Microstructure Descriptor Fields as Supervision for Scientific Images

Mohamed Wahib[1,2] **Du Wu**[1] **Cong Ma**[1,2] **Isaac Lyngaas**[3]
Xiao Wang[3] **Enzhi Zhang**[1,2] **Peng Chen**[1] * **Satoshi Matsuoka**[1]

[1]RIKEN Center for Computational Science, Kobe, Japan
[2]Hokkaido University, Sapporo, Japan
[3]Oak Ridge National Laboratory, Tennessee, USA
{mohamed.attia, du.wu, cong.ma, peng.chen}@riken.jp
zhangenzhi@iic.hokudai.ac.jp
{lyngaasir, wangx2}@ornl.gov
matsu@acm.org

Abstract

Scientific images often depend not only on exact pixel fidelity but also on the preservation of microstructure, including morphology, connectivity, anisotropy, interface complexity, phase organization, and characteristic length scales. Yet common image-learning objectives, including masked image modeling and reconstruction losses, supervise models in pixel, token, or latent feature space rather than through domain-relevant structural measurements. We introduce MICROFIELD, an objective-level supervision framework that applies established microstructure descriptors locally over spatial regions, producing a spatially indexed and optionally multiscale descriptor field. MICROFIELD augments a base learning objective with local descriptor matching, level-wise descriptor-distribution alignment, and, when a hierarchy is available, cross-scale transition preservation, without modifying the model architecture. We evaluate this supervision principle through masked self-supervised pretraining on one controlled synthetic benchmark and three real-world imaging regimes. MICROFIELD outperforms matched self-supervised baselines; it improves segmentation by 8.9% and regression/classification by 8.3%, while reducing connected-component error by 52.1% at essentially unchanged PSNR. These results show that spatial descriptor-field supervision can improve structure-sensitive representation learning without modifying the encoder or adding inference-time computation.

1 Introduction

Scientific imaging increasingly operates in regimes where standard computer vision assumptions are misaligned with scientific use. In X-ray computed tomography (XCT), electron microscopy, materials characterization, pathology whole-slide imaging, and biological microscopy, an image is not merely a natural visual scene. It is a measurement of underlying physical, material, or biological organization. What matters scientifically is often not exact pixel identity, but preservation of microstructure: pores, grains, interfaces, anisotropic textures, phase boundaries, connected components, tissue morphology, and scale-dependent organization. Such a structure is routinely characterized through statistical and morphological descriptors rather than through raw intensities alone [42, 21, 30, 29]. These descriptors are observable, task-conditioned summaries of image structure; they should not be interpreted as uniquely recovering the latent physical or biological state that generated the image.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Despite this, most modern image-learning objectives remain weakly aligned with scientific fidelity. Masked autoencoders reconstruct missing patches in pixel space [16]; related masked image modeling methods predict discrete tokens, teacher features, or latent representations [1, 47, 50, 4, 32]. These objectives are powerful and generic, but their supervision target is not designed around scientific observables. Two reconstructions may differ at the pixel level while preserving the same pore-network statistics, interface density, or anisotropy. Conversely, visually small changes may alter topology, connectivity, or phase organization in ways that are scientifically significant. In scientific imaging, pixel fidelity is therefore an imperfect surrogate for structural fidelity.

This paper asks how established microstructure descriptors can be used as a supervision space for representation learning. We introduce MICROFIELD, which represents an image by a discrete descriptor field: a spatially indexed collection in which every region R_i is associated with a descriptor vector $s_i(x)$. For example, on a uniform grid, each grid cell may be associated with its local intensity moments, interface density, wavelet energy, and anisotropy. Unlike a single global descriptor vector, the resulting field retains where these measurements occur. When the region system is hierarchical, it additionally represents how the measurements vary across scale. In the masked-pretraining instantiation studied here, the model predicts an image-like reconstruction $\hat{x}$. The same fixed descriptor extractor is then applied to corresponding regions of x and $\hat{x}$, and the resulting descriptor fields are compared in the loss. Thus, the default model does not directly predict a standalone descriptor vector or descriptor field.

Recent high-resolution vision methods use hierarchy mainly as an architectural or tokenization device for making large images tractable [26, 5, 41, 45, 15, 49, 19]. MICROFIELD uses hierarchy differently: when available, it serves as a multiscale measurement domain for descriptor-field supervision. The contribution is therefore objective-level rather than architectural, and can be attached to standard ViTs, convolutional autoencoders, hierarchical transformers, reconstruction networks, or supervised segmentation models.

MICROFIELD defines three complementary descriptor-field constraints: local preservation of region-wise descriptors, preservation of parent–child descriptor transitions when a hierarchy is available, and level-wise alignment of descriptor distributions. These terms allow the method to supervise scientific structure locally, across scale, and statistically. The method is physics-inspired in a deliberately practical sense. Unlike physics-informed neural networks, MICROFIELD does not require governing PDEs, differentiable simulators, or boundary conditions [34]. Instead, it preserves observables that correlate with latent material or biological organization. This is useful in imaging domains where physical structure matters, but direct equation-based supervision is unavailable, unknown, or not naturally expressed in the image variables.

Importantly, microstructure descriptors are not used as a replacement for learned representations. They are training-time measurements of structural fidelity: the network does not consume descriptor vectors as input, and downstream prediction is not performed on handcrafted descriptor features. Instead, descriptors specify structural observables that should be preserved while the neural model remains free to learn high-dimensional representations from image data. *This places* MICROFIELD *closer to knowledge-constrained learning, such as physics-informed or equivariant learning, than to classical handcrafted-feature pipelines.* Our contributions are:

- Building on established microstructure descriptors and structure-aware objectives, we formulate a descriptor field as a spatially indexed collection of local descriptor vectors, optionally organized across a multiscale region hierarchy, and use it as a training-time supervision object for scientific representation learning.
- We instantiate this formulation in MICROFIELD, an architecture-agnostic auxiliary objective combining local descriptor matching, level-wise descriptor-distribution alignment, and optional cross-scale transition preservation. MICROFIELD modifies neither the encoder architecture nor the inference model.
- We evaluate the same objective-level mechanism on one controlled synthetic benchmark and three distinct real-world imaging regimes. Matched baselines and ablations separate the effect of descriptor-field supervision from global descriptor matching, hierarchy alone, generic non-pixel regularization, and non-scientific regional targets.

2 Related Work

Image pretraining and high-resolution scientific vision. Modern vision pretraining objectives supervise models by predicting or matching pixels, tokens, teacher features, or latent representations, with a masked image modeling and related self-supervised methods being prominent examples [8, 9, 16, 1, 47, 50, 4, 32, 25, 20, 35]. High-resolution vision methods address the cost of large images through hierarchical tokenization, windowed attention, quadtree or octree representations, graph aggregation, and adaptive patching [26, 41, 19, 45, 5, 38, 15, 49]. MICROFIELD is orthogonal to these architecture-centric approaches: it asks what structural quantities should be preserved during training. Hierarchy, when available, is used only as a measurement domain for descriptor-field supervision, not as the defining contribution.

Structure-aware objectives and microstructure descriptors. Many image-learning objectives go beyond raw pixel regression. Perceptual losses compare deep feature activations [22]; style losses match Gram-matrix texture statistics [12, 13]; wavelet and frequency-domain losses encourage multiscale spectral fidelity [10]; and distributional objectives such as maximum mean discrepancy compare empirical feature distributions [14]. These methods establish that useful supervision can be defined outside pixel space. However, their target spaces are usually generic visual features, frequency statistics, or global distributions, rather than scientific observables defined over a spatial measurement domain.

A related line of work uses structural, geometric, or topological measurements directly as losses. In medical image segmentation, boundary losses optimize distances between predicted and target contours [24]; topology-aware losses use persistent homology or related summaries to penalize topological errors in predicted masks [17, 7]; clDice preserves connectivity of tubular structures such as vessels, neurons, and roads through differentiable skeleton-based supervision [39]; and global shape descriptors have been used as weak supervision when dense pixel labels are unavailable [23]. These methods are important precedents because they show that non-pixel structural measurements can provide effective training signals. Their scope, however, is primarily output-specific: they are designed to improve a particular supervised prediction, such as a segmentation mask, boundary, topology, connectivity, or global shape property.

In scientific imaging and materials analysis, microstructure is commonly characterized through phase fractions, correlation functions, interface measures, chord-length statistics, connected components, Euler characteristic, anisotropy statistics, and Minkowski functionals [42, 21, 30, 29]. These descriptors are central to microstructure characterization, reconstruction, and inverse materials design [2, 3, 36]. Descriptor-based reconstruction frameworks such as MCRpy formulate microstructure reconstruction as an optimization problem in which arbitrary descriptor losses can be minimized, including with gradient-based optimizers when descriptors are differentiable [37]. These works are close in spirit to MICROFIELD because they optimize images using scientific descriptors.

MICROFIELD differs from prior structure-aware objectives in where descriptors are measured and how they are used. Perceptual, style, and spectral losses target generic visual features or frequency statistics; boundary, topology, and shape losses are typically specialized to supervised masks or contours; and microstructure reconstruction methods optimize images or generators to match target statistics. MICROFIELD instead treats scientific descriptors as a spatially indexed supervision field for representation learning. This field supports local descriptor matching, level-wise distribution alignment, and cross-scale transition preservation, while remaining independent of the underlying architecture. Accordingly, we do not claim novelty for the individual microstructure descriptors, spatial subdivisions, hierarchical representations, or structural distance functions. The contribution is their formulation as a spatially indexed, optionally multiscale supervision object for representation pretraining, together with a controlled cross-domain evaluation of this formulation. A compact positioning relative to related structure-aware objectives is provided in Appendix I, Table 13.

Physics-informed/physics-inspired learning. Physics-informed models embed governing equations, boundary conditions, or conservation laws into training objectives, typically through PDE residuals [34]. This is powerful when equations are known, differentiable, and expressed in the model variables. Scientific images often violate these assumptions: they may be projections, reconstructions, stain responses, or microscopy intensity fields rather than PDE states. MICROFIELD is therefore physics-inspired rather than physics-informed: it preserves observable microstructure quantities correlated with latent material or biological organization, without requiring explicit governing equations.

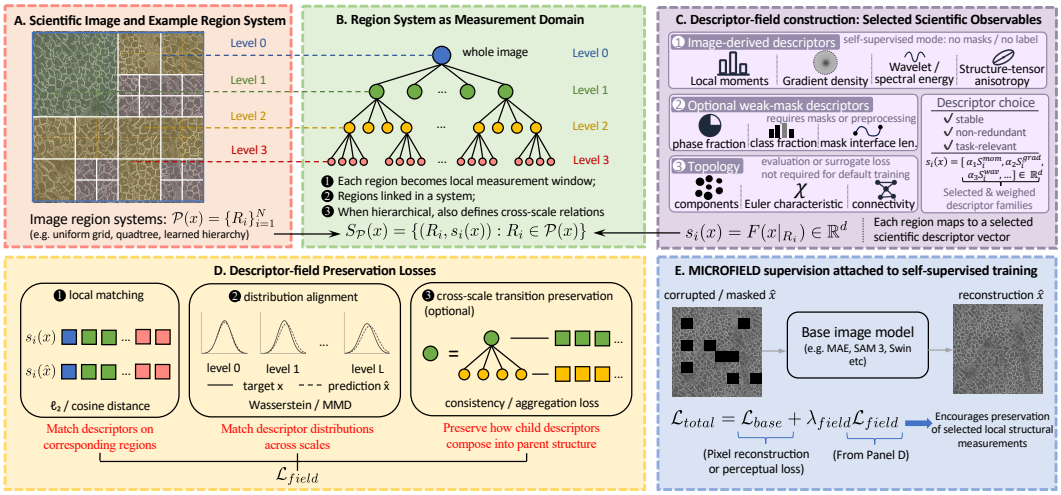


Figure 1: Overview of MICROFIELD. An image is decomposed into regions that serve as local measurement windows. A fixed descriptor extractor maps every region to a vector of selected structural measurements, forming a discrete descriptor field. In the masked-pretraining instantiation evaluated here, the base model predicts an image-like reconstruction $\hat{x}$; descriptor fields are then extracted from both x and $\hat{x}$. The auxiliary objective matches descriptors on corresponding regions, aligns their level-wise distributions, and, when a hierarchy is used, preserves cross-scale relations. Only the image-derived differentiable descriptors are used in the default unlabeled configuration.

3 Methodology

MICROFIELD is a training-time supervision framework that augments a base learning objective with a descriptor-field loss. Given a base vision model and learning objective, MICROFIELD constructs a spatial measurement domain over the target image, computes per-region microstructure descriptors, and penalizes discrepancies between the descriptor field of the target and that of the model output. The framework does not change the encoder, decoder, attention mechanism, or tokenizer. It acts as an objective-level layer that can be attached to masked image modeling, reconstruction, restoration, segmentation, classification, or regression.

The objective shifts supervision from raw pixels to scientific observables. In the default unlabeled setting, MICROFIELD matches image-derived signal, texture, interface, and anisotropy descriptors across spatial regions and, when a hierarchy is used, across scales. Phase- and mask-derived descriptors provide optional weak supervision, while non-differentiable topology quantities are used only for evaluation. Improvements in connectivity and topology are therefore transfer effects, not quantities optimized by the default loss. Figure 1 summarizes the framework.

Although MICROFIELD uses reconstruction during self-supervised training, the reconstruction is only a means to learn better representations. The goal is not image restoration quality itself, but to make the learned representation preserve scientific structure through descriptor-field supervision.

3.1 Problem Setup

Let $x \in \mathbb{R}^{H\times W\times C}$ denote a scientific image and let f_θ be a base model. In the masked self-supervised setting, a masking or corruption operator M produces $x_m = M(x)$, and the model predicts

$$\hat{x} = f_\theta(x_m). \tag{1}$$

Here, $\hat{x}$ is an image-like reconstruction, not a predicted descriptor field. Given a common region system $\mathcal{P}$, the descriptor extractor F is subsequently applied to both x and $\hat{x}$:

$$s_i(x) = F(x|_{R_i}), \qquad s_i(\hat{x}) = F(\hat{x}|_{R_i}).$$

Gradients from the descriptor-field loss propagate through $s_i(\hat{x})$ to the reconstruction model. The base objective $\mathcal{L}_{\text{base}}$ may be a pixel, token, perceptual, latent, supervised, or reconstruction loss. MICROFIELD adds a descriptor-field regularizer:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda_{\text{field}}\mathcal{L}_{\text{field}}. \tag{2}$$

The motivation is that pixel similarity does not necessarily imply scientific similarity. Let $\Psi(x)$ denote a task-relevant structural functional, such as pore connectivity, interface density, anisotropy, tissue organization, or characteristic length scale. In general,

$$\|x - \hat{x}\| \text{ small} \not\Longrightarrow \|\Psi(x) - \Psi(\hat{x})\| \text{ small}. \tag{3}$$

MICROFIELD therefore adds a loss that directly measures preservation of local and multiscale microstructure.

3.2 Descriptor Fields

MICROFIELD first defines a spatial measurement domain

$$\mathcal{P}(x) = \{R_i\}_{i=1}^{N}, \tag{4}$$

where each region R_i is a local measurement window. Generally speaking, the region system may be a uniform grid, quadtree or octree, salience-adaptive partition, learned tokenizer, superpixel map, segmentation-derived graph, or hierarchy produced by an existing high-resolution image model. The method does not require hierarchy: a uniform grid is sufficient for local descriptor-field supervision. When hierarchy is available, it provides parent–child relations for cross-scale losses.

Let F be a descriptor extractor. For each region R_i, define

$$s_i(x) = F(x|_{R_i}) \in \mathbb{R}^d, \tag{5}$$

where $x|_{R_i}$ is the restriction of the image to R_i. The descriptor field induced by $\mathcal{P}$ is

$$S_{\mathcal{P}}(x) = \big((R_i, s_i(x))\big)_{i=1}^{N}, \tag{6}$$

For a hierarchical region system $\mathcal{T}$, each node n corresponds to a spatial region R_n, and the descriptor field becomes

$$S_{\mathcal{T}}(x) = \{s(n; x) : n \in \mathcal{N}(\mathcal{T})\}, \qquad s(n; x) = F(x|_{R_n}) \tag{7}$$

were $\mathcal{N}_\ell(T)$ is the set of nodes at level ℓ and $\mathcal{N}_{\text{int}}(T)$ is the set of internal nodes. Descriptor dimensions have heterogeneous units and numerical ranges. For example, wavelet energy may be substantially larger than a bounded anisotropy score. Without normalization, high-magnitude dimensions would dominate the auxiliary objective. Therefore, descriptor dimensions are normalized with running statistics

$$\tilde{s}_{ij}(x) = \frac{s_{ij}(x) - \mu_j}{\sigma_j + \epsilon}, \tag{8}$$

The index j denotes the descriptor dimension. The statistics μ_j and σ_j are estimated from target-image regions in the training split and then held fixed during optimization. They are unrelated to the exponential moving averages used later to normalize loss-component magnitudes. No ordering of regions is assumed in estimating these statistics. All descriptor-field losses in the remainder of this section are computed using normalized descriptors.

3.3 MICROFIELD Training

MICROFIELD is implemented as a measurement-and-regularization loop around a base model. For masked self-supervised pretraining, the target image defines the measurement domain; the same regions are used to evaluate both x and $\hat{x}$, avoiding a matching problem between different partitions. The complete training step is summarized in Algorithm 1 in Appendix F.

The same loop applies to supervised learning or restoration by replacing $\mathcal{L}_{\text{base}}$ with the corresponding task loss. If the model does not produce an image-like output, MICROFIELD can be applied through an auxiliary training-only reconstruction or projection head.

3.4 Descriptor-Field Losses

MICROFIELD uses three complementary losses. The first is the mean squared error over regions and normalized descriptor dimensions. This local loss matches descriptors at corresponding regions

$$\mathcal{L}_{\text{local}}(x, \hat{x}) = \frac{1}{d} \sum_{i=1}^{N} w_i \, \|\tilde{s}_i(x) - \tilde{s}_i(\hat{x})\|_2^2, \qquad \sum_{i=1}^{N} w_i = 1. \tag{9}$$

where w_i is a region weight, set to $1/N$ by default or chosen proportional to area, salience, or mask visibility. When $\mathcal{P}$ is hierarchical, MICROFIELD also matches cross-scale descriptor transitions. For each internal node p, let $\mathrm{ch}(p)$ denote its children and let A_p be the descriptor-specific aggregation operator (described in detail in Appendix K.4). We define

$$r(p; x) = \tilde{s}(p; x) - A_p(\{\tilde{s}(c; x) \,|\, c \in \mathrm{ch}(p)\}) \,. \tag{10}$$

The cross-scale transition loss is then

$$\mathcal{L}_{\mathrm{trans}}(x, \hat{x}) = \frac{1}{|\mathcal{N}_{\mathrm{int}}(T)|} \sum_{p \in \mathcal{N}_{\mathrm{int}}(T)} \|r(p; x) - r(p; \hat{x})\|_2^2 \,, \tag{11}$$

where $\mathcal{N}_{\mathrm{int}}(T)$ denotes the set of internal nodes in the hierarchy. This term preserves how coarse structures decompose into finer structures.

Finally, MICROFIELD aligns descriptor distributions at each level. Let

$$\mathcal{S}_\ell(x) = \{\tilde{s}(n; x) : n \in \mathcal{N}_\ell\} \tag{12}$$

be the multiset of descriptors at level ℓ. We define

$$\mathcal{L}_{\mathrm{dist}}(x, \hat{x}) = \sum_\ell D\left(\mathcal{S}_\ell(x), \mathcal{S}_\ell(\hat{x})\right) , \tag{13}$$

where, in all default experiments, D is mean–covariance matching

$$D(S_\ell(x), S_\ell(\hat{x})) = \|\mu_\ell(x) - \mu_\ell(\hat{x})\|_2^2 + \|\Sigma_\ell(x) - \Sigma_\ell(\hat{x})\|_F^2,$$

where μ_ℓ and Σ_ℓ are the empirical mean and covariance of the normalized regional descriptors at level ℓ. MMD, sliced Wasserstein distance, and histogram distance are possible alternatives but are not used for the main reported results. In the non-hierarchical case, the same distributional loss can be applied to the single region set $\mathcal{P}$.

The full descriptor-field objective is

$$\mathcal{L}_{\mathrm{field}} = \lambda_{\mathrm{local}}\mathcal{L}_{\mathrm{local}} + \lambda_{\mathrm{trans}}\mathcal{L}_{\mathrm{trans}} + \lambda_{\mathrm{dist}}\mathcal{L}_{\mathrm{dist}} \tag{14}$$

Setting individual coefficients to zero gives the ablations used in our experiments: local-only MICROFIELD, non-hierarchical MICROFIELD, and distribution-only descriptor supervision.

3.5 Descriptor Families

MICROFIELD is not tied to a single descriptor set. We use compact descriptor families that are scientifically interpretable, stable across scales, and differentiable when used as training losses. Table 1 summarizes the default families; full definitions are given in Appendix K.2.

Descriptor-supervision modes. We distinguish three modes. First, the default unlabeled mode uses only differentiable image-derived measurements, including intensity moments, gradient and Laplacian energy, wavelet energy, and structure-tensor anisotropy. This is the mode used for the reported self-supervised pretraining results and requires no masks during optimization. Second, an optional weak-mask mode may use phase- or object-level measurements derived from fixed target masks. A hard threshold may be used to construct a target mask outside the gradient path, but prediction-side descriptors must be computed from soft probabilities or another differentiable surrogate. Hard thresholding of the model output is not used inside the optimization path. Third, non-differentiable quantities such as connected-component count and Euler characteristic are used only for evaluation in this work. They are not components of the default training objective.

For training, we use differentiable descriptors such as moments, gradients, wavelet energy, and structure-tensor anisotropy. Non-differentiable topology descriptors, such as connected-component counts or Euler characteristic, are used for evaluation or through differentiable surrogates. Descriptor-family weights may be fixed or learned; we describe the learnable-weight variant in Appendix F.

Table 1: Descriptor families in MICROFIELD. The default self-supervised setting uses only image-derived descriptors that require no masks or labels. Mask-derived phase and topology descriptors are optional domain-specific extensions or evaluation metrics.

Mode	Family	Examples	Use
Unlabeled	Signal / texture	mean, variance, entropy, wavelet energy	training
Unlabeled	Interface / boundary	gradient density, Laplacian energy	training
Unlabeled	Anisotropy	structure-tensor anisotropy	training
Weak-mask	Phase / composition	phase fraction, class fraction, interface length	optional training
Evaluation-only	Topology / connectivity	components, Euler characteristic, Betti numbers	evaluation

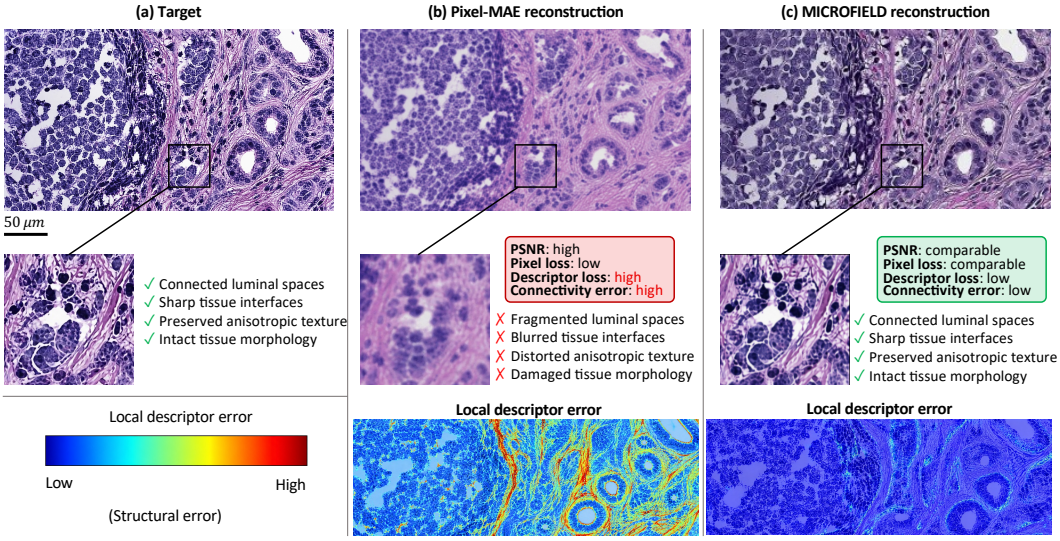


Figure 2: Pixel fidelity does not guarantee structural fidelity. In this representative example, Pixel-MAE [16] and MICROFIELD achieve comparable pixel fidelity, but Pixel-MAE introduces larger descriptor-field and connectivity errors. Aggregate PSNR/SSIM and structure metrics in Table 3.

3.6 Practical Instantiations and Cost

MICROFIELD is attached by replacing or augmenting the base loss in Eq. (2). In masked pretraining, it compares the descriptor field of the decoded output to that of the target image; in reconstruction or inverse problems, it acts as a structure-aware fidelity term; and in supervised tasks, it can be applied through an image-like prediction or auxiliary training head. The additional cost is incurred only during training. Most descriptors can be implemented with pooled convolutions, integral images, histograms, or batched tensor operations, and region sampling can reduce overhead for large images. At inference, the base model is unchanged. Additional details on region construction, descriptor definitions, distributional losses, and implementation are provided in Appendix K.

4 Why Descriptor Fields?

MICROFIELD targets a mismatch between pixel fidelity and scientific fidelity where scientific equivalence is often morphological rather than pointwise. As illustrated in Figure 2, reconstructions with similar pixel fidelity can differ substantially in descriptor-field error. MICROFIELD addresses this by using default objective supervises image-derived texture, interface, and anisotropy measurements; phase descriptors are optional weak-mask extensions, while topology is evaluated as a transfer effect.

MICROFIELD *is not a handcrafted-feature pipeline*. Classical pipelines feed expert-designed descriptors to the downstream predictor. Here, the representation remains fully learned; descriptors are used only as training-time measurements of structural fidelity in Eq. (2). They act as auxiliary structural constraints, analogous to physics-informed residuals or equivariance constraints when explicit governing equations are unavailable in image space.

Descriptor choice is a structural prior rather than a fixed recipe. MICROFIELD does not assume a universal descriptor set because scientific equivalence is domain-dependent: pore networks, membranes, tissue morphology, and fibrous materials preserve different invariants. We therefore select descriptors that are stable under nuisance variation, non-redundant with each other, and predictive of the scientific quantity of interest. This makes descriptor choice a testable modeling assumption, analogous to choosing a residual in physics-informed learning or a symmetry group in equivariant learning, and we evaluate it through descriptor-family ablations.

Finally, the field formulation is essential. Global descriptors can match image-level statistics while destroying spatial organization. Region-wise descriptor fields preserve where structural measurements occur, not only their global averages. When a hierarchy is available, it additionally exposes how fine-scale descriptors compose into coarser morphology. A formal quotient-space and sufficiency interpretation is provided in Appendix L.

5 Evaluation

We evaluate whether MICROFIELD improves scientific representation learning by preserving microstructure rather than only pixel-level appearance. The experiments are designed to answer four questions: **(i)** Does descriptor-field supervision improve downstream scientific image tasks? **(ii)** Does it improve structural fidelity under metrics that reflect morphology, connectivity, anisotropy, and interfaces? **(iii)** Are the gains due to descriptor-field supervision rather than hierarchy, generic frequency losses, or additional regularization? **(iv)** What training overhead does the objective introduce while leaving the inference model unchanged?

5.1 Experimental Setup

Datasets. We evaluate four image regimes chosen to stress different microstructural properties.

Synthetic-μStruct is a controlled dataset generated from stochastic two-phase fields with known phase fraction, interface density, anisotropy, connected components, Euler characteristic, and correlation length. This dataset tests whether pretraining preserves known structural parameters rather than merely improving natural-image-style accuracy.

Porous-XCT contains X-ray computed tomography images of porous or heterogeneous materials [33, 43]. We use it for phase segmentation, porosity regression, and pore-network preservation. Porous media are a natural test case because phase fraction, interface density, connectivity, and anisotropy are central scientific quantities.

MitoEM/EM contains electron microscopy volumes for mitochondria or membrane-rich biological structures [46]. We use 2D slices for masked pretraining and evaluate on segmentation and topology-sensitive morphology metrics. This setting stresses fine interfaces, elongated structures, and object connectivity.

HuBMAP-Vasc contains PAS-stained human kidney histology images with microvasculature annotations [18, 40]. We use it for vessel segmentation and structure-sensitive transfer. This setting stresses thin connected objects, luminal spaces, tissue texture, and anisotropic morphology.

Unless otherwise stated, all methods are pretrained without labels and then evaluated through linear probing, fine-tuning, and structure-preservation metrics. Labels and masks are used only for downstream evaluation and for evaluation-only topology metrics.

Architectures and baselines. Our controlled experiments use ViT-B/16 with matched in-domain data, masking, optimization, and training schedules. We compare against standard MIM/SSL objectives, generic structure losses, hierarchy-only controls, and MICROFIELD ablations. External foundation models are included only as domain-compatible references and are not used for matched-data causal comparisons. The complete baseline taxonomy and tuning protocol are provided in Appendix F.2 (Table 11). We also report MICROFIELD ablations that enable local matching, level-wise distribution alignment, and cross-scale transition preservation. External foundation models are included only as reference baselines on compatible domains and are not used for matched-data causal claims. All controlled self-supervised baselines are pretrained on the same in-domain unlabeled images from the four evaluation regimes; ImageNet-supervised initialization is included only as a standard transfer baseline.

Table 2: Main downstream results. External foundation models are marked with † because they use substantially larger external pretraining corpora. Results are averaged over Synthetic-μStruct, Porous-XCT, MitoEM/EM, and HuBMAP-Vasc where the corresponding method is applicable. External foundation models are evaluated only on compatible domains and are reported as reference baselines, not matched-data causal comparisons.

Domain	Method	Seg. ↑	Reg./Cls. ↑	Struct. err. ↓	Data setting
Controlled	Pixel-MAE [16]	78.4	81.6	0.236	matched in-domain
	SimMIM/Swin [47, 26]	79.1	82.3	0.224	matched in-domain
	iBOT [50]	80.2	83.1	0.211	matched in-domain
	DINOv2 [32]	81.0	83.8	0.206	matched in-domain
	Hierarchy-only	81.7	84.2	0.192	matched in-domain
	MICROFIELD-Grid	85.1	86.9	0.132	matched in-domain
	MICROFIELD-Full	**87.3**	**89.9**	**0.121**	matched in-domain
External FM	UNI / Phikon / Virchow [6, 11, 44] †	83.8	87.4	0.181	external WSI
	Prov-GigaPath / HIPT[48, 5] †	84.4	87.8	0.176	external WSI
	CA-MAE [25] †	82.7	85.9	0.188	external microscopy
	CellSAM / MedSAM [28, 27] †	84.9	82.6	0.169	external masks
	MatSSL / SEM SSL [31, 35] †	82.9	85.4	0.172	external/domain SSL

Seg. is the unweighted macro-average of dataset-level Dice scores reported in percentage points. Reg./Cls. is the unweighted macro-average of the available dataset-level $R^2 \times 100$ scores. Struct. err. averages normalized descriptor/topology errors: PFE, IDE, AD, CLE, CCE, ECE, and DDD.

Metrics. We distinguish three forms of evidence. First, downstream segmentation, regression, and classification objectives are not components of the pretraining loss and constitute the primary transfer evaluation. We report Dice, mIoU, boundary F1, object-level average precision, accuracy, AUROC, R^2, and MAE as applicable. Second, connected-component error (CCE) and Euler-characteristic error (ECE) are computed only for evaluation and are not optimized in the reported default pretraining objective. They provide topology-sensitive evaluation, although they are not completely orthogonal to the trained interface and texture measurements. Third, descriptor-distribution distance (DDD), interface-density error (IDE), and anisotropy discrepancy (AD) directly reuse descriptor families aligned with the training objective. We therefore report them as mechanism diagnostics, not as independent validation of scientific fidelity.

Efficiency. MICROFIELD adds only training-time descriptor computation and leaves the inference model unchanged; detailed training-time and memory overheads are reported in Appendix F.1.

5.2 Main Results: Downstream Scientific Tasks

Table 2 summarizes average downstream performance over Synthetic-μStruct, Porous-XCT, MitoEM/EM, and HuBMAP-Vasc; external foundation models are evaluated only on compatible domains and are included as reference baselines rather than controlled same-data comparisons. Per-dataset means and standard deviations over three complete runs are reported in Appendix D, Table 5. MICROFIELD improves downstream accuracy across material, electron microscopy, and histology tasks. The largest gains occur on tasks where the target depends on connectivity or morphology rather than only local texture, such as vessel segmentation, pore-network segmentation, and mitochondria instance segmentation. These results indicate that descriptor-field supervision yields representations better aligned with morphology and connectivity-sensitive scientific tasks.

5.3 Scientific Fidelity Beyond PSNR and SSIM

Pixel metrics can be misleading for scientific images: similar PSNR/SSIM can hide large differences in topology, interfaces, anisotropy, and connectivity. We therefore evaluate masked reconstructions with descriptor-level and topology-sensitive metrics alongside pixel fidelity. Table 3 shows that the strongest baselines have comparable PSNR, while MICROFIELD-Full substantially reduces descriptor-distribution distance, interface-density error, anisotropy discrepancy, and connected-component error relative to Pixel-MAE. This matches the failure mode in Figure 2: high pixel fidelity can coexist with degraded scientific structure.

Table 3: Scientific fidelity of masked reconstructions. DDD, IDE, and AD reuse descriptor families aligned with the training objective and are therefore reported as mechanism diagnostics rather than independent evidence. CCE is a topology-sensitive metric computed only for evaluation in the reported default experiments. Results are averaged over the reconstruction datasets for which each metric is applicable. Higher is better for PSNR and SSIM; lower is better for all other metrics.

	Pixel fidelity		**Descriptor-aligned diagnostics**			**Evaluation-only topology**
Method	PSNR ↑	SSIM ↑	DDD ↓	IDE ↓	AD ↓	CCE ↓
Pixel-MAE [16]	31.8	0.902	0.246	0.218	0.174	0.286
MAE + perceptual	31.5	0.907	0.231	0.207	0.166	0.271
MAE + spectral	**32.1**	0.911	0.214	0.184	0.151	0.249
MAE + global descriptors	31.9	0.909	0.197	0.171	0.143	0.224
MICROFIELD-Grid	32.0	0.913	0.156	0.132	0.109	0.181
MICROFIELD-Full	31.9	**0.916**	**0.121**	**0.098**	**0.083**	**0.137**

DDD: descriptor-distribution distance; IDE: interface-density error; AD: anisotropy discrepancy; CCE: connected-component error.

Table 4: Ablation study. The key comparisons separate descriptor supervision, spatial indexing, hierarchy, distribution alignment, and cross-scale transition preservation. Downstream is the average of the normalized Seg. and Reg./Cls. scores in Table 2. Struct. error is the same normalized descriptor/topology error used in Table 2.

Variant	Local	Dist.	Trans.	Downstream ↑	Struct. error ↓
Pixel-MAE [16]	–	–	–	80.0	0.236
MAE + global descriptors	–	–	–	82.4	0.197
Hierarchy-only MAE	–	–	–	83.0	0.192
MICROFIELD-Grid	✓	✓	–	86.0	0.132
MICROFIELD-Hier, local only	✓	–	–	86.4	0.139
MICROFIELD-Hier, local + dist.	✓	✓	–	87.5	0.127
MICROFIELD-Full	✓	✓	✓	**88.6**	**0.121**

5.4 Ablation: What Makes MICROFIELD Work?

Table 4 isolates the source of improvement. A global descriptor loss improves over Pixel-MAE but remains weaker than local descriptor fields, showing that global statistics are insufficient. A hierarchy-only model improves compute organization but does not match MICROFIELD, showing that the gain is not merely due to hierarchical processing. MICROFIELD-Grid already improves structure fidelity, confirming that descriptor supervision is useful even without hierarchy. The full hierarchical version performs best, indicating that cross-scale transition preservation provides additional gains when multiscale organization is meaningful. Descriptor-family ablations in Appendix L.5 show that interface, anisotropy, and texture descriptors are complementary, while shuffled descriptors fail to improve transfer, ruling out generic regularization as the explanation.

6 Conclusion

We introduced MICROFIELD, an objective-level supervision framework that represents scientific images as spatial fields of microstructure descriptors and trains models to preserve them. MICROFIELD supports local descriptor matching, distributional alignment, and optional cross-scale transition preservation when hierarchy is available. It can be attached to self-supervised, supervised, or reconstruction models, providing a practical path toward microstructure-aware, physics-inspired representation learning without explicit governing equations.

Acknowledgments and Disclosure of Funding

This research used resources of the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory, which is supported by the DOE Office of Science, ASCR, under Contract No. DE-AC05-00OR22725.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[2] Ramin Bostanabad, Yichi Zhang, Xiaolin Li, Tucker Kearney, L. Catherine Brinson, Daniel W. Apley, Wing Kam Liu, and Wei Chen. Computational microstructure characterization and reconstruction: Review of the state-of-the-art techniques. *Progress in Materials Science*, 95:1–41, 2018.

[3] Ruijin Cang, Yaopengxiao Xu, Shaohua Chen, Yongming Liu, Yang Jiao, and Max Yi Ren. Microstructure representation and reconstruction of heterogeneous materials via deep belief network for computational material design. *Journal of Mechanical Design*, 139(7):071404, 2017.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[5] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.

[6] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[7] James R. Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A. Zimmer, Julia A. Schnabel, and Andrew P. King. A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8766–8778, 2022.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, 2019.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[10] Mario A. T. Figueiredo and Robert D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906–916, 2003.

[11] Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2: A large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024.

[12] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[13] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[14] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[15] Ziyu Guo, Weiqin Zhao, Shujun Wang, and Lequan Yu. HIGT: Hierarchical interaction graph-transformer for whole slide image analysis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI*, pages 755–764. Springer, 2023.

[16] 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*, pages 16000–16009, 2022.

[17] Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[18] HuBMAP Consortium. HuBMAP – hacking the human vasculature. `https://www.kaggle.com/competitions/hubmap-hacking-the-human-vasculature`, 2023. Kaggle competition dataset; PAS-stained human kidney histology tiles with microvasculature annotations.

[19] Moritz Ibing, Gregor Kobsik, and Leif Kobbelt. Octree Transformer: Autoregressive 3D Shape Generation on Hierarchically Structured Sequences . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2698–2707, Los Alamitos, CA, USA, June 2023. IEEE Computer Society.

[20] Shuai Jiang, Liesbeth Hondelink, Arief A. Suriawinata, and Saeed Hassanpour. Masked pre-training of transformers for histology image analysis. *Journal of Pathology Informatics*, 15:100386, 2024.

[21] Yang Jiao, Frank H. Stillinger, and Salvatore Torquato. Modeling heterogeneous materials via two-point correlation functions: Basic principles. *Physical Review E*, 76(3):031110, 2007.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711, 2016.

[23] Hoel Kervadec, Houda Bahig, Laurent Letourneau-Guillon, Jose Dolz, and Ismail Ben Ayed. Beyond pixel-wise supervision for segmentation: A few global shape descriptors might be surprisingly good! In *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, volume 143 of *Proceedings of Machine Learning Research*, pages 354–368. PMLR, 2021.

[24] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pages 285–296. PMLR, 2019.

[25] Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, Dominique Beaini, Maciej Sypetkowski, Chi Vicky Cheng, Kristen Morse, Maureen Makes, Ben Mabey, and Berton Earnshaw. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11757–11768, 2024.

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[27] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[28] Markus Marks, Uriah Israel, Rohit Dilip, Qilin Li, Changhua Yu, Emily Laubscher, Ahamed Iqbal, Elora Pradhan, Ada Ates, Martin Abt, Caitlin Brown, Edward Pao, Shenyi Li, Alexander Pearson-Goulart, Pietro Perona, Georgia Gkioxari, Ross Barnowski, Yisong Yue, and David Van Valen. CellSAM: a foundation model for cell segmentation. *Nature Methods*, 22(12):2585–2593, 2025.

[29] Klaus R. Mecke. Additivity, convexity, and beyond: Applications of minkowski functionals in statistical physics. In Klaus R. Mecke and Dietrich Stoyan, editors, *Statistical Physics and Spatial Statistics*, pages 111–184. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[30] Kristel Michielsen and Hans De Raedt. Integral-geometry morphological image analysis. *Physics Reports*, 347(6):461–538, 2001.

[31] Hoang Hai Nam Nguyen, Phan Nguyen Duc Hieu, and Ho Won Lee. MatSSL: Robust self-supervised representation learning for metallographic image segmentation. *arXiv preprint arXiv:2507.18184*, 2025.

[32] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[33] Masa Prodanovic, Maria Esteva, Richard Ketcham, Bernard Chang, Cinar Turhan, James Gentle, Saad Khan, and Victoria Belcher. Digital porous media portal (dpmp) for publication, analysis, and simulation of porous media images. Digital Porous Media Portal, 2025.

[34] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[35] Luca Rettenberger, Nathan J. Szymanski, Andrea Giunto, Olympia Dartsi, Anubhav Jain, Gerbrand Ceder, Veit Hagenmeyer, and Markus Reischl. Leveraging unlabeled SEM datasets with self-supervised learning for enhanced particle segmentation. *Computational Materials*, 11(289), 2025.

[36] Muhammad Sahimi and Pejman Tahmasebi. Reconstruction, optimization, and design of heterogeneous materials and media: Basic principles, computational algorithms, and applications. *Physics Reports*, 939:1–82, 2021.

[37] Paul Seibert, Alexander Raßloff, Karl Kalina, Marreddy Ambati, and Markus Kästner. Microstructure characterization and reconstruction in python: MCRpy. *arXiv preprint arXiv:2207.04652*, 2022.

[38] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147, 2021.

[39] Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. clDice: A novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16560–16569, 2021.

[40] Michael P. Snyder, Sheng Lin, Amanda Posgai, Mark Atkinson, Aviv Regev, Jennifer Rood, et al. The human body at cellular resolution: the NIH human biomolecular atlas program. *Nature*, 574(7777):187–192, 2019.

[41] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. In *International Conference on Learning Representations*, 2022.

[42] Salvatore Torquato. *Random Heterogeneous Materials: Microstructure and Macroscopic Properties*. Springer, 2002.

[43] Cinar Turhan, Bernard Chang, Ali Mohamed, Maria Esteva, Richard Ketcham, James McClure, and Masa Prodanovic. Digital porous media portal for image curation, characterization, visualization, and transport simulation in porous media. In *International Symposium of the Society of Core Analysts*, 2024. SCA2024-1056.

[44] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, 2024.

[45] Peng-Shuai Wang. Octformer: Octree-based transformers for 3d point clouds. *ACM Trans. Graph.*, 42(4), July 2023.

[46] Donglai Wei, Zudi Lin, Daniel Franco-Barranco, Nils Wendt, Xingyu Liu, Wenjie Yin, Xin Huang, Aarush Gupta, Won-Dong Jang, Xueying Wang, Ignacio Arganda-Carreras, Jeff W. Lichtman, and Hanspeter Pfister. MitoEM dataset: Large-scale 3d mitochondria instance segmentation from em images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2020.

[47] 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*, pages 9653–9663, 2022.

[48] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024.

[49] Enzhi Zhang, Isaac Lyngaas, Peng Chen, Xiao Wang, Jun Igarashi, Yuankai Huo, Masaharu Munetomo, and Mohamed Wahib. Adaptive patching for high-resolution image segmentation with transformers. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2024.

[50] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.

Supplementary Material

A Broader Impacts

Improved representation learning for scientific images could reduce annotation cost and improve analysis in materials science, microscopy, computational biology, infrastructure inspection, and medical imaging. The main risk is that structure-preserving objectives may create overconfidence if the chosen descriptors encode incomplete or biased notions of scientific fidelity. Care is required when applying such methods to clinical, safety-critical, or policy-relevant settings.

B Limitations

A failure mode of MICROFIELD is descriptor misspecification: if the selected descriptor field is insufficient for the downstream scientific structure, the objective may faithfully preserve the wrong invariants. Descriptor choice is therefore a modeling assumption, analogous to choosing the residual in physics-informed learning or the symmetry group in equivariant learning, and should be validated through ablations or downstream structure metrics.

C Compute platform and training cost.

Experiments were run on a system with 400 NVIDIA Grace Blackwell nodes and x1,600 GPUs in total. Each compute node contains one NVIDIA GB200 NVL4 unit with two Grace CPUs and four NVIDIA B200 GPUs. Each node provides 960 GiB CPU memory and 692.8 GiB GPU HBM3e memory. Nodes are connected by NVIDIA InfiniBand XDR at 800 Gbps, with four XDR links per compute node.

We report the full platform specification for reproducibility; individual experiments used only a small subset of the machine. Unless otherwise stated, self-supervised pretraining runs used x64 NVIDIA B200 GPUs. A MICROFIELD-Full pretraining run with the default ViT-B/16 backbone took approximately 22 hours, corresponding to 1,408 GPU-hours per run. The matched Pixel-MAE baseline took approximately 18.5 hours under the same backbone, batch size, masking ratio, optimizer, and training schedule. Thus, descriptor-field supervision increased pretraining wall-clock time by approximately 19%, while leaving the inference model unchanged.

Main results are averaged over three random seeds. Including controlled pretraining baselines, MICROFIELD variants, downstream fine-tuning, linear probing, and ablations, the total compute used for the reported experiments was approximately 31,000 GPU-hours. Downstream fine-tuning and linear-probe evaluations were substantially cheaper than pretraining: a typical downstream run used x4 B200 GPUs for 2–3 hours per dataset. No experiment used the full x1,600-GPU system.

MICROFIELD adds only training-time descriptor computation. In our implementation, the descriptor-field objective increased peak GPU memory by approximately 8–10% relative to the matched Pixel-MAE baseline. At inference time, the base encoder is used without descriptor extraction, descriptor losses, or additional model parameters, so MICROFIELD incurs no additional inference computation.

D Statistical reliability

Unless otherwise stated, all reported task and structure metrics are averaged over three independent runs with different random seeds. We vary the random seed for pretraining initialization, masking/corruption sampling, fine-tuning initialization, data order, and downstream evaluation. We report aggregate means for readability; per-dataset means and standard deviations over three complete runs in Table 5. Across the main comparisons, the gains from MICROFIELD are consistently larger than the observed run-to-run standard deviation.

For the controlled comparisons, aggregate task scores are computed as unweighted macro-averages of one primary metric per dataset. Metrics reported on the $[0, 1]$ scale are multiplied by 100, and regression performance is reported as $100R^2$; we do not apply min–max or z-score normalization

Table 5: Per-dataset downstream performance, reported as mean ± standard deviation over three complete runs. Each run varies the pretraining initialization, masking pattern, data order, downstream initialization, and fine-tuning.

Dataset and metric	Pixel-MAE	Hierarchy-only	MICROFIELD-Grid	MICROFIELD-Full
Synthetic-μStruct, structural $R^2 \times 100 \uparrow$	84.0 ± 0.7	86.0 ± 0.6	88.5 ± 0.5	91.3 ± 0.4
Synthetic-μStruct, Dice $\uparrow$	72.0 ± 0.4	75.8 ± 0.3	79.8 ± 0.2	82.7 ± 0.2
Porous-XCT, Dice $\uparrow$	80.2 ± 0.8	83.4 ± 0.7	86.7 ± 0.5	88.8 ± 0.5
Porous-XCT, porosity $R^2 \times 100 \uparrow$	79.2 ± 0.9	82.4 ± 0.8	85.3 ± 0.6	88.5 ± 0.5
MitoEM, Dice $\uparrow$	79.3 ± 0.9	82.6 ± 0.8	85.8 ± 0.6	87.7 ± 0.6
MitoEM, object AP $\uparrow$	64.5 ± 1.2	68.8 ± 1.1	74.0 ± 0.9	77.4 ± 0.8
HuBMAP-Vasc, Dice $\uparrow$	82.1 ± 0.8	85.0 ± 0.7	88.1 ± 0.6	90.0 ± 0.5
HuBMAP-Vasc, boundary F1 $\uparrow$	76.0 ± 1.0	80.9 ± 0.9	85.0 ± 0.7	87.9 ± 0.6

across methods. Specifically, letting m denote a method,

$$\mathrm{Seg}(m) = \frac{1}{4} \sum_{d \in \mathcal{D}_{\mathrm{seg}}} 100 \ \mathrm{Dice}_{m,d}, \tag{15}$$

where $\mathcal{D}_{\mathrm{seg}} = \{\text{Synthetic-}\mu\text{Struct}, \text{Porous-XCT}, \text{MitoEM}, \text{HuBMAP-Vasc}\}$, and

$$\mathrm{Reg.}\,/\mathrm{Cls.}(m) = \frac{1}{2}\left(100 R^2_{m,\mathrm{Synthetic}} + 100 R^2_{m,\mathrm{Porous}}\right). \tag{16}$$

The Synthetic-μStruct structural R^2 is the dataset-level structural-regression score, while the Porous-XCT value is the porosity-regression score. Secondary metrics such as object AP and boundary F1 are reported separately and are not included in these aggregate scores.

E Datasets

E.1 Synthetic-μStruct generation

Synthetic-μStruct is a controlled two-phase microstructure dataset designed to test whether representation learning preserves known structural parameters rather than only pixel appearance. Each sample consists of an intensity image x, a binary phase mask $m \in \{0,1\}^{H \times W}$, and ground-truth structural descriptors: phase fraction, interface density, anisotropy, connected-component count, Euler characteristic, and correlation length.

We generate each sample from a stochastic Gaussian-field model. First, we sample a latent random field

$$z(u) = \mathcal{F}^{-1}\left(\sqrt{S_{\ell,\rho,\theta}(k)}\,\epsilon(k)\right),$$

where u indexes spatial position, $\epsilon(k)$ is complex Gaussian white noise with Hermitian symmetry, and $S_{\ell,\rho,\theta}$ is an anisotropic power spectrum. The parameter ℓ controls the characteristic length scale, ρ controls anisotropy, and θ controls the dominant orientation. In our implementation,

$$S_{\ell,\rho,\theta}(k) = \exp\left(-\frac{1}{2} k_\theta^\top \Sigma_{\ell,\rho}^{-1} k_\theta\right),$$

where k_θ denotes Fourier coordinates rotated by angle θ, and $\Sigma_{\ell,\rho}$ sets the major and minor correlation lengths. The field is standardized to zero mean and unit variance.

A target phase fraction $\phi \sim \mathcal{U}(\phi_{\min}, \phi_{\max})$ is then sampled, and the binary phase mask is obtained by thresholding:

$$m(u) = \mathbb{1}\{z(u) > \tau_\phi\},$$

where τ_ϕ is chosen so that $\frac{1}{HW}\sum_u m(u) = \phi$. This gives direct control over phase fraction while preserving the sampled spatial correlation and anisotropy. To increase diversity, we optionally apply one to three morphological operations sampled from dilation, erosion, opening, and closing with small disk structuring elements, followed by rejection if the resulting phase fraction deviates from the target range.

The final image is produced by assigning phase-dependent intensities and adding realistic acquisition variation:

$$x(u) = \mu_0(1 - m(u)) + \mu_1 m(u) + \eta(u),$$

where μ_0, μ_1 are sampled phase intensities and η combines Gaussian noise, low-frequency illumination variation, and mild blur. Images are normalized to $[0, 1]$. The mask m is not used during self-supervised pretraining; it is used only to compute controlled structural ground truth and evaluation metrics.

For each generated mask, we compute: (i) phase fraction, (ii) interface density from the perimeter of the binary phase boundary normalized by image area, (iii) anisotropy from the eigenvalue ratio of the structure tensor, (iv) connected-component count, (v) Euler characteristic, and (vi) correlation length estimated from the radial autocorrelation function. Samples are rejected if they are nearly empty, nearly full, have degenerate components, or have a correlation length outside the predefined range. This produces a controlled dataset in which the structural parameters are known by construction.

Dataset size and splits. We generate 100,000 unlabeled Synthetic-μStruct images for self-supervised pretraining and 10,000 labeled images for downstream controlled evaluation. All images are generated at 512×512 resolution. The labeled subset is split into 7,000/1,000/2,000 train/validation/test images. The binary phase masks and known structural parameters are retained for evaluation only; they are not used during self-supervised pretraining. To avoid leakage, the split is performed before any augmentation, and no augmented views of the same latent random field appear in more than one split. We use fixed random seeds for field generation, threshold sampling, morphology perturbations, and train/validation/test assignment.

E.2 Porous-XCT dataset

Porous-XCT is our curated porous-media X-ray computed tomography dataset, constructed from public porous-media micro-CT sources, including datasets hosted through the Digital Porous Media Portal / Digital Rocks Portal [33, 43]. These repositories provide 2D and 3D images of porous microstructures, including X-ray CT and micro-CT scans of geomaterials and related porous media. We use Porous-XCT to evaluate whether pretraining preserves phase fraction, interface density, pore-network connectivity, anisotropy, and correlation length. Table 6 lists the dataset settings.

We extract 2D slices and non-overlapping or weakly overlapping 512×512 patches from the available 3D volumes. To avoid leakage, train/validation/test splits are performed at the source-volume level whenever multiple volumes are available; otherwise, spatially separated block-level splits are used before patch extraction. No augmented views of the same spatial block are allowed to appear in different splits.

In our experiments, Porous-XCT contains 120,000 unlabeled patches for self-supervised pretraining and 12,000 patches for downstream evaluation. All patches are extracted at 512×512 resolution from public porous-media micro-CT/XCT volumes. Unless otherwise stated, the evaluation set is split into $70/10/20$ train/validation/test partitions. When multiple source volumes are available, splits are performed at the source-volume level; otherwise, spatially separated block-level splits are used before patch extraction to reduce leakage.

Binary phase masks, when provided by the source dataset or obtained through standard thresholding and cleanup, are used only for downstream phase segmentation and structural evaluation; they are not used during self-supervised pretraining. For evaluation, we report task metrics for phase segmentation and porosity regression, together with structural metrics: phase-fraction error (PFE), interface-density error (IDE), anisotropy discrepancy (AD), correlation-length error (CLE), connected-component error (CCE), Euler-characteristic error (ECE), and descriptor-distribution distance (DDD).

E.3 MitoEM/EM dataset

MitoEM/EM is based on the MitoEM dataset, a large-scale 3D electron microscopy dataset for mitochondria instance segmentation [46]. The dataset contains two annotated $30\mu\mathrm{m}^3$ EM volumes, one from adult human brain tissue and one from adult rat brain tissue. Each volume has spatial size $1000 \times 4096 \times 4096$ voxels at $30 \times 8 \times 8$ nm resolution. The official benchmark partitions the volumes along the axial dimension; we follow the official split when using labeled mitochondria masks. Table 7 lists the dataset settings.

Table 6: Porous-XCT dataset settings.

Property	Setting
Source	Public porous-media micro-CT / XCT volumes
Primary use	Porosity, phase segmentation, pore-network preservation
Patch size	512×512
Pretraining patches	$N_{\mathrm{pre}} = 120{,}000$ unlabeled patches
Evaluation patches	$N_{\mathrm{eval}} = 12{,}000$ patches
Evaluation split	$70/10/20$ train/val/test
Split policy	Source-volume split when possible; otherwise block-level split
Labels used in SSL	None
Evaluation labels	Phase masks or threshold-derived masks, when available
Structure metrics	PFE, IDE, AD, CLE, CCE, ECE, DDD

Table 7: MitoEM/EM dataset settings.

Property	Setting
Source	MitoEM human and rat cortex EM volumes
Volume size	$1000 \times 4096 \times 4096$ voxels per volume
Voxel resolution	$30 \times 8 \times 8$ nm
Primary use	Mitochondria / membrane-rich EM segmentation
Crop size	512×512
Pretraining crops	$N_{\mathrm{pre}} = 51{,}200$ unlabeled crops
Evaluation crops	$N_{\mathrm{eval}} = 12{,}800$ labeled crops
Evaluation split	Official train/validation split; test evaluated only if challenge-server labels are used
Labels used in SSL	None
Evaluation labels	Mitochondria instance masks
Structure metrics	IDE, AD, CCE, ECE, DDD

For self-supervised pretraining, we use 2D EM slices from the training portions of the human and rat volumes and extract non-overlapping 512×512 crops. Since each 4096×4096 slice yields $8 \times 8 = 64$ crops, the two 400-slice training partitions provide $N_{\mathrm{pre}} = 51{,}200$ unlabeled crops for self-supervised pretraining. Mitochondria masks are not used during self-supervised pretraining.

For downstream evaluation, we use labeled crops from the official validation partitions. The two 100-slice validation partitions provide $N_{\mathrm{eval}} = 12{,}800$ labeled 512×512 crops. Crops are sampled only after assigning each slice or volume block to train, validation, or test, and no augmented crop from the same spatial location appears in more than one split. When challenge-server evaluation is used, the held-out test partitions are evaluated separately and are not counted as locally labeled evaluation crops.

We evaluate segmentation performance using Dice, mIoU, boundary F1, and object-level average precision. We evaluate structural preservation using interface-density error, anisotropy discrepancy, connected-component error, Euler-characteristic error, and descriptor-distribution distance.

E.4 HuBMAP-Vasc dataset

HuBMAP-Vasc is based on the HuBMAP–Hacking the Human Vasculature dataset, which contains 2D PAS-stained human kidney histology tiles with microvasculature annotations [18, 40]. The public dataset contains $7{,}033$ 512×512 PAS-stained TIFF tiles from healthy human kidney tissue, of which $1{,}633$ tiles include polygon annotations for microvascular structures. We use this dataset to test whether MICROFIELD improves transfer to thin connected structures, luminal spaces, tissue morphology, and anisotropic histology texture. Table 8 lists the dataset settings.

For self-supervised pretraining, we use all $7{,}033$ image tiles as unlabeled data and ignore all annotations. Vessel annotations are used only for downstream segmentation evaluation and for structure-sensitive metrics. For downstream evaluation, we use the $1{,}633$ annotated tiles and split them into $1{,}143/163/327$ train/validation/test tiles under the $70/10/20$ split. When whole-slide identifiers are available, we split by slide; otherwise, we use a fixed tile-level split with no augmented views shared across partitions.

Table 8: HuBMAP-Vasc dataset settings.

Property	Setting
Source	HuBMAP–Hacking the Human Vasculature
Image type	PAS-stained human kidney histology tiles
Tile size	512×512
Primary use	Microvasculature segmentation and structure-sensitive transfer
Pretraining tiles	$N_{\mathrm{pre}} = 7{,}033$ unlabeled tiles
Evaluation tiles	$N_{\mathrm{eval}} = 1{,}633$ annotated tiles
Evaluation split	$70/10/20$ train/val/test over annotated tiles
Labels used in SSL	None
Evaluation labels	Microvasculature polygon / instance masks
Structure metrics	IDE, AD, CCE, ECE, DDD

Table 9: Default pretraining configuration. The component weights $(\lambda_{\mathrm{local}}, \lambda_{\mathrm{dist}}, \lambda_{\mathrm{trans}}) = (1.0, 0.5, 0.5)$ multiply the normalized loss components and specify their relative contributions; they need not sum to one. The overall coefficient $\lambda_{\mathrm{field}} = 0.1$ scales the resulting weighted sum relative to the base loss.

Setting	Value
Model and base objective	ViT-B/16 at 512×512; normalized-pixel MAE
Optimizer	AdamW, $(\beta_1, \beta_2) = (0.9, 0.95)$; weight decay 0.05
Learning rate and batch	Base LR 1.5×10^{-4}, scaled to 1.2×10^{-3}; global batch 2,048
Schedule	800 epochs; 40 warm-up epochs; cosine decay
Masking and augmentation	75% masking; random resized crop; horizontal and vertical flips
Loss weights	$(\lambda_{\mathrm{field}}, \lambda_{\mathrm{local}}, \lambda_{\mathrm{dist}}, \lambda_{\mathrm{trans}}) = (0.1, 1.0, 0.5, 0.5)$, applied after normalization
Uniform field	8×8 grid
Hierarchical field	Quadtree depth 4; minimum 32×32 region; 64 leaves
Distribution term	Mean–covariance matching
Descriptor settings	16 histogram bins; wavelet scales $\{1, 2, 4\}$
Descriptor normalization	Per-dim. training-set μ and σ, estimated once from target regions and held
Loss normalization	EMA magnitude norm. of the three loss terms with $\rho = 0.99$ and $\epsilon = 10^{-8}$
Replicates	Three complete runs for main results

We evaluate vessel segmentation using Dice, mIoU, boundary F1, and object-level average precision. We evaluate structural preservation using connected-component error, interface-density error, anisotropy discrepancy, Euler-characteristic error, and descriptor-distribution distance.

F Hyperparameters and Implementation Details

The proposed method adds a descriptor-field loss to a base training objective. In the self-supervised setting, the base objective is a masked image modeling loss, but the descriptor-field loss is not tied to any specific architecture. Let

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{field}} \mathcal{L}_{\mathrm{field}}, \tag{17}$$

For masked self-supervised pretraining, $\mathcal{L}_{\mathrm{base}}$ is typically a pixel, perceptual, token, or latent reconstruction loss. For supervised learning, $\mathcal{L}_{\mathrm{base}}$ is the task loss. For reconstruction and restoration, $\mathcal{L}_{\mathrm{base}}$ is the reconstruction fidelity term.

Unless otherwise stated, the descriptor-field objective is

$$\mathcal{L}_{\mathrm{field}} = \lambda_{\mathrm{local}} \mathcal{L}_{\mathrm{local}} + \lambda_{\mathrm{trans}} \mathcal{L}_{\mathrm{trans}} + \lambda_{\mathrm{dist}} \mathcal{L}_{\mathrm{dist}}, \tag{18}$$

where $\mathcal{L}_{\mathrm{trans}}$ and $\mathcal{L}_{\mathrm{dist}}$ are used when the region system is hierarchical. For a non-hierarchical region system such as a uniform patch grid, the default objective reduces to

$$\mathcal{L}_{\mathrm{field}} = \lambda_{\mathrm{local}} \mathcal{L}_{\mathrm{local}} + \lambda_{\mathrm{dist}} \mathcal{L}_{\mathrm{dist}}, \tag{19}$$

Descriptor dimensions are normalized using statistics computed once from target-image regions in the training split. Specifically,

$$\tilde{s}_{i,j}(x) = \frac{s_{i,j}(x) - \mu_j}{\sigma_j + \epsilon}, \tag{20}$$

Algorithm 1 MICROFIELD training step

Require: image x, model f_θ, masking operator M, region builder B, descriptor extractor F
Construct measurement domain $\mathcal{P} \leftarrow B(x)$
Corrupt input $x_m \leftarrow M(x)$ and predict $\hat{x} \leftarrow f_\theta(x_m)$
Compute descriptor fields $S_\mathcal{P}(x)$ and $S_\mathcal{P}(\hat{x})$
Compute $\mathcal{L}_{\text{local}}$
if $\mathcal{P}$ is hierarchical **then**
    Compute $\mathcal{L}_{\text{trans}}$ and $\mathcal{L}_{\text{dist}}$
else
    Set $\mathcal{L}_{\text{trans}} = 0$ and compute $\mathcal{L}_{\text{dist}}$ if used
end if
$\mathcal{L}_{\text{field}} = \lambda_{\text{local}}\mathcal{L}_{\text{local}} + \lambda_{\text{trans}}\mathcal{L}_{\text{trans}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}}$
Update θ using $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda_{\text{field}}\mathcal{L}_{\text{field}}$

where μ_j and σ_j are the mean and standard deviation of descriptor dimension j, estimated over target-image regions in the training split. These statistics are then held fixed throughout optimization and are applied to descriptors extracted from both the target x and the prediction $\hat{x}$. They are distinct from any exponential moving averages used to normalize loss-component magnitudes.

The loss weights can be tuned directly or normalized by exponential moving averages of the component losses:

$$\bar{\mathcal{L}}_k^{(t)} = \rho\bar{\mathcal{L}}_k^{(t-1)} + (1-\rho)\mathcal{L}_k^{(t)}. \tag{21}$$

The normalized component $\mathcal{L}_k/(\bar{\mathcal{L}}_k + \epsilon)$ reduces sensitivity to descriptor scale and makes the weighting more stable across datasets.

Algorithm 1 lists the end-to-end training step of MICROFIELD

Learnable descriptor-family weights. In addition to fixed expert-selected descriptor families, MICROFIELD can learn nonnegative weights over descriptor groups. Let $s_i^{(k)}(x)$ denote the descriptor subvector for family k, such as texture, interface, anisotropy, or phase descriptors. We define

$$L_{\text{local}} = \sum_{R_i \in P} w_i \sum_{k=1}^{K} \alpha_k \left\| \tilde{s}_i^{(k)}(x) - \tilde{s}_i^{(k)}(\hat{x}) \right\|_1, \qquad \alpha_k \geq 0.$$

The weights α_k may be fixed, tuned on validation data, or learned with sparsity or entropy regularization. This variant allows the model to adapt descriptor relevance while keeping the descriptor families interpretable.

Loss-weight selection. We will evaluate $\lambda_{\text{field}} \in \{0, 0.01, 0.03, 0.1, 0.3, 1.0\}$, where zero denotes the base-loss control. Selection uses held-out validation splits of all downstream datasets in the main evaluation, with specimen- or patient-level separation where applicable. Test results do not inform selection.

We rank configurations within each downstream task using its primary validation metric: Dice for segmentation, R^2 for regression, and the designated primary metric for classification. We average these ranks with equal weight per task.

During the initial sweep, we fix $(\lambda_{\text{local}}, \lambda_{\text{dist}}, \lambda_{\text{trans}}) = (1.0, 0.5, 0.5)$. These heuristic weights prioritize local correspondence and multiply EMA-normalized loss components. Candidates are first screened using 100-epoch pretraining runs with a common random seed and evaluation protocol. The two best nonzero candidates are then evaluated at the full pretraining budget using three seeds, together with the base-loss control. We select the candidate with the best mean validation rank across seeds, preferring the smaller λ_{field} when differences are within run-to-run variability.

To assess the component-weight heuristic, we additionally compared local-to-distribution-to-transition ratios of $1 : 0.25 : 0.25$, $1 : 0.5 : 0.5$, and $1 : 1 : 1$ in preliminary validation experiments. Each triplet is rescaled to sum to two, preserving the overall coefficient scale while varying the relative contributions. This procedure uses downstream labels for hyperparameter selection, while pretraining itself remains self-supervised. Competing methods received the same validation access and tuning budget.

Table 10: Training and inference cost. Overhead is measured relative to the corresponding Pixel-MAE baseline with the same backbone. MICROFIELD adds only training-time descriptor computation; the inference model is unchanged.

Method	Train time / epoch	Peak mem.	Inference FLOPs	Inference latency
Pixel-MAE [16]	1.00×	1.00×	1.00×	1.00×
MICROFIELD-Grid	1.11×	1.07×	1.00×	1.00×
MICROFIELD-Full	1.19×	1.09×	1.00×	1.00×

Table 11: Baseline families used in evaluation. Controlled baselines use matched in-domain pretraining data and matched backbones whenever applicable. External foundation models are reported only as reference baselines on compatible domains and are not used for causal claims.

Family	Methods	Purpose
Transfer / scratch	Scratch; ImageNet-supervised init.	Standard training and transfer references
Controlled MIM/SSL	Pixel-MAE [16]; SimMIM/Swin [47, 26]; iBOT [50]; DINOv2 [32]	Compare SSL objectives under matched data/capacity
Generic structure losses	MAE + perceptual; MAE + spectral/wavelet; MAE + global descriptors	Rule out generic non-pixel regularization
Architecture controls	Hierarchy-only MAE; MICROFIELD-Grid; MICROFIELD-Hier	Separate descriptor fields from hierarchy
MICROFIELD ablations	Local only; local + distribution; local + distribution + transition	Isolate the contribution of each descriptor-field term
Pathology / WSI FMs	UNI [6] / Phikon [11] / Virchow [44]; Prov-GigaPath [48] / HIPT [5]	External references for histology/WSI
Bio/medical FMs	CA-MAE [25]; CellSAM [28] / MedSAM [27]	External references for microscopy and segmentation
Materials / SEM SSL	MatSSL [31] / SEM SSL [35]	External references for materials microscopy

F.1 Efficiency and Implementation Cost

MICROFIELD adds training-time descriptor computation but does not change the inference model. Table 10 reports the overhead relative to Pixel-MAE. The overhead is modest because the descriptors are implemented with pooled convolutions, wavelet filters, integral-image statistics, and batched region aggregation. At inference, MICROFIELD incurs no additional cost because the descriptor-field loss and reconstruction head are discarded.

F.2 Baseline Tuning and Selection

Table 11 summarizes the role of each baseline family. The following protocol was used to tune and select the controlled baselines. DINOv2, iBOT, SimMIM, and MICROFIELD received the same search budget: eight candidate configurations were trained for 100 epochs with one seed. The selected configuration maximized the macro-average of normalized downstream validation performance, without access to test results. The selected configuration was then trained for 800 epochs and repeated using two additional seeds. All methods used the same data splits, input resolution, effective batch size, and downstream evaluation protocol. Gradient accumulation was used where necessary.

G Zero-shot Downstream Tasks Evaluation

In addition to linear probing and fine-tuning, we evaluate whether descriptor-field pretraining organizes images according to scientific structure without task-specific training. These evaluations are training-free: the pretrained encoder is frozen, no labels are used to fit a downstream classifier or regressor, and predictions are made using distances in representation space. This setting tests whether the representation itself encodes morphology, connectivity, anisotropy, interface density, and related microstructural observables.

Protocol. For each pretrained model, we extract frozen image-level embeddings by average-pooling encoder tokens. Unless otherwise stated, embeddings are ℓ_2 normalized and compared using cosine similarity. We evaluate four zero-shot tasks.

- **Structure retrieval.** Given a query patch, retrieve the top-k nearest patches in representation space. Relevance is defined by descriptor similarity in phase fraction, interface density, anisotropy, correlation length, and topology-sensitive metrics where available.
- **Morphology ranking.** Given a query and two candidate patches, the model is correct if the candidate closer in representation space is also closer in a target morphology variable, such as porosity, interface density, anisotropy, or characteristic length scale.

Table 12: Zero-shot structure-sensitive evaluation with frozen representations. Retrieval and ranking are performed directly in representation space, without fitting a downstream classifier or regressor. MICROFIELD improves structure retrieval, morphology ranking, and nearest-neighbor morphology estimation, indicating that descriptor-field supervision organizes embeddings by scientific structure rather than only pixel appearance.

Method	Struct. R@10 ↑	Rank acc. ↑	NN morph. err. ↓	Shift R@10 ↑
Pixel-MAE	42.8	67.1	0.318	35.4
SimMIM/Swin	44.6	68.4	0.302	37.1
iBOT	46.9	70.2	0.286	39.5
DINOv2	48.1	71.0	0.279	40.6
Hierarchy-only	50.3	72.4	0.264	42.7
MICROFIELD-Grid	57.8	77.9	0.213	50.8
MICROFIELD-Full	**61.4**	**80.6**	**0.188**	**54.2**

- **Nearest-neighbor transfer.** We estimate continuous morphology variables by averaging the descriptor values of the k nearest neighbors in representation space. For segmentation-derived quantities, labels or masks are used only to compute the evaluation target, not to train the encoder.
- **Domain-shift retrieval.** We evaluate whether a query retrieves structurally similar samples under nuisance changes such as contrast, blur, acquisition noise, or resolution changes. This tests whether the representation is organized by structure rather than low-level appearance.

We report retrieval Recall@10, pairwise ranking accuracy, nearest-neighbor morphology error, and domain-shift Recall@10. For the morphology error, values are normalized by the standard deviation of the corresponding target variable within each dataset and then averaged across variables. Lower is better.

Table 12 shows that MICROFIELD improves all zero-shot structure-sensitive metrics. The gains are largest for structure retrieval and nearest-neighbor morphology estimation, where success depends directly on whether embeddings preserve descriptor-level organization. MICROFIELD-Grid already outperforms hierarchy-only pretraining, showing that the improvement is not explained merely by multiscale tokenization or hierarchical processing. MICROFIELD-Full further improves over the grid variant, consistent with the main results showing that cross-scale descriptor transition preservation is useful when scientific structure is organized hierarchically.

These results support the same conclusion as the supervised downstream evaluation: descriptor-field pretraining improves the scientific organization of the learned representation. Importantly, the zero-shot setting removes task-specific optimization as an explanation. Since no downstream model is trained, improved retrieval and ranking indicate that the encoder embedding space itself better reflects microstructure. This is especially visible under domain-shift retrieval, where MICROFIELD is less sensitive to contrast and acquisition changes than pixel-centric pretraining methods.

Limitations. Zero-shot retrieval is not a replacement for downstream task evaluation. It depends on the choice of descriptor-defined relevance, and some scientific tasks may require features that are not captured by the selected descriptor families. We therefore treat these results as complementary evidence: they test whether the representation space is structurally organized, while the main paper evaluates whether that organization improves task performance after probing or fine-tuning.

H Additional Qualitative Results

Figure 3 provides additional qualitative evidence for the failure mode discussed in Section 4. Across electron microscopy, histology, and porous XCT, Pixel-MAE produces reconstructions that remain close to the target under pixel-level criteria but alter structures that are scientifically meaningful in each domain. In electron microscopy, membranes and fine interfaces become fragmented or blurred; in histology, luminal spaces and tissue boundaries lose connectivity and sharpness; and in porous XCT, grain boundaries and pore-like structures are smoothed or broken.

MICROFIELD produces reconstructions with comparable pixel fidelity but visibly lower descriptor-field error. The descriptor-error maps emphasize structural discrepancies rather than raw intensity

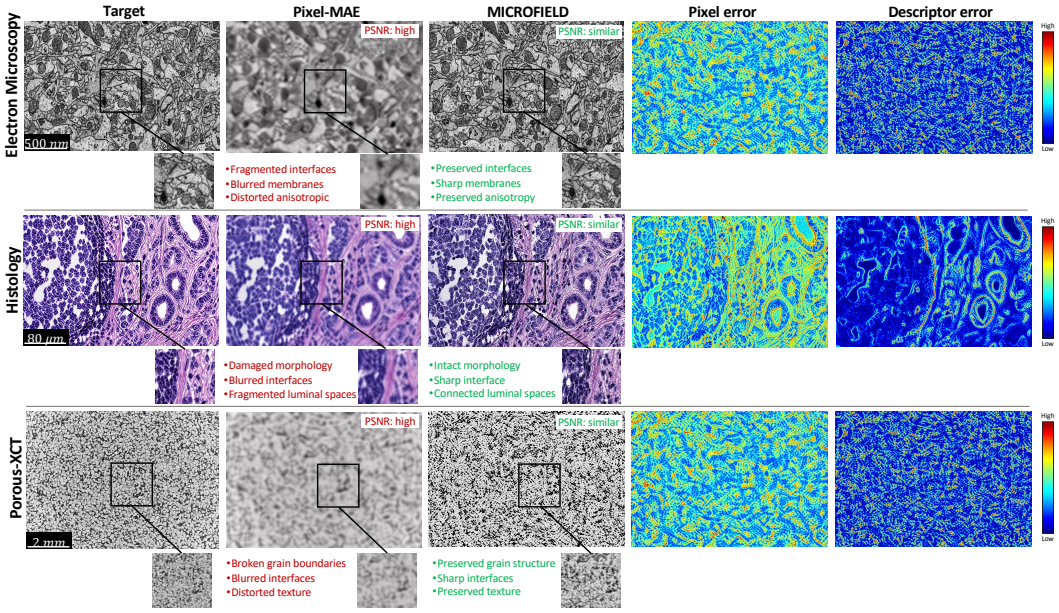


Figure 3: Additional qualitative evidence across scientific imaging domains. For electron microscopy, histology, and porous XCT, Pixel-MAE reconstructions can retain high pixel fidelity while altering scientifically relevant structure, including interfaces, membranes, luminal spaces, anisotropic texture, and grain boundaries. MICROFIELD achieves comparable pixel fidelity but lower descriptor-field error, preserving the local structural observables targeted by descriptor-field supervision. Pixel-error maps show intensity-level discrepancies, whereas descriptor-error maps highlight structural mismatches that are more aligned with morphology, connectivity, anisotropy, and interface preservation.

differences, highlighting failures in morphology, interface organization, anisotropy, and connectivity that are weakly captured by pixel-error maps. These examples support the central motivation of MICROFIELD: for scientific images, reconstruction quality should not be judged only by pixel similarity, because similar PSNR can correspond to substantially different preservation of scientific observables.

I Positioning Relative to Structure-Aware Objectives

This appendix positions MICROFIELD relative to existing objectives that move beyond pixel-wise supervision. The goal is not to claim that descriptor-based or structure-aware losses are new in isolation. Rather, the distinction is where the structural measurements are taken, how they are indexed, and what role they play during learning.

Pixel, perceptual, and spectral losses. Pixel reconstruction losses supervise pointwise intensity agreement. Perceptual losses compare activations in a pretrained network, while style and spectral losses match generic feature, texture, or frequency statistics. These objectives are useful because they define supervision outside raw pixel space, but their target spaces are not usually scientific observables. A low pixel, perceptual, or spectral loss does not necessarily imply preservation of pore connectivity, interface density, anisotropy, phase organization, or topology. MICROFIELD instead defines the supervision target as a field of domain-relevant structural measurements.

Boundary, topology, connectivity, and shape losses. Boundary losses, topology-aware losses, skeleton losses such as clDice, and global shape-descriptor losses provide important precedents for using non-pixel structural measurements as supervision. Their scope, however, is usually output-specific: they are designed for a particular supervised prediction target, such as a segmentation mask, vessel tree, object boundary, or global shape constraint. MICROFIELD is different in two ways. First, it is used as a training-time supervision layer for representation learning, including self-supervised pretraining. Second, it does not define a single structural constraint; it constructs a spatially

indexed descriptor field and can combine texture, interface, anisotropy, phase, and topology-related observables depending on the scientific domain.

Microstructure reconstruction and descriptor matching. Classical and modern microstructure reconstruction methods optimize images, random fields, or generators so that their global or local statistics match those of a target microstructure. These methods are close in spirit because they treat scientific descriptors as meaningful observables. The difference is the learning role. Microstructure reconstruction usually seeks to synthesize or recover an image whose statistics match a target. MICROFIELD instead uses descriptors as auxiliary supervision for a neural model whose representation remains learned and whose downstream use is not restricted to descriptor vectors. The descriptors are not the final representation; they are measurements of structural fidelity during training.

Architecture-centric hierarchy. Many high-resolution image models use hierarchy to reduce computational cost: windowed attention, patch aggregation, graph pooling, quadtree or octree tokenization, and adaptive patching. In these methods, hierarchy is primarily an architectural or tokenization mechanism. MICROFIELD uses hierarchy differently. When a hierarchy is available, it becomes a multiscale measurement domain on which descriptor fields are computed. The method does not require hierarchy: the same local descriptor-field objective can be applied on a uniform grid. Conversely, when a hierarchy is available, MICROFIELD can additionally preserve parent–child descriptor transitions and level-wise descriptor distributions.

Physics-informed and knowledge-constrained learning. Physics-informed neural networks impose governing equations, residuals, boundary conditions, or conservation laws. MICROFIELD is weaker but more broadly applicable in imaging settings where the image is not itself a PDE state variable, or where the governing equations are unknown, unavailable, or not differentiable in the observed image variables. It preserves observable structural quantities that are known to matter scientifically. In this sense, MICROFIELD is closer to a knowledge-constrained objective than to a handcrafted-feature pipeline: the neural representation remains learned, while descriptors specify which structural observables should be preserved.

Table 13: Positioning of MICROFIELD relative to structure-aware objectives. The key distinction is not whether structural measurements are used, but whether they form a spatially indexed supervision field for representation learning.

Method family	Supervision target	Typical role	Difference from MICROFIELD
Pixel reconstruction	Pointwise intensity or patch values	Reconstruction or masked image modeling	Does not directly preserve scientific observables such as connectivity, anisotropy, or interface geometry.
Perceptual / feature losses	Activations of pretrained networks	Generic visual fidelity or restoration	Targets generic feature similarity rather than domain-relevant microstructure measurements.
Style / texture / spectral losses	Gram statistics, frequency content, wavelet or spectral energy	Texture or multiscale appearance fidelity	Captures generic texture statistics, but usually not spatially indexed scientific descriptors or topology-sensitive structure.
Boundary / shape losses	Contours, distance transforms, global shape descriptors	Supervised segmentation or weakly supervised shape control	Usually tied to a specific output mask or shape target rather than self-supervised representation learning.
Topology / connectivity losses	Persistent homology, skeleton overlap, connectedness penalties	Preserve topology in segmentation outputs	Strong for specific topological constraints, but not a general descriptor-field supervision space.
Microstructure reconstruction	Global or local material descriptors	Synthesize or reconstruct microstructures matching target statistics	Optimizes images or generators; MICROFIELD uses descriptors as auxiliary training-time supervision for learned representations.
Hierarchical high-resolution models	Tokens, windows, regions, graphs, quadtrees, or octrees	Reduce sequence length or aggregate information	Hierarchy is architectural; in MICROFIELD, hierarchy is optional and serves as a multiscale measurement domain.
MICROFIELD	Spatially indexed microstructure descriptor fields	Objective-level supervision for representation learning, reconstruction, or supervised learning	Preserves local descriptors, level-wise descriptor distributions, and optional cross-scale descriptor transitions while leaving the base architecture unchanged.

Summary. The closest prior lines are structure-aware losses and microstructure reconstruction. MICROFIELD combines their central insight that scientific structure should be measured through meaningful observables, with a different learning interface: descriptors are organized as a spatial field and used as training-time supervision for neural representation learning. This makes MICROFIELD complementary to high-resolution architectures and applicable to self-supervised, supervised, and reconstruction settings.

J Region-System Construction Efficiency

Let $\mathcal{P}(x) = \{R_i\}_{i=1}^N$ denote a region system. It may be a uniform grid, a salience-adaptive partition, a quadtree, an octree, a learned partition, or a hierarchy produced by an existing high-resolution image model. The cost of constructing $\mathcal{P}(x)$ depends on the region-system choice.

For a uniform grid, construction is negligible. For a quadtree or octree, construction is typically dominated by computing a salience score such as local variance, gradient magnitude, entropy, or wavelet energy. If salience scores are computed using integral images or pooled feature maps, construction is approximately linear in the number of pixels. For learned region systems, construction cost depends on the tokenizer or partitioning model and should be reported separately from descriptor computation.

In the main method, the same region system is used to evaluate the target and the prediction. This avoids solving a matching problem between different partitions.

K Additional Methodological Details

This appendix provides details omitted from the main methodology for space. These details are not required to understand the core MICROFIELD objective, but specify practical choices for constructing region systems, computing descriptors, defining distributional losses, and implementing the training objective.

K.1 Region-System Construction

MICROFIELD can operate on any spatial measurement domain $\mathcal{P}(x) = \{R_i\}_{i=1}^N$, including uniform grids, adaptive partitions, quadtrees, octrees, learned tokenizers, superpixels, segmentation maps, or domain-specific region graphs. When a hierarchical measurement domain is used, we write it as a rooted tree $\mathcal{T}(x)$, where each node $n \in \mathcal{N}(\mathcal{T})$ corresponds to a spatial region R_n. The root covers the full image, internal nodes partition into children, and $\mathcal{N}_\ell$ denotes the nodes at level ℓ.

A practical hierarchy builder recursively refines regions according to a salience functional $V(R; x)$. The salience function may be based on local variance, entropy, edge density, gradient magnitude, wavelet energy, or a learned score. In deterministic refinement, the next region to split is

$$R^* = \arg\max_{R \in \mathcal{A}} V(R; x), \tag{22}$$

where $\mathcal{A}$ is the active leaf set. A stochastic variant samples active leaves according to

$$p(R \mid \mathcal{A}, \tau) = \frac{\exp(V(R; x)/\tau)}{\sum_{Q \in \mathcal{A}} \exp(V(Q; x)/\tau)}, \tag{23}$$

where τ controls the sharpness of the selection. As $\tau \to 0$, the sampling approaches deterministic refinement; as $\tau \to \infty$, it approaches uniform random refinement. The hierarchy is refined until a maximum depth, minimum region size, or fixed leaf budget is reached.

Unless otherwise stated, the region system is constructed from the target image x and then reused for both x and the model output $\hat{x}$. This gives one-to-one correspondence between target and predicted regions and avoids solving a matching problem between two independently constructed partitions.

K.2 Descriptor Definitions

The descriptor family should be compact, stable, scientifically interpretable, and sensitive to microstructure changes that matter for downstream tasks. MICROFIELD is not tied to a single descriptor set; the following families are the default choices used in this paper.

Signal and texture descriptors. These descriptors require no segmentation and apply broadly across microscopy, XCT, pathology, and materials images. For a grayscale channel and region R_i, we compute the local mean,

$$\mu_i = \frac{1}{|R_i|} \sum_{p \in R_i} x(p), \tag{24}$$

local variance,

$$\sigma_i^2 = \frac{1}{|R_i|} \sum_{p \in R_i} (x(p) - \mu_i)^2, \tag{25}$$

intensity entropy,

$$H_i = -\sum_{b=1}^{B} p_b^{(i)} \log(p_b^{(i)} + \epsilon), \tag{26}$$

where during training, $p_b^{(i)}$ is computed using differentiable soft binning, in which each pixel contributes to neighboring fixed bin centers with triangular-kernel weights; hard histogram assignments are used only for evaluation.

band-pass or wavelet energy,

$$w_{ik} = \frac{1}{|R_i|} \sum_{p \in R_i} |(\psi_k * x)(p)|^2, \tag{27}$$

where $p_b^{(i)}$ is the intensity-histogram probability for bin b in region R_i, and ψ_k is a wavelet or band-pass filter. These descriptors measure local contrast, intensity diversity, heterogeneity, and multiscale texture energy.

Interface and anisotropy descriptors. Interfaces, boundaries, and directional organization are often more important than absolute intensity in scientific images. We compute gradient density,

$$g_i = \frac{1}{|R_i|} \sum_{p \in R_i} \|\nabla x(p)\|, \tag{28}$$

and Laplacian energy,

$$e_{\Delta,i} = \frac{1}{|R_i|} \sum_{p \in R_i} |\Delta x(p)|. \tag{29}$$

To measure local anisotropy, we compute the structure tensor

$$J_i = \sum_{p \in R_i} \begin{bmatrix} I_x(p)^2 & I_x(p)I_y(p) \\ I_x(p)I_y(p) & I_y(p)^2 \end{bmatrix}. \tag{30}$$

Let $\lambda_1(i) \geq \lambda_2(i) \geq 0$ be its eigenvalues. The anisotropy score is

$$a_i = \frac{\lambda_1(i) - \lambda_2(i)}{\lambda_1(i) + \lambda_2(i) + \epsilon}. \tag{31}$$

These descriptors capture boundary density, local orientation, and directional organization.

Phase and topology descriptors. When a reliable phase, tissue, object, or pore mask is available, richer morphology-aware descriptors can be included. Let $Q(p) \in \{0, 1\}$ denote a binary phase mask. For region R_i, define the phase fraction

$$\phi_i = \frac{1}{|R_i|} \sum_{p \in R_i} Q(p), \tag{32}$$

the connected-component count

$$N_{c,i} = \text{number of connected components in } Q|_{R_i}, \tag{33}$$

the normalized interface length

$$\rho_{I,i} = \frac{\text{interface length within } R_i}{|R_i|}, \tag{34}$$

and the Euler characteristic

$$\chi_i = \text{Euler characteristic of } Q|_{R_i}. \tag{35}$$

For multiclass phase maps, ϕ_i is replaced by class-wise phase fractions. These descriptors are useful for porous XCT, materials microstructure, biological microscopy, and pathology when approximate object or phase masks are available.

Descriptor normalization. Descriptor dimensions have heterogeneous scales. We normalize each descriptor dimension using training-set statistics estimated once and held fixed

$$\tilde{s}_{ij}(x) = \frac{s_{ij}(x) - \mu_j}{\sigma_j + \epsilon}, \tag{36}$$

where μ_j and σ_j are running estimates for descriptor dimension j. Unless otherwise stated, all descriptor-field losses use normalized descriptors.

K.3 Distributional Descriptor Losses

The level-wise distribution term in the main paper is written generally as

$$\mathcal{L}_{\text{dist}}(x, \hat{x}) = \sum_{\ell} D\left(\mathcal{S}_\ell(x), \mathcal{S}_\ell(\hat{x})\right), \tag{37}$$

where $\mathcal{S}_\ell(x) = \{\tilde{s}(n; x) : n \in \mathcal{N}_\ell\}$ is the multiset of normalized descriptors at level ℓ. The discrepancy D can be chosen according to the desired accuracy–cost tradeoff.

A simple and efficient choice is moment matching:

$$D_{\text{mom}}\left(\mathcal{S}_\ell(x), \mathcal{S}_\ell(\hat{x})\right) = \|\mu_\ell(x) - \mu_\ell(\hat{x})\|_2^2 + \|\Sigma_\ell(x) - \Sigma_\ell(\hat{x})\|_F^2, \tag{38}$$

where μ_ℓ and Σ_ℓ are the descriptor mean and covariance at level ℓ.

A richer nonparametric alternative is the squared maximum mean discrepancy:

$$D^2_{\text{MMD}}(A, B) = \frac{1}{|A|^2} \sum_{a,a' \in A} k(a, a') + \frac{1}{|B|^2} \sum_{b,b' \in B} k(b, b') - \frac{2}{|A||B|} \sum_{a \in A, b \in B} k(a, b), \tag{39}$$

where k is a positive-definite kernel. Moment matching is inexpensive and stable, while MMD captures higher-order discrepancies in descriptor distributions. Other choices, such as sliced Wasserstein distance or histogram distances, can also be used when appropriate.

K.4 Implementation Details

Region correspondence. The default implementation constructs the region system from the target image and evaluates both the target x and model output $\hat{x}$ on this same region system. This ensures exact descriptor correspondence. Constructing separate region systems for x and $\hat{x}$ is possible, but would require solving a matching problem between regions and is not used in the main method.

Cross-scale aggregation. For the cross-scale transition loss in Eqs. (10) and (11), the reported implementation uses the same componentwise area-weighted aggregation operator for all default descriptor dimensions. Specifically, for an internal node p with children $\text{ch}(p)$, we define

$$A_p(\{\tilde{s}(c; x) \,|\, c \in \text{ch}(p)\}) = \sum_{c \in \text{ch}(p)} \alpha_{p,c}\, \tilde{s}(c; x), \qquad \alpha_{p,c} = \frac{|R_c|}{\sum_{c' \in \text{ch}(p)} |R_{c'}|}. \tag{40}$$

Because the children form a disjoint partition of the parent region, $\sum_{c' \in \text{ch}(p)} |R_{c'}| = |R_p|$. For equal-size quadtree children, Eq. (40) reduces to the arithmetic mean of the child descriptor vectors.

The operator is applied componentwise to the normalized descriptors used in the loss. In the reported experiments, this includes the local intensity mean, variance, histogram-derived entropy, gradient density, Laplacian energy, wavelet-energy components at each selected scale, and structure-tensor

anisotropy. We do not use a learned aggregation operator or a separately fitted operator for different descriptor families.

For additive or area-normalized descriptors, such as the local intensity mean, the area-weighted child aggregation has a direct composition interpretation. For nonlinear descriptors, such as variance, entropy, wavelet energy, and structure-tensor anisotropy, the aggregated child value need not equal the descriptor computed directly on the parent. This difference is intentional: the residual

$$r(p; x) = \tilde{s}(p; x) - A_p(\{\tilde{s}(c; x) \mid c \in \mathrm{ch}(p)\}) \tag{41}$$

measures the descriptor-specific information present at the parent scale that is not explained by a simple composition of its children. The transition loss preserves this residual between the target and the reconstruction.

Differentiability. Descriptors computed from differentiable image operations, such as moments, gradients, wavelet energy, and structure tensors, allow gradients to flow through $\hat{x}$. Non-differentiable descriptors, such as connected-component counts or Euler characteristic, can be used as evaluation metrics, through differentiable surrogates, or with stop-gradient approximations. In the main training objective, we use differentiable descriptors and reserve non-differentiable topology statistics for evaluation or optional domain-specific variants.

Loss scaling. Because descriptor losses may have different magnitudes from the base objective, we normalize descriptor dimensions and tune λ_{field} on validation data. In our experiments, we additionally normalize component losses using running magnitudes

$$\widehat{\mathcal{L}}_k = \frac{\mathcal{L}_k}{\mathrm{EMA}(\mathcal{L}_k) + \epsilon}. \tag{42}$$

where $k \in \{\mathrm{local}, \mathrm{trans}, \mathrm{dist}\}$ and EMA denotes an exponential moving average.

Computational cost. Let N be the number of regions and M the number of hierarchy nodes. For balanced trees, $M = O(N)$. Most signal, interface, histogram, and tensor descriptors can be computed in time linear in the number of pixels plus linear in the number of regions using pooled convolutions, integral images, histograms, or batched tensor operations. Distribution alignment is the most expensive term if exact optimal transport is used; moment matching, MMD, sliced Wasserstein distance, or histogram distances provide more efficient alternatives. For large images, node sampling can reduce overhead by evaluating descriptor losses on a stratified subset of regions per iteration.

At inference time, MICROFIELD introduces no additional model-side computation. The descriptor-field losses are disabled because there is no target-output comparison or gradient update. Descriptor extraction is only needed after training if the same quantities are reported as evaluation metrics or used in downstream scientific analysis.

L Additional Rationale and Theoretical Interpretation

This section clarifies why descriptor-field supervision is better aligned with scientific images than pixel reconstruction alone. The argument is not that pixel losses are useless: local appearance fidelity remains important, and in our default instantiation it is retained through $\mathcal{L}_{\mathrm{base}}$. The claim is narrower and more precise: for many scientific images, the geometry induced by pixel losses is poorly matched to the geometry of scientific equivalence. Descriptor fields provide an intermediate supervision space between raw pixels and full physical models: they preserve scientifically meaningful observables without requiring governing equations.

L.1 A task-conditioned descriptor pseudometric

A descriptor extractor F maps a region R_i to a vector of observables

$$s_i(x) = F(x|_{R_i}). \tag{43}$$

This map is intentionally many-to-one: many different pixel configurations can have the same phase fraction, interface density, anisotropy, correlation length, or topology. Thus, descriptor-field

supervision does not ask the model to reproduce one exact raster. It asks the model to remain within an equivalence class of images that preserve selected scientific observables.

For a fixed indexed region system $\mathcal{P}$ shared by both images, nonnegative fixed weights w_i, and a fixed descriptor extractor F, the local descriptor discrepancy induces the pseudometric

$$d_{\mathcal{P}}(x, \hat{x}) = \left(\sum_{R_i \in \mathcal{P}} w_i \| s_i(x) - s_i(\hat{x}) \|_2^2 \right)^{1/2} . \tag{44}$$

This pseudometric interpretation applies only when $\mathcal{P}$ is fixed. In the adaptive implementation, the partition may be constructed from the target x and reused for $\hat{x}$. The resulting training loss is then a target-conditioned discrepancy and need not be symmetric over arbitrary pairs of images. This is a pseudometric rather than a metric because $d_{\mathcal{P}}(x, \hat{x}) = 0$ does not imply $x = \hat{x}$ pixelwise. This is a feature, not a defect. The null space of $d_{\mathcal{P}}$ contains pixel variations that do not alter the chosen scientific observables. In contrast, pixel loss has a trivial null space: it is minimized only by exact pointwise agreement.

This quotient view explains why descriptor-field supervision is useful for self-supervised learning. Pixel reconstruction forces the model to spend capacity on recovering all visible details, including nuisance variation. Descriptor-field supervision biases the representation toward preserving the invariants that downstream scientific tasks are expected to depend on.

L.2 A structural sufficiency condition

Descriptor-field supervision is useful when the downstream structural quantity is approximately determined by the descriptor field. Let

$$S_{\mathcal{P}}(x) = ((R_i, s_i(x)))_{i=1}^{N} . \tag{45}$$

denote the descriptor field. Suppose there exists a map G such that $G(S_{\mathcal{P}}(x))$ approximates the task-relevant structure $\Psi(x)$. This is not an arbitrary assumption: it is precisely the modeling assumption behind classical microstructure characterization, where physical or biological structure is summarized by local composition, interfaces, connectivity, anisotropy, and scale statistics.

Proposition 1 (Descriptor-field preservation controls task-relevant structural error)**.** *Let G be L_G-Lipschitz with respect to a vectorized representation of $S_{\mathcal{P}}$. Then for any x and $\hat{x}$,*

$$\begin{aligned} \|\Psi(x) - \Psi(\hat{x})\| \le \; & \underbrace{\|\Psi(x) - G(S_{\mathcal{P}}(x))\|}_{\text{descriptor sufficiency error for } x} + \underbrace{L_G \|S_{\mathcal{P}}(x) - S_{\mathcal{P}}(\hat{x})\|}_{\text{descriptor-field mismatch}} \\ & + \underbrace{\|G(S_{\mathcal{P}}(\hat{x})) - \Psi(\hat{x})\|}_{\text{descriptor sufficiency error for } \hat{x}} . \end{aligned} \tag{46}$$

Proof. Add and subtract $G(S_{\mathcal{P}}(x))$ and $G(S_{\mathcal{P}}(\hat{x}))$, apply the triangle inequality, and use the Lipschitz property of G. □

The proposition separates two issues. The first and third terms measure whether the chosen descriptors are sufficient for the scientific task. The middle term is the quantity optimized by descriptor-field supervision. Thus, the method is useful exactly when the descriptor family captures the structural aspects needed by the downstream task. It also clarifies a failure mode: if the descriptor family omits the relevant structure, minimizing descriptor loss alone cannot recover it.

L.3 Stability to nuisance variation and sensitivity to structural change

Descriptor-field supervision also changes the sensitivity profile of the loss. Pixel losses assign penalty to every pointwise intensity difference. Descriptor losses respond to differences only insofar as they alter the chosen observables. If a descriptor extractor is locally Lipschitz on region R_i,

$$\|F(x|_{R_i}) - F(y|_{R_i})\| \le L_F(R_i) \|x|_{R_i} - y|_{R_i}\|, \tag{47}$$

then small nuisance perturbations produce bounded descriptor changes. More importantly, different descriptors amplify different structural changes: interface density responds to boundary fragmentation,

anisotropy responds to orientation distortion, topology responds to component merging or splitting, and correlation length responds to changes in characteristic scale.

Thus, descriptor-field supervision is not merely a smoothed pixel loss. It reweights the training signal toward scientifically meaningful failure modes:

- preserving phase amount without preserving interface complexity is insufficient;
- preserving texture without preserving connectivity is insufficient;
- preserving local patches without preserving cross-scale organization is insufficient;
- preserving global statistics without preserving spatial heterogeneity is insufficient.

L.4 Why hierarchy can matter

A uniform grid descriptor field captures local structure, but it does not explicitly represent how fine-scale structures compose into coarse-scale morphology. This matters because many scientific structures are defined by scale coupling. Pore networks, cracks, grains, vascular structures, glands, and fibrous tissues are not merely collections of local patches; their scientific meaning depends on how small components organize into larger connected or oriented structures.

When $\mathcal{P}$ is a hierarchy $\mathcal{T}$, each parent-child relation provides a scale transition. For an internal node n, define

$$r(n;x) = \tilde{s}(n;x) - A_n(\{\tilde{s}(c;x) \,|\, c \in \mathrm{ch}(n)\}), \tag{48}$$

where A aggregates child descriptors. The residual $r(n;x)$ measures the part of the parent descriptor not explained by simple child aggregation. Matching this residual,

$$\mathcal{L}_{\mathrm{trans}}(x,\hat{x}) = \frac{1}{|\mathcal{N}_{\mathrm{int}}(T)|} \sum_{n \in \mathcal{N}_{\mathrm{int}}(T)} \|r(n;x) - r(n;\hat{x})\|_2^2, \tag{49}$$

preserves cross-scale transition behavior.

For additive descriptors, this has a conservation interpretation. For example, phase fraction satisfies

$$\phi_n = \sum_{c \in \mathrm{ch}(n)} \frac{|R_c|}{|R_n|} \phi_c, \tag{50}$$

so cross-scale consistency preserves local composition across refinement. For non-additive descriptors such as anisotropy, interface density, correlation length, and topology, the residual is even more informative: it captures the failure of coarse structure to be reduced to a simple average of fine structures. This is precisely where multiscale morphology lives.

L.5 Why a field rather than global descriptors

A global descriptor vector can match image-level statistics while destroying local organization. For example, two images can have the same global phase fraction and interface density but differ in whether pores are clustered or uniformly distributed, whether fibers align locally or globally, or whether tissue compartments form coherent structures. Scientific images are often heterogeneous: the spatial distribution of microstructure matters.

Descriptor fields retain spatial indexing. For a region system $\mathcal{P} = \{R_i\}_{i=1}^N$, the descriptor field is the indexed collection

$$S_{\mathcal{P}}(x) = ((R_i, s_i(x)))_{i=1}^N, \tag{51}$$

This preserves where structural measurements occur, not only their global averages. The local loss preserves descriptor correspondence between matching regions

$$\mathcal{L}_{\mathrm{local}}(x,\hat{x}) = \frac{1}{d} \sum_{i=1}^N w_i \, \|\tilde{s}_i(x) - \tilde{s}_i(\hat{x})\|_2^2, \qquad w_i \geq 0, \qquad \sum_{i=1}^N w_i = 1, \tag{52}$$

The distributional loss intentionally discards region correspondence and instead compares the multi-sets of normalized regional descriptors

$$\mathcal{L}_{\mathrm{dist}}(x,\hat{x}) = D\Big(\{\tilde{s}_i(x)\}_{i=1}^N, \{\tilde{s}_i(\hat{x})\}_{i=1}^N\Big), \tag{53}$$

Table 14: Practical guidance for selecting descriptor families. The recommended descriptors are starting points rather than universal choices and should be validated using independent downstream or structural metrics. Phase-derived descriptors require masks or soft phase estimates and are therefore optional weak-mask extensions. Topology descriptors are used only for evaluation in our default experiments, although differentiable surrogates may be used in domain-specific extensions.

Scientific objective	**Recommended starting descriptors**	**Mode**	**Principal limitation**
Phase amount or porosity	Local phase fraction or composition	Weak-mask	Preserves the amount of each phase but not its spatial arrangement.
Membranes and interfaces	Gradient density and interface or Laplacian energy	Unlabeled	Captures boundary content but does not fully determine connectivity.
Oriented structures	Structure-tensor anisotropy and dominant orientation	Unlabeled	Captures directional organization but not phase amount or topology.
Characteristic feature scale	Multiscale wavelet energy or autocorrelation length	Unlabeled	Captures dominant spatial scale but does not determine connectivity.
Connectivity and organization	Interface descriptors combined with coarse-scale topology evaluation or a differentiable topology surrogate	Optional or evaluation-only	Fine-scale topology is sensitive to noise and threshold selection.
Spatial heterogeneity	Local descriptor fields combined with level-wise distribution alignment	Unlabeled or weak-mask	Distribution matching can preserve regional statistics while misplacing individual structures.

We use $w_i = 1/N$ by default. Thus, the local term preserves where descriptor values occur, whereas the distributional term preserves their regional statistics without requiring exact spatial correspondence.

L.6 Descriptor choice as a structural prior

MICROFIELD does not assume that a universal descriptor set is optimal. The descriptor family should be viewed as a task-conditioned structural prior: descriptors are useful when they are stable, non-redundant, and predictive of the scientific quantity of interest.

Let $\Psi(x)$ denote a downstream structural quantity, such as pore connectivity, anisotropy, interface density, or tissue organization, and let $S_P(x)$ be the descriptor field induced by region system P. Descriptor-field supervision is useful when $\Psi(x)$ is approximately determined by $S_P(x)$; that is, when there exists a map G such that $G(S_P(x)) \approx \Psi(x)$. Then the error in preserving Ψ can be decomposed into descriptor-sufficiency error and descriptor-field mismatch. This clarifies both the value and the limitation of MICROFIELD: minimizing descriptor-field loss can control task-relevant structural error only when the chosen descriptors capture the relevant scientific structure.

Thus, descriptor selection is not an arbitrary enumeration of handcrafted features. It is analogous to choosing physical residuals in physics-informed learning or symmetry groups in equivariant learning: the selected observables encode assumptions about which structures should be preserved. Descriptor choice should therefore be treated as a testable modeling decision rather than as a universal prescription. Table 14 provides practical starting points for common scientific objectives.

For a new application, we recommend the following procedure. First, identify the scientific quantity that should remain stable, such as phase amount, interface density, orientation, characteristic scale, or connectivity. Second, select the smallest stable descriptor family that is sensitive to that quantity and determine whether it can be computed directly from the image or requires masks or domain-specific annotations. Third, choose region sizes that match the phenomenon: finer regions for interfaces and texture, and coarser regions for connectivity and organization. Finally, validate the choice through descriptor-family ablations and independent downstream or structural metrics. Shuffled targets and equal-dimensional random projections provide negative controls for determining whether

Table 15: Matched target-control experiment. The fixed random patch projection is a frozen, equal-dimensional, non-scientific transformation of each regional image patch. It controls for target dimensionality and auxiliary prediction capacity without encoding scientific descriptors. Standard deviations are reported for the controls evaluated over three complete runs; the shuffled-target result is the aggregate mean reported in the corresponding ablation.

Regional supervision target	**Seg.** ↑	**Reg./Cls.** ↑	**Topology error** ↓
Shuffled descriptor targets	78.8	81.9	0.279
Fixed random patch projection	80.6 ± 0.7	83.0 ± 0.6	0.226 ± 0.009
Valid MICROFIELD descriptor field	$\mathbf{86.6 \pm 0.5}$	$\mathbf{88.1 \pm 0.5}$	$\mathbf{0.151 \pm 0.006}$

Table 16: Descriptor-family ablation. "Random" shuffles descriptor targets across images and serves as a negative control. Results are averaged over the same downstream settings as Table 2.

Descriptor set	Seg. ↑	Reg. ↑	DDD ↓	Topology err. ↓
Signal / texture only	83.2	85.1	0.169	0.204
Interface only	84.0	84.7	0.181	0.187
Anisotropy only	82.7	85.5	0.176	0.211
Signal + interface + anisotropy	85.8	87.4	0.139	0.164
All differentiable descriptors	**86.6**	**88.1**	**0.128**	**0.151**
Shuffled descriptor targets	78.8	81.9	0.232	0.279

an improvement comes from scientifically meaningful supervision rather than from the presence or capacity of an auxiliary objective.

L.7 Descriptor Family Analysis

The shuffled-target control tests whether improvement arises merely from adding an auxiliary loss. The fixed random projection is a stricter matched-capacity control: it preserves the target dimensionality and regional prediction structure but removes scientific meaning. Although the fixed projection performs somewhat better than shuffled targets, it remains substantially weaker than the valid descriptor field. This indicates that the improvement cannot be explained by auxiliary target capacity alone and depends on the scientific content of the regional targets.

We next evaluate which descriptor families contribute to performance. Table 16 shows that no single family explains the full gain. Signal and texture descriptors improve descriptor-distribution alignment, interface descriptors improve boundary-sensitive tasks, and anisotropy descriptors improve texture- and orientation-sensitive tasks. Combining the three differentiable descriptor families gives the strongest performance, while randomly shuffled descriptors do not improve over Pixel-MAE, indicating that the gains are not due to generic regularization.

M Microstructure Descriptor Theory and Implementation Details

This appendix provides a detailed justification for the descriptor families used in the proposed descriptor-field objective. The goal is not to claim that a small descriptor vector fully characterizes a scientific image. Rather, the aim is to select local observables that are: (i) scientifically interpretable, (ii) stable under small irrelevant pixel perturbations, (iii) sensitive to changes in morphology and organization, and (iv) computable over a region system.

Let $x : \Omega \to \mathbb{R}^C$ denote an image over a discrete domain $\Omega \subset \mathbb{Z}^2$, and let

$$\mathcal{P}(x) = \{R_i\}_{i=1}^N \tag{54}$$

be a collection of spatial regions. The regions may form a uniform grid, an adaptive partition, a hierarchy, a learned region system, or a segmentation-induced set of regions. For each region R_i, a descriptor extractor F defines

$$s_i(x) = F(x|_{R_i}) \in \mathbb{R}^d. \tag{55}$$

The corresponding microstructure descriptor field is

$$S_{\mathcal{P}}(x) = \{(R_i, s_i(x)) : R_i \in \mathcal{P}(x)\}. \tag{56}$$

When $\mathcal{P}(x)$ is hierarchical, we write $\mathcal{T}(x)$ for the hierarchy, $\mathcal{N}$ for its nodes, R_n for the spatial region associated with node n, and $s(n;x)$ for the descriptor at node n. The hierarchical descriptor field is then

$$S_{\mathcal{T}}(x) = \{s(n;x) : n \in \mathcal{N}(\mathcal{T})\}. \tag{57}$$

When a binary or multiclass phase map is available or can be estimated, we denote it by

$$m : \Omega \to \{0, 1, \ldots, K-1\}. \tag{58}$$

For binary notation, $m(p) = 1$ may denote a pore, foreground object, cell nucleus, extracellular matrix component, or other phase of interest, depending on the application.

M.1 Phase Fraction and Local Composition

The simplest morphology descriptor is local phase fraction. For a binary phase map, the phase fraction inside region R_i is

$$\phi_i = \frac{1}{|R_i|} \sum_{p \in R_i} m(p). \tag{59}$$

In porous materials, ϕ_i corresponds to local porosity when $m = 1$ denotes the pore phase. In biological images, the same statistic can represent local cell density, tissue occupancy, nuclear fraction, or abundance of a segmented compartment.

Although ϕ_i is low-dimensional, it is a physically meaningful observable. In porous media, porosity controls transport, permeability, and mechanical response at the continuum scale. In tissue imaging, local area fractions are widely used as summaries of cellularity, glandular density, fibrosis, necrosis, and tissue compartment organization. The same global phase fraction can be produced by very different microstructures, so ϕ_i alone is insufficient, but it provides a necessary first-order constraint.

For a multiclass phase map, we use

$$\boldsymbol{\phi}_i = [\phi_{i,0}, \ldots, \phi_{i,K-1}], \qquad \phi_{i,k} = \frac{1}{|R_i|} \sum_{p \in R_i} \mathbb{1}[m(p) = k]. \tag{60}$$

This descriptor lies on the probability simplex and captures local composition. It is useful when scientific images contain multiple material phases, tissue classes, or cellular compartments.

Cross-scale relation. If the region system is hierarchical and the children of node n form a disjoint cover of R_n, then phase fraction obeys an exact additive relation:

$$\phi_n = \sum_{c \in \mathrm{ch}(n)} \frac{|R_c|}{|R_n|} \phi_c. \tag{61}$$

Thus, for phase-fraction descriptors, the cross-scale transition term has a conservation interpretation. Deviations indicate that the model has changed the amount of material or biological phase represented across scales.

M.2 Interface Density

Phase fraction measures how much of a phase exists, but not how it is spatially arranged. Interface density captures the amount of boundary between phases. For a binary phase map, a discrete interface-density descriptor can be defined as

$$\rho_{I,i} = \frac{1}{|R_i|} \sum_{p \in R_i} \|\nabla m(p)\|_1, \tag{62}$$

where ∇m is computed with finite differences. Equivalently, for a 4-neighborhood graph E, one may define

$$\rho_{I,i} = \frac{1}{|R_i|} \sum_{\substack{(p,q) \in E \\ p,q \in R_i}} \mathbb{1}[m(p) \neq m(q)]. \tag{63}$$

This counts phase transitions per unit area and is often more robust for binary masks.

In materials imaging, interface density is related to grain-boundary density, pore-solid surface area, fracture density, and surface-to-volume ratio. These quantities affect transport, reaction rates, mechanical strength, and failure behavior. In biological imaging, interface density captures membrane complexity, glandular boundary irregularity, tissue compartment fragmentation, or cell-matrix interface structure.

Boundary correction. When computing $\rho_{I,i}$ on small regions, interfaces crossing region boundaries may be undercounted. We therefore distinguish internal and boundary-crossing edges:

$$E_i^{\mathrm{int}} = \{(p,q) \in E : p, q \in R_i\}, \qquad E_i^{\partial} = \{(p,q) \in E : p \in R_i,\ q \notin R_i\}. \tag{64}$$

A boundary-corrected descriptor is

$$\rho_{I,i}^{\mathrm{bc}} = \frac{\sum_{(p,q)\in E_i^{\mathrm{int}}} \mathbb{1}[m(p) \neq m(q)] + \gamma \sum_{(p,q)\in E_i^{\partial}} \mathbb{1}[m(p) \neq m(q)]}{|R_i|}, \tag{65}$$

where $\gamma \in [0,1]$ controls how much boundary-crossing interfaces contribute. A natural default is $\gamma = 1/2$ because each cross-boundary edge is shared by two adjacent regions.

Relation to total variation. For binary masks, interface density is a discrete approximation of the total variation of the phase indicator:

$$\mathrm{TV}(m; R_i) = \int_{R_i} |\nabla m(u)|\, du. \tag{66}$$

Thus, matching $\rho_{I,i}$ discourages predictions that preserve phase fraction while introducing spurious fragmentation, artificial smoothing, or excessive boundary roughness.

M.3 Local Anisotropy and Orientation

Many scientific images contain oriented structure: fibers, lamellae, cracks, elongated pores, aligned cells, collagen bundles, trabecular structures, vessels, or directional texture. To measure this, we use the local structure tensor. For a scalar image x, define image gradients $x_x(p)$ and $x_y(p)$, and compute

$$J_i = \sum_{p\in R_i} w_i(p) \begin{bmatrix} x_x(p)^2 & x_x(p)x_y(p) \\ x_x(p)x_y(p) & x_y(p)^2 \end{bmatrix}, \tag{67}$$

where $w_i(p)$ is a nonnegative weighting window supported on R_i. Uniform weighting is sufficient for most regions, while Gaussian weighting can reduce boundary artifacts.

Let $\lambda_{1,i} \geq \lambda_{2,i} \geq 0$ be the eigenvalues of J_i. We define the anisotropy index

$$a_i = \frac{\lambda_{1,i} - \lambda_{2,i}}{\lambda_{1,i} + \lambda_{2,i} + \epsilon}. \tag{68}$$

The value a_i lies in $[0,1]$ up to numerical error. If gradients are equally distributed in all directions, then $\lambda_{1,i} \approx \lambda_{2,i}$ and $a_i \approx 0$. If the region contains strongly oriented structure, then $\lambda_{1,i} \gg \lambda_{2,i}$ and $a_i \approx 1$.

The dominant orientation can be represented by

$$\theta_i = \frac{1}{2} \tan^{-1}\left(\frac{2J_{12,i}}{J_{11,i} - J_{22,i} + \epsilon}\right). \tag{69}$$

Because orientation is axial rather than directional, θ_i and $\theta_i + \pi$ represent the same physical direction. For stable learning, we encode orientation using

$$o_i = [\cos(2\theta_i), \sin(2\theta_i)]. \tag{70}$$

Material and biological interpretation. In materials, local anisotropy captures preferential grain orientation, fiber alignment, crack directionality, lamellar structure, and anisotropic pore geometry. In biological images, it captures oriented cell organization, collagen alignment, nerve or vessel directionality, muscle fibers, and elongated nuclear arrangements. These are not merely visual features; they often correspond to mechanical, functional, or pathological structure.

M.4 Connectivity and Euler Characteristic

Phase fraction and interface density still do not fully capture topology. Two regions may contain the same amount of pore phase and similar boundary length while differing in whether pores form connected channels or isolated islands. To capture this, we include simple topological descriptors.

For a binary phase m restricted to R_i, let $\beta_{0,i}$ denote the number of connected components of the foreground phase and $\beta_{1,i}$ denote the number of holes. In 2D, the Euler characteristic is

$$\chi_i = \beta_{0,i} - \beta_{1,i}. \tag{71}$$

Connectivity can also be summarized by the size of the largest connected component:

$$\eta_i = \frac{1}{|R_i|} \max_{C \in \mathcal{C}(m|_{R_i})} |C|, \tag{72}$$

where $\mathcal{C}(m|_{R_i})$ denotes the connected components of the foreground phase inside R_i.

Why topology matters. In porous media, connectivity controls percolation, permeability, diffusion paths, and failure modes. In biological images, connectivity describes glandular structure, vascular networks, cellular clusters, membrane continuity, and tissue compartment organization. A model that reconstructs plausible-looking texture but breaks connected channels or merges distinct components may achieve low pixel error while corrupting scientific structure.

Locality caveat. Topological descriptors computed on small regions are local and may not capture long-range percolation. This is a scale issue rather than a failure of the descriptor. A multiscale region system addresses this by computing topology at several scales: fine regions capture local fragmentation, while coarse regions capture larger connected structures.

M.5 Two-Point Correlation and Correlation Length

Spatial correlation descriptors capture how phases or intensities are arranged beyond first-order statistics. For a binary phase map, the local two-point correlation function can be estimated as

$$S_{2,i}(r) = \frac{1}{|\mathcal{Q}_{i,r}|} \sum_{(p,q) \in \mathcal{Q}_{i,r}} m(p)m(q), \qquad \mathcal{Q}_{i,r} = \{(p,q) : p, q \in R_i,\ \|p - q\| \approx r\}. \tag{73}$$

This measures the probability that two points separated by distance r both belong to the phase of interest.

A normalized autocorrelation descriptor for scalar images can similarly be defined as

$$C_i(r) = \frac{\sum_{(p,q) \in \mathcal{Q}_{i,r}} (x(p) - \mu_i)(x(q) - \mu_i)}{\sum_{p \in R_i} (x(p) - \mu_i)^2 + \epsilon}. \tag{74}$$

A compact correlation-length proxy is then

$$\ell_{c,i} = \min\{r : C_i(r) \le \tau C_i(0)\}, \tag{75}$$

for a fixed threshold $\tau \in (0, 1)$.

Interpretation. Correlation length measures characteristic feature size. In materials, it relates to pore size, grain scale, phase clustering, and representative volume effects. In biological images, it captures tissue texture scale, cell spacing, gland size, repeated structures, and local organization. Unlike pixel losses, correlation descriptors are tolerant to small translations while remaining sensitive to changes in characteristic scale.

Efficient implementation. Computing full $S_2(r)$ for every region is expensive. In practice, we use one of three approximations:

1. compute $C_i(r)$ only for a small set of radii $r \in \mathcal{R}$;
2. estimate correlations using random point pairs within each region;
3. compute local power-spectrum summaries and use them as frequency-domain proxies.

The descriptor vector therefore contains either sampled correlations

$$[C_i(r_1), \ldots, C_i(r_M)] \tag{76}$$

or a compact scale statistic such as $\ell_{c,i}$.

N Descriptor Stability and Normalization

Descriptors differ in numerical scale and robustness. Directly combining them without normalization can cause the loss to be dominated by high-magnitude quantities. We therefore normalize each descriptor dimension using training-set statistics estimated once and held fixed

$$\tilde{s}_{i,j}(x) = \frac{s_{i,j}(x) - \mu_j}{\sigma_j + \epsilon}, \tag{77}$$

where μ_j and σ_j are estimated over regions sampled from the training set.

Some descriptors are more sensitive to noise than others. Interface density and Euler characteristic, for example, can change under isolated pixel flips. We therefore use one or more of the following stabilizers:

1. compute descriptors after light Gaussian smoothing or morphological opening;
2. ignore connected components smaller than a fixed area threshold;
3. compute descriptors on soft masks when phase probabilities are available;
4. downweight topology descriptors at very fine scales where they are unstable;
5. aggregate descriptor loss over scales rather than relying on individual small regions.

N.1 Soft Differentiable Descriptor Variants

When descriptors are computed from model outputs, hard thresholding can make gradients unstable or unavailable. A practical alternative is to compute descriptors on soft phase probabilities $q(p) \in [0, 1]$. The soft phase fraction is

$$\phi_i^{\mathrm{soft}} = \frac{1}{|R_i|} \sum_{p \in R_i} q(p). \tag{78}$$

A soft interface density can be defined as

$$\rho_{I,i}^{\mathrm{soft}} = \frac{1}{|R_i|} \sum_{(p,q) \in E_i} |q(p) - q(q)|. \tag{79}$$

Intensity-based anisotropy and correlation descriptors are naturally differentiable when computed from continuous outputs.

In this work, exact topology descriptors are used only as non-differentiable evaluation metrics. A domain-specific extension could instead apply a differentiable topology surrogate to the prediction, but such a surrogate is not used in the reported experiments.

A fully differentiable topology loss is possible but not required for the proposed framework. The descriptor-field objective only requires that descriptor discrepancies provide useful supervision; not every descriptor must be used in every training configuration.

O Recommended Descriptor Sets by Domain

The descriptor vector should be selected according to the scientific domain.

Porous materials and XCT. For porous materials, we recommend

$$s_i = [\phi_i,\, \rho_{I,i},\, \eta_i,\, \chi_i,\, a_i,\, \ell_{c,i}]. \tag{80}$$

This set captures porosity, pore-solid interface density, connected pore structure, topology, anisotropy, and characteristic pore scale.

Polycrystalline or granular materials. For grain-structured images, useful descriptors include

$$s_i = [\mu_i,\, \sigma_i^2,\, \rho_{I,i},\, a_i,\, o_i,\, \ell_{c,i}]. \tag{81}$$

Here interface density approximates local grain-boundary density, while anisotropy and orientation encode preferred grain or lamellar direction.

Pathology and tissue microscopy. For tissue images, we recommend

$$s_i = [\boldsymbol{\phi}_i,\, \rho_{I,i},\, \eta_i,\, a_i,\, o_i,\, \ell_{c,i},\, H_i]. \tag{82}$$

Here $\boldsymbol{\phi}_i$ may represent tissue compartments or cell types, $\rho_{I,i}$ captures boundary complexity, η_i captures clustering or glandular connectivity, a_i and o_i capture oriented tissue organization, and H_i captures local appearance heterogeneity.

Electron microscopy and cellular ultrastructure. For EM data, where membranes and elongated organelles are important, useful descriptors include

$$s_i = [g_i,\, \rho_{I,i},\, a_i,\, o_i,\, C_i(r_1), \ldots, C_i(r_M)]. \tag{83}$$

These descriptors emphasize edges, membrane-like structures, anisotropic organization, and repeated subcellular texture.

P Practical Computation on Region Systems

For each training image, we construct or retrieve a region system $\mathcal{P}(x)$ and use the same regions to evaluate the target x and the prediction $\hat{x}$. The local descriptor loss is

$$\mathcal{L}_{\mathrm{local}}(x, \hat{x}) = \frac{1}{d} \sum_{i=1}^{N} w_i \, \|\tilde{s}_i(x) - \tilde{s}_i(\hat{x})\|_2^2 \,, \qquad \sum_{i=1}^{N} w_i = 1, \tag{84}$$

We use $w_i = 1/N$ in the reported experiments. When equal weighting across hierarchy levels is desired, an optional level-balanced choice for node n is

$$w_n = \frac{1}{(L+1)\,\left|\mathcal{N}_{\ell(n)}(T)\right|}, \tag{85}$$

where $\ell(n)$ is the level of node n and the hierarchy has levels $\ell = 0, \ldots, L$. This definition satisfies $\sum_{n \in \mathcal{N}(T)} w_n = 1$.

For hierarchical descriptor fields, the cross-scale transition loss is

$$\mathcal{L}_{\mathrm{trans}}(x, \hat{x}) = \frac{1}{|\mathcal{N}_{\mathrm{int}}(T)|} \sum_{n \in \mathcal{N}_{\mathrm{int}}(T)} \|r(n; x) - r(n; \hat{x})\|_2^2 \,, \tag{86}$$

where

$$r(n; x) = \tilde{s}(n; x) - A_n(\{\tilde{s}(c; x) \,|\, c \in \mathrm{ch}(n)\}) \,, \tag{87}$$

For additive descriptors, A_n is the corresponding area-weighted aggregation operator. For non-additive descriptors, it is the fixed aggregation operator used in the implementation, and the residual measures empirical parent-versus-children scale-transition behavior.

Finally, define

$$S_\ell(x) = \{\tilde{s}(n; x) \,|\, n \in \mathcal{N}_\ell(T)\} \,.$$

The level-wise distribution loss is

$$\mathcal{L}_{\mathrm{dist}}(x, \hat{x}) = \sum_{\ell=0}^{L} D(S_\ell(x), S_\ell(\hat{x})) \,. \tag{88}$$

In the reported experiments, D is mean–covariance matching. Alternative discrepancies such as MMD, sliced Wasserstein distance, or histogram distance are possible extensions but are not used for the main results.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state that MICROFIELD is an objective-level supervision framework for scientific images, rather than a new masked autoencoder architecture. The stated contributions–microstructure descriptor fields, local descriptor matching, level-wise descriptor-distribution alignment, optional cross-scale transition preservation, and architecture-agnostic attachment to scientific image models–are formalized in Section 3. The empirical claims are evaluated in Section 5, with downstream results in Table 2, scientific-fidelity results in Table 3, overhead and cost results in Table 10, and ablations in Table 4.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Limitations section (Appendix B) explicitly identifies descriptor misspecification as a central failure mode: if the selected descriptor field is insufficient for the downstream scientific structure, MICROFIELD may preserve the wrong invariants. Section 4 further states that descriptor choice is a task-conditioned structural prior rather than a universal recipe, and Appendix L frames this through descriptor sufficiency. Section F.1 and Table 10 discuss the additional training-time overhead, while Appendix K.4 describes implementation cost and scaling tradeoffs.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete and correct proof?

   Answer: [Yes]

   Justification: Appendix L provides the theoretical interpretation of descriptor-field supervision. In particular, Appendix L defines the descriptor-field pseudometric and quotient-space view, and states the descriptor-sufficiency result showing when preservation of descriptor fields controls task-relevant structural error. The assumptions are stated in terms of the descriptor field being sufficient for the downstream structural functional up to an approximation error, and the proof follows directly from the triangle inequality and the Lipschitz property of the sufficiency map.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper, regardless of whether the code and data are provided or not?

   Answer: [Yes]

   Justification: Section 3 specifies the full MICROFIELD objective, including the total objective in Eq. (2), descriptor fields in Eqs. (6)–(7), local descriptor matching, cross-scale transition preservation, level-wise distribution alignment, and the complete descriptor-field loss. Section 5 specifies the datasets, baseline families, matched-backbone comparisons, external foundation-model reference baselines, and evaluation metrics. Appendix F gives Algorithm 1 for the training step, and Appendix K.4 gives implementation details for region construction, descriptor definitions, distributional losses, differentiability, loss scaling, and computational cost.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

Justification: The paper provides the implementation details needed to reproduce MICROFIELD, including the descriptor-field objective in Section 3, Algorithm 1 in Appendix F, and the practical implementation details in Appendix K.4, including region construction, descriptor definitions, distributional losses, differentiability, loss scaling, and computational cost. The evaluated datasets are either public or generated from documented procedures: Synthetic-μStruct is generated from the controlled stochastic two-phase-field process described in Appendix E; Porous-XCT, MitoEM/EM, and HuBMAP-Vasc are accessed from their original public sources or standard dataset releases. The released instructions describe data preparation, pretraining, fine-tuning, and metric computation for the main results in Appendix E, Tables 6, 7, and 8.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details, e.g., data splits, hyperparameters, how they were chosen, type of optimizer, necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix E describes the evaluated regimes–Synthetic-μStruct, Porous-XCT, MitoEM/EM, and HuBMAP-Vasc–and states that methods are pretrained without labels and evaluated through linear probing, fine-tuning, and structure-preservation metrics. It also specifies the controlled baseline families, matched-backbone comparisons, external foundation-model references, and task and structure metrics. Appendix K.4 provides hyperparameter and implementation details for attaching the descriptor-field loss to the base masked-image-modeling objective.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main downstream results in Table 2, scientific-fidelity results in Table 3, and ablations in Table 4 are reported as averages over multiple independent runs with different random seeds for pretraining, fine-tuning, and evaluation. For readability, the main tables report mean values and the corresponding standard deviations are provided in Appendix D. The observed improvements of MICROFIELD over Pixel-MAE, hierarchy-only MAE, global descriptor supervision, and generic perceptual or spectral losses are consistently larger than the run-to-run variation, supporting the statistical reliability of the reported trends.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources, type of compute workers, memory, time of execution, needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C describes the compute platform and compute cost.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper studies a supervision objective for scientific image representation learning using pre-existing scientific imaging datasets, synthetic controlled dataset, and standard machine-learning evaluation protocols. It does not introduce human-subject experiments, scraped personal data, surveillance data, or generative content with obvious misuse risk. The submission is anonymized.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Appendix A discusses broader impacts. On the positive side, improved representation learning for scientific images could reduce annotation cost and improve analysis in materials science, microscopy, computational biology, infrastructure inspection, and medical imaging. The main risk identified is that structure-preserving objectives may create overconfidence if the chosen descriptors encode incomplete or biased notions of scientific fidelity, especially in clinical, safety-critical, or policy-relevant settings.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse, e.g., pre-trained language models, image generators, or scraped datasets?

    Answer: [N/A]

    Justification: The paper introduces a training-time supervision framework for scientific images. It does not release a language model, image generator, scraped dataset, or other asset with high misuse risk. The main risks are scientific or clinical overconfidence from descriptor misspecification, which are discussed under limitations and broader impacts rather than requiring model-release safeguards.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets, e.g., code, data, models, used in the paper properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing methods, datasets, and foundation models used for comparison are credited in Section 5 and Table 11, including Pixel-MAE, SimMIM/Swin, iBOT, DINOv2, pathology foundation models such as UNI, Virchow, Prov-GigaPath, Phikon, and HIPT, bio/medical models such as CA-MAE, CellSAM, MedSAM, SAM-Med2D, and materials/microscopy SSL baselines. Dataset regimes are described in Appendix E. The supplementary material states the access conditions and licenses for the datasets and code assets used in the experiments.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces MICROFIELD as a new objective-level framework and documents the method in Section 3, Algorithm 1, and Appendices F and K.4. The Synthetic-μStruct controlled dataset is described in Section E as a generated two-phase dataset with known phase fraction, interface density, anisotropy, connected components, Euler characteristic, and correlation length. The released supplementary material includes documentation, generation instructions, and usage notes for the dataset and implementation.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation, if any?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, participant studies, or newly collected human-subject data. The experiments use pre-existing scientific imaging datasets, public or standard dataset releases, and synthetic controlled scientific image regimes.

15. **Institutional review board approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board approvals, or an equivalent approval/review based on the requirements of your country or institution, were obtained?

    Answer: [N/A]

Justification: The work does not conduct human-subject experiments or collect new data from participants. Therefore, IRB approval or equivalent human-subject review is not applicable.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method is a descriptor-field supervision framework for scientific image representation learning and does not use LLMs as an important, original, or non-standard component of the method, experiments, or evaluation.