to a specified perturbation family and estimates a low-rank stable subspace from cross-view student–teacher representations via mini-batch geometric projection. The complementary component is defined only as the orthogonal residual and is not assumed to represent pure nuisance or style. During cross-reconstruction, the same masked student representation h_A is kept fixed while only the residual-derived conditioning is changed, making perturbation-sensitive information less useful in the shared representation pathway. Thus, HyGDL should be understood as a perturbation-relative, invariance-aligned intervention on generative pretraining dynamics rather than as a general semantic–variation disentanglement framework.

To complement this conceptual comparison, we additionally reproduce DisMAE and MS-UDG under the same source-only/from-scratch PACS and DomainNet evaluation protocol used for HyGDL; the matched results are reported in Table 2.

A.2 Experimental Settings and Evaluation Protocols

This section summarizes the datasets, training configurations, and evaluation protocols used throughout our experiments.

Table 5: **Settings for the trajectory diagnostic experiments.** The pre-training column reports the MAE configuration, while the same frozen linear-probing protocol is used for checkpoint-wise evaluation.

Setting	Pre-training	Linear probing
Backbone	ViT-Base	Frozen encoder
Mask ratio	75%	–
Batch size	256	2048
Epochs	1600	90
Optimizer	AdamW	SGD
AdamW betas	$(0.9, 0.999)$	–
Momentum	–	0.9
Learning rate	1.5×10^{-4}	5×10^{-3}
LR schedule	Warm-up + cosine	Warm-up + cosine
Weight decay	0.05	0
Warm-up epochs	40	10
Checkpoint interval	200 epochs	–

A.2.1 Trajectory Diagnostics

We conduct the trajectory analysis on ImageNet-100. Following Geirhos et al. (2019), we construct ImageNet-Sty by applying AdaIN with a pre-trained VGG-19 to introduce controlled appearance variation.

For the MAE diagnostic, we pre-train a ViT-Base model for 1600 epochs and save checkpoints every 200 epochs. Each checkpoint is evaluated using the same frozen linear-probing protocol, so that changes in downstream performance reflect the evolution of the pre-trained representation rather than differences in evaluation.

In addition to ImageNet-100 and ImageNet-Sty, we evaluate the same representations on ImageNet-Sketch and ImageNet-R. We further introduce ImageNet-R_{overlap}, which contains the 27 ImageNet-R categories shared with the ImageNet-100 label space. The subset is defined solely according to label overlap and is independent of model performance.

To examine whether the observed dynamics depend on a particular generative objective, we additionally evaluate SimMIM using the same checkpoint-wise frozen-probing protocol. HyGDL is evaluated with the same downstream probing procedure for direct comparison of its late-stage representation dynamics. All checkpoint-wise trajectory results are reported as mean$\pm$standard deviation over three runs. The MAE pre-training configuration and the common linear-probing setup are summarized in Table 5. Unless otherwise stated, the SimMIM and HyGDL trajectory experiments use the same backbone, training horizon, checkpoint interval, and downstream linear-probing protocol as the MAE diagnostic; objective-specific hyperparameters follow their corresponding implementations.

A.2.2 Domain Generalization Protocols

We evaluate domain generalization on PACS and DomainNet following the comparison settings of Scalbert et al. (2024); Peng et al. (2019). For PACS, we adopt the standard leave-one-domain-out protocol: one domain is used as the unseen target and the remaining three domains are used for source-domain training. All available source-domain training data are used for downstream fine-tuning.

For DomainNet, our main experiment adopts an asymmetric multi-source protocol. Real, Painting, and Sketch are used as source domains, while Clipart, Infograph, and Quickdraw serve as unseen target domains. Following the original comparison setting, 10% of the available source-domain data are used for downstream training. We additionally evaluate the reverse source–target configuration in the supplementary experiments.

For perturbation-only comparisons, MAE+BSS applies Fourier-amplitude manipulation, whereas MAE+Sty applies online AdaIN-based style perturbation. Both retain the underlying MAE-style generative learning objective, allowing us to isolate the effect of nuisance variation from the ad-

Table 6: **Settings for the domain generalization experiments.**

Setting	Pre-training	Fine-tuning
Backbone	ViT-Small	–
Mask ratio	75%	–
Batch size	128	256
Epochs	1000	150
Optimizer	AdamW	AdamW
AdamW betas	$(0.9, 0.95)$	$(0.9, 0.95)$
Learning rate	3×10^{-6}	5×10^{-4}
LR schedule	Warm-up + cosine	Warm-up + cosine
Weight decay	0.05	0.05
Warm-up epochs	30	10
Augmentation	ResizeCrop + HFlip	ResizeCrop + HFlip
Curriculum phases	0–10 / 10–100 / 100–1000	–
α_{style}	$[0, 0.5]$	–
λ_{align}	$[0, 0.15]$	–
λ_{cross}	$[0, 0.1]$	–

ditional invariance-aligned components of HyGDL. Methods that rely on external ImageNet pre-training are reported separately from source-only methods.

Unless otherwise specified, HyGDL results are averaged over three random seeds. The pre-training and downstream fine-tuning configurations are summarized in Table 6. During the progressive activation phase, α_{style} and λ_{align} are linearly increased to 0.5 and 0.15, respectively. λ_{cross} is set to 0.1 when cross-reconstruction is activated in the final curriculum phase and is zero beforehand.

A.2.3 Representation Probing

We further evaluate the information encoded by different components of the HyGDL representation on PACS. For each trained model, we extract the frozen original representation z, the stable projection z_{stab}, and the orthogonal residual z_{res}. Independent linear classifiers are then trained for semantic class prediction and domain prediction.

The same probe configuration is used for all three representations, and each probe is repeated with three random seeds. We evaluate representations obtained from the four PACS leave-one-domain-out models and aggregate the results across the corresponding settings. All representation probes use a single linear classification layer trained with the same optimizer, learning rate, batch size, and number of epochs; only the frozen input representation is changed across Original, Stable, and Residual. The optimization settings follow the linear-probing configuration in Table 5.

Our main analysis operates directly on the native HyGDL feature geometry. To examine whether the observed separation depends on feature magnitude, we additionally evaluate a variant in which each representation is ℓ_2-normalized before linear probing. The majority-class accuracy for PACS domain prediction is 39.26% and is reported as a reference baseline.

A.2.4 Ablation Protocol

Unless otherwise stated, all ablation experiments follow the PACS configuration in Table 6. We vary one design choice at a time while keeping the remaining training and evaluation settings unchanged. The core-objective ablations remove self-distillation, self-reconstruction, or cross-reconstruction individually. We additionally compare alternative alignment objectives and residual conditioning mechanisms.

For the geometric decomposition, we evaluate stable-subspace ranks $k \in \{1, 2, 4\}$, with $k = 2$ used as the default configuration. Additional sensitivity experiments vary the curriculum schedule, optimization choices, and perturbation-related settings while retaining the same downstream evaluation protocol.

Table 7: **Complete MAE training trajectory.** Frozen linear-probing accuracy (%) is reported at every 200 training epochs. All results are reported as mean±standard deviation over three runs. Best results in each column are shown in **bold**.

Epoch	IN-100	IN-Sty	IN-Sketch	IN-$R_{overlap}$	IN-R
200	67.68±0.17	22.06±0.51	20.26±0.29	30.29±0.11	17.97±0.35
400	75.15±0.22	26.11±0.39	23.44±0.14	33.32±0.31	21.02±0.27
600	78.68±0.33	29.22±0.20	24.96±0.24	34.66±0.16	22.78±0.09
800	80.70±0.30	30.18±0.20	27.39±0.13	36.32±0.24	24.10±0.34
1000	81.55±0.18	**30.25±0.18**	**28.19±0.13**	**36.73±0.22**	24.75±0.28
1200	82.15±0.15	29.98±0.29	27.91±0.10	35.85±0.04	24.70±0.17
1400	**82.23±0.36**	29.50±0.54	27.41±0.12	35.41±0.19	**25.03±0.14**
1600	82.18±0.21	28.86±0.12	27.63±0.15	35.46±0.09	24.80±0.14

Table 8: **Additional checkpoint-wise trajectories.** Frozen linear-probing accuracy (%) for SimMIM and HyGDL on ImageNet-100, ImageNet-Sty, and ImageNet-Sketch. All results are reported as mean±standard deviation over three runs. Best results in each column are shown in **bold**.IN-S means IN-Sketch.

Epoch	SimMIM IN-100	SimMIM IN-Sty	SimMIM IN-S	HyGDL IN-Sty	HyGDL IN-S
200	51.37±0.08	10.07±0.01	18.14±0.13	21.44±0.41	21.07±0.13
400	56.37±0.01	10.44±0.05	19.47±0.21	25.41±0.12	23.49±0.11
600	60.68±0.17	11.26±0.11	23.04±0.16	28.37±0.35	24.07±0.23
800	63.42±0.05	11.57±0.04	23.97±0.23	29.47±0.41	26.36±0.21
1000	63.49±0.07	11.80±0.02	25.33±0.16	30.14±0.17	27.43±0.27
1200	64.45±0.18	**12.56±0.04**	**27.94±0.31**	30.24±0.20	28.31±0.26
1400	**65.19±0.06**	12.02±0.56	26.83±0.21	30.27±0.37	**28.93±0.31**
1600	65.09±0.13	11.44±0.55	25.19±0.31	**30.51±0.17**	28.87±0.13

A.3 Full Training Trajectories

We provide the complete checkpoint-wise results underlying the trajectory analysis in the main paper. All checkpoints are evaluated using the frozen linear-probing protocol described in Sec. A.2.1. Unless otherwise stated, results are reported as mean±standard deviation over three runs.

MAE trajectories. Table 7 reports the complete MAE trajectory from epoch 200 to 1600. In-distribution ImageNet-100 accuracy continues to improve until late in training, whereas several shifted distributions reach their best transfer performance earlier. In particular, ImageNet-Sty, ImageNet-Sketch, and ImageNet-$R_{overlap}$ all peak at epoch 1000, while ImageNet-100 reaches its maximum at epoch 1400. Their subsequent declines differ in magnitude, with the clearest degradation observed on ImageNet-Sty and ImageNet-$R_{overlap}$. By contrast, the full ImageNet-R trajectory remains comparatively flat, again illustrating that late-stage robustness divergence is distribution-dependent rather than universal.

Cross-objective reproducibility and mitigation. Table 8 extends the checkpoint-wise analysis to SimMIM and HyGDL on both ImageNet-Sty and ImageNet-Sketch. SimMIM exhibits late-stage degradation on both shifts: ImageNet-Sty peaks at epoch 1200 and decreases from $12.56 \pm 0.04\%$ to $11.44 \pm 0.55\%$ by epoch 1600, while ImageNet-Sketch peaks at $27.94 \pm 0.31\%$ at epoch 1200 and declines to $25.19 \pm 0.31\%$ at the final checkpoint. Meanwhile, its ImageNet-100 accuracy continues to improve until epoch 1400.

HyGDL substantially changes these late-stage trajectories. On ImageNet-Sty, accuracy continues to increase through epoch 1600, reaching $30.51 \pm 0.17\%$. On ImageNet-Sketch, performance increases from $27.43 \pm 0.27\%$ at epoch 1000 to $28.93 \pm 0.31\%$ at epoch 1400 and remains nearly unchanged at $28.87 \pm 0.13\%$ at epoch 1600. Thus, under both the controlled style shift and the natural sketch shift, HyGDL avoids the pronounced late-stage degradation observed with the standard generative objectives over the evaluated training horizon.

Table 9: **Additional asymmetric domain generalization results on DomainNet.** The left block reports the main protocol used in the paper (Painting/Real/Sketch → Clipart/Infograph/Quickdraw), while the right block reports the reverse transfer direction. Results are accuracy (%). Best results within each comparison group are shown in **bold**.

	P/R/S → C/I/Q				C/I/Q → P/R/S			
Method	Clipart	Infograph	Quickdraw	Avg.	Painting	Real	Sketch	Avg.
Source-only / in-domain pre-training								
ERM	15.10	9.39	7.11	10.53	9.90	9.19	5.12	8.07
MoCo v2	32.46	18.54	8.05	19.68	25.35	29.91	23.71	26.32
SimCLR v2	37.11	19.87	12.33	23.10	24.01	30.17	**31.58**	28.59
BYOL	14.55	8.71	5.95	9.74	9.50	10.38	4.45	8.11
AdCo	32.25	17.96	11.56	20.59	23.35	29.98	27.57	26.97
DARLING	35.15	20.88	**15.69**	23.91	**25.90**	**33.29**	30.77	**29.99**
MAE+BSS	46.32	20.91	9.42	25.55	18.38	23.32	24.60	22.10
MAE+Sty	41.92	17.53	8.41	22.62	19.45	23.70	22.30	21.81
HyGDL	**51.49**	**23.86**	12.76	**29.37**	21.71	28.11	28.60	26.14
External ImageNet pre-training								
DIMAE	70.78	38.06	27.39	45.41	50.73	64.89	55.41	57.01
BrAD	68.27	26.60	**34.03**	42.97	31.08	38.48	48.71	39.24
CycleMAE	**74.87**	**38.42**	28.32	**47.20**	**52.81**	**67.13**	**56.37**	**58.77**

Note. P/R/S denotes Painting, Real, and Sketch; C/I/Q denotes Clipart, Infograph, and Quickdraw. MAE+BSS applies Fourier-amplitude style manipulation, while MAE+Sty applies AdaIN-based appearance perturbation to the standard MAE objective. DIMAE, BrAD, and CycleMAE use external ImageNet pre-training and are therefore reported separately from the source-only setting.

Summary. The full trajectories reinforce two aspects of the main findings. First, the mismatch between in-distribution improvement and shifted-domain transfer is reproduced across MAE and SimMIM and is visible under both controlled and natural appearance shifts. Second, HyGDL substantially reduces the corresponding late-stage degradation on the two shifts evaluated here. At the same time, the comparatively stable full ImageNet-R trajectory demonstrates that the magnitude of this phenomenon remains dependent on the target distribution.

A.4 Additional Domain Generalization Results

To deeply investigate the behavior of our generative-based hybrid framework on complex, multi-domain benchmarks, we conducted two symmetric transfer experiments on a subset of DomainNet. Setting A (Schematic-to-Naturalistic): We used Clipart, Infograph, Quickdraw as the multi-source training domains and evaluated on Real, Painting, Sketch. Setting B (Naturalistic-to-Schematic): We used Real, Painting, Sketch as the source domains and evaluated on Clipart, Infograph, Quickdraw.

The results in Table 9 reveal a clear asymmetry between the two transfer directions. HyGDL achieves the strongest average performance among the source-only methods in the main P/R/S→C/I/Q setting, whereas this advantage does not persist in the reverse C/I/Q→P/R/S direction. This indicates that the effectiveness of the proposed approach depends on the source–target configuration.

One possible explanation is the substantial difference in visual statistics between schematic domains such as Quickdraw and the more visually detailed Real or Painting domains, as qualitatively illustrated in Fig. 3. However, the current experiment does not isolate the contribution of Quickdraw or establish a causal mechanism. We therefore regard the reverse direction as a boundary case demonstrating the domain dependence of the observed generalization gains.

A.5 Additional Ablations and Sensitivity Analysis

We provide additional analyses of optimization choices and hyperparameter sensitivity on PACS. Unless otherwise specified, the auxiliary experiments on teacher augmentation, loss weights, gradient flow, and style-source selection use the $k = 1$ configuration for computational efficiency. These

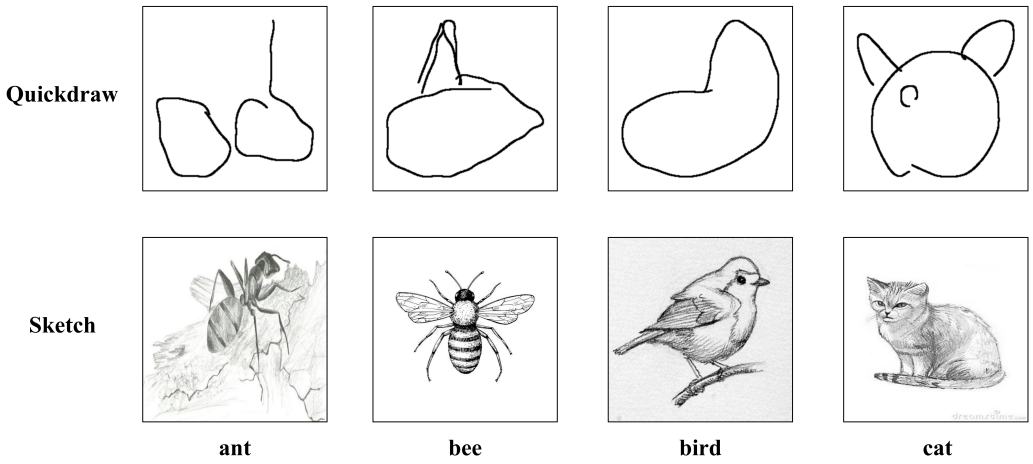

Figure 3: The comparison of four object classes (ant, bee, bird and cat) in subset "Quickdraw" and "Sketch".

experiments are intended to study relative trends and should therefore not be directly compared with the $k = 2$ main results in Table 3(a). The curriculum and subspace-rank analyses use the configurations specified below.

Teacher Input Augmentation. We examine the role of input asymmetry in the student–teacher pathway by comparing the standard setting, where the teacher receives a randomly augmented view of the original image, with a variant that uses the un-augmented image as teacher input. As shown in Table 10, removing teacher-side augmentation leads to a clear reduction in average PACS accuracy. This result suggests that cross-view asymmetry provides a useful training signal for learning perturbation-stable representations, rather than the student and teacher operating on nearly identical inputs. We do not interpret this experiment as evidence that teacher augmentation alone prevents representation collapse.

Sensitivity to Loss Weights. We evaluate the sensitivity of HyGDL to the weights associated with self-distillation and cross-reconstruction. We vary one weight at a time while keeping the remaining configuration fixed. As summarized in Table 11, $\lambda_{\text{align}} = 0.15$ and $\lambda_{\text{cross}} = 0.1$ give the strongest average performance among the tested values. Performance changes moderately within the evaluated ranges, suggesting that the method does not depend on a single narrowly tuned value for either objective. These results are limited to the tested hyperparameter ranges and do not imply global insensitivity to the loss weights.

Gradient Flow from Cross-Reconstruction. We further examine whether the cross-reconstruction objective benefits the encoder itself or mainly trains the residual projector and decoder. We compare the standard configuration with a variant in which gradients from $\mathcal{L}_{\text{cross}}$ are stopped before reaching the encoder. As shown in Table 12, stopping this gradient decreases the average PACS accuracy from 56.88% to 55.83%. This result supports allowing the cross-reconstruction signal to update the encoder, indicating that its contribution is not restricted to decoder-side optimization. However, the experiment does not establish that cross-reconstruction produces an exact semantic–nuisance disentanglement.

Sensitivity to Style-Source Selection. Finally, we investigate how the choice of AdaIN reference-style distribution affects downstream generalization. For each PACS leave-one-domain-out split, we designate one PACS domain (Photo, Cartoon, or Art) as the *preferred* reference-style source. When the designated domain belongs to the current source domains, reference images are sampled from that PACS domain. When the designated domain coincides with the held-out target domain, we do not access any images from that target domain and instead use WikiArt, the external style-image

collection used in the original AdaIN setting, as the reference-style pool. Thus, no held-out PACS target images are used during pre-training in any leave-one-domain-out split.

Concretely, for the Photo configuration, the reference pool is WikiArt when Photo is the target and PACS-Photo for the other three target splits. Likewise, for the Cartoon configuration, WikiArt is used when Cartoon is the target and PACS-Cartoon otherwise; for the Art configuration, WikiArt is used when Art is the target and PACS-Art otherwise. We exclude Sketch as a preferred PACS style source because its sparse line-drawing statistics induce a qualitatively different type of AdaIN perturbation, making it less comparable to the other reference distributions.

As shown in Table 15, among the evaluated configurations, using Art as the preferred PACS reference source yields the highest average accuracy. These results indicate that downstream robustness is sensitive to the reference-style distribution used during pre-training. One possible explanation is that different reference pools induce different magnitudes and types of appearance variation under AdaIN, thereby changing the difficulty of the cross-view alignment and reconstruction objectives. We treat this explanation as a hypothesis rather than a causal conclusion.

More broadly, this sensitivity highlights an important limitation of the proposed invariance-aligned formulation: the learned invariance is defined relative to the perturbation distribution encountered during training. Consequently, the choice and diversity of the reference-style distribution can influence downstream robustness.

Table 10: Ablation studies of input augmentation on PACS. Best methods are highlighted in **bold**, and all reported accuracies are in percent (%).

	PACS				
	Target Domain				
Method	**photo**	**art**	**cartoon**	**sketch**	**avg.**
un-augmented	59.88	42.23	46.2	35.89	46.05
augmented	**76.17**	**54.0**	**59.51**	**37.82**	**56.88**

Table 11: Ablation studies of Loss Weights on PACS. Best methods are highlighted in **bold**, and all reported accuracies are in percent (%).

	PACS				
	Target Domain				
Loss Weights	**photo**	**art**	**cartoon**	**sketch**	**avg.**
$\lambda_{align} = 0.1$	74.85	52.58	59.17	**38.71**	56.33
$\lambda_{align} = 0.15$	74.97	53.02	**61.56**	37.92	56.87
$\lambda_{align} = 0.2$	75.39	52.0	60.11	35.15	55.66
$\lambda_{cross} = 0.05$	74.79	53.71	61.56	37.2	56.82
$\lambda_{cross} = 0.1$	**76.17**	**54.0**	59.51	37.82	**56.88**
$\lambda_{cross} = 0.15$	73.47	53.9	60.58	37.08	56.26

Table 12: Ablation studies of gradient flow from cross-reconstruction on PACS. Best methods are highlighted in **bold**, and all reported accuracies are in percent (%).

	PACS				
	Target Domain				
Method	**photo**	**art**	**cartoon**	**sketch**	**avg.**
stop gradient	73.77	51.46	58.3	**39.78**	55.83
normal gradient	**76.17**	**54.0**	**59.51**	37.82	**56.88**

Table 13: **Sensitivity to curriculum scheduling on PACS.** We report target-domain accuracy (%) under different curriculum schedules.

Curriculum Schedule	Photo	Art	Cartoon	Sketch	Avg.
0–10 / 10–100 / 100–1000	**77.31**	55.37	**59.13**	40.72	58.13
0–0 / 0–0 / 0–1000	76.16	**55.90**	58.83	**43.03**	**58.48**
0–100 / 0–0 / 0–1000	76.79	55.84	58.21	40.49	57.83

Table 14: **Sensitivity to the stable-subspace rank on PACS.** Target-domain accuracy (%) is reported for different values of k.

Rank k	Photo	Art	Cartoon	Sketch	Avg.
$k = 1$	76.17	54.00	**59.51**	37.82	56.88
$k = 2$	**77.31**	**55.37**	59.13	**40.72**	**58.13**
$k = 4$	75.69	52.83	59.01	38.10	56.41

Table 15: **Sensitivity to the reference-style distribution on PACS.** Each row specifies the preferred PACS domain used as the AdaIN reference-style source. When the preferred domain is itself the held-out target, the reference pool is replaced by WikiArt; no target-domain images are accessed during pre-training. All results are target-domain accuracy (%).

	PACS				
	Target Domain				
Preferred PACS Style Source	**photo**	**art**	**cartoon**	**sketch**	**avg.**
photo	62.51†	45.94	56.19	36.96	50.40
cartoon	57.54	47.85	55.85†	**40.49**	50.43
art	**76.17**	**54.0**†	**59.51**	37.82	**56.88**

Note. † indicates a leave-one-domain-out split in which the preferred PACS style source coincides with the held-out target. For these entries, AdaIN reference images are sampled from WikiArt instead of the target PACS domain. Thus, the exact reference pools are: Photo row = {WikiArt, Photo, Photo, Photo}, Cartoon row = {Cartoon, Cartoon, WikiArt, Cartoon}, and Art row = {Art, WikiArt, Art, Art}, where the columns are ordered as Photo, Art, Cartoon, and Sketch targets, respectively.

A.6 Additional Representation Analysis

We provide additional analyses of the representation geometry induced by HyGDL. The main paper evaluates the original representation, stable projection, and orthogonal residual using linear probes in their native feature geometry. Here, we further examine the sensitivity of this observation to feature normalization and verify the geometric properties of the decomposition.

A.6.1 Sensitivity to Feature Normalization

The probing results in the main paper are obtained directly from the native HyGDL representations. Since the stable projection and residual may differ substantially in feature magnitude, we additionally apply per-sample ℓ_2 normalization before training the linear probes. Table 16 compares the two settings.

Under the native geometry, the stable projection retains substantially more semantic predictability than the residual (26.43% vs. 17.31%), while the residual still contains non-trivial domain information above the majority baseline. However, this separation weakens substantially after per-sample ℓ_2 normalization. In particular, domain prediction approaches the majority baseline and the semantic advantage of the stable projection over the residual disappears.

These results indicate that the probing behavior observed in the native representation is partly scale-sensitive. We therefore interpret the decomposition as an *operational geometric separation* induced by the training objective, rather than evidence of a scale-invariant or exact semantic–style factorization.

A.6.2 Geometric Properties of the Decomposition

We further verify that the implementation satisfies the geometric properties defined by the projection. For an orthonormal stable basis V_{stab}, the two components are constructed as

$$z_{\mathrm{stab}} = V_{\mathrm{stab}} V_{\mathrm{stab}}^{\top} z, \qquad z_{\mathrm{res}} = z - z_{\mathrm{stab}}. \tag{8}$$

Consequently,

$$z = z_{\mathrm{stab}} + z_{\mathrm{res}}, \qquad z_{\mathrm{stab}}^{\top} z_{\mathrm{res}} = 0, \tag{9}$$

up to floating-point numerical precision. We verify both properties numerically throughout training.

We additionally observe that the residual has a substantially smaller norm than the original representation, with $\|z_{\mathrm{res}}\|_2 / \|z\|_2$ approximately 1.1% on average. Together with the normalization experiment above, this suggests that feature magnitude contributes meaningfully to the asymmetric probing behavior of the stable and residual components.

Importantly, these observations do not imply that the stable component is a pure semantic representation or that the residual contains only style. The stable projection remains predictive of domain information in its native geometry, while the residual retains measurable semantic information. The decomposition should therefore be understood as separating dominant cross-view-consistent directions from their orthogonal remainder under the specified perturbation process.

A.7 Computational Cost Analysis

To quantify the computational overhead of our proposed framework, we compare its training efficiency with the standard MAE baseline. Both models were pre-trained on the PACS dataset (using 3 domains as source) for 1000 epochs on two NVIDIA RTX 4090 GPUs with a total batch size of 128. The results, calculated directly from the training logs, are summarized in Table 17.

As shown in Table 17, our HyGDL framework introduces a notable computational overhead compared to the standard MAE. The total training time increases by approximately 40%, with a corresponding decrease in throughput. The peak GPU memory usage is also significantly higher. This overhead is an expected consequence of our training strategy, which involves multiple additional computations per step: a full forward pass through the teacher encoder for distillation, a second decoder pass for the cross-reconstruction objective, style-transfer based on AdaIN and the SVD computation for subspace identification.

Table 16: **Effect of feature normalization on representation probing.** Linear-probe accuracy (%) for semantic class and domain prediction on PACS. Results are averaged over three probe seeds. The native results are included for direct comparison with the ℓ_2-normalized setting. The majority baseline for domain prediction is 39.26%.

Geometry	Representation	Semantic	Domain
Native	Original	27.90±1.15	52.80±1.65
	Stable	26.43±1.30	48.75±1.50
	Residual	17.31±0.43	42.84±0.25
ℓ_2-normalized	Original	14.70±2.43	39.26±0.00
	Stable	14.82±3.81	39.26±0.00
	Residual	17.40±0.59	37.82±0.36

Table 17: The comparison of computational cost, the primary overhead in HyGDL comes from the additional forward pass of the teacher encoder, cross-reconstruction and the SVD computation.

Model	**Hours**	**GPU-Hours**	**samples/sec**	**Peak Memory (GB/GPU)**
MAE	4.17	8.34	263.6	4.35
HyGDL	5.8	11.6	174	18.3

It is crucial to note that this is a one-time pre-training cost. For all downstream applications and inference tasks, only the trained student encoder is utilized. Therefore, the inference cost of HyGDL, in terms of latency, throughput, and memory footprint, is identical to that of a standard MAE model of the same architecture. We believe this investment in a more sophisticated pre-training process is a reasonable trade-off for the observed gains in out-of-domain generalization and the improved stability of the learned representations.