# Co-adaptive Graph Learning through Coupled Spectral Refinement for 3D Anomaly Detection

Supplementary Material

## 1. Additional Dataset Details

### 1.1. MVTec-3D AD Dataset

All anomaly localization experiments in this work are conducted on the MVTec-3D Anomaly Detection (MVTec-3D AD) [6] dataset, which is the first benchmark designed specifically for 3D industrial anomaly detection. The dataset contains ten object categories with a total of 2,656 training samples and 1,137 testing samples. Each sample is obtained using an industrial structured-light sensor, and the geometry is stored as three coordinate channels corresponding to the $x$, $y$, and $z$ positions of each surface point. These three channels can be directly converted into point-cloud representations, and in our work, we use only the point-cloud data. The dataset provides high-resolution 3D geometry along with ground-truth annotations for local surface defects such as cracks, dents, and missing material, making it a comprehensive benchmark for evaluating 3D geometry-based anomaly localization models.

### 1.2. Bone Side Estimation (BSE) Dataset

For the Bone Side Estimation (BSE) task, we construct a benchmark by collecting bone surface point clouds from several publicly available datasets. These datasets include left and right instances of major long bones, covering different subjects and acquisition setups. The benchmark spans Femur, Hip, Tibia, and Fibula bone surfaces and reflects natural variations in shape and scanning conditions, providing a diverse and challenging testbed for bilateral classification.

Specifically, Femur and Hip surfaces are sourced from the dataset of Fisher et al. [12]; right Tibia and Fibula surfaces are taken from the SSM-Tibia dataset [19],; and additional Femur, Tibia, and Fibula samples are included from the Imperial College London (ICL) dataset [29]. The complete composition of the benchmark is summarized in Table 6.

| Dataset | Femur (L/R) | Hip (L/R) | Tibia (L/R) | Fibula (L/R) |
|---|---|---|---|---|
| Fisher et al. | 18 / 19 | 20 / 20 | - | - |
| SSM-Tibia | - | - | - / 30 | - / 30 |
| ICL | 35 / 35 | - | 35 / 35 | 35 / 35 |
| **Total** | 53 / 54 | 20 / 20 | 35 / 65 | 35 / 65 |

Table 6. Summary of point-cloud bone structures collected for the BSE benchmark. L and R denote left and right bones, respectively, and S refers to scans from the internal source.

## 2. Detailed Training Details

The training of CoLE-LSA is carried out on the MVTec 3D-AD dataset, which provides both normal and defective 3D point clouds. For each cloud, local neighborhoods are constructed using a radius-based search with $r = 2 \times \overline{d}_{\text{nn}}$, where $\overline{d}_{\text{nn}}$ denotes the average nearest-neighbor distance. This adaptive radius normalizes geometric scale across differently sized or sampled objects. Neighborhood sizes vary naturally with local point density but are capped at $k_{\max} = 48$ to ensure efficiency. Each point is treated as a potential subgraph center, and the total number of subgraphs scales proportionally with object size, subject to a maximum of $4096$ to prevent memory overflow. Farthest-point sampling is used to select subgraph centers uniformly, and overlapping neighborhoods promote geometric smoothness during anomaly localization.

The CoLE-LSA encoder employs a looped transformer architecture in which a single transformer block is executed repeatedly for $T = 3$ refinement loops. Unlike a standard transformer applied on a fixed graph, each loop in our model performs *co-evolution* of both graph structure and node embeddings. At loop $t$, the transformer updates embeddings from $H^{(t-1)}$ to $H^{(t)}$. Using these refined embeddings, we recompute pairwise distances, angle encodings, and a new attention matrix, yielding a dynamic adjacency $A^{(t)}$. This enables the graph topology and the feature representations to evolve jointly: normal regions tend to converge to stable embeddings across loops, whereas anomalous regions produce unstable or divergent behavior. The full refinement trajectory is unrolled during training, and gradients propagate through all $T$ iterations, analogous to backpropagation-through-time. As a result, the model must be trained and evaluated with the same number of loops; training without loops would not produce meaningful co-evolution during inference.

Optimization is performed using AdamW with an initial learning rate of $1 \times 10^{-4}$, cosine annealing decay, and gradient clipping with a maximum norm of 5.0. The model is trained for 100 epochs with a batch size of one object per batch on a 40GB A100 GPU. We use the composite co-evolution loss described in *Spectral Analysis with Looped Self-Attention (CoLE-LSA)* section of the main paper, with weights $\lambda_1 = 0.2$, $\lambda_2 = 1.0$, and $\lambda_3 = 0.3$, and the angular balance coefficient in $\mathcal{L}_{\text{geo}}$ is fixed to $\beta = 1.0$.

During inference, anomaly scores are computed from the final refinement loop using pointwise residuals between em-

beddings of a normal reference and a test cloud:

$$s_i = \|H_{n,i}^{(T)} - H_{a,i}^{(T)}\|_2.$$

These residuals are smoothed and normalized to produce anomaly heatmaps, where higher values correspond to localized structural deviations. Evaluation follows the MVTec 3D-AD protocol using the AUPRO metric, which measures region-level anomaly localization accuracy.

# 3. Additional Experiments and Ablation Studies

## 3.1. Ablations for CoLE-SR

CoLE-SR relies on spectral embeddings and iterative refinement, making its behavior sensitive to both the spectral basis size and the number of refinement loops. To disentangle the contribution of each component, we perform detailed ablations across all object categories, examining how increasing the number of eigenvectors enhances spectral resolution, how iterative refinement affects correspondence stability, and how performance changes when replacing the baseline RANSAC registration with a stronger Gaussian Mixture Model–Coherent Point Drift (GMM–CPD) alignment [28]. The following results provide a comprehensive class-wise analysis of these design choices.

### 3.1.1. Extended Ablation on Spectral Basis Size and Loop Count

Table 7 provides the full class-wise evaluation of CoLE-SR for different spectral basis sizes ($t$) and refinement loops ($s$). As observed in the main paper, performance improves steadily from $s$=0 to $s$=3 due to the progressive refinement of spectral embeddings, while additional iterations lead to mild oversmoothing. Increasing the number of eigenvectors from $t$=100 to $t$=200 consistently enhances accuracy across all categories, confirming that higher spectral resolution improves correspondence stability.

### 3.1.2. CoLE-SR with Gaussian Mixture Model + Coherent Point Drift Registration

To evaluate the effect of a stronger registration backend, we replace the RANSAC-based alignment in CoLE-SR with a Gaussian Mixture Model (GMM) initialization followed by Coherent Point Drift (CPD) [28] registration. Table 8 reports class-wise results for multiple spectral basis sizes ($t$) and loop counts ($s$). As expected, GMM+CPD generally improves performance across most categories due to its tighter probabilistic alignment. The best performance is obtained for $t$=200, $s$=3, which closely approaches the behavior observed in CoLE-LSA (pos+angle), while still preserving the characteristic trend of iterative refinement peaking at three loops.

## 3.2. Ablations for CoLE-LSA

CoLE-LSA relies on iterative geometric refinement, making its performance sensitive to neighborhood construction, adjacency modeling, sampling density, and loss design. To isolate the contribution of each component, we perform detailed ablations on the major architectural and algorithmic choices, including neighborhood size, dynamic attention-based graph formation, subgraph sampling, subgraph budget, and loss terms. The following experiments quantify how each factor influences correspondence stability and overall detection accuracy.

### 3.2.1. Neighborhood Size ($k$)

CoLE-LSA builds local neighborhoods using a radius-based search with a cap of $k$ nearest points. The value of $k$ controls the trade-off between geometric detail and noise: too few neighbors under-represent local structure, while too many introduce unstable or irrelevant interactions. We evaluate four settings, $k \in \{16, 32, 48, 64\}$, summarized in Table 9. Performance peaks at $k = 48$, which provides enough local support to capture surface geometry while avoiding the oversmoothing that appears at larger $k$. This value offers the most stable behavior across all object categories and is used in all main evaluations.

### 3.2.2. Dynamic Attention vs. Fixed kNN Adjacency

In CoLE-LSA, the attention weights are used to construct a learned adjacency matrix that is recomputed at every refinement loop, producing a dynamic and feature-dependent neighborhood structure. To assess its importance, we compare against a fixed $k$NN graph ($k = 32$), where neighbors are chosen purely by Euclidean distance and remain unchanged. As shown in Table 10, the dynamic adjacency consistently outperforms fixed $k$NN across all categories, confirming that adaptive neighborhood reasoning is crucial for capturing fine geometric variations.

### 3.2.3. Subgraph Sampling Strategy

The choice of subgraph centers directly affects the spatial coverage of the object surface and the stability of the refinement loops. Random sampling often produces uneven or clustered coverage, while uniform grid-based sampling reduces variance but is sensitive to local point density and object shape. Farthest point sampling (FPS), in contrast, progressively selects points that maximize pairwise separation, leading to a more homogeneous set of centers and better representation of both smooth and high-curvature regions. As reported in Table 11, FPS consistently achieves the highest accuracy across categories, demonstrating that uniform geometric coverage is critical for reliable correspondence refinement.

| Method (RANSAC) | Bagel | Cable | Carrot | Cookie | Dowel | Foam | Peach | Potato | Rope | Tire |
|---|---|---|---|---|---|---|---|---|---|---|
| Ours (100, 0) | 0.654 | 0.645 | 0.802 | 0.715 | 0.811 | 0.584 | 0.747 | 0.899 | 0.752 | 0.733 |
| Ours (100, 1) | 0.715 | 0.701 | 0.842 | 0.781 | 0.846 | 0.741 | 0.781 | 0.928 | 0.804 | 0.775 |
| Ours (100, 2) | 0.787 | 0.735 | 0.871 | 0.829 | 0.874 | 0.796 | 0.811 | 0.946 | 0.819 | 0.792 |
| Ours (100, 3) | 0.851 | 0.782 | 0.893 | 0.867 | 0.897 | 0.835 | 0.834 | 0.962 | 0.824 | 0.811 |
| Ours (100, 4) | 0.832 | 0.761 | 0.869 | 0.841 | 0.877 | 0.821 | 0.808 | 0.947 | 0.802 | 0.791 |
| Ours (100, 5) | 0.827 | 0.752 | 0.846 | 0.824 | 0.869 | 0.813 | 0.791 | 0.939 | 0.794 | 0.783 |
| Ours (200, 0) | 0.704 | 0.630 | 0.728 | 0.732 | 0.812 | 0.705 | 0.781 | 0.915 | 0.764 | 0.710 |
| Ours (200, 1) | 0.741 | 0.675 | 0.812 | 0.781 | 0.842 | 0.812 | 0.826 | 0.941 | 0.809 | 0.763 |
| Ours (200, 2) | 0.834 | 0.745 | 0.884 | 0.832 | 0.879 | 0.861 | 0.863 | 0.957 | 0.828 | 0.801 |
| **Ours (200, 3)** | **0.932** | **0.802** | **0.913** | **0.869** | **0.909** | **0.895** | **0.887** | **0.967** | **0.932** | **0.845** |
| Ours (200, 4) | 0.928 | 0.791 | 0.898 | 0.855 | 0.901 | 0.876 | 0.871 | 0.898 | 0.821 | 0.831 |
| Ours (200, 5) | 0.928 | 0.782 | 0.891 | 0.849 | 0.892 | 0.865 | 0.867 | 0.889 | 0.814 | 0.817 |

Table 7. Full class-wise ablation of CoLE-SR for different spectral basis sizes ($t$) and refinement loops ($s$). Accuracy increases consistently up to three loops, with $t$=200, $s$=3 achieving the best overall performance.

| Method (GMM+CPD) | Bagel | Cable | Carrot | Cookie | Dowel | Foam | Peach | Potato | Rope | Tire |
|---|---|---|---|---|---|---|---|---|---|---|
| Ours (100, 3) | 0.885 | 0.794 | 0.912 | 0.884 | 0.909 | 0.861 | 0.865 | 0.968 | 0.842 | 0.829 |
| Ours (100, 4) | 0.852 | 0.777 | 0.888 | 0.861 | 0.893 | 0.840 | 0.829 | 0.958 | 0.821 | 0.808 |
| Ours (100, 5) | 0.842 | 0.769 | 0.871 | 0.846 | 0.887 | 0.832 | 0.813 | 0.953 | 0.812 | 0.801 |
| Ours (200, 3) | **0.947** | **0.809** | **0.932** | **0.905** | **0.922** | **0.902** | **0.908** | **0.972** | **0.939** | **0.875** |
| Ours (200, 4) | 0.940 | 0.802 | 0.912 | 0.892 | 0.916 | 0.892 | 0.898 | 0.962 | 0.913 | 0.863 |
| Ours (200, 5) | 0.938 | 0.796 | 0.904 | 0.885 | 0.909 | 0.884 | 0.892 | 0.955 | 0.903 | 0.854 |

Table 8. Full class-wise CoLE-SR performance using GMM+CPD registration for different spectral basis sizes ($t$) and refinement loops ($s$). GMM+CPD consistently improves alignment quality over RANSAC, with $t$=200, $s$=3 yielding the best overall results.

| $k$ | Bagel | Cable | Carrot | Cookie | Dowel | Foam | Peach | Potato | Rope | Tire | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | 0.941 | 0.874 | 0.852 | 0.886 | 0.871 | 0.918 | 0.902 | 0.893 | 0.872 | 0.908 | 0.892 |
| 32 | 0.968 | 0.921 | 0.903 | 0.933 | 0.912 | 0.945 | 0.931 | 0.924 | 0.902 | 0.934 | 0.918 |
| **48** | **0.988** | **0.947** | **0.932** | **0.958** | **0.941** | **0.969** | **0.954** | **0.948** | **0.923** | **0.951** | **0.951** |
| 64 | 0.972 | 0.936 | 0.921 | 0.949 | 0.935 | 0.961 | 0.948 | 0.939 | 0.917 | 0.944 | 0.942 |

Table 9. Effect of neighborhood size $k$ on performance across all ten MVTec 3D-AD categories. A cap of $k = 48$ yields the best overall results.

| Method | Bagel | Cable | Carrot | Cookie | Dowel | Foam | Peach | Potato | Rope | Tire | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Static $k$NN | 0.952 | 0.889 | 0.871 | 0.904 | 0.889 | 0.928 | 0.915 | 0.907 | 0.892 | 0.921 | 0.898 |
| **Dynamic Attention** | **0.988** | **0.947** | **0.932** | **0.958** | **0.941** | **0.969** | **0.954** | **0.948** | **0.923** | **0.951** | **0.951** |

Table 10. Class-wise comparison of attention-based adaptive adjacency vs. fixed $k$NN connectivity across all ten categories. Dynamic attention consistently outperforms static $k$NN neighborhoods, demonstrating the importance of adaptive relational modeling.

#### 3.2.4. Maximum Number of Subgraphs ($N_{s,\max}$)

The number of extracted subgraphs controls how densely the object surface is partitioned during refinement. When $N_{s,\max}$ is too small, large or geometrically complex objects are under-sampled, causing important regions-especially high-curvature areas and fine defect structures-to be missed. Increasing $N_{s,\max}$ improves coverage by generating more

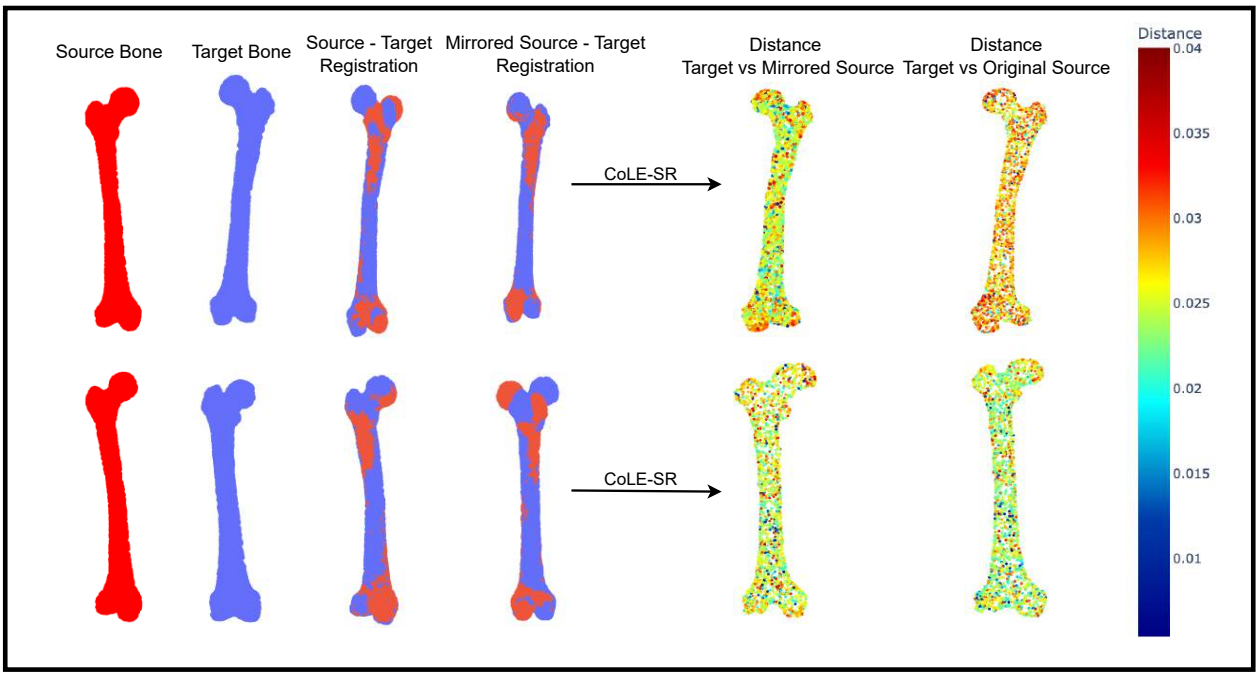


Figure 4. **Qualitative visualization of bone side estimation on human Femur samples.** From left to right: source bone, target bone, their rigid alignment, mirrored-source alignment, and the distance maps comparing the target with the mirrored and original sources. Cooler colors denote lower geometric error, while warmer colors indicate mismatches. From the heat maps, the first row corresponds to a source-target pair from opposite sides (mirroring produces lower error), whereas the second row shows a same-side pair (original source yields lower error). These visual cues demonstrate how mirroring helps discriminate left-right bone identity.

| **Sampling Strategy** | **Avg. Score** |
|---|---|
| Random | 0.934 |
| Uniform | 0.939 |
| **FPS** | **0.951** |

Table 11. Comparison of subgraph center sampling strategies. FPS achieves the most uniform spatial coverage and consistently outperforms random and uniform grid sampling in terms of average detection accuracy.

| $N_{s,\max}$ | **Avg. Score** |
|---|---|
| 2048 | 0.934 |
| **4096** | **0.951** |
| 8192 | 0.953 |

Table 12. Impact of the subgraph budget $N_{s,\max}$ on average performance. A limit of 4096 achieves the best trade-off between coverage and computational cost.

local neighborhoods, but beyond a certain point the redundancy between adjacent subgraphs grows, leading to higher memory and computation costs without meaningful performance gains. As summarized in Table 12, the best trade-off is achieved at $N_{s,\max} = 4096$, which provides sufficient spatial density while keeping the overall model lightweight and efficient.

### 3.2.5. Extended Loss Component Study

To better understand the role of each objective, we evaluate the model after removing individual loss components. The geometric consistency loss $\mathcal{L}_{\text{geo}}$ encourages local neighborhoods to preserve relative distances and angles across refinement loops; removing it weakens the structural constraints and leads to less stable subgraph embeddings. The alignment loss $\mathcal{L}_{\text{align}}$ promotes global correspondence between normal and test embeddings, and its removal reduces the model's ability to maintain consistent global structure. The regularization term $\mathcal{L}_{\text{reg}}$ prevents degenerate embeddings and suppresses overly smooth or collapsed representations, which becomes important in later refinement loops. As shown in Table 13, each component contributes to overall accuracy, and the full combination yields the best performance, indicating that all three objectives provide complementary geometric guidance during refinement.

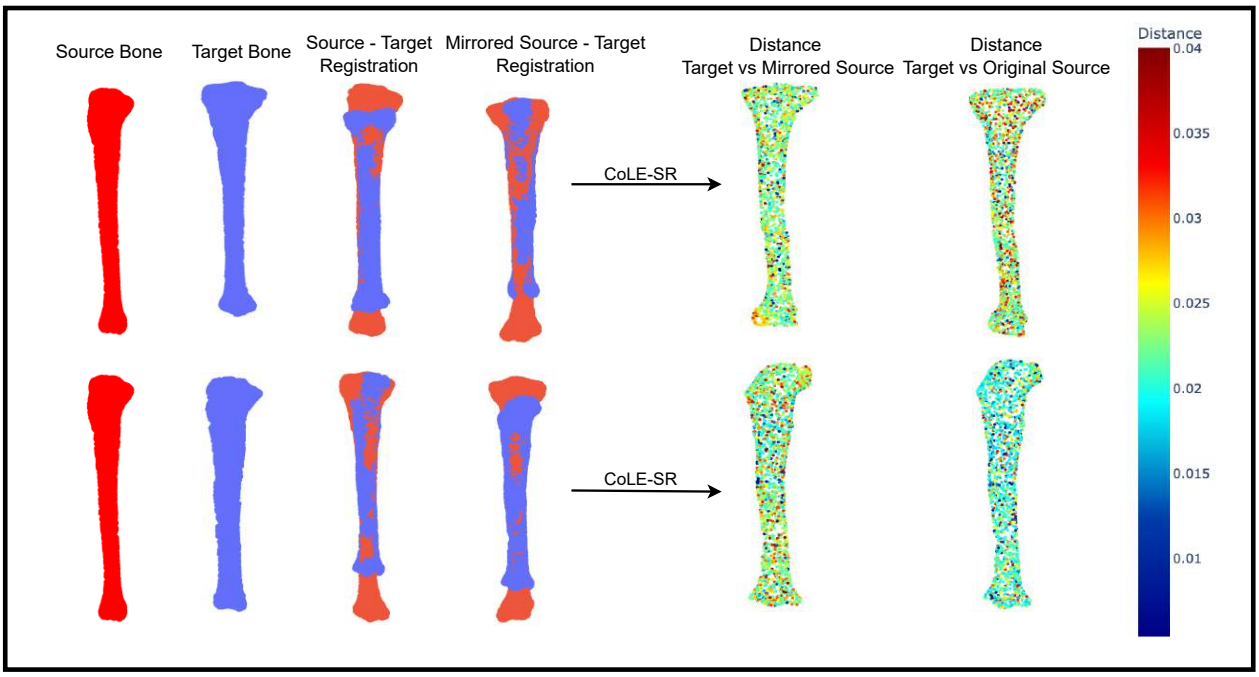


Figure 5. **Qualitative visualization of bone side estimation on human Tibia samples.** From left to right: source bone, target bone, their rigid alignment, mirrored-source alignment, and the distance maps comparing the target with the mirrored and original sources. Cooler colors denote lower geometric error, while warmer colors indicate mismatches. As seen in the heat maps, the first row corresponds to a source-target pair from opposite sides (mirroring produces lower error), whereas the second row shows a same-side pair where the original source yields lower error. These visual cues further demonstrate the effectiveness of mirroring in distinguishing left-right bone identity.

| $\mathcal{L}_{\text{geo}}$ | $\mathcal{L}_{\text{align}}$ | $\mathcal{L}_{\text{reg}}$ | **Avg. Score** |
|---|---|---|---|
| ✓ | ✗ | ✗ | 0.901 |
| ✗ | ✓ | ✗ | 0.925 |
| ✗ | ✗ | ✓ | 0.933 |
| ✓ | ✓ | ✗ | 0.945 |
| ✓ | ✗ | ✓ | 0.939 |
| ✓ | ✓ | ✓ | **0.951** |

Table 13. Ablation on loss components. Pairwise combinations highlight the complementary role of geometric consistency and alignment, with the full model providing the highest accuracy.

#### 3.2.6. Extended Ablation on Geometric Cues

To complement the two-class analysis in the main paper, Table 14 reports the full class-wise impact of positional and angular geometric cues across all ten MVTec 3D-AD categories. As in the main ablation, CoLE-LSA (pos) applies only the positional consistency loss, CoLE-LSA (angle) relies solely on angular supervision, and CoLE-LSA (pos + angle) integrates both cues within each refinement loop. Consistent with the trends observed in the main paper, the combined formulation achieves the best performance across all categories, confirming that positional and angular constraints provide complementary geometric information critical for stable multi-view correspondence refinement.

### 3.3. Qualitative Results

To complement the quantitative results presented in the main paper, we provide qualitative visualizations for the Femur and Tibia classes, two of the four bone categories in the BSE benchmark(Figures. 4 and 5). For each source-target pair, we first apply a rigid alignment to bring the two bone surfaces into a shared coordinate frame. We then mirror the source bone about the medial-lateral axis, derived from the second principal component, to obtain a synthetic contralateral candidate that reflects the expected left-right symmetry of human long bones. Each figure displays the original source and target surfaces, followed by the registration results for the unmirrored and mirrored sources, and the corresponding point-wise distance maps relative to the target. These qualitative views illustrate how geometric agreement shifts depending on whether the bones originate from the same or opposite sides, which forms the basis of our left-right discrimination strategy.

| Method | Bagel | Cable | Carrot | Cookie | Dowel | Foam | Peach | Potato | Rope | Tire |
|---|---|---|---|---|---|---|---|---|---|---|
| CoLE-LSA (pos) | 0.975 | 0.809 | 0.945 | 0.942 | 0.934 | 0.902 | 0.972 | 0.965 | 0.912 | 0.906 |
| CoLE-LSA (angle) | 0.948 | 0.794 | 0.921 | 0.961 | 0.911 | 0.893 | 0.961 | 0.931 | 0.891 | 0.932 |
| **CoLE-LSA (pos + angle)** | **0.988** | **0.861** | **0.978** | **0.991** | **0.956** | **0.914** | **0.982** | **0.978** | **0.923** | **0.956** |

Table 14. Ablation of positional and angular geometric cues. Integrating both cues (pos + angle) yields the best class-wise performance across all MVTec 3D-AD categories.

**Femur.** Figure 4 shows two representative Femur examples. In the first row, the source and target bones originate from opposite sides, which is evident from the distance maps: the mirrored source yields markedly lower geometric discrepancy (cooler colors), indicating that the reflected configuration better matches the target anatomy. In contrast, the second row corresponds to a same-side pair, where the original (unmirrored) source aligns more naturally with the target, and mirroring introduces additional inconsistencies (warmer colors). These observations align with the underlying principle of our approach, wherein global shape similarity under mirrored versus original configurations reveals side identity. The embeddings produced by CoLE-SR capture these structural cues reliably, supporting the strong quantitative performance reported in the main paper.

**Tibia.** A similar trend is observed for the Tibia examples in Fig. 5. As in the Femur case, opposite-side Tibia pairs show substantially reduced error after mirroring, especially around the distal and proximal ends where asymmetry is most pronounced. Same-side pairs, on the other hand, achieve lower residuals with the original configuration. These consistent patterns across bone classes demonstrate that CoLE-SR effectively models bilateral structure and exploits mirrored correspondences for robust bone side estimation.