Cell Identification from Partial Observation using Spatiotemporal Attention in *C. elegans* Embryogenesis

Henry Xue

Abstract

Automated cell identification in developing embryos represents a critical bottleneck in developmental biology, where manual annotation of genetic screens requires hundreds of hours of expert labor [10]. Real experimental workflows capture only 5–20 cells through selective labeling or optical constraints [1], yet existing computational methods require complete embryo observation [41]. We hypothesized that transformer-based architectures with joint attention could identify cells from partial observations. We developed an architecture processing paired cellular neighborhoods simultaneously through multi-head attention, incorporating geometric features encoding local spatial relationships, learnable no-match tokens for missing correspondences, and curriculum learning. Training used 270 semi-synthetic *C. elegans* embryos; evaluation used 30 held-out simulated embryos and 3 independent real embryos. The model achieved 90.3% identification accuracy on simulated data and 87.4% on real embryos using k-nearest neighbor classification in the learned embedding space. Performance remained robust under perturbations: 86.4% with 20% missing cells. The approach substantially outperformed geometric registration (ICP: 29.2%, CPD: 35.9%) and Siamese architectures (59.3%). Ablation studies confirmed that joint attention (+31.0 percentage points over Siamese), no-match modeling (+14.7 pp), and geometric features (+21.9 pp over raw coordinates) each contribute critically. Error analysis revealed biologically structured failures concentrated among siblings (49.3%) and spatial neighbors (30.1%), indicating learned representations respect developmental relationships. We conclude that local cellular neighborhoods contain sufficient information for reliable identification, enabling automated annotation for large-scale genetic screens and real-time microscopy applications.

1 Introduction

1.1 Rationale

Understanding how organisms develop from a single cell into complex multicellular forms requires tracking the behavior of individual cells as they divide, migrate, and differentiate. Modern live-imaging techniques can now capture these cellular dynamics at single-cell resolution across entire developing embryos [10, 20, 30], transforming our ability to observe the fundamental processes of morphogenesis. However, the biological insight promised by these imaging advances remains largely unrealized due to a critical computational bottleneck: identifying which specific cells are present and which are which in microscopy images.

1.2 Biological Context

The nematode *Caenorhabditis elegans* provides a uniquely powerful platform for addressing this challenge. Its transparent embryo develops through an invariant lineage from the zygote through 558 cells [32], providing gold-standard ground truth rarely available in other systems. Moreover, approximately 60–80% of *C. elegans* genes have human orthologs with particularly high conservation among developmental regulators [17, 16], ensuring that insights translate to human biology and disease. Complete 4D datasets can be acquired in 30–60 minutes using standard confocal microscopy with automated segmentation pipelines [1], while extensive molecular atlases enable integration of spatial dynamics with gene expression patterns [25, 8].

1.3 Annotation Bottleneck

Despite these advantages, manual cell annotation represents a severe experimental bottleneck. Curating *C. elegans* embryos from a single genetic screen requires hundreds hours of expert labor [10]. This overwhelming effort prevents the large-scale quantitative studies necessary to identify therapeutic targets for developmental disorders and cancer [34]. Across biological research, the inability to automatically identify cells from positional information fundamentally limits progress: researchers imaging hundreds of mutant embryos cannot determine which cells are affected or when trajectories diverge without manual curation [10, 23].

1.4 Partial Observation

The computational challenge becomes substantially more difficult under real experimental constraints. Researchers rarely capture complete tissue samples but instead observe small local neighborhoods of 5–20 cells, constrained by photobleaching, limited optical access, selective fluorescent labeling, and temporal resolution requirements [1, 21]. Existing computational approaches require observing complete embryos to establish cell identities through global spatial context and template matching [18, 41], fundamentally limiting their applicability to real experimental workflows.

Cell identification in developing tissues represents a challenging spatiotemporal representation problem. Unlike static pattern recognition, cells must be identified within spatiotemporal contexts that constantly shift through divisions, migrations, and tissue deformations [32, 20]. Identity is encoded not just in absolute position but in relationships to surrounding neighbors, relationships evolving continuously as development progresses. The ideal representation must be local, context-sensitive, and richly descriptive, uniquely representing diverse configurations across development where all spatiotemporal events occupy distinct regions in a unified latent space [9].

Three biological phenomena make this particularly challenging under partial observation. Neighborhood volatility: local configurations change dramatically as cells divide and move, with 10 cells at time t sharing only 5–7 cells with the same spatial region minutes later [30]. Developmental heterochrony: stochastic division timing causes different cells to be present at nominally equivalent stages, preventing simple template-matching [10, 21]. Partial observation constraints: researchers capture only subsets of 5–20 cells rather than complete embryos, requiring methods robust to limited spatial context [1, 34].

1.5 Existing Methods

Previous computational methods tackle related yet distinct challenges. Template-based registration establishes correspondences through iterative alignment [18, 1] but requires pose standardization, synchronized timing, and complete observation—assumptions that fail under natural biological variation. These methods fail catastrophically on small subsets because they require geometric landmarks for global alignment, unavailable when observing 5–20 cells from 100+ cell embryos. Hand-crafted geometric features using rotation-invariant descriptors [18] or iterative closest point (ICP) algorithms [5] capture simple spatial relationships but cannot express higher-order relational structure necessary for distinguishing biologically meaningful arrangements. Similarly, probabilistic point cloud registration methods like Coherent Point Drift (CPD) [24] assume global geometric structure unavailable in sparse partial observations. Methods analyzing molecular composition [14, 28] or global tissue shape [33, 6] operate at different scales, averaging away the local spatial arrangements encoding individual cell identities [37].

1.6 Relevant Technical Advances

Recent advances in machine learning suggest a path forward. Transformer-based architectures have revolutionized 3D shape analysis, succeeding in object classification, detection, and segmentation through multi-scale spatial attention [35, 12, 26]. These methods learn rich point descriptors from data rather than relying on hand-crafted features. Limited biological applications have explored learned descriptors for neural circuit registration [41], though whether such methods could handle developmental dynamics and neighborhood volatility remained unclear. Spatiotemporal transformers have advanced video understanding and sequential point cloud processing through attention over space and time [4, 40]. Recent work by Santella and colleagues demonstrated the

effectiveness of joint attention strategies—processing pairs of cellular configurations simultaneously and attempting to learn from creating matches in between the pairs—for whole-tissue matching in complete *C. elegans* embryos [27]. This Twin Attention (joint attention) approach contrasts with Siamese networks where computation remains independent per input [7], enabling the model to learn complex relational features within single training instances.

1.7 Purpose

However, no existing architecture enables reliable cell identification from the small, partially-observed neighborhoods typical of real experimental workflows. Without this method, images of small cellular neighborhoods, as seen through real microscopes, simply show an image of a few highlighted dots. Without identification, nothing can be learned from said anonymized dots. Furthermore, the transition from whole-tissue to local neighborhood analysis presents unique difficulties: fewer geometric reference points complicate position encoding [18], temporal dynamics of subsets can diverge from whole-tissue movements [33], and methods requiring comprehensive context may not scale to sparse observations [38]. Furthermore, existing methods lack explicit mechanisms for handling no-correspondence cases ubiquitous in partial observation—cells present in one timepoint or embryo but absent from another due to sampling, birth, or death.

This study demonstrates that transformer-based architectures with joint attention can be effectively adapted for automated cell identification from small, partially-observed cellular neighborhoods and that partial cellular neighborhoods contain sufficient context for identification accuracies of above 90%. We develop a specialized architecture processing paired neighborhoods of 5–20 cells simultaneously through multi-head attention, learning rich spatiotemporal representations without requiring complete embryonic context. The architecture incorporates geometric features encoding local spatial relationships, learnable no-match tokens for handling missing correspondences, and biologically-informed training strategies. By achieving robust identification from partial observations, this work addresses the critical experimental bottleneck limiting large-scale developmental studies, genetic screens, and therapeutic discovery.

[This space was intentionally left blank]

2 Methods

2.1 Student vs. Mentor Role

Student: Independently designed and implemented all architectural components. Conducted all data preprocessing, model training, and result interpretation.

Mentor: No direct involvement in coding, data analysis, or experimental decisions.

2.2 Overview

Automated cell identification from partially observed neighborhoods addresses a critical experimental bottleneck where manual annotation of complete embryos prevents large-scale studies. Traditional computational approaches require observing all $\tilde{2}00$ cells to establish identities through global template matching [18, 1], fundamentally limiting experimental flexibility. This study demonstrates that small cellular neighborhoods of 5–20 cells contain sufficient spatiotemporal information for identification when processed through joint attention mechanisms that directly compare paired observations rather than classifying in isolation. The architecture incorporates four geometric features encoding spatial relationships, no-match modeling for missing correspondences, and other training strategies addressing neighborhood volatility and developmental heterochrony.

2.3 Data Sources and Preprocessing

Training data comprised 270 semi-synthetic embryos generated using a validated agent-based simulator [38] that models stochastic division timing, realistic cell migration, and physical collision constraints while providing ground truth identities unavailable in experimental data at scale. This approach enabled controlled evaluation of robustness while maintaining biological realism. Simulated embryos spanned 4-cell through 194-cell stages, represented as time series of 3D point clouds with complete cell identity labels and lineage tracking.

Real embryo validation utilized three complete developmental time series from embryos imaged and segmented with automated lineage reconstruction via graph optimization methods [22]. These embryos served as an independent held-out test set, providing validation on real experimental data not seen during training.

All coordinates were centered to remove global translation ($\mathbf{x}'_i = \mathbf{x}_i - \bar{\mathbf{x}}$) and scaled to unit variance per dimension (minimum variance floor 10^{-6}) to ensure dimensional isotropy. Offline augmentation generated 10 random SO(3) rotations per timepoint, increasing effective dataset size 10-fold. Online augmentation during training applied progressive perturbations calibrated to curriculum stage: random rotations ($\pi/36$ to $\pi/12$ radians), Gaussian coordinate noise (standard deviation 0.01–0.03 times mean nearest-neighbor distance), and optional translations.

2.4 Architecture Design

2.4.1 Geometric Feature Extraction for Sparse Neighborhoods

Small neighborhoods (5–20 cells) present limited spatial context compared to complete embryos. Raw xyz coordinates provide insufficient information to distinguish biologically meaningful configurations, as absolute position carries less identity information than relative arrangement [18]. We, thus, designed a feature extraction module incorporating four geometric representations.

Relative position encoding computed as $\mathbf{r}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ removed global translation while preserving local spatial arrangement, ensuring cell identity depends on configuration rather than embryo orientation. Centroid distance $d_i = \|\mathbf{r}_i\|_2$ distinguished interior from boundary cells, providing radial information critical for identifying cells by tissue location. Local density quantified spatial crowding through mean distance to $k = 3$ nearest neighbors, computed using KDTree spatial indexing [36], identifying regions of recent or impending division where identification is most challenging. Learned point-count embeddings $e(n)$ for $n \in [5, 20]$ conditioned feature interpretation on total observed cells, allowing the model to adjust processing strategies between sparse (5–7 cells) and dense (17–20 cells) observations.

Each feature was projected through separate linear layers to $d/4 = 32$ dimensions, concatenated to $d = 128$–d, and processed through a two-layer MLP with ReLU activation and dropout (0.1).

2.4.2 Joint Attention with No-Correspondence Modeling

The architecture processes paired neighborhoods simultaneously through joint self-attention rather than independent encoding followed by comparison. This design choice directly addresses the core challenge: learning to identify cells requires understanding both what makes a cell distinctive and how it relates to potential matches in another observation [27].

Given anchor neighborhood $\mathbf{A}$ with N_A cells and comparison neighborhood $\mathbf{B}$ with N_B cells, geometric features were computed independently then concatenated into a single sequence of length $N_A + N_B + 1$. The additional position contained a learnable no-match token $\mathbf{z}_{\text{no-match}} \in \mathbb{R}^{128}$, explicitly representing cells with no valid correspondence due to heterochrony (stochastic division timing causing different cells to be present at nominally equivalent stages), cell birth or death between timepoints, or incomplete sampling. This feature stops the model from forcing assignments even when no biologically valid matches exists.

The concatenated sequence is processed through a six-layer Transformer Encoder [35]: 128-dimensional embeddings, eight attention heads, feed-forward dimension 512, GELU activation [**?**], dropout 0.1, and pre-layer normalization [39]. Key-padding masks excluded positions corresponding to batch padding from attention computation. Multi-head self-attention enabled the model to learn complex relational patterns across both neighborhoods simultaneously. Output embeddings were L2-normalized before similarity computation to ensure numerical stability.

Temperature-scaled cosine similarity matrices $\mathbf{S} = \mathbf{Z}_A\mathbf{Z}_B^T/T$ were computed with learnable temperature $T = \exp(\tau)$ initialized to 1.0. Temperature underwent five-epoch linear warmup from 1.0 to learned value, with L2 regularization ($\lambda = 10^{-3}$) preventing extreme scaling. Row-wise softmax yielded match probabilities $p(j|i) = \text{softmax}(\mathbf{S}_{i,:})_j$ where $p(N_B|i)$ represents probability that anchor cell i matches the no-match token.

Training minimized negative log-likelihood: $\mathcal{L}_{\text{match}} = -\sum_{i=1}^{N_A} \mathbb{I}[m_i < N_B] \log p(m_i|i)$ for true correspondences plus reduced-weight outlier loss $\mathcal{L}_{\text{outlier}} = -0.5 \sum_{i=1}^{N_A} \mathbb{I}[m_i = N_B] \log p(N_B|i)$ for no-match cases. Temporal consistency loss $\mathcal{L}_{\text{temporal}} = 0.1 \cdot \mathbb{E}_{(p,c)}[1 - \cos(\mathbf{z}_p, \mathbf{z}_c)]$ encouraged smooth embedding changes across parent-child pairs in lineage. Total loss: $\mathcal{L} = \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{outlier}} + \mathcal{L}_{\text{temporal}} + 10^{-3}\tau^2$.

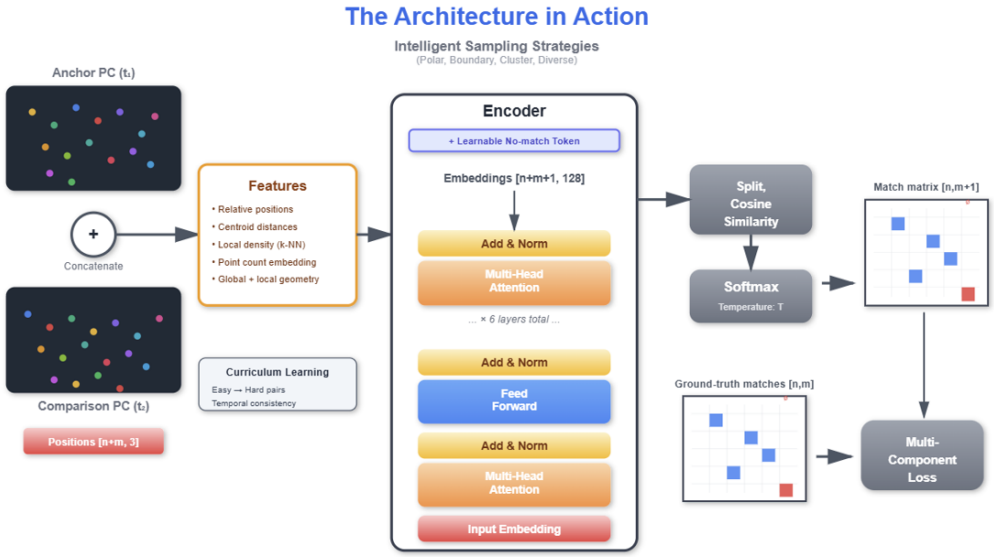

Figure 1: Architecture for partial-observation cell identification. Two neighborhoods (anchor at t_0, comparison at t_1) are processed through geometric feature extraction encoding relative position, centroid distance, local density, and point count. A learnable no-match token handles missing correspondences. The Transformer encoder performs joint self-attention across the concatenated sequence, enabling cross-neighborhood reasoning. L2-normalized embeddings produce a temperature-scaled similarity matrix; row-wise softmax yields match probabilities including no-match. Multi-component loss combines match negative log-likelihood, reduced-weight no-match loss, temperature regularization, and temporal consistency. Curriculum learning progresses from easy to challenging pairs. Figure created by student using PowerPoint.

2.5 Training Strategy

2.5.1 Curriculum Learning

Direct training on challenging inter-embryo pairs with minimal overlap failed to converge well. Four-stage curriculum learning [2] addressed this by progressively increasing task difficulty, allowing the model to build robust spatial representations before confronting complex cases.

Totally model trained upon 100 epochs. Epochs 0–19 consisted of nearly identical pairs sampled from the same embryo within 1–2 timepoints. this stage established basic spatial encoding without tempoeral confounds. Epochs 20–39 consisted of temporal pairs 1–3 timepoints apart from the same

embryo. Epochs 40–59 mixed temporal and inter-embryo matching pairs up to five timepoints apart with 30% drawn across embryos. Epochs 60+ maximally challenged the model with only one shared cell between embryos required.

2.5.2 Biologically–Informed Sampling

Uniform random sampling would oversample stable developmental plateaus while undersampling critical transition periods with high division rates where identification is most challenging. Adaptive sampling addressed this by emphasizing challenging stages while maintaining diversity.

Developmental stages were defined by total cell count (4–194 cells). For each stage serving as anchor, an adaptive sliding window determined available timepoints. Within each window, four sampling strategies were randomly mixed during training: (1) Spatial clustering selected a random seed cell and its k nearest neighbors, maintaining local neighborhoods; (2) Boundary sampling extracted convex hull vertices then added random interior points if needed, prioritizing tissue surfaces. (3) Polar sampling projected cells onto the principal axis via singular value decomposition, selecting extremes along anterior-posterior orientation plus random intermediate cells. (4) Diverse sampling performed farthest point sampling, iteratively selecting the point maximally distant from all previously selected points. Random mixing exposed the model to varied spatial configurations reflecting different experimental scenarios.

2.5.3 Optimization

Model parameters were optimized using AdamW [19] with weight decay 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$. Learning rate was determined via Leslie Smith's cyclical learning rate range test. OneCycle learning rate scheduling [31] provided 10% warmup, cosine annealing over remaining epochs, and momentum cycling between 0.85 and 0.95. Batch size was set to 16, and gradient clipping applied at norm 1.0. Training on CPU (Intel Core i7, 32GB RAM) required approximately 10 hours per curriculum stage.

2.6 Evaluation

2.6.1 Held-Out Test Set

The primary evaluation task is identifying individual cells within a query neighborhood of 5–20 cells by retrieving their identities from the learned embedding space via k-nearest neighbor classification. Evaluation used 30 held-out simulated embryos (not included in the 270 training embryos) and 3 independent publicly–available real embryos [22]. All reported accuracies derive from these held-out sets unless otherwise specified.

2.6.2 Practical Deployment

For practical cell identification, query neighborhoods of 5–20 cells are processed through the trained model paired with a randomly selected stage-matched reference embryo from training data. The model produces 128-dimensional embeddings for each query cell. Cell identities are then assigned via k-nearest neighbor (k=30) classification using cosine similarity against the full training embedding space, with predictions determined by majority vote. The FAISS library [15] enables efficient approximate nearest neighbor search, achieving approximately 25 milliseconds per cell prediction on standard hardware (Intel Core i7, 32GB RAM), enabling real-time identification during live microscopy.

2.6.3 Statistical Analysis

Embryo-level bootstrap resampling with 1,000 resamples computed 95% confidence intervals accounting for within-embryo correlation [11]. Paired permutation tests with 10,000 permutations compared methods, with effect sizes quantified via Cohen's d. Multiple comparison correction applied the Benjamini-Hochberg procedure [3] controlling false discovery rate at $\alpha = 0.05$.

2.7 Challenges and Limitations

Training on semi-synthetic data [38] may limit generalization to real experimental conditions, though the simulator was extensively validated and built upon real experimental embryos. Evaluation constrained to early-to-mid embryogenesis (up to 194 cells out of 558 cells total during embryogenesis) limiting use cases and scenarios. Method requires pre-segmented positions, inheriting segmentation errors [1] which could potentially affect the model's generalization.

2.8 Implementation

All code was implemented in Python 3.9 using PyTorch 2.0 for neural network components, NumPy 1.24 and SciPy 1.10 for numerical operations [36]. Dimensionality reduction for visualization used t-SNE. Code and trained models are available in supplementary materials.

3 Results and Discussion

This study addresses whether small cellular neighborhoods of 5–20 cells contain sufficient spatiotemporal information for reliable cell identification in *C. elegans* embryogenesis. We trained a joint-attention transformer on 270 simulated embryos and evaluated identification accuracy using k-nearest neighbor classification (k=30) in the learned embedding space on 30 held-out simulated embryos and 3 independent real embryos.

3.1 Core Identification Performance

Cell identities were predicted by processing query neighborhoods through the trained model and retrieving matches via k-nearest neighbor classification in the learned embedding space.

The model achieved 90.3% overall accuracy (95% CI: [88.1%, 92.4%]) on 30 held-out simulated embryos and 87.4% (95% CI: [84.2%, 90.3%]) on three independent real embryos (Figure 1A). Hierarchical analysis showed accuracy increased at coarser biological resolutions: 90.3% for exact cell identity, 96.1% for sublineage, 97.1% for founder lineage, and 98.2% for binary classification (Figure 1B). Performance varied with neighborhood size (Figure 1C).

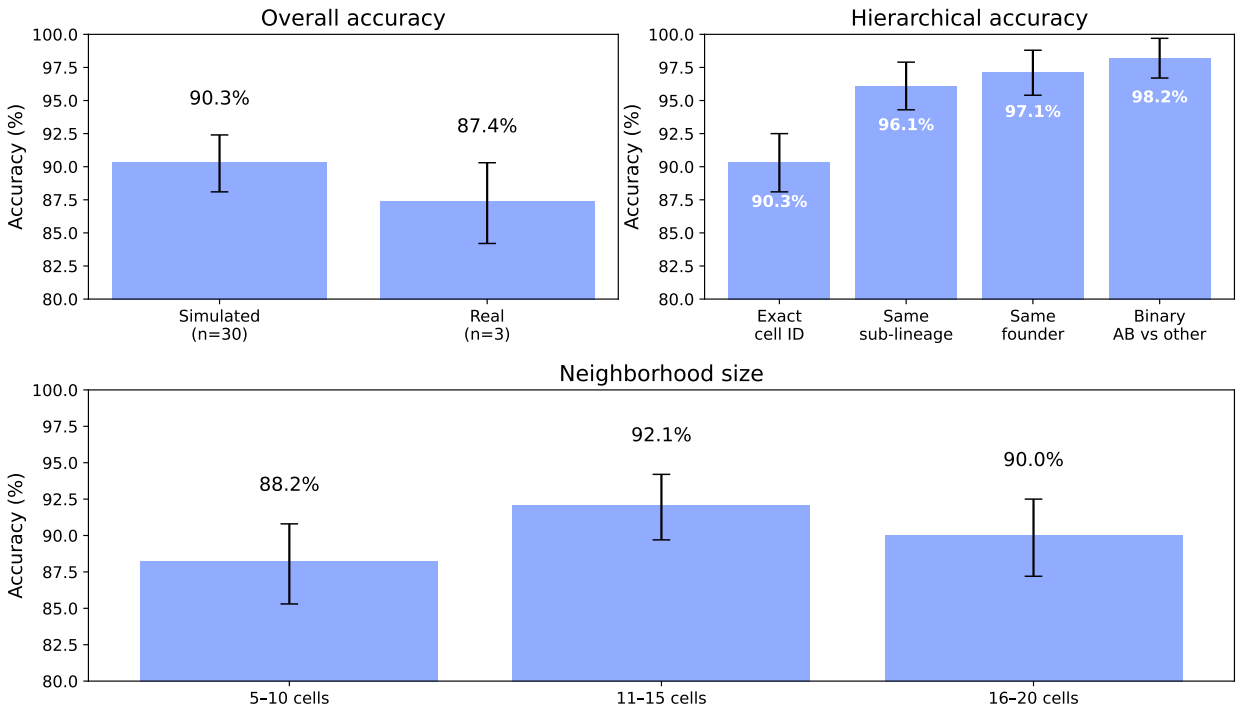

Figure 2: Core identification performance from partial neighborhoods. (A) Overall accuracy comparing simulated embryos (n=30; 90.3%, 95% CI: [88.1%, 92.4%]) and real embryos (n=3; 87.4%, 95% CI: [84.2%, 90.3%]). (B) Hierarchical accuracy at multiple biological resolutions: exact identity (90.3%), sublineage (96.1%), founder lineage (97.1%), binary (98.2%). (C) Accuracy by neighborhood size: 5–10 cells (88.2%), 11–15 cells (92.1%), 16–20 cells (90.0%). Error bars show 95% bootstrap confidence intervals. Figure created by student using Python/matplotlib.

These results demonstrate that local cellular neighborhoods contain sufficient relational information for reliable identification without complete embryonic context. The 2.9 percentage-point gap between simulated and real embryo performance likely reflects imperfect segmentation and biological variation not fully captured in simulation [1], yet 87.4% accuracy on real data validates practical applicability. The hierarchical structure of accuracy, 97.1% at founder lineage level, indicates the model reliably encodes major developmental compartments established during early cleavages [32]; when exact identity is uncertain, developmental lineage remains correct. The modest variation with neighborhood size (88.2% to 92.1%) contrasts sharply with template-matching approaches requiring near-complete embryo observation [18], confirming that identity can be found and is encoded primarily in local relational geometry rather than global position [9].

3.2 Baseline Comparisons

To contextualize performance, we compared the joint-attention approach against geometric registration methods and an alternative neural architecture on identical held-out test pairs.

Traditional geometric methods performed poorly: Iterative Closest Point (ICP) achieved 29.2%, Coherent Point Drift (CPD) 35.9% , and Hungarian assignment 37.9%. A Siamese transformer using identical architecture but independent neighborhood encoding achieved 59.3%. The joint-attention model reached 90.3%, outperforming all baselines (Figure 2).

[This space was intentionally left blank]

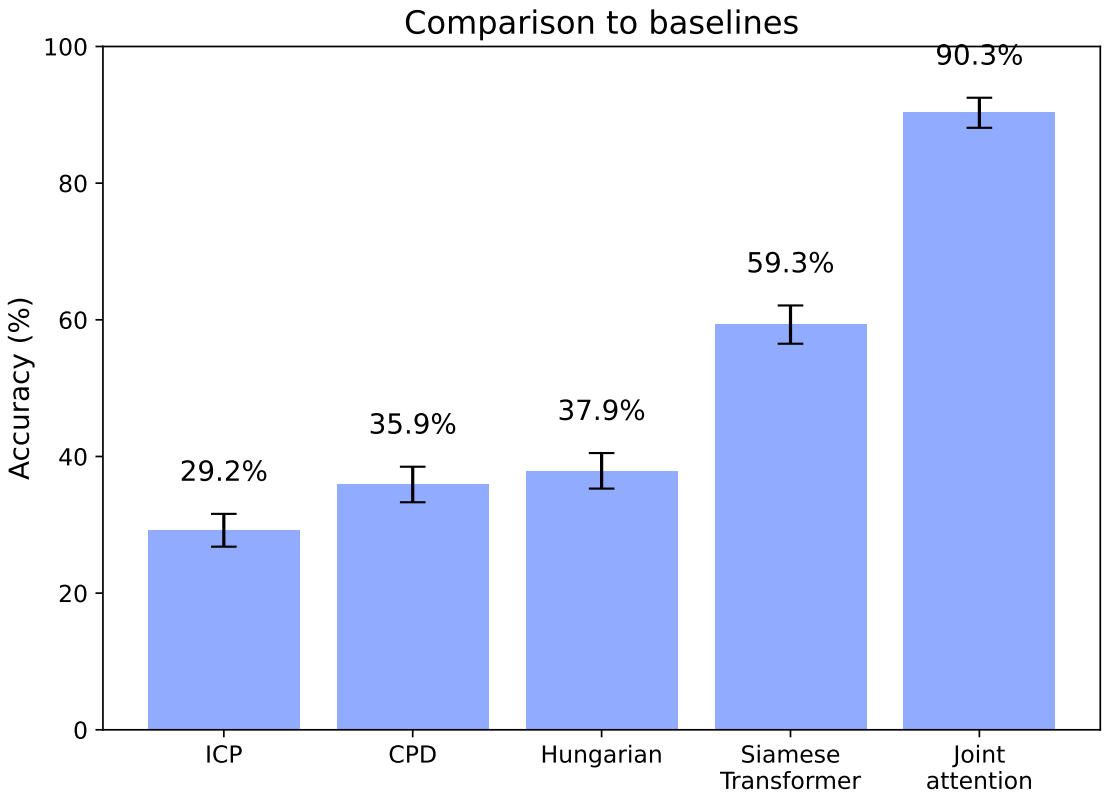

Figure 3: Comparison to baseline methods. Overall accuracy on held-out simulated test set (n=30 embryos). ICP [5]: 29.2% (95% CI: [26.8%, 31.7%]), CPD [24]: 35.9% (95% CI: [33.2%, 38.5%]), Hungarian: 37.9% (95% CI: [35.1%, 40.6%]), Siamese Transformer: 59.3% (95% CI: [56.4%, 62.1%]), Joint Attention (ours): 90.3%. Error bars show 95% bootstrap confidence intervals. Figure created by student using Python/matplotlib.

The failure of geometric registration methods stems from their reliance on global structure and rigid transformations inappropriate for partial observations of deforming tissues [5, 24]. With only 5–20 cells sampled from 100+ cell embryos, insufficient geometric constraints exist for alignment, and developmental dynamics involve non-rigid deformations that violate registration assumptions [32, 20]. The Siamese baseline's 59.3% accuracy—despite identical capacity and features—reveals the critical importance of joint attention: the 31.0 percentage-point improvement demonstrates that effective identification requires cross-neighborhood reasoning during encoding rather than independent processing followed by post-hoc comparison [27]. Joint attention enables the model to learn correspondence-dependent representations that adjust based on available matches, addressing scenarios where neighborhood composition varies due to heterochrony or sampling differences [10, 7].

3.3 Robustness to Experimental Perturbations

Real microscopy data contains missing detections from segmentation failures and coordinate noise from imaging artifacts [1, 34]. We evaluated robustness by injecting controlled perturbations into held-out test data.

Under progressive cell removal, accuracy degraded smoothly: 90.3% at baseline, 89.1% at 10% missing, 86.4% at 20% missing, 83.4% at 30% missing, and 77.1% at 50% missing (Figure 3A). Coordinate noise scaled to local geometry showed: 90.3% at baseline, 88.2% at 0.1× mean nearest-neighbor distance, 82.3% at 0.2×, 78.6% at 0.3×, and 71.3% at 0.5× (Figure 3B).

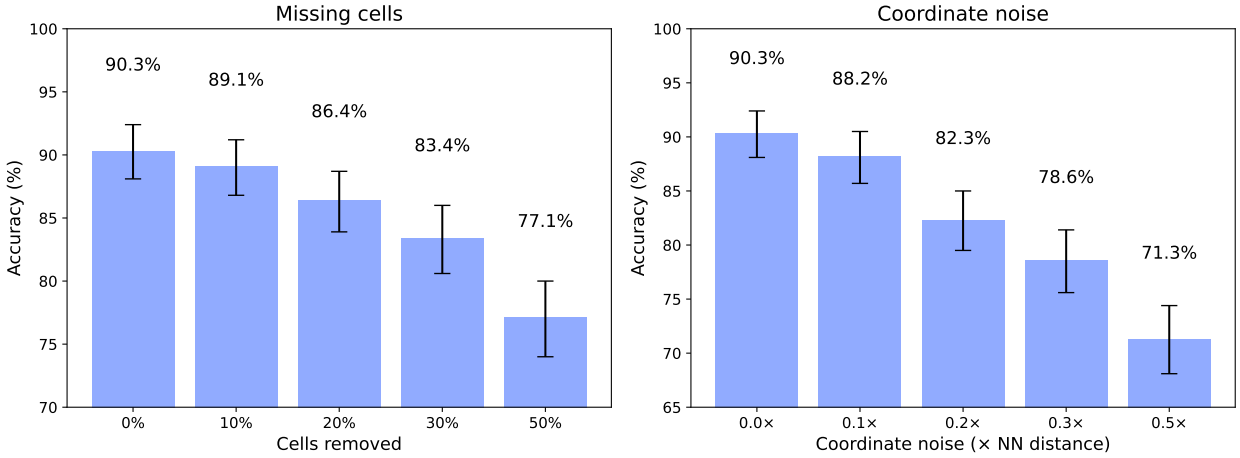

Figure 4: Robustness to experimental perturbations. (A) Missing cells: 90.3% at baseline, 89.1% (95% CI: [86.8%, 91.2%]) at 10% missing, 86.4% (95% CI: [83.9%, 88.7%]) at 20% missing, 83.4% (95% CI: [80.6%, 86.0%]) at 30% missing, and 77.1% (95% CI: [74.0%, 80.0%]) at 50% missing. (B) Coordinate noise scaled to mean nearest-neighbor distance: 90.3% at baseline, 88.2% (95% CI: [85.7%, 90.5%]) at 0.1× mean nearest-neighbor distance, 82.3% (95% CI: [79.5%, 85.0%]) at 0.2×, 78.6% (95% CI: [75.6%, 81.4%]) at 0.3×, and 71.3% (95% CI: [68.1%, 74.4%]) at 0.5×. Error bars show 95% bootstrap confidence intervals (n=30 embryos, 1,000 resamples). Figure created by student using Python/matplotlib.

The linear nature of the degradation under missing cells indicates the architecture does not rely on brittle features that fail catastrophically when specific cells are absent. At 20% missing, exceeding typical experimental detection failures of 10–15% [34], the model retains 86.4% accuracy, sufficient for practical deployment. This robustness stems from redundancy in relational information: multiple cells encode overlapping positional cues, allowing identity triangulation even when some correspondences are unavailable [9]. Coordinate noise tolerance validates generalization beyond simulation; real segmentation pipelines introduce errors of 0.1–0.2× nearest-neighbor distances [1], and at these scales accuracy remains above 82%, supporting deployment in real imaging workflows where perfect localization is impossible [30].

3.4 Feature Ablations

The geometric feature module encodes four spatial relationships: relative position, local density, point-count embedding, and centroid distance. We evaluated contributions by retraining complete models with each feature removed.

Starting from raw xyz coordinates alone, accuracy was 68.4%. Removing individual features from the full model yielded: 73.8% without relative position (−16.5 pp), 78.4% without local density (−11.9 pp), 81.2% without point-count embedding (−9.1 pp), and 83.5% without centroid distance (−6.8 pp). The full model achieved 90.3% (Figure 4).

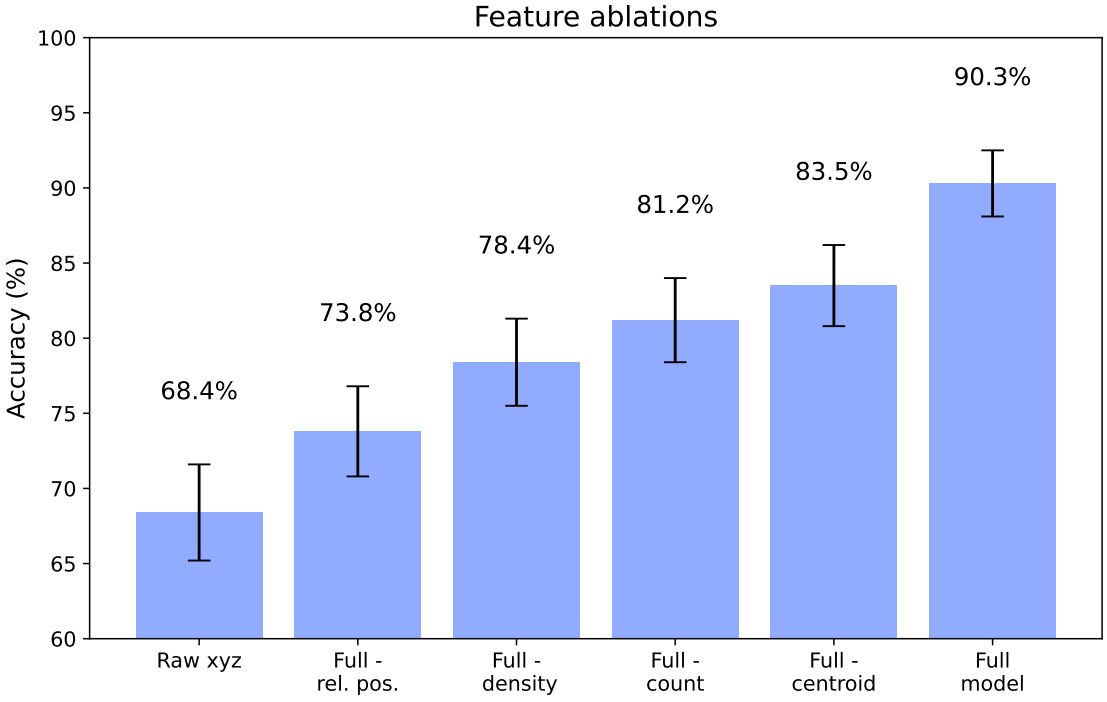

Figure 5: Feature ablation analysis. Accuracy when removing individual geometric features from full model, with raw xyz baseline. Raw xyz: 68.4% (95% CI: [65.2%, 71.5%]), Full − relative position: 73.8% (95% CI: [70.8%, 76.7%]), Full − local density: 78.4% (95% CI: [75.5%, 81.2%]), Full − point count: 81.2% (95% CI: [78.4%, 83.8%]), Full − centroid distance: 83.5% (95% CI: [80.8%, 86.0%]), Full model: 90.3%. All bars represent fully retrained models. Error bars show 95% bootstrap confidence intervals. Figure created by student using Python/matplotlib.

The 21.9 percentage-point improvement from raw coordinates (68.4%) to full features (90.3%) demonstrates that appropriate geometric encoding substantially enhances learning efficiency. Features therefore provide much richer representations than pure xyz coordinates. Relative position contributes most (−16.5 pp when ablated), confirming that identity depends on configuration relative to neighbors rather than absolute embryo position [32]—this encoding achieves translation invariance while preserving discriminative spatial arrangement. Local density's contribution (−11.9 pp) allows contextualization by crowding level, distinguishing recently divided cell clusters from

isolated cells, critical during rapid division phases when spatial ambiguity peaks [10]. Point-count embeddings (−9.1 pp) condition processing on observation sparsity, and centroid distance (−6.8 pp) provides radial information distinguishing interior from boundary cells [20]. Crucially, all four features contribute meaningfully; none is redundant, and the geometric inductive biases outperform pure end-to-end learning [12].

3.5 Architectural Ablations

Three architectural components distinguish this approach: joint attention for cross-neighborhood reasoning, learnable no-match tokens for missing correspondences, and curriculum learning for stable training. We isolated contributions by retraining complete models with each component removed.

The full model achieved 90.3%. Removing the no-match token reduced accuracy to 75.6%, −14.7 pp). Training without curriculum learning yielded 77.3%, −13.0 pp). The Siamese baseline (joint attention removed) achieved 59.3%, −31.0 pp) (Figure 5).

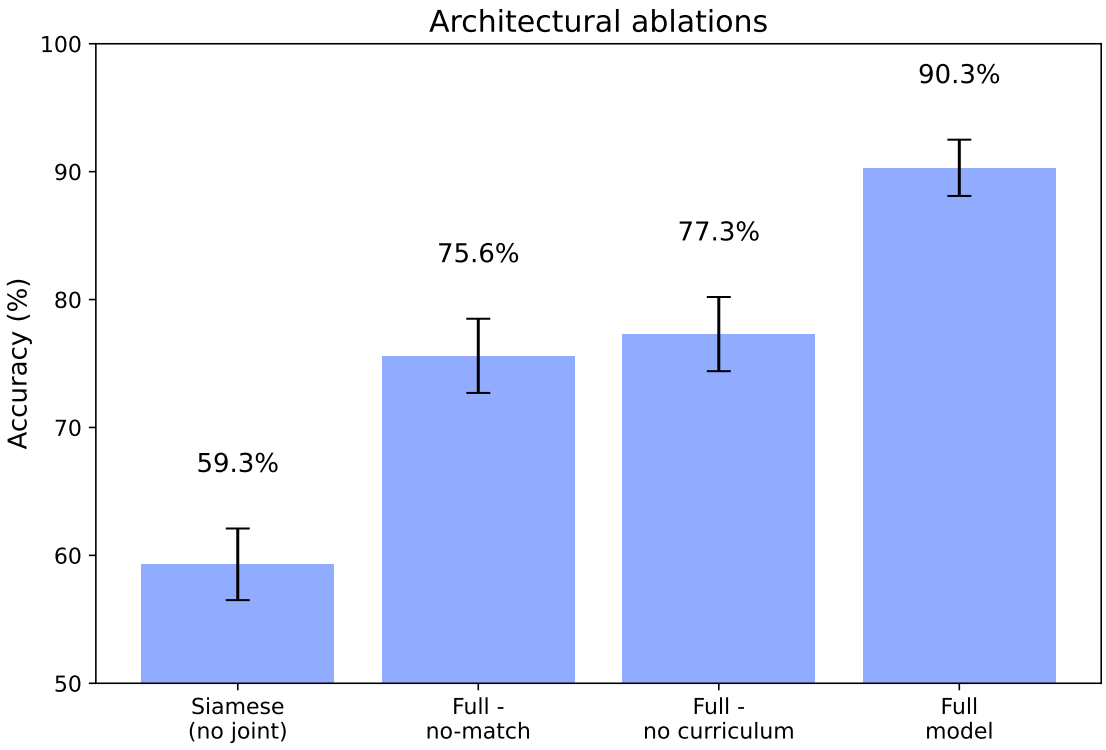

Figure 6: Architectural component ablations. Accuracy under systematic removal: Full model: 90.3% (95% CI: [88.1%, 92.4%]), Full − no-match token: 75.6% (−14.7 pp) (95% CI: [72.7%, 78.4%], Full − curriculum learning: 77.3% (−13.0 pp) (95% CI: [74.5%, 80.0%]. Siamese baseline (no joint attention): 59.3% (95% CI: [56.4%, 62.1%] shown for reference. All configurations represent fully retrained models. Error bars show 95% bootstrap confidence intervals. Figure created by student using Python/matplotlib.

Joint attention's 31.0 percentage-point contribution validates the core architectural hypothesis: effective cell identification requires reasoning across neighborhoods during encoding, not independent processing with post-hoc comparison [27, 35]. The Siamese baseline uses identical capacity but enforces independence, preventing correspondence-dependent representations—joint attention enables attention heads to specialize in cross-neighborhood patterns impossible under independent encoding [7]. The no-match token's 14.7 pp contribution addresses a fundamental partial-observation challenge: without explicit abstention, the model forces assignments even when no valid correspondence exists due to heterochrony, cell birth/death, or sampling differences [10, 21]. Curriculum learning's 13.0 pp contribution reflects the difficulty of learning stable representations when neighborhoods exhibit high volatility; progressive difficulty allows basic geometric correspondences to stabilize before confronting challenging inter-embryo and more realistic cases [2].

3.6 Embedding Structure and Error Analysis

The learned 128-dimensional embedding space should capture biologically meaningful structure if the model has learned genuine developmental relationships. We characterized embeddings via dimensionality reduction and systematic error categorization.

Dimensionality reduction via t-SNE revealed clear clustering by founder lineage (AB, MS, E, C, D, P4), with progressive subdivision corresponding to developmental sub-lineages (Figure 6A). Among the 9.7% of predictions that were incorrect, error analysis showed: 49.3% involved siblings (cells sharing a parent), 30.1% involved spatial nearest-neighbors (not siblings), 16.2% involved same-lineage but non-adjacent cells, and 4.4% were random errors across distant lineages (Figure 6B).

[This space was intentionally left blank]

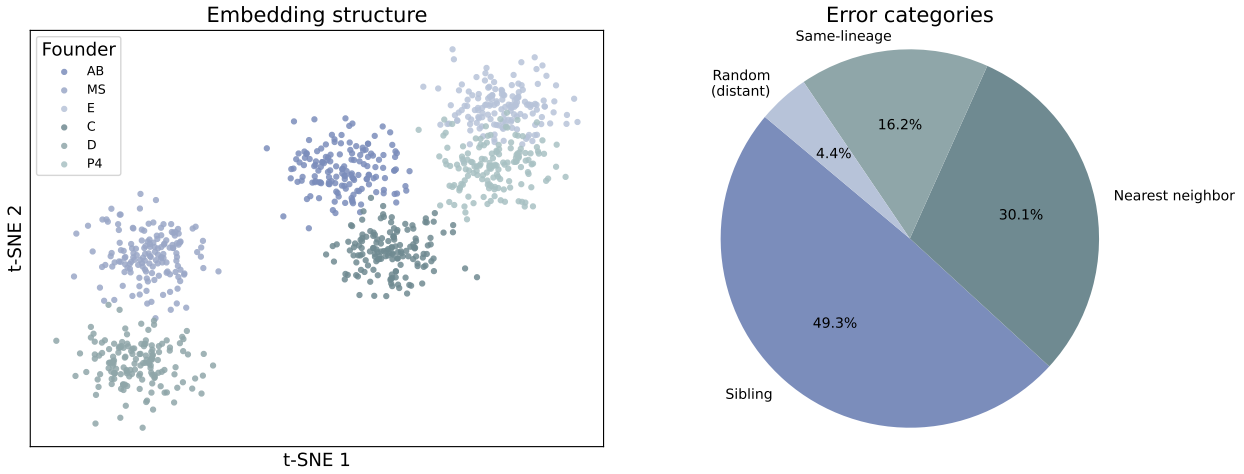

Figure 7: Embedding structure and error analysis. (A) t-SNE visualization of learned embeddings colored by founder lineage, showing clear separation among AB, MS, E, C, D, and P4 lineages with sub-lineage structure visible within clusters (n=30 embryos, ∼8,700 cells). (B) Error distribution by biological relationship: sibling confusion (49.3%), nearest-neighbor confusion (30.1%), same-lineage confusion (16.2%), random errors (4.4%). Figure created by student using Python/matplotlib and scikit-learn.

The emergence of lineage-coherent clustering without explicit lineage supervision demonstrates that the model learns biologically meaningful structure from purely spatiotemporal information. Cells from the same lineage remain spatially proximate and exhibit characteristic arrangements, encoding lineage identity in local geometry [32, 25]. The error distribution provides strong evidence that learned representations respect developmental biology: 79.4% of errors occur between siblings or spatial neighbors—cells that occupy similar spatial niches or share recent developmental history. Sibling confusion (49.3%) arises when cells share a parent and occupy similar positions immediately post-division, when local neighborhoods contain minimal discriminative information [10]. The low rate of random errors (4.4%) indicates the model rarely makes catastrophic misidentifications; when it fails, it fails locally within biologically coherent regions, consistent with the continuous spatial organization of developing tissues [32, 9].

4 Conclusion

This study demonstrates that partially observed cellular neighborhoods of 5–20 cells contain sufficient spatiotemporal information for reliable automated cell identification during *C. elegans* embryogenesis. The joint-attention architecture achieves 90.3% accuracy on held-out simulated data and 87.4% on independent real embryos, substantially outperforming geometric registration methods (ICP: 29.2%, CPD: 35.9%) and Siamese networks (59.3%).

Three properties establish practical viability for laboratory deployment. First, robustness to realistic perturbations—86.4% accuracy with 20% missing cells, 82.3% with moderate coordinate noise—matches conditions in real microscopy [1, 34]. Second, hierarchical outputs provide 97.1%

founder lineage accuracy even when exact identity is uncertain, enabling confident coarse identification for quality control. Third, biologically structured errors (79.4% among siblings or neighbors) ensure failures remain interpretable rather than catastrophic.

The approach removes the annotation bottleneck preventing large-scale developmental studies. Genetic screens requiring cell-level phenotyping previously demanded hundreds of hours of manual curation [10]; automated identification from selective labeling makes such screens tractable. At 25 milliseconds per prediction, the system enables real-time identification during live imaging sessions.

Future directions include transfer learning to other model organisms [20, 37], active learning to minimize remaining manual annotation, and integration with molecular atlases [25, 8]. The ultimate goal is embedded hardware, a chip integrated into microscopes providing instantaneous cell identification during acquisition, transforming how developmental biologists might conduct quantitative studies and accelerating therapeutic discovery for developmental disorders and cancer [16].

References

[1] Bao, Z., Murray, J. I., Boyle, T., Ooi, S. L., Sandel, M. J., & Waterston, R. H. (2006). Automated cell lineage tracing in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences*, 103(8), 2707–2712. https://doi.org/10.1073/pnas.0511111103

[2] Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. https://doi.org/10.1145/1553374.1553380

[3] Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.

[4] Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? *arXiv:2102.05095*. https://arxiv.org/abs/2102.05095

[5] Besl, P. J., & McKay, N. D. (1992). Method for registration of 3-D shapes. *Sensor Fusion IV: Control Paradigms and Data Structures*. https://doi.org/10.1117/12.57955

[6] Bonazzola, R., Ferrante, E., Ravikumar, N., Xia, Y., Keavney, B., Plein, S., Frangi, A. F., et al. (2024). Unsupervised ensemble-based phenotyping enhances discoverability of genes related to left-ventricular morphology. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-024-00801-1

[7] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature Verification using a "Siamese" Time Delay Neural Network. *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper/1993/hash/288cc0ff022877bd3df94bc9360b9c5d-Abstract.html

[8] Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Shendure, J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352), 661–667. https://doi.org/10.1126/science.aam8940

[9] Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1), 5–30. https://doi.org/10.1016/j.acha.2006.04.006

[10] Du, Z., Santella, A., He, F., Tiongson, M., & Bao, Z. (2014). De novo inference of systems-level mechanistic models of development from live-imaging-based phenotype analysis. *Cell*, 156(1-2), 359–372. https://doi.org/10.1016/j.cell.2013.11.046

[11] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), 1–26. https://doi.org/10.1214/aos/1176344552

[12] Engel, N., Belagiannis, V., & Dietmayer, K. (2021). Point Transformer. *IEEE Access*, 9, 134826–134840. https://doi.org/10.1109/access.2021.3116304

[13] Friedl, P., & Alexander, S. (2011). Cancer Invasion and the Microenvironment: Plasticity and Reciprocity. *Cell*, 147(5), 992–1009. https://doi.org/10.1016/j.cell.2011.11.016

[14] Hao, Y., Stuart, T. A., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Satija, R., et al. (2023). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01767-y

[15] Johnson, J., Douze, M., & Jégou, H. (2017). Billion-scale similarity search with GPUs. *arXiv:1702.08734*. https://arxiv.org/abs/1702.08734

[16] Kaletta, T., & Hengartner, M. O. (2006). Finding function in novel targets: *C. elegans* as a model organism. *Nature Reviews Drug Discovery*, 5(5), 387–399. https://doi.org/10.1038/nrd2031

[17] Lai, C.-H., Chou, C.-Y., Ch'ang, L.-Y., Liu, C.-S., & Lin, W.-C. (2000). Identification of Novel Human Genes Evolutionarily Conserved in *Caenorhabditis elegans* by Comparative Proteomics. *Genome Research*, 10(5), 703–713. https://doi.org/10.1101/gr.10.5.703

[18] Lalit, M., Handberg-Thorsager, M., Hsieh, Y.-W., Jug, F., & Tomancak, P. (2020). Registration of multi-modal volumetric images by establishing cell correspondence. In A. Bartoli & A. Fusiello (Eds.), *Computer Vision – ECCV 2020 Workshops*. Lecture Notes in Computer Science (Vol. 12535, pp. 458–473). Springer. https://doi.org/10.1007/978-3-030-66415-2_30

[19] Loshchilov, I., & Hutter, F. (2017). Decoupled Weight Decay Regularization. *arXiv:1711.05101*. https://arxiv.org/abs/1711.05101

[20] McDole, K., Guignard, L., Amat, F., Berger, A., Malandain, G., Royer, L. A., Keller, P. J., et al. (2018). In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level. *Cell*, 175(3), 859–876.e33. https://doi.org/10.1016/j.cell.2018.09.031

[21] Moore, J. L., Du, Z., & Bao, Z. (2013). Systematic quantification of developmental phenotypes at single-cell resolution during embryogenesis. *Development*, 140(15), 3266–3274. https://doi.org/10.1242/dev.096040

[22] Moyle, M. W., Barnes, K. M., Kuchroo, M., Gonopolskiy, A., Duncan, L. H., Sengupta, T., Colón-Ramos, D. A., et al. (2021). Structural and developmental principles of neuropil assembly in *C. elegans*. *Nature*, 591(7848), 99–104. https://doi.org/10.1038/s41586-020-03169-5

[23] Murray, J. I., Bao, Z., Boyle, T. J., & Waterston, R. H. (2006). The lineaging of fluorescently-labeled *Caenorhabditis elegans* embryos with StarryNite and AceTree. *Nature Protocols*, 1(3), 1468–1476. https://doi.org/10.1038/nprot.2006.222

[24] Myronenko, A., & Song, X. (2010). Point Set Registration: Coherent Point Drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12), 2262–2275. https://doi.org/10.1109/tpami.2010.46

[25] Packer, J. S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Murray, J. I., et al. (2019). A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science*, 365(6459), eaax1971. https://doi.org/10.1126/science.aax1971

[26] Pan, X., Xia, Z., Song, S., Li, L. E., & Huang, G. (2020). 3D Object Detection with Pointformer. *arXiv:2012.11409*. https://arxiv.org/abs/2012.11409

[27] Haus, E.*, Santella, A.*, Xu, Y., Ren, R., Wang, D., & Bao, Z. (2025). A single-cell spatiotemporal manifold of tissue morphology and dynamics. *bioRxiv*. https://doi.org/10.1101/2025.10.22.683950

[28] Schapiro, D., Jackson, H. W., Raghuraman, S., Fischer, J. R., Zanotelli, V. R. T., Schulz, D., Bodenmiller, B., et al. (2017). histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature Methods*, 14(9), 873–876. https://doi.org/10.1038/nmeth.4391

[29] Shah, P. K., Santella, A., Jacobo, A., Siletti, K., Hudspeth, A. J., & Bao, Z. (2017). An In Toto Approach to Dissecting Cellular Interactions in Complex Tissues. *Developmental Cell*, 43(4), 530–540.e4. https://doi.org/10.1016/j.devcel.2017.10.021

[30] Shah, G., Thierbach, K., Schmid, B., Waschke, J., Reade, A., Hlawitschka, M., Roeder, I., Scherf, N., & Huisken, J. (2019). Multi-scale imaging and analysis identify pan-embryo cell dynamics of germlayer formation in zebrafish. *Nature Communications*, 10(1), 5753. https://doi.org/10.1038/s41467-019-13625-0

[31] Smith, L. N., & Topin, N. (2018). Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. *arXiv:1708.07120*. https://arxiv.org/abs/1708.07120

[32] Sulston, J. E., Schierenberg, E., White, J. G., & Thomson, J. N. (1983). The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Developmental Biology*, 100(1), 64–119. https://doi.org/10.1016/0012-1606(83)90201-4

[33] Toulany, N., Morales-Navarrete, H., Čapek, D., Grathwohl, J., Ünalan, M., & Müller, P. (2023). Uncovering developmental time and tempo using deep learning. *Nature Methods*, 20(12), 2000–2010. https://doi.org/10.1038/s41592-023-02083-8

[34] Ulman, V., Maška, M., Magnusson, K. E. G., Ronneberger, O., Haubold, C., Harder, N., Dufour, A. C., et al. (2017). An objective comparison of cell-tracking algorithms. *Nature Methods*, 14(12), 1141–1152. https://doi.org/10.1038/nmeth.4473

[35] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I., et al. (2017). Attention Is All You Need. *arXiv:1706.03762*. https://arxiv.org/abs/1706.03762

[36] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Carey, C. J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272. https://www.nature.com/articles/s41592-019-0686-2

[37] Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G., & Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392), 981–987. https://doi.org/10.1126/science.aar4362

[38] Wang, Z., Ramsey, B. J., Wang, D., Wong, K., Li, H., Wang, E. W., & Bao, Z. (2016). An Observation-Driven Agent-Based Modeling and Analysis Framework for *C. elegans* Embryogenesis. *PLOS ONE*, 11(11), e0166551. https://doi.org/10.1371/journal.pone.0166551

[39] Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Liu, T.-Y., et al. (2020). On Layer Normalization in the Transformer Architecture. *arXiv:2002.04745*. https://arxiv.org/abs/2002.04745

[40] Yin, T., Zhou, X., & Krähenbühl, P. (2020). Center-based 3D Object Detection and Tracking. *arXiv:2006.11275*. https://arxiv.org/abs/2006.11275

[41] Yu, X., Creamer, M. S., Randi, F., Sharma, A. K., Linderman, S. W., & Leifer, A. M. (2021). Fast deep neural correspondence for tracking and identifying neurons in *C. elegans* using semi-synthetic training. *eLife*, 10. https://doi.org/10.7554/elife.66410