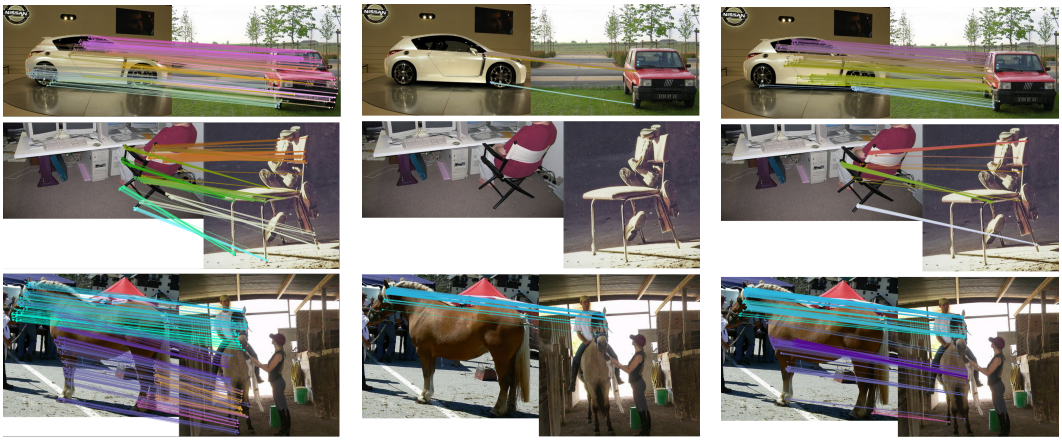
GEOMETRY MATTERS: 3D FOUNDATION PRIORS FOR LEARNING SEMANTIC CORRESPONDENCE

Anonymous authors
Paper under double-blind review



(a) **SD+DINO.** (b) **SD+DINO + Geodesic Filtering.** (c) **SD+DINO+PartField + Geodesic Filtering.**

Figure 1: **3D priors improve both correspondence proposals and verification.** Each row shows the same image pair across three configurations. (a) SD+DINO produces left–right and repeated-part confusions. (b) Geodesic filtering correctly rejects inconsistent proposals. (c) Adding rendered PartField features before filtering retains more matches. These are training pseudo-labels.

ABSTRACT

Foundation features from self-supervised vision models and text-to-image diffusion models transfer well to semantic correspondence, but often confuse symmetric sides and repeated object parts. We introduce 3D-SC, a post-training framework that uses instance-specific 3D geometry to improve both correspondence proposals and their verification. Starting from an image and its category label, we reconstruct an object mesh with SAM3D, refine its image alignment, and canonicalize its orientation. We then render PartField descriptors into the image plane to complement 2D features when proposing matches. A bidirectional geodesic consistency check on the reconstructed meshes filters these proposals into pseudo-labels for training lightweight adapters. The resulting image-based matcher requires no 3D reconstruction at inference and no manual pose or keypoint labels for training. On SPair-71k and its Geo-Aware subset, 3D-SC improves PCK@0.1 by 2.8 and 7.3 points over the strongest weakly supervised method, which additionally relies on manual pose labels. On SPair-U, which evaluates generalization to unseen keypoints, our method outperforms keypoint-supervised methods. These results support using 3D foundation models as geometric teachers for semantic correspondence, with the largest gains on geometrically ambiguous matches.

1 INTRODUCTION

Semantic correspondence establishes matches between semantically equivalent object parts across images, supporting applications in 3D understanding (Wang et al., 2024) and robotics (Zhu et al., 2025). Unlike low-level image matching, semantic correspondence requires robustness to changes in appearance, viewpoint, articulation, intra-class shape variation, and background clutter. As a result, it

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

remains challenging to match object parts that are visually different but semantically equivalent, or visually similar but semantically distinct.

Recent progress has been driven by foundation features, with self-supervised vision transformers (DINOv2) and text-to-image diffusion models (Stable Diffusion) producing representations that transfer surprisingly well to dense semantic matching (Caron et al., 2021; Amir et al., 2022; Oquab et al., 2023; Rombach et al., 2022; Tumanyan et al., 2023). Their fusion has become a strong zero-shot baseline on benchmarks such as SPair-71k, PF-PASCAL, and TSS (Min et al., 2019; Ham et al., 2017; Taniai et al., 2016; Zhang et al., 2023), combining DINOv2's semantic discrimination with the smoother spatial cues of diffusion models. However, their image-based training objectives do not explicitly enforce 3D consistency, leaving systematic failure modes (Mariotti et al., 2024; Dünkel et al., 2025). For symmetric objects, such as cars, buses, and animals, 2D features may confuse left and right object sides (Zhang et al., 2024). For objects with repeated parts, such as wheels, legs, windows, or chair legs, visually similar regions may collapse to nearly identical feature representations despite corresponding to different object parts (see Fig. 1a). These ambiguities motivate using object geometry to distinguish visually similar structures. Some recent methods address these ambiguities by injecting a weak 3D prior to guide feature learning and correspondence filtering (Mariotti et al., 2024; Dünkel et al., 2025). While effective, both approaches *require human pose annotations and approximate object geometry with a coarse spherical proxy*, which cannot represent the geometric structure of an actual instance. The reliance on manual pose annotations also limits scalability, as extending to new object categories requires additional labeling effort.

We propose 3D-SC, a post-training framework that leverages 3D foundation models as geometric teachers without pose annotations or keypoint training labels. Given an image, SAM3D estimates object geometry and pose (SAM 3D Team, 2025). We refine the mesh's image alignment, then render PartField descriptors (Liu et al., 2025) back into the image plane. The reconstructed geometry serves two complementary roles. First, PartField complements DINO and SD when proposing matches, helping separate repeated parts such as front and rear wheels. Second, surface geodesics measure agreement between each image-space proposal and a 3D-grounded match. Filtering by this agreement yields pseudo-labels for training lightweight adapters, whose features are used at inference.

This distinction between *proposing better matches* and *verifying them on instance geometry* is central to our approach. Filtering SD+DINO proposals alone can discard most matches; adding PartField before filtering improves the retained coverage at a similar error rate (Fig. 1 and Tab. 2). The resulting model reaches 74.4 PCK@0.1 on SPair-71k and 74.8 on its Geo-Aware subset, exceeding the pose-supervised DIY-SC by 2.8 and 7.3 points. Our contributions are:

- *(i)* a pipeline for reconstructing, aligning, and canonicalizing instance meshes to render 3D descriptors as complementary correspondence features;
- *(ii)* bidirectional geodesic verification that relies on geometry to select pseudo-labels for an image-based matcher; and
- *(iii)* geometry-aware refined features that outperform prior weakly supervised semantic correspondence methods, including those trained with manual pose labels.

2 Related Work

Semantic correspondence with foundation features. Semantic correspondence aims to match semantically equivalent parts across object instances, which is substantially harder than low-level image matching because appearance, shape, pose, articulation, and visibility all vary. Early approaches relied on hand-crafted descriptors and learned matching networks (Lowe, 2004; Liu et al., 2011; Ham et al., 2017; Yi et al., 2016), and because dense annotations are costly, later work explored weak supervision, cycle-consistency losses, and pseudo-label expansion from sparse labels (Zhou et al., 2016; Kim et al., 2022; Li et al., 2021; Huang et al., 2023). Recent progress has shifted to foundation features: self-supervised vision transformers such as DINO and DINOv2 encode transferable semantic concepts (Caron et al., 2021; Amir et al., 2022; Oquab et al., 2023), while text-to-image diffusion features provide complementary spatial and semantic cues (Rombach et al., 2022; Hedlin et al., 2023; Tang et al., 2023; Luo et al., 2023; Li et al., 2024). Their fusion has become a strong zero-shot baseline (Zhang et al., 2023), and distillation or adapter-based refinement further improves them when supervision is available (Zhang et al., 2024; Fundel et al., 2025; Xue et al., 2025). Such

supervised models, however, generalize poorly to keypoints outside the annotated set (Mariotti et al., 2025). MARCO (Cuttano et al., 2026) narrows this gap with a DINOv2-only model trained with a coarse-to-fine objective and self-distillation, which expands sparse keypoint annotations into dense correspondences. Like MARCO, we seek dense supervision beyond the annotated keypoints and adopt its AdaptFormer-based backbone adaptation (Chen et al., 2022); unlike MARCO, we derive this supervision from reconstructed 3D geometry rather than from manual keypoints. Features learned from images alone, however, remain prone to geometry-sensitive failures such as left-right confusion, front-back ambiguity, and repeated parts (Zhang et al., 2024; Mariotti et al., 2024; Dünkel et al., 2025; Mariotti et al., 2025). Recent benchmarks (Sommer et al., 2026; Zhu et al., 2025) also address these ambiguities directly in the 3D domain.

Geometric priors and 3D-aware features. A complementary line of work introduces geometric structure to disambiguate the failures of purely image-based correspondence. CAD-based cycle consistency and canonical surface mappings link image pixels to a shared object surface (Zhou et al., 2016; Kulkarni et al., 2019; Neverova et al., 2020), while category-level templates, atlases, and learned 3D representations capture correspondences via a shared geometric frame (Jesslen et al., 2024; Shtedritski et al., 2024; Sommer et al., 2025; Wandel & Wang, 2025; Chi et al., 2026). These methods show the value of 3D structure but typically require mesh templates, precise pose, or category-level reconstruction pipelines. Closer to our setting, Spherical Maps inject a weak 3D prior by mapping image features to a category-conditioned sphere with viewpoint supervision (Mariotti et al., 2024), and DIY-SC produces pseudo-labels from DINOv2 and Stable Diffusion features, then filters them against a spherical 3D prototype before training a lightweight adapter (Dünkel et al., 2025). In parallel, 3D foundation models make instance-level geometry practical from a single image: SAM3D reconstructs object-centric 3D shape (SAM 3D Team, 2025), orientation models help resolve canonical-frame ambiguities (Wang et al., 2025), and 3D feature fields or functional-map methods provide geometry-aware descriptors on surfaces (Liu et al., 2025; Ovsjanikov et al., 2012; Donati et al., 2020; Dutt et al., 2024; Zhu et al., 2025; Wang et al., 2024). In contrast to spherical-prior approaches, we combine instance-specific meshes with PartField descriptors to both *generate* and *filter* pseudo-labels using estimated instance surfaces in place of a spherical proxy.

3 METHOD

We estimate semantic correspondences by combining 2D foundation features with 3D geometric priors obtained from reconstructed meshes. Our pipeline has three stages: (i) we first reconstruct and canonicalize 3D meshes for each instance; (ii) we then project 3D-aware features into the image plane and use them together with DINOv2 and Stable Diffusion features to propose semantic correspondences; (iii) finally, we reject geometrically inconsistent matches using geodesic consistency on the reconstructed meshes and train lightweight adapters on the retained pseudo-labels. The 3D models are used to construct training supervision; inference uses only the adapted DINOv2 features.

3.1 CANONICALIZED 3D OBJECT RECONSTRUCTION

Our correspondence pipeline relies on a 3D mesh for each object instance, expressed in a canonical frame that is consistent across instances of the same category, since symmetric parts such as left and right legs can only be told apart relative to the object's orientation. We obtain such meshes from a single image without manual pose annotation by combining recent foundation models for segmentation and single-image 3D reconstruction with two refinement stages. While these foundation models provide a strong geometric prior, their outputs exhibit two systematic issues: the predicted scale and translation can be inaccurate, causing the rendered mesh to misalign with the image, and the canonical orientation is ambiguous up to discrete yaw rotations across instances. We address the first issue with a render-and-compare optimization that aligns the rendered silhouette to the observed mask, and the second with a yaw canonicalization step based on multi-view orientation estimation. The full process is illustrated in Fig. 2.

2D Mask and 3D Mesh Initialization. We extract a 2D instance mask $\mathbf{M} \in \{0,1\}^{H\times W}$ with SAM3 (Carion et al., 2025), using the image together with the dataset-provided category label. Apart from the category-level grouping of image pairs (Sec. 3.2), this label is the only dataset annotation used to generate pseudo-labels; no keypoint or pose annotations are used. Given this mask, SAM3D (SAM 3D Team, 2025) reconstructs an object-centric mesh from the masked image in

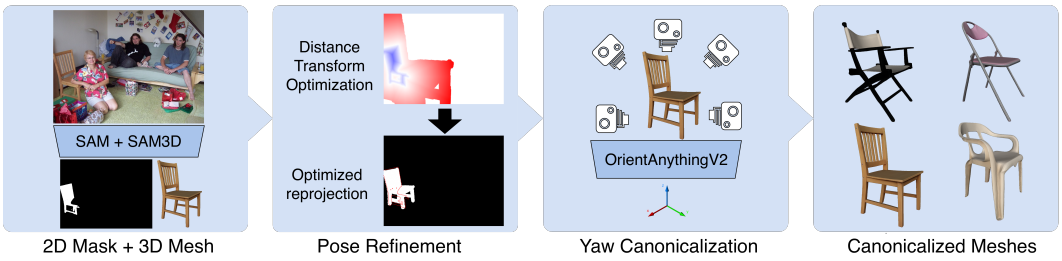


Figure 2: **Canonicalized 3D object reconstruction pipeline.** Given an image, we obtain an instance mask and a mesh from foundation models. We then refine the mesh scale and translation via render-and-compare optimization of a distance-transform objective. Finally, we resolve the residual four-fold yaw ambiguity by rendering the mesh at eight known orientations and applying OrientAnything V2 with majority voting to select the canonical yaw correction $\Delta\psi^*$.

a feed-forward manner and additionally predicts the camera parameters used for rendering. In the following, we show how we refine and canonicalize this initial reconstruction.

Render-and-Compare Pose Refinement. To correct the residual scale and translation error in the SAM3D reconstruction, we apply a render-and-compare optimization on top of the predicted camera. Concretely, we optimize a scale factor $s = e^{\ell} \in \mathbb{R}_{>0}$ (parameterized in log-space to remain strictly positive) and a translation $\mathbf{t} \in \mathbb{R}^3$ applied to the mesh, by minimizing the discrepancy between the rendered silhouette $\hat{\mathbf{M}}(s, \mathbf{t})$ and the observed mask $\mathbf{M}$. We use a distance-transform (DT) objective, which provides distance-dependent attraction within the rendered image.

Distance-transform attraction. From the observed mask $\mathbf{M}$, we compute two squared distance fields normalized by d^2, where d is the image diagonal:

$$\mathcal{D}_{\text{out}}(p) = \tfrac{1}{d^2} \min_{p' :\, \mathbf{M}(p')=1} \|p - p'\|_2^2, \qquad \mathcal{D}_{\text{in}}(p) = \tfrac{1}{d^2} \min_{p' :\, \mathbf{M}(p')=0} \|p - p'\|_2^2. \tag{1}$$

$\mathcal{D}_{\text{out}}$ is zero inside $\mathbf{M}$ and grows with distance to the mask; $\mathcal{D}_{\text{in}}$ is zero outside $\mathbf{M}$ and grows with depth into its interior. The DT loss combines these into a mask-alignment objective:

$$\mathcal{L}_{\text{DT}} = \frac{1}{HW} \sum\nolimits_p \Big[\hat{\mathbf{M}}_p \, \mathcal{D}_{\text{out}}(p) + \mathcal{D}_{\text{in}}(p) \big(1 - \lambda \hat{\mathbf{M}}_p\big) \Big]. \tag{2}$$

The first term penalizes rendered mass outside the observed mask. The second rewards coverage of its interior, with $\lambda > 1$ increasing this incentive relative to the exterior penalty. It discourages solutions that cover only a small visible region. Scale and depth are not separately identifiable from the silhouette, but nothing downstream depends on them: PartField rasterization uses only the image projection, and geodesic errors are normalized by the mesh diagonal.

Yaw Canonicalization. SAM3D meshes of the same category do not always share a canonical orientation: our procedure applies a multiple-of-90° yaw correction to about 6% of meshes (Sec. A). Manual visual inspection confirms the voted correction for 99.0% of meshes with an unambiguous orientation. We estimate this correction without manual annotation using OrientAnything V2 (Wang et al., 2025). For each mesh, we render eight views at known yaw angles $\psi_{\text{known}} \in \{0°, 45°, \ldots, 315°\}$ and we estimate the apparent orientation ψ_{est} of each rendering. If the mesh is correctly canonicalized, ψ_{est} should match ψ_{known} up to estimator noise; otherwise, the two differ by a multiple of 90°. For each rendered view we therefore pick the discrete correction that best closes this gap,

$$\Delta\psi^* = \operatorname*{arg\,min}_{\Delta\psi \in \{0°, 90°, 180°, 270°\}} \operatorname{dist}_{\mathbb{S}^1}\big(\psi_{\text{est}} + \Delta\psi,\ \psi_{\text{known}}\big), \tag{3}$$

where $\operatorname{dist}_{\mathbb{S}^1}(a, b) = \min_{k \in \mathbb{Z}} |a - b + 360°k|$ is circular angular distance. We aggregate the eight view-wise corrections by majority vote and rotate the mesh by the selected correction. Voting reduces sensitivity to individual orientation estimates.

3.2 Pseudo-Label Semantic Correspondences

Given two images of the same category, we propose candidate matches from fused DINO, SD, and PartField features and discard obvious mismatches with relaxed cyclic consistency, a cheap image-space filter. We then verify each candidate on the reconstructed meshes by comparing it with a

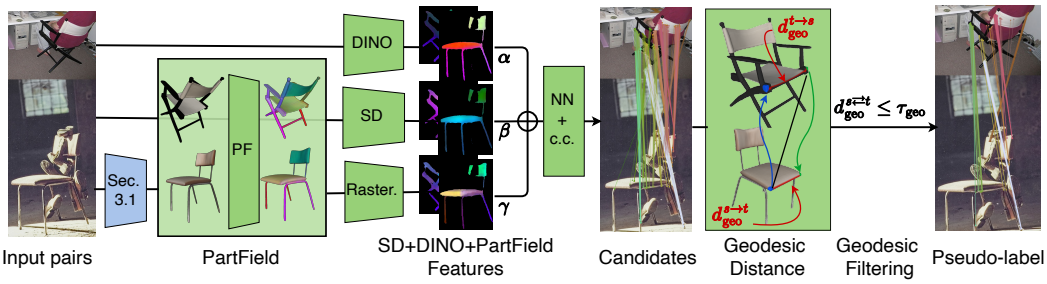


Figure 3: **Pseudo-label correspondence pipeline.** Given two images, we fuse DINO, SD, and PartField features (rasterized from the meshes of Sec. 3.1) and propose candidate matches via nearest-neighbor (NN) search with relaxed cyclic consistency (c.c.). Each candidate is then geometrically verified by lifting the matched pixels onto the reconstructed meshes and computing the geodesic error $d_{\text{geo}}^{s \rightleftarrows t}$; candidates exceeding threshold τ_{geo} are rejected. The retained pseudo-labels $\mathcal{P}$ are used to train lightweight adapters on DINO features only (Sec. 3.2).

PartField cross-mesh match through within-mesh geodesic distances (Fig. 3). Superscripts $\square^s$ and $\square^t$ denote source and target; p is an image point and $\mathbf{v}$ a 3D point.

PartField Features. PartField (Liu et al., 2025) (PF) predicts a continuous per-vertex feature field encoding geometric and part-level structure directly from the 3D shape S. These descriptors can distinguish parts that are visually similar but geometrically distinct (*e.g.*, front vs. rear wheels, left vs. right legs), using the surrounding shape as context. They provide a complementary cue rather than a guaranteed semantic correspondence. To use PartField in image space, we render the reconstructed mesh with the SAM3D camera and the refined pose from Sec. 3.1, map each pixel to its surface point, and assign it the PartField descriptor of that point. Pixels outside the SAM3 foreground mask are discarded, and mask pixels not covered by the rendered mesh are filled by nearest-neighbor propagation. The result is a 2D PartField map aligned with the RGB image, which can be combined with 2D features for semantic correspondence estimation. More details about visualizations and rasterization in Sec. B.1.

Candidate Generation. Given the fused image-space DINO+SD+PF features, we propose candidate matches via nearest-neighbor search, retaining only those that pass a cyclic consistency check.

Feature fusion. Following Zhang et al. (2023), we fuse our three feature sources by independently L2-normalizing each and concatenating them with category-agnostic weights. We denote the normalized feature vectors as $\widehat{\mathcal{F}}_m = \frac{\mathcal{F}_m}{\|\mathcal{F}_m\|_2}$ for each feature source $m \in \{\text{SD}, \text{DINO}, \text{PF}\}$. The fused representation is then defined as

$$\mathcal{F}_{\text{fused}} = \left(\sqrt{\alpha}\,\widehat{\mathcal{F}}_{\text{SD}},\ \sqrt{\beta}\,\widehat{\mathcal{F}}_{\text{DINO}},\ \sqrt{\gamma}\,\widehat{\mathcal{F}}_{\text{PF}}\right), \quad \text{with } \gamma = 1 - \alpha - \beta. \tag{4}$$

We use the weights $\alpha = 1/2$, $\beta = 1/3$, and $\gamma = 1/6$, selected using the SPair-71k validation set; a weight sweep is provided in Sec. B.2. Candidate matches are then proposed by nearest-neighbor search in the fused space.

Relaxed cyclic consistency. While the 3D-aware PartField features improve the matching quality, some candidates remain wrongly matched. To filter these mismatches, we apply a relaxed cyclic consistency check inspired by Aberman et al. (2018). As observed in Dünkel et al. (2025), strict cyclic consistency rejects a large fraction of correct matches due to sub-pixel noise; we therefore relax the criterion to require only that the backward match lies within a small spatial tolerance of the source. A candidate $(p^s, \hat{p}^t)$ with $\hat{p}^t = \text{NN}_{\text{fused}}^{s \to t}(p^s)$ is retained if

$$\left\| \text{NN}_{\text{fused}}^{t \to s}(\hat{p}^t) - p^s \right\|_2 < \max(\tau_{cc} \cdot \max(h_s, w_s),\ \delta_{\text{patch}}), \tag{5}$$

where NN_{fused} denotes nearest-neighbor search in the fused feature space, h_s, w_s are the dimensions of the bounding box of the source SAM3 mask, and δ_{patch} is one feature-map patch, expressed in the same image coordinates. The dimensionless τ_{cc} sets the relative tolerance.

Candidate Verification via Geodesic Filtering. Cyclic consistency operates purely in feature space and keeps mutually consistent but geometrically wrong matches, such as left–right swaps. We therefore add a geodesic consistency stage: matched locations lifted onto canonically posed meshes must land in nearby surface regions.

Lifting matches to 3D. Given a candidate match (p^s, p^t), we cast a ray from each camera through the corresponding pixel and intersect it with the respective mesh, obtaining the unprojected points $\mathbf{v}^s$ and $\mathbf{v}^t$ together with their containing triangles and barycentric coordinates. Because geodesic distances are computed between mesh vertices, we snap each unprojected point to the dominant vertex of its triangle (the vertex with the largest barycentric weight), giving $\bar{\mathbf{v}}^s$ and $\bar{\mathbf{v}}^t$.

Cross-mesh correspondence via PartField. The previous step places each candidate match onto the source and target meshes individually. The meshes share a canonical orientation but no vertex correspondence. To compare the lifted source and target points, we therefore estimate a 3D correspondence between the meshes themselves. Hence, we use PartField nearest-neighbor as the cross-mesh correspondence: we interpolate the PartField descriptor at $\mathbf{v}^s$ on the source mesh and search for its nearest neighbor among the PartField descriptors on the target mesh, $\hat{\mathbf{v}}^t = \mathrm{NN}_{\mathrm{PF}}^{s\to t}(\mathbf{v}^s)$, yielding a target vertex $\hat{\mathbf{v}}^t$ that represents the cross-mesh counterpart of $\mathbf{v}^s$. A candidate is then geometrically consistent if this PartField-predicted target $\hat{\mathbf{v}}^t$ is geodesically close to the target obtained from the image-space match, $\bar{\mathbf{v}}^t$.

Bidirectional geodesic error. We measure the disagreement between the two target predictions as a *bidirectional* geodesic distance, combining a forward and a backward geodesic error on the source and target meshes. The forward error measures, on the target mesh, the geodesic distance between the cross-mesh prediction $\hat{\mathbf{v}}^t$ and the target $\bar{\mathbf{v}}^t$ obtained from the image-space match:

$$d_{\mathrm{geo}}^{s\to t} = d_{\mathcal{M}_t}\big(\hat{\mathbf{v}}^t,\, \bar{\mathbf{v}}^t\big). \tag{6}$$

A symmetric computation in the reverse direction yields a backward error $d_{\mathrm{geo}}^{t\to s} = d_{\mathcal{M}_s}(\hat{\mathbf{v}}^s,\, \bar{\mathbf{v}}^s)$, where $\hat{\mathbf{v}}^s = \mathrm{NN}_{\mathrm{PF}}^{t\to s}(\mathbf{v}^t)$. We average the two and normalize by the mesh bounding-box diagonals so that the score is comparable across instances and categories of varying scale:

$$d_{\mathrm{geo}}^{s\rightleftarrows t} = \frac{1}{2}\left(\frac{d_{\mathrm{geo}}^{s\to t}}{\mathrm{diag}(\mathcal{M}_t)} + \frac{d_{\mathrm{geo}}^{t\to s}}{\mathrm{diag}(\mathcal{M}_s)}\right). \tag{7}$$

Intuitively, $d_{\mathrm{geo}}^{s\rightleftarrows t}$ is small when the image-space candidate and the PartField cross-mesh correspondence agree on the same surface location, and large when they disagree.

Rejection of wrong pseudo-labels. We use the bidirectional geodesic error as a per-candidate quality score and threshold it to reject inconsistent pseudo-labels. A candidate (p^s, p^t) is retained if and only if its error falls below a threshold τ_{geo}:

$$\mathcal{P} = \left\{(p^s, p^t) \;\middle|\; d_{\mathrm{geo}}^{s\rightleftarrows t} \le \tau_{\mathrm{geo}}\right\}. \tag{8}$$

Because $d_{\mathrm{geo}}^{s\rightleftarrows t}$ is normalized by the mesh bounding-box diagonals, a single value of τ_{geo} applies across object instances and categories of varying scale. This criterion trades coverage for agreement with the 3D teacher. Since PartField contributes to both proposal generation and verification, errors shared by the two stages can survive; geodesic consistency is not independent ground truth. The same criterion guards against defective reconstructions. Distorted shapes, missing parts, or merged objects make the 3D cross-mesh match disagree with the 2D match, so the affected candidates are rejected rather than kept. Section 4.3 quantifies the effect of reconstruction failures.

Supervised Training with Pseudo-Labels.

We use $\mathcal{P}$ to train a correspondence adapter $f_p(\cdot)$ over DINOv2 features only, following prior correspondence adaptation approaches (Zhang et al., 2024; Dünkel et al., 2025); unlike DIY-SC, which adapts fused SD and DINOv2 features, we use SD and PartField only to generate pseudo-labels. Our final model also trains AdaptFormer modules within DINOv2's last four blocks (Sec. 4.1); the original backbone parameters remain frozen, but the DINOv2 representations can change through these added modules. We denote the resulting source and target features by $\mathcal{F}^s$ and $\mathcal{F}^t$. We supervise $f_p(\cdot)$ with two complementary losses. A sparse contrastive loss (Luo et al., 2023) acts on the labeled pseudo-correspondences, maximizing similarity between matched points and minimizing it against non-matching points:

$$\mathcal{L}_{\mathrm{sparse}} = \mathrm{CL}\big(\mathcal{F}^s(\mathcal{P}^s),\, \mathcal{F}^t(\mathcal{P}^t)\big). \tag{9}$$

A localization regression loss (Zhang et al., 2024) predicts the target location with a window soft-argmax over the feature similarity map and penalizes its deviation from the pseudo-label. Although

Table 1: **Benchmark evaluation.** Per-image PCK (%, ↑) at multiple thresholds on SPair-71k (test and Geo-Aware subset), AP-10K, SPair-U, and SOCO. '–' indicates unreported results. †Results obtained from the official checkpoint. ‡ Trained with manual pose annotations. § Extracts SD features via a method trained on semantic segmentation annotations. Best per group is shown in **bold**. I.S./C.S./C.F.: intra-species/cross-species/cross-family. SOC/X-SOC: semantic object correspondence within/across categories. The full table, including additional *Supervised* methods, is provided in Tab. C1.

		SPair-71k			SPair-Geo-Aware			AP-10K (0.10)			SPair-U			SOCO (0.10)	
Method		0.01	0.05	0.10	0.01	0.05	0.10	I.S.	C.S.	C.F.	0.01	0.05	0.10	SOC	X-SOC
Supervised															
MARCO	Cuttano et al. (2026)	**27.0**	**77.6**	**87.2**	**22.8**†	**76.8**†	**87.5**†	**89.1**	**88.3**	**83.4**	**5.0**†	**42.7**†	**67.5**	**79.3**†	**76.1**†
Unsupervised															
DINOv2+NN	Zhang et al. (2023)	6.3	38.4	**53.9**	**3.4**	**28.2**	42.0	**60.9**	**57.3**	**47.4**	–	–	**54.9**	**62.1**†	**56.9**†
DIFT	Tang et al. (2023)	**7.2**	**39.7**	52.9	**3.4**	**28.2**	**42.5**	50.3	46.0	35.0	–	–	47.4	57.4†	49.5†
Weakly supervised															
Spherical Map.‡	Mariotti et al. (2024)	8.4	48.2	64.4	–	–	–	65.4	63.1	51.0	–	–	61.0	67.0†	62.5†
SD+DINOv2§	Zhang et al. (2023)	7.9	44.7	59.9	5.3	34.5	49.3	62.9	59.3	48.3	–	–	59.4	65.1†	60.2†
DIY-SC‡	Dünkel et al. (2025)	10.1	53.8	71.6	7.7	47.7	67.5	70.6	**69.8**	57.8	5.4	44.0	67.9	70.8†	66.5†
DIY-SC+OriAny	Dünkel et al. (2025)	9.5	51.2	69.6	6.9	45.7	65.8	69.3	66.8	54.0	5.2	43.1	66.3	69.6†	65.3†
Shape-of-You	Im et al. (2026)	8.3	45.1	63.1	5.8	40.0	60.0	68.1	65.5	54.8	5.0	39.8	61.8	66.2†	61.6†
3D-SC (Ours)		**10.8**	**54.1**	**74.4**	**10.0**	**53.0**	**74.8**	**71.9**	68.7	**59.4**	**5.5**	**45.5**	**70.1**	**75.7**	**70.2**

evaluated at pseudo-labeled source points, it provides gradients through the target similarity window:

$$\mathcal{L}_{\text{dense}} = \sum_{(p^s,p^t)\in\mathcal{P}} \left\|\hat{p}^t - (p^t+\epsilon)\right\|_2, \qquad \hat{p}^t = \textsc{WindowSoftArgmax}\big(\mathcal{F}^s(p^s)^\top \mathcal{F}^t\big), \tag{10}$$

where ϵ is small Gaussian noise that regularizes the predicted location at sub-pixel scale. The trainable modules are optimized with the sum $\mathcal{L} = \mathcal{L}_{\text{sparse}} + \mathcal{L}_{\text{dense}}$.

4 Experiments

We evaluate correspondence accuracy on SPair-71k, its Geo-Aware subset, SPair-U, AP-10K, and SOCO, and separate the effects of pseudo-label construction and feature adaptation.

4.1 Implementation Details

Following Zhang et al. (2023), we resize images to 960^2 for SD and 840^2 for DINOv2. PartField descriptors are rasterized at the shared 60^2 correspondence-map resolution. We set $\tau_{cc} = 0.05$ for the cyclic consistency, with a minimum tolerance of one feature-map patch, and use $\tau_{\text{geo}} = 0.05$ for geometric verification. We use a four-layer correspondence adapter (Zhang et al., 2024), and following Cuttano et al. (2026), we insert AdaptFormer (Chen et al., 2022) modules into the *last four* transformer blocks of DINOv2 ViT-B/14, adding up to approximately 3.6M parameters. We train the model with AdamW, a learning rate of $5{\cdot}10^{-3}$, weight decay of 10^{-3}, and a one-cycle schedule for 200k iterations. We sample 50 pseudo-correspondences per image pair per iteration to limit the influence of pairs with many labels. Additional details are in Sec. C.1.

Benchmarks and Metrics. SPair-71k (Min et al., 2019) contains approximately 71k image pairs across 18 categories. Its *Geo-Aware* subset (Zhang et al., 2024) emphasizes symmetric or repeated parts. *SPair-U* (Mariotti et al., 2025) adds approximately four unseen keypoint definitions per category, testing transfer beyond the landmarks annotated in SPair-71k rather than transfer to new object categories. *AP-10K* (Yu et al., 2021) contains 17 shared keypoints across 54 animal species; we report intra-species, cross-species, and cross-family matching. The AP-10K results measure how well the learned 2D features transfer to articulated animals, most of them from species absent from SPair-71k. *SOCO* (Dünkel et al., 2026) consists of 100 categories with 40 images each. We report two settings: *SOC* matches keypoints within a category, and *X-SOC* matches such keypoints across related categories through shared part concepts (*e.g.*, a wheel on a car and on a bus). A prediction is correct under PCK@α if its distance to the target keypoint is at most $\alpha \max(h, w)$, where h, w are the target object's bounding-box dimensions. The results use per-image aggregated PCK.

Baselines and supervision. We compare foundation-feature baselines, weakly supervised methods, and the supervised MARCO reference (Cuttano et al., 2026). DINOv2 and DIFT (Tang et al.,

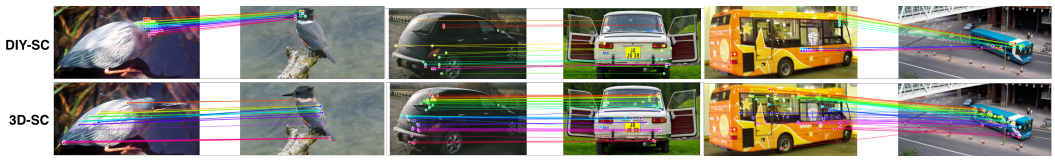


Figure 4: **Training pseudo-labels on identical image pairs.** Top: DIY-SC. Bottom: 3D-SC. Colored lines connect retained matches for bird, car, and bus pairs. The examples illustrate increased coverage, particularly over vehicle structure.

2023) use nearest-neighbor feature matching; SD+DINOv2 (Zhang et al., 2023) combines complementary image features. Spherical Maps and DIY-SC use pose annotations (Mariotti et al., 2024; Dünkel et al., 2025), while DIY-SC+OriAny replaces them with orientations estimated by OrientAnything V2 (Wang et al., 2025). Shape-of-You (Im et al., 2026) matches features with fused Gromov–Wasserstein optimal transport. All methods trained on SPair-71k use its category labels, since training pairs are formed within a category. Our method additionally uses the category label to prompt SAM3 (Sec. 3.1), but uses no manual pose nor keypoint labels. We therefore group all these methods as weakly supervised.

4.2 Experimental Results

SPair-71k and geometric ambiguity. Table 1 shows that 3D-SC reaches 74.4 PCK@0.1 on SPair-71k, improving over DIY-SC+OriAny by 4.8 points and pose-supervised DIY-SC by 2.8 points. On SPair-Geo-Aware, the corresponding gains are larger: 74.8 versus 65.8 and 67.5, respectively. This pattern is consistent with geometry being particularly useful for symmetric and repeated parts, although the full-model comparison also includes AdaptFormer; we separate its contribution below. The per-category results in Tab. C2 show large gains over DIY-SC+OriAny for bus (+13.0), tv/monitor (+20.6), chair (+12.5), and car (+8.3), with mixed gains on animals.

SPair-U: unseen keypoints. 3D-SC reaches 5.5/45.5/70.1 at PCK@0.01/PCK@0.05/PCK@0.1, ahead of DIY-SC (5.4/44.0/67.9) and the keypoint-supervised MARCO (5.0/42.7/67.5) at every threshold. From SPair-71k to SPair-U, weakly supervised methods lose at most 4.3 points at PCK@0.1, whereas keypoint-supervised methods lose 18.4 to 26.3 (Tab. C1): geometry-derived pseudo-labels transfer better to unseen keypoints than keypoint supervision.

Transfer beyond SPair-71k. On AP-10K, 3D-SC obtains 71.9/68.7/59.4 PCK@0.1 (intra-species/cross-species/cross-family), improving over DIY-SC+OriAny by 2.6/1.9/5.4 points and over pose-supervised DIY-SC except on cross-species matching (68.7 versus 69.8). On SOCO, whose 100 categories go far beyond the 18 training categories, 3D-SC reaches 75.7 (SOC) and 70.2 (X-SOC), 4.9 and 3.7 points above DIY-SC. Keypoint-supervised MARCO remains ahead on both benchmarks.

Qualitative pseudo-labels. Figure 4 compares the training supervision generated by 3D-SC and DIY-SC. In these examples, 3D-SC retains more correspondences over the visible object, including repeated structures. Quantitative filtering results follow below; additional examples appear in Fig. C1.

4.3 Ablations

Proposal features and retained coverage. Table 2 evaluates filters on validation keypoints: a retained fused-feature nearest-neighbor prediction is a false discovery if it is incorrect under PCK@0.1. On the same SD+DINO proposals, geodesic verification lowers the FDR to 1.82%, from 10.95% and 13.15% for the spherical-map and triplane filters, at the cost of fewer retained labels (103 versus 331 and 157 per pair). Adding PartField to the proposals more than doubles the retained labels (103 to 234) at a similar FDR (1.78%), while PartField feature similarity as a filter is worse on both axes.

Geodesic filtering and sampling. In Tab. 3, adding geodesic filtering to cyclic-consistent SD+DINO+PartField pseudo-labels improves downstream PCK from 68.8 to 73.1 (+4.3 points), with sampling enabled and AdaptFormer disabled in both rows. Capping the sampled labels improves the otherwise matching configuration from 72.4 to 73.1 (+0.7).

AdaptFormer. Whereas the correspondence adapter only transforms the backbone's output features, AdaptFormer lets the last four DINOv2 blocks adapt their representations to the geometric pseudo-

Table 2: **Filtering on SPair-71k validation.** FDR (%, ↓) is the fraction of retained keypoint predictions that are incorrect under PCK@0.1. Retained = median number of dense pseudo-labels per pair.

Filter	FDR	Retained
Features: *SD+DINO*		
Spherical Map.	10.95	331
Triplane	13.15	157
PF Feature sim.	2.81	6
Geodesic	1.82	103
Features: *SD+DINO+PartField*		
Spherical Map.	10.75	$394_{\uparrow 63}$
Triplane	13.07	$442_{\uparrow 285}$
PF Feature sim.	2.47	$149_{\uparrow 143}$
Geodesic	1.78	$234_{\uparrow 131}$

Table 3: **SPair-71k ablations (PCK@0.1).** PL: train with pseudo-labels; PF: PartField proposals; c.c.: cyclic consistency; filter: geodesic verification; cap: sample at most 50 labels per pair; AF: AdaptFormer (Chen et al., 2022). The first row is zero-shot SD+DINO with window soft-argmax.

PL	PF	c.c.	filter	cap	AF	PCK
				✓		64.9
✓				✓		67.0
✓		✓		✓		67.6
✓		✓	✓	✓		71.6
✓		✓	✓	✓		72.4
✓	✓			✓		66.9
✓	✓	✓		✓		68.8
✓	✓	✓	✓			72.4
✓	✓	✓	✓	✓		73.1
✓	✓	✓	✓	✓	✓	**74.4**

labels. This proves useful: SPair-71k PCK@0.1 improves from 73.1 to 74.4 (+1.3 points). Without AdaptFormer, 3D-SC already exceeds DIY-SC+OriAny (69.6) and pose-supervised DIY-SC (71.6), which separates the benefit of our pseudo-labels from the added adaptation capacity.

Robustness to reconstruction failures. Our pseudo-labels depend on SAM3D reconstructions, which occasionally fail. We visually inspected all 1,319 reconstructed SPair-71k images and marked 23 (1.7%) as failures: the object is unrecognizable, only partially reconstructed, or merged with another object. Most failures are curled-up cats, crawling babies, and long trains. For image pairs involving a failed mesh, our pipeline retains a median of only 2 pseudo-labels per pair, compared with 234 for pairs of correctly reconstructed meshes. Grouping validation pairs by silhouette IoU shows the same pattern beyond outright failures: from the highest to the lowest IoU quartile, the median number of retained labels per pair falls from 409 to 123, while the FDR stays below 3% in every quartile (Tab. A1). **Geodesic verification thus mostly turns reconstruction failures into missing supervision rather than wrong supervision**, keeping the overall FDR low (1.78%, Tab. 2).

5 Limitations and Future Work

The quality of our pseudo-labels is bounded by the 3D foundation models we build on. Reconstruction failures are rare and mainly reduce the number of retained labels rather than introducing wrong ones, but strongly articulated or self-occluding objects, such as curled-up animals, remain harder to reconstruct, which is reflected in smaller gains on some animal categories. Since these models are used only offline to generate supervision, our framework directly benefits from progress in single-image reconstruction and 3D descriptors, and more accurate cross-mesh registration, *e.g.*, via functional maps (Ovsjanikov et al., 2012), could further increase label coverage.

6 Conclusion

We introduced 3D-SC, which uses reconstructed instance geometry to generate and verify semantic correspondence pseudo-labels. Rendered PartField descriptors complement DINO+SD when proposing matches, while geodesic consistency on the reconstructed surfaces filters disagreements before training. Lightweight adapters transfer this supervision to an image-based matcher that requires no 3D reconstruction at inference. The results show improved PCK@0.1 over the compared weakly supervised methods on SPair-71k, SPair-U, and SOCO, without using pose or keypoint training labels, with particularly large gains on SPair-Geo-Aware and strong performance on unseen SPair-U keypoints. They support using 3D foundation models as geometric teachers for 2D correspondence, with further gains expected as 3D reconstruction continues to improve.

AI USE STATEMENT

Generative AI tools assisted with research-methodology feedback, implementation, literature discovery, brainstorming, manuscript editing, and checks of numerical and narrative consistency. We have not used generative AI tools for helping to develop theoretical models or conceptual frameworks, assist with translation, support qualitative and thematic data analysis, interpret results. Furthermore, tasks such as generating synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, clean and reformat dataset are not applicable to this work. The authors are responsible for verifying all AI-assisted changes and for the final scientific claims, experimental results, and manuscript content.

REPRODUCIBILITY STATEMENT

We plan to release the processing, training, and inference code, together with experiment configurations and instructions for reproducing the reported results. The planned release includes documentation of dataset preparation, mesh reconstruction and canonicalization, pseudo-label generation, and the final adapter configuration.

REFERENCES

Kfir Aberman, Jing Liao, Mingyi Shi, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Neural best-buddies: Sparse cross-domain correspondence. *ACM Transactions on Graphics (TOG)*, 2018.

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. In *European Conference on Computer Vision Workshop (ECCVW)*, 2022.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. arXiv:2511.16719, 2025.

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 (ICCV)*, 2021.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Yu Chi, Leonhard Sommer, Olaf Dünkel, Dominik Muhle, Daniel Cremers, Christian Theobalt, and Adam Kortylewski. C3PO: Canonicalization of 3D pose from partial views with generalizable correspondence features. In *International Conference on 3D Vision (3DV)*, 2026.

Claudia Cuttano, Gabriele Trivigno, Carlo Masone, and Stefan Roth. MARCO: Navigating the unseen space of semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Nicolas Donati, Abhishek Sharma, and Maks Ovsjanikov. Deep geometric functional maps: Robust feature learning for shape correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Olaf Dünkel, Thomas Wimmer, Christian Theobalt, Christian Rupprecht, and Adam Kortylewski. Do it yourself: Learning semantic correspondence from pseudo-labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Olaf Dünkel, Basavaraj Sunagad, Haoran Wang, David T. Hoffmann, Christian Theobalt, and Adam Kortylewski. SOCO: Benchmarking semantic object correspondence in vision foundation models. In *European Conference on Computer Vision (ECCV)*, 2026.

Niladri Shekhar Dutt, Sanjeev Muralikrishnan, and Niloy J Mitra. Diffusion 3D features (Diff3F): Decorating untextured shapes with distilled semantic features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Frank Fundel, Johannes Schusterbauer, Vincent Tao Hu, and Bjorn Ommer. Distillation of diffusion features for semantic correspondence. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(7):1711–1725, 2017.

Regine Hartwig, Dominik Muhle, Riccardo Marin, and Daniel Cremers. GECO: Geometrically consistent embedding with lightspeed inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using Stable Diffusion. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Yiwen Huang, Yixuan Sun, Chenghang Lai, Qing Xu, Xiaomei Wang, Xuli Shen, and Weifeng Ge. Weakly supervised learning of semantic correspondence through cascaded online correspondence refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Jiin Im, Sisung Liu, and Je Hyeong Hong. Shape-of-You: Fused Gromov-Wasserstein optimal transport for semantic correspondence in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Artur Jesslen, Guofeng Zhang, Angtian Wang, Wufei Ma, Alan Yuille, and Adam Kortylewski. NOVUM: Neural object volumes for robust object classification. In *European Conference on Computer Vision (ECCV)*, 2024.

Jiwon Kim, Kwangrok Ryoo, Junyoung Seo, Gyuseong Lee, Daehwan Kim, Hansang Cho, and Seungryong Kim. Semi-supervised learning of semantic correspondence with pseudo-labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Nilesh Kulkarni, Shubham Tulsiani, and Abhinav Gupta. Canonical surface mapping via geometric cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Xin Li, Deng-Ping Fan, Fan Yang, Ao Luo, Hong Cheng, and Zicheng Liu. Probabilistic model distillation for semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Xinghui Li, Jingyi Lu, Kai Han, and Victor Adrian Prisacariu. SD4Match: Learning to prompt Stable Diffusion model for semantic matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(5): 978–994, 2011.

Minghua Liu, Mikaela Angelina Uy, Donglai Xiang, Hao Su, Sanja Fidler, Nicholas Sharp, and Jun Gao. PartField: Learning 3D feature fields for part segmentation and beyond. *arXiv preprint arXiv:2504.11451*, 2025.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Octave Mariotti, Oisin Mac Aodha, and Hakan Bilen. Improving semantic correspondence with viewpoint-guided spherical maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Octave Mariotti, Zhipeng Du, Yash Bhalgat, Oisin Mac Aodha, and Hakan Bilen. Jamais Vu: Exposing the generalization gap in supervised semantic correspondence. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. SPair-71k: A large-scale benchmark for semantic correspondence. arXiv:1908.10543, 2019.

Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings for deformable shape correspondence. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics (TOG)*, 2012.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. SAM 3D: 3Dfy anything in images. arXiv:2511.16624, 2025.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. SHIC: Shape-image correspondences with no keypoint supervision. In *European Conference on Computer Vision (ECCV)*, 2024.

Leonhard Sommer, Olaf Dünkel, Christian Theobalt, and Adam Kortylewski. Common3D: Self-supervised learning of 3D morphable models for common objects in neural feature space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Leonhard Sommer, Artur Jesslen, Basavaraj Sunagad, and Adam Kortylewski. Category-level 3D correspondence in camera space via morphable object priors. In *European Conference on Computer Vision (ECCV)*, 2026.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Tatsunori Taniai, Sudipta N. Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Krispin Wandel and Hesheng Wang. SemAlign3D: Semantic correspondence between RGB-images through aligning 3D object-class representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Pengyuan Wang, Takuya Ikeda, Robert Lee, and Koichi Nishiwaki. GS-Pose: Category-level object pose estimation via geometric and semantic correspondence. In *European Conference on Computer Vision (ECCV)*, 2024.

Zehan Wang, Ziang Zhang, Jiayang Xu, Jialei Wang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient Anything V2: Unifying orientation and rotation understanding. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Fei Xue, Sven Elflein, Laura Leal-Taixe, and Qunjie Zhou. MATCHA: Towards matching anything. *arXiv preprint arXiv:2501.14945*, 2025.

Li Yi, Vladimir G Kim, Duygu Ceylan, Wei Shen, Mengyuan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3D shape collections. In *Proceedings of ACM SIGGRAPH Asia)*, 2016.

Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. AP-10K: A benchmark for animal pose estimation in the wild. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa F. Polanía, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable Diffusion complements DINO for zero-shot semantic correspondence. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros. Learning dense correspondence via 3D-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Junzhe Zhu, Yuanchen Ju, Junyi Zhang, Muhan Wang, Zhecheng Yuan, Kaizhe Hu, and Huazhe Xu. DenseMatcher: Learning 3D semantic correspondence for category-level manipulation from a single demo. *International Conference on Learning Representations (ICLR)*, 2025.