SEMANTIC CORRESPONDENCE IS A DENSE PREDICTION TASK: IT SHOULD BE EVALUATED AS ONE

Anonymous authors
Paper under double-blind review

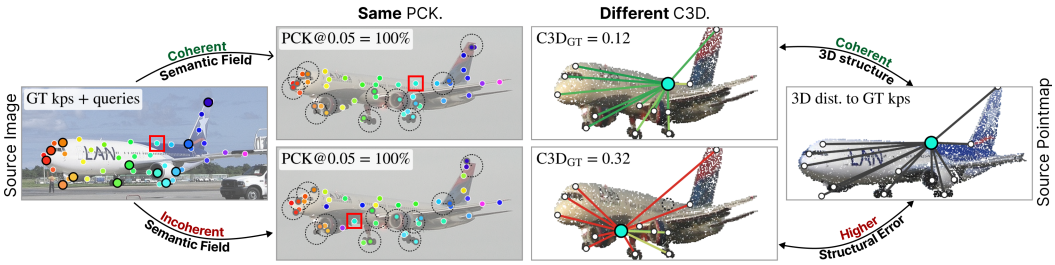

Figure 1: **Different Semantic Field, same PCK, different C3D.** *Left:* source image with its annotated keypoints (outlined) and sampled query points. Two predicted correspondence fields on the target. The bottom remains correct for every prediction that falls inside a PCK tolerance region (dashed line). PCK evaluates sparse GT keypoints and is unaffected by any change outside their tolerance regions: both fields reach 100%, although one semantic field is not coherent. *Right:* C3D lifts each query and its match into 3D (we only visualize a single query, highlighted with a red □) and compares their distances to the annotated keypoints in the source (grey) and in the target (*i.e.* preserved, distorted). We see that the incoherent semantic field breaks the 3D structure.

ABSTRACT

Semantic correspondence establishes matches between semantically equivalent locations on different object instances. The desired output is a correspondence field that can be queried anywhere on an object, yet the task remains evaluated at sparse annotated landmarks. Current metrics (*e.g.*, PCK) leave predictions between those landmarks unmeasured, leaving the spatial generality of the predicted correspondence field unassessed. We introduce 2D-3D Correspondence Consistency (C3D), an evaluation protocol that probes a fixed correspondence map at independently sampled foreground locations. C3D lifts sampled queries and their predicted matches into 3D via monocular point maps and measures how much the correspondence distorts normalized intra-object distances, either among the samples themselves (annotation-free) or with respect to annotated anchors (annotation-assisted). We evaluate each method's fixed correspondence map to probe its representation quality, without post-processing or test-time optimization. Computed on SPair-71k, C3D predicts how methods rank on the unseen keypoints of SPair-U better than SPair-71k PCK does, showing that it is a good generalization indicator. Finally, we turn the annotation-assisted structural residual into supervision: we train K2D using pseudo-correspondences obtained by fused semantic–structural transport against ground-truth anchors. This dense supervision improves transfer to unseen keypoint definitions, providing a concrete setting in which dense supervision and dense evaluation can be studied together.

1 INTRODUCTION

Semantic correspondence establishes matches between semantically equivalent locations on different object instances. Unlike geometric correspondence, it must accommodate differences in shape, appearance, articulation, and visibility. The desired output is a correspondence field that can be queried across an object, rather than a detector restricted to a predefined set of landmarks. Robot manipulation systems (Di Palo & Johns, 2024; Seker et al., 2025) often still rely on self-supervised

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.

features to transfer interactions to novel objects. These applications need matches at task-relevant locations that might not coincide with benchmark landmarks. Their continued use of general-purpose features may reflect a need for transfer beyond training landmarks, where supervised correspondence models have shown generalization gaps (Mariotti et al., 2025). Yet major correspondence benchmarks annotate only sparse landmarks. Benchmarks such as SPair-71k (Min et al., 2019) quantify how often predicted landmarks fall within a chosen distance of their annotated targets, typically using Percentage of Correct Keypoints (PCK). These annotations provide valuable semantic reference points, but they cover only a small part of the prediction domain, and landmark definitions differ across datasets. Two correspondence fields that agree at every evaluated landmark receive the same PCK, even if they behave differently everywhere else. Furthermore, a single threshold does not distinguish errors within its acceptance radius or characterize the distribution of larger errors. Evaluating multiple thresholds improves localization analysis, but it does not extend spatial coverage.

Recent work has exposed complementary limitations of this evaluation regime. Keypoint Average Precision (KAP) accounts for spurious high-similarity matches, including those proposed for source keypoints without a visible target counterpart (Mariotti et al., 2024). Jamais Vu demonstrates that strong supervised correspondence models can generalize poorly beyond their training keypoints and introduces SPair-U with newly annotated landmarks (Mariotti et al., 2025). MARCO evaluates transfer to unseen categories on MP-100 (Cuttano et al., 2026), a more category-diverse keypoint dataset originally introduced for category-agnostic pose estimation (Xu et al., 2022). Such additional annotations broaden semantic tests of generalization, but retain dataset-specific category and landmark selection biases and still leave most foreground locations untested.

Training methods have also begun to use supervision beyond annotated landmarks. DINOv2 provides transferable features (Oquab et al., 2023), and combining DINO with diffusion features improves zero-shot semantic matching (Zhang et al., 2023). DIY-SC and 3D-SC mine geometry-informed pseudo-matches, MARCO extends sparse labels through dense self-distillation, and Shape-of-You uses structural transport to generate soft supervision (Dünkel et al., 2025; Jesslen et al., 2026; Cuttano et al., 2026; Im et al., 2026). These efforts move training toward denser correspondence fields, while benchmark improvements alone cannot establish how predictions behave between annotated landmarks. This motivates evaluating the predicted field at independently sampled object locations, alongside its accuracy at known keypoints.

We address this need with *2D-3D Correspondence Consistency (C3D)*. Given a fixed predicted 2D map, we lift sampled foreground queries and their predicted targets into frozen monocular point maps. We then measure how much the correspondence distorts normalized intra-object distances. The annotation-free variant compares relationships among sampled points, while the annotation-assisted variant compares their relationships to known semantic anchors. No correspondence optimization or geometric alignment is performed to improve a method's score. C3D measures a geometric property of the field; it does not establish dense semantic correctness by itself.

We evaluate eight released baselines and three variants of our model with C3D on the SPair-71k test pairs and with PCK on eleven keypoint benchmarks. Computed on SPair-71k alone, C3D orders the methods closer to their accuracy on the unseen landmarks of SPair-U than SPair-71k PCK does (Kendall τ_b of 0.75–0.82 against 0.60), and it exposes supervised models whose keypoint accuracy does not transfer. We also describe a dense-supervision model, K2D, that uses structurally informed pseudo-correspondences to adapt a frozen feature extractor. Over training on the keypoints alone, these labels add 9.1 points of PCK@0.10 on SPair-U, and K2D is the best or tied-best method on all eleven benchmarks.

In summary, our contributions are as follows:

(i) A protocol for evaluating the structural consistency of sampled dense correspondence fields, with annotation-free and annotation-assisted C3D errors.

(ii) A comparative study relating C3D to sparse metrics, which exposes models whose accuracy does not transfer without evaluating on out-of-distribution benchmarks.

(iii) A dense-supervision method with matched controls for studying how geometry-guided pseudo-correspondences affect predictions beyond training landmarks.

2 Related work

Semantic correspondence. Handcrafted descriptors (Liu et al., 2010) gave way to learned matching networks (Rocco et al., 2017) and pretrained features that support matching without correspondence supervision (Aberman et al., 2018; Amir et al., 2022). Zhang et al. (2023) combine complementary DINO and diffusion features, while Stracke et al. (2025) extract diffusion features from clean images. Limited annotations motivate weak supervision (Chen et al., 2020; Lan et al., 2021) and pseudo-label learning (Kim et al., 2022; Huang et al., 2023). Dünkel et al. (2025) filter pseudo-matches using geometric constraints, and Cuttano et al. (2026) extend sparse labels through dense self-distillation.

Geometric priors. Pretrained features can confuse symmetric parts and object orientations (Zhang et al., 2024). Mariotti et al. (2024) address these ambiguities with spherical priors, while Mariotti et al. (2025) learn canonical 3D correspondences. Wang et al. (2024); Leroy et al. (2024); Sommer et al. (2026) connect point-map with image matching, and Wang et al. (2025) jointly predict camera and scene geometry. These priors support geometric filtering of pseudo-labels (Jesslen et al., 2026) and soft supervision through structural transport (Im et al., 2026). Previous works have also shows that VGGT partially encodes correspondences (Bratulić et al., 2026), and can be trained to predict them explicitly (Yang et al., 2026). Our evaluation uses VGGT point-maps to assess dense correspondence.

Semantic correspondence evaluation. Standard PCK evaluation on SPair-71k (Min et al., 2019) queries only landmarks annotated and visible in both images. Mariotti et al. (2024) introduce KAP to account for spurious matches and absent target counterparts. Mariotti et al. (2025) test unseen landmarks through SPair-U, while Cuttano et al. (2026) repurpose MP-100 pose annotations (Xu et al., 2022) to test new categories and landmark definitions. These benchmarks broaden the queried categories and keypoints, but leave predictions elsewhere unmeasured. We address this limitation by evaluating structural consistency at independently sampled foreground locations.

Structural and cycle consistency. Mémoli (2011) formalizes object matching using Gromov–Wasserstein distances, which will later be used in Im et al. (2026) for structural consistency. Ovsjanikov et al. (2012) represent shape correspondences as maps between function spaces; Litany et al. (2017) learn this representation, and Cheng et al. (2024) extend it to image matching. Cycle and warp consistency also supervise dense matching (Zhou et al., 2016; Truong et al., 2022) and correspondence with canonical 3D shapes (Zhang et al., 2022).

3 How to evaluate Semantic correspondence as a dense task?

3.1 Limitations of existing metrics

Let I_s, I_t be a source and target image, and let $f_{s\to t}$ map a source pixel to a predicted target location. Throughout, subscripts s and t mark quantities of the source and target image; when a statement holds for either, we write $\square$, as in $I_\square$. We write $\mathcal{P}_t$ for the target image domain, *i.e.* the set of positions in I_t that a match can point to. Both standard metrics build on the same tolerance region: for an annotated keypoint pair (p_s^k, p_t^k), the correct-match region $\mathcal{B}_{k,\alpha} \subseteq \mathcal{P}_t$ collects the positions within αL_t of p_t^k, where L_t is the normalization length, typically the largest side of the target bounding box.

Percentage of Correct Keypoints (PCK) at threshold α measures how often the prediction lands in that region, averaged over the K keypoints annotated as visible in both images:

$$\mathrm{PCK}_\alpha(f) = \tfrac{1}{K}\textstyle\sum_{k=1}^{K} \mathbb{1}\big[f_{s\to t}(p_s^k) \in \mathcal{B}_{k,\alpha}\big], \tag{1}$$

with missing predictions counted as failures. PCK is interpretable and directly measures semantic localization at known landmarks, but it probes f only at limited keypoints: any change elsewhere in the image leaves Eq. (1) unchanged, as shown in Fig. 1. Threshold sweeps and PCK area under the curve (AUC) vary α rather than the queried locations, and therefore inherit the same spatial limitation.

Keypoint Average Precision (KAP) adds a notion of confidence and covers keypoints that are visible in the source but not in the target. Instead of the predicted coordinate alone, it consumes the full similarity map and compares the best score inside the correct-match region (same as in PCK), u_k^+, against the best score outside it, u_k^-. Pooling u_k^+ as a positive and u_k^- as a negative over all source-visible keypoints and taking the AP of these labeled scores yields KAP_α. A keypoint that is absent from the target contributes a negative only, so a confidently matched missing part is penalized. Full definition in App. B. However, KAP depends on the annotated keypoints and remains sparse.

3.2 SEMANTIC CORRESPONDENCE AS A DENSE PREDICTION TASK

Frozen inputs and sampling. For each ordered image pair, we use fixed foreground masks $\mathcal{M}_\square$ from SAM 3 (Carion et al., 2025) and point maps $X_\square$ inferred independently per image by VGGT (Wang et al., 2025), where $X_\square(p) \in \mathbb{R}^3$ is the 3D point assigned to pixel p. We uniformly sample (consistently across methods) up to $N = 256$ source-foreground pixels, denoted $p_s^{1,\dots,N}$. The two point maps are not guaranteed to share a coordinate frame, so we normalize each one independently, $\widehat{X}_\square = X_\square / \operatorname{scale}(\mathcal{X}_\square)$, with $\mathcal{X}_\square$ the foreground point cloud. More details in App. C.

Predicted correspondence map. Every method is queried by sampling a batch of source pixels, mapped to target pixels. For the feature-based methods (including K2D), the map is induced by a dense descriptor map F_s. The descriptor of a source pixel is bilinearly sampled from F_s and compared by cosine similarity with every target cell that lies on the image. The best cell is refined by a soft-argmax over its 3×3 neighbourhood at temperature $\tau = 0.05$ (Zhang et al., 2024) and mapped back to target-image pixels; this location is $f_{s\to t}(p)$.

Annotation-free structural error ($\mathrm{C3D}_{\mathrm{Free}}$). Each sampled source pixel p_s^i is mapped to $p_t^i = f_{s\to t}(p_s^i)$. After excluding predictions without finite geometry, n correspondences remain. For $n \geq 2$, the annotation-free error compares, over all unordered pairs, the distance $d_{ij}^\square$ between samples i and j in the normalized source and target point maps, weighting each pair by a Gaussian kernel on its source distance,

$$\mathrm{C3D}_{\mathrm{Free}} = \frac{\sum_{i<j} w_{ij} \left| d_{ij}^s - d_{ij}^t \right|}{\sum_{i<j} w_{ij}}, \quad \text{with } d_{ij}^\square = \|\widehat{X}_\square(p_\square^i) - \widehat{X}_\square(p_\square^j)\|_2, \;\; w_{ij} = e^{-(d_{ij}^s)^2/2\sigma^2}. \tag{2}$$

The weights make the score local. The number of pairs at source distance d grows with d up to the size of the object, so an unweighted average would be dominated by long-range pairs, whose distances largely reflect the global shape of each instance: two objects with different topology differ at long range even under a perfect correspondence. A wrong match, in contrast, is not amplified by range: by the reverse triangle inequality, moving one endpoint by ε changes any pair distance by at most ε. Down-weighting long-range pairs therefore keeps the matching signal and removes most of the shape variation. The weights depend only on the source side, which is method-independent. We set $\sigma = 0.25$ in normalized units, half the median radius of the normalized foreground pointcloud.

Annotation-assisted structural error ($\mathrm{C3D}_{\mathrm{GT}}$). When semantic keypoints are available, we retain every jointly visible anchor pair (p_s^k, p_t^k) with finite geometry. For $K_a \geq 3$ valid anchors, the annotation-assisted error compares the distances $d_{ik}^\square$ from each sample i to each anchor k, defined as in Eq. (2) with the anchor $p_\square^k$ in place of sample j,

$$\mathrm{C3D}_{\mathrm{GT}} = \frac{1}{n} \sum_{i=1}^{n} r_i^{\mathrm{GT}}, \quad \text{with } r_i^{\mathrm{GT}} = \frac{1}{K_a} \sum_{k=1}^{K_a} \left| d_{ik}^s - d_{ik}^t \right|. \tag{3}$$

Anchors can expose incorrect assignments between geometrically similar parts, but three anchors do not guarantee that all symmetries are resolved.

Combined score (C3D). Local weighting has a cost. A map that is globally wrong but locally smooth, such as a left–right flip or an object mapped onto a shifted region, preserves every short-range distance, and a many-to-one collapse is penalized mainly through long pairs, which the kernel discounts. The anchored error keeps exactly this sensitivity: an anchor on the far side of the object exposes a flip or a shift however smooth the map is locally. We therefore keep $\mathrm{C3D}_{\mathrm{GT}}$ unweighted and combine it, per image pair, with $\mathrm{C3D}_{\mathrm{Free}}$,

$$\mathrm{C3D} = \tfrac{1}{2}\,\mathrm{C3D}_{\mathrm{Free}} + \tfrac{1}{2}\,\mathrm{C3D}_{\mathrm{GT}}. \tag{4}$$

Both terms are mean absolute differences between distances in the same normalized units. The equal weights act on these units rather than on influence, and $\mathrm{C3D}_{\mathrm{GT}}$ is typically the larger term. C3D is defined for pairs with at least three valid anchors and is not annotation-free, so $\mathrm{C3D}_{\mathrm{Free}}$ is always reported alongside it. Applying the kernel of Eq. (2) to the source query-to-anchor distances gives a weighted anchored error $\mathrm{C3D}_{\mathrm{GT}}^w$, which we report for completeness but leave out of Eq. (4), since it discounts the distant anchors that detect global errors. All C3D errors are lower-is-better structural discrepancies rather than calibrated probabilities of semantic correctness, and the annotation-free terms cannot detect reflections, which preserve every intra-image distance.

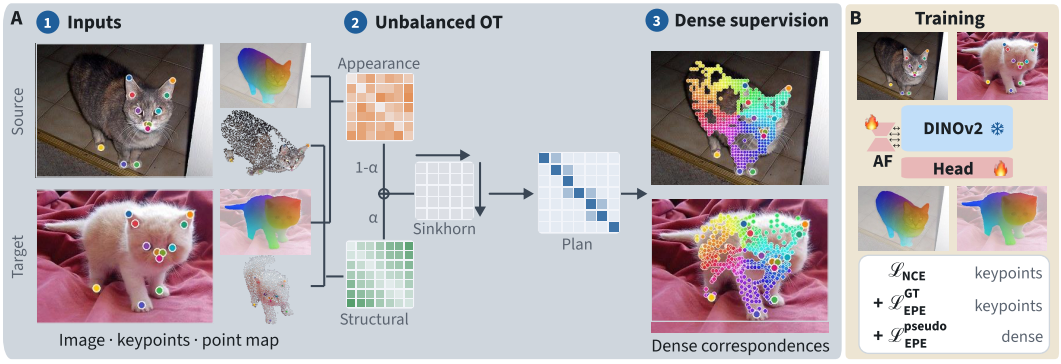

Figure 2: **K2D pipeline.** *A. Offline labeling.* ❶ Each image comes with its annotated keypoints, extracte features (DINOv2) and a VGGT point map. ❷ An appearance cost from the backbone features and a structural cost, which compares 3D distances to the annotated keypoints, are blended with weight α and solved by unbalanced optimal transport. ❸ The mutual best matches of the resulting plan give dense pseudo-correspondences over the whole object.
B. Training. AdaptFormer adapters (AF) and a light head on a frozen DINOv2 are trained with InfoNCE and end-point losses on the keypoints, and an end-point loss on the dense pseudo-correspondences obtained from the previous labeling pipeline..

4 K2D: From Keypoints To Dense Supervision

Sec. 3 scores a correspondence field by how well it preserves normalized intra-object 3D structure relative to semantic correspondences. Inspired by this idea, we test whether the same structure is also useful to learn from, by turning it into supervision. Given an image pair with a few annotated keypoints, we densify them: we generate pseudo-correspondences whose targets agree with the keypoints in normalized 3D structure as well as in appearance, and use them as pseudo-annotations to train a correspondence model, K2D. K2D is not a new architecture: apart from one added loss term, every component is shared with a control trained on the annotated keypoints alone, so comparing the two isolates the effect of the dense structural supervision. Geometry enters through the labels only; the network never sees point maps, and no term optimizes C3D directly. Structural consistency thus acts as a proxy task for training, and Sec. 5 shows that it leads to strong keypoint accuracy on held-out benchmarks, measured independently of the metric that motivated it.

4.1 Anchored structural transport

Inputs. Labels are generated offline from the same frozen inputs as the evaluation: SAM 3 masks (Carion et al., 2025), normalized per-image VGGT point maps (Wang et al., 2025) (normalization details provided in App. C), and the features of the frozen backbone, $F_\square$. Let $\mathcal{G}_\square^\circ$ be the foreground cells with finite geometry of each image. The anchor set $\mathcal{A} = \{(p_s^k, p_t^k)\}_{k=1}^{K_a}$ contains every jointly visible annotated keypoint pair with finite geometry. A pair with $K_a < 3$ is left unlabeled.

Correspondence Field. A pretrained backbone maps each image to a dense feature map, $F_\square$; $F_\square(p)$ is the bilinearly sampled feature at pixel p, and $\mathcal{G}_\square$ are the feature cells that lie on the image rather than on the padding. $f_{s\to t}(p)$ is the cell of $\mathcal{G}_t$ most similar to $F_s(p)$, refined by a local soft-argmax (Sec. 3.2); training uses the differentiable soft-argmax over all of $\mathcal{G}_t$, $\hat{f}_{s\to t}$, with temperature τ.

Costs. A candidate match between a source cell $p_s^i \in \mathcal{G}_s^\circ$ and a target cell $q^j \in \mathcal{G}_t^\circ$ is assigned an appearance cost (semantic) and a structural cost (geometric),

$$C_{ij}^{\text{sem}} = 1 - \langle F_s(p_s^i), F_t(q^j) \rangle, \qquad C_{ij}^{\text{geo}} = \frac{1}{K_a} \sum_{k=1}^{K_a} \left| d_{ik}^s - d_{jk}^t \right|. \tag{5}$$

Here d_{ik}^s is the source distance of Eq. (3) and $d_{jk}^t = \|\widehat{X}_t(q^j) - \widehat{X}_t(p_t^k)\|_2$ the distance from the candidate target cell to anchor k, so the structural cost is the residual r_i^{GT} of Eq. (3) for the hypothetical correspondence $p_s^i \mapsto q^j$: a candidate pays exactly the annotation-assisted structural error it would

incur under C3D. This is the anchor linearization with which Shape-of-You (Im et al., 2026) makes a fused Gromov–Wasserstein objective (Mémoli, 2011) tractable, except that our anchors are annotated keypoints rather than pairs mined from the plan: their quality is known, and since the cost depends on them alone, it is built once instead of being refreshed at every iteration. Each cost matrix is rescaled so that its 1st and 99th percentiles map to 0 and 1, which keeps the mixing weight meaningful across pairs, and the two are blended into $C = (1-\alpha)\,C^{\text{sem}} + \alpha\,C^{\text{geo}}$ with $\alpha = 0.3$, as in Im et al. (2026).

Transport. As in Im et al. (2026), we solve an entropic unbalanced transport problem (Cuturi, 2013; Chizat et al., 2018), here a single one on the fixed cost, with uniform reference marginals a and b over the two supports, we obtain the optimal plan:

$$\pi^\star = \arg\min_{\pi \geq 0} \langle C, \pi \rangle + \varepsilon\,\mathrm{KL}(\pi \,\|\, a \otimes b) + \rho\,\mathrm{KL}(\pi \mathbf{1} \,\|\, a) + \rho\,\mathrm{KL}(\pi^\top \mathbf{1} \,\|\, b), \tag{6}$$

with 100 log-domain Sinkhorn iterations, $\varepsilon = 1$, and $\rho = 0.75$. The relaxed marginals let cells without a counterpart, such as occluded or truncated parts, carry little mass instead of being forced into a spurious match. Correspondences are the mutual best matches of the plan: p_s^i and q^j are matched if each is the other's highest-mass partner, $j = \arg\max_{j'} \pi^\star_{ij'}$ and $i = \arg\max_{i'} \pi^\star_{i'j}$. Unlike Shape-of-You, which trains on a soft target blended with the network's current plan, we keep these hard matches and use them offline.

4.2 Training objective

Each ordered training pair contributes three terms. The first two supervise the K jointly visible annotated keypoints with a symmetric InfoNCE loss and a soft-argmax end-point error (Zhang et al., 2024), and are identical in every arm. The InfoNCE loss uses the other keypoints of the same pair as negatives, with a learned inverse temperature β, and $\mathrm{CE}(\cdot, \mathrm{I})$ the row-wise cross-entropy against the identity,

$$\mathcal{L}_{\text{NCE}} = \tfrac{1}{2}\Big[\mathrm{CE}\big(S, \mathrm{I}\big) + \mathrm{CE}\big(S^\top, \mathrm{I}\big)\Big], \qquad \text{with } S_{kl} = \beta\,\big\langle F_s(p_s^k), F_t(p_t^l)\big\rangle, \tag{7}$$

and the end-point error is measured at the annotated keypoints. The third term applies the same end-point error to $\mathcal{D}' \subseteq \mathcal{D}$, a random subset of the pair's pseudo-correspondences redrawn every epoch (App. F.2),

$$\mathcal{L}_{\text{EPE}}^{\text{GT}} = \frac{1}{K}\sum_{k=1}^{K} \big\|\hat{f}_{s\to t}(p_s^k) - p_t^k\big\|_2, \qquad \mathcal{L}_{\text{EPE}}^{\text{pseudo}} = \frac{1}{|\mathcal{D}'|}\sum_{(p,p')\in\mathcal{D}'} \big\|\hat{f}_{s\to t}(p) - p'\big\|_2. \tag{8}$$

Hence, the training loss is the sum of the three terms: $\mathcal{L} = \mathcal{L}_{\text{NCE}} + \mathcal{L}_{\text{EPE}}^{\text{GT}} + \mathcal{L}_{\text{EPE}}^{\text{pseudo}}$. (9)

5 Experiments

5.1 Evaluation protocol

Baselines. We compare against the released checkpoints of eight methods that span the supervision spectrum: the self-supervised DINOv2 (Oquab et al., 2023) and SD+DINO (Zhang et al., 2023), the weakly supervised Shape-of-You (Im et al., 2026), Spherical Maps (Mariotti et al., 2024) and DIY-SC (Dünkel et al., 2025), and the supervised GeoAware-SC (Zhang et al., 2024), GECO (Hartwig et al., 2025) and MARCO (Cuttano et al., 2026). Every method exposes a dense feature map and is queried through the one decoder of Sec. 3.2 (cosine nearest neighbour over on-image cells, refined by a 3×3 soft-argmax), so that a row of every table *measures a representation* rather than a post-processing step. All six weakly and fully supervised baselines are trained on SPair-71k; GECO additionally trains on PF-Pascal and AP-10K.

Benchmarks. Keypoint accuracy is measured on eleven benchmarks: SPair-71k (Min et al., 2019), SPair-U (Mariotti et al., 2025), PF-Pascal and PF-Willow (Ham et al., 2017), the SOC and Cross-SOC (abbreviated X-SOC) test sets of SOCO (Dünkel et al., 2026), CUB-200 (Wah et al., 2011), AP-10K (Yu et al., 2021), MP-100 (Cuttano et al., 2026), Animal3D (Xu et al., 2023) and COCO-DensePose (Güler et al., 2018) (per-dataset pair counts in Table A7). Two of them probe generalization

Table 1: **Relation between PCK and C3D.** PCK@0.10 (per image, ↑), and different variations of C3D (↓) are scored on SPair-71k test pairs with $N = 256$ queries. Free° is the *unweighted* annotation-free error; C3D_{Free} is its distance-weighted form. Cell shading gives each method's rank per column, from white (best) to dark blue (worst); the best and second-best values are also shown in **bold** and underlined. We observe a similar ranking patterns on each side of the dotted line.

Method	Sup.	# pairs	PCK@0.10 ↑		Free° ↓	C3D_{Free} ↓	C3D_{GT} ↓	C3D ↓
			SPair-71k	SPair-U	SPair-71k	SPair-71k	SPair-71k	SPair-71k
DINOv2	sS	N/A	55.4	53.5	0.652	0.497	0.620	0.550
SD+DINO	sS	N/A	61.0	57.5	0.579	0.403	0.570	0.478
Shape-of-You	wS	53k†	63.0	60.2	0.610	0.446	0.576	0.500
Spherical Maps	wS	53k†	65.8	59.4	0.714	0.475	0.668	0.561
DIY-SC	wS	53k†	70.9	65.0	0.512	0.333	0.506	0.412
GeoAware-SC	S	53k†	84.9	56.1	0.741	0.607	0.621	0.601
GECO	S	76k‡	74.2	57.4	0.740	0.536	0.689	0.602
MARCO	S	53k†	87.2	67.8	0.481	0.304	0.488	0.389
$\text{K2D}_{\text{SPAIR}}$	S	53k†	85.7	71.6	0.397	0.233	0.441	0.332
$\text{K2D}_{\text{DIVERSE}}$	S	54k	88.5	**74.4**	0.373	0.217	0.422	0.316
K2D_{FULL}	S	155k	**88.7**	**74.4**	**0.370**	**0.214**	**0.421**	**0.314**
$\text{K2D}_{\text{ORACLE}}$	S	413k	99.2	94.8	0.341	0.197	0.394	0.291

beyond the training annotations by construction. SPair-U with new landmark definitions, so it measures transfer to *unseen landmarks*. MP-100 shares no category between its test and train/val splits, so it measures transfer to *unseen categories*; we evaluate only 18 of its 20 test categories (see App. D). No test split of any benchmark enters the training; except the oracle (details in App. G).

Metrics. PCK@0.10 averaged per image is the headline keypoint metric (Table 2); PCK@0.05 is reported in Table A8. C3D is scored following Sec. 3.2: SAM 3 masks, VGGT point maps, $N = 256$ queries per pair drawn with fixed pair-based seed (consistent across methods). We report the three versions: the annotation-free C3D_{Free}, the annotation-assisted C3D_{GT} and the combined C3D. Note that the labeler of Sec. 4.1 and C3D use the same point maps, so a low structural error of our own rows is not, by itself, independent evidence of semantic correctness; that evidence comes from the keypoint benchmarks, which never enter the labeler.

5.2 What C3D measures

Before using C3D to compare methods we ask three questions of the metric itself: *(i)* whether it agrees with human annotation it never sees, *(ii)* whether it detects changes in the field that PCK cannot, and *(iii)* whether its ranking depends on its own settings.

Agreement with denser annotations. DensePose (Güler et al., 2018) gives, on 489 COCO-DensePose test pairs, 29 surface correspondences per pair on limb and torso interiors: three times the density of the keypoint benchmark and away from its annoated keypoints (App. I). Per pair, C3D_{Free} tracks the error at these surface points for every method (Spearman $+0.58$ pooled) without seeing any annotation; the keypoint score of the same pair reaches -0.74, as it scores annotated points of the same instance. At the method level the two are equally faithful, each ordering the six methods with a single inversion against the surface scores (Table A9, Fig. A4): the keypoints overrate GeoAware-SC, ten points above DIY-SC on the COCO landmarks (55.2 vs 45.1) but below it at the surface points (47.3 vs 49.7), where C3D_{Free} places it. This shows that an annotation-free score (C3D) can track the accuracy of the field where no keypoints is annotated, on the most articulated category.

Sensitivity to noise. To measure sensitivity, we perturb the sampled field: 10, 25 and 50% of the sampled correspondences are shuffled among the queries or displaced locally, for all evaluated methods (details in Table A12). Every component of C3D rises monotonically with the corrupted fraction for every method and both corruptions (see Fig. A7). This is the expected behavior, and the reason the sweep is worth reporting: both corruptions destroy the local layout of the transferred field, which is exactly what the pairwise residual of Eq. (2) measures, while PCK is blind to them by construction. A structural score that did not rise here could not be used to compare fields at all. Importantly, **the ranking is preserved**.

Table 2: **Evaluation on standard benchmarks.** PCK@0.10 (per image, ↑) on eleven 2D correspondence benchmarks. The oracle (gray) is also trained on test splits. Supervision: sS = self-Supervised, wS = weakly Supervised, S = Supervised. †Trained on SPair-71k. Methods trained on pseudo-labels may use a increased pair count. ‡GECO also trains on PF-Pascal and AP-10K; its pair count is not disclosed, current estimation is conservative. Best highlighted in **bold**, second best underlined.

Method	Sup.	# pairs	PCK@0.10 ↑										
			SPair-71k	SPair-U	PF-Pascal	PF-Willow	SOC	X-SOC	CUB-200	AP-10K	MP-100	Animal3D	COCO-DP
DINOv2	sS	N/A	55.4	53.5	75.9	64.5	62.1	56.2	85.2	62.5	53.3	67.6	38.4
SD+DINO	sS	N/A	61.0	57.5	84.7	75.7	64.7	59.0	85.7	64.6	55.6	67.6	41.4
Shape-of-You	wS	53k†	63.0	60.2	88.5	77.6	65.8	61.0	86.9	67.6	58.3	69.9	42.8
Spherical Maps	wS	53k†	65.8	59.4	85.9	79.0	66.7	61.9	85.8	65.6	55.9	66.5	35.3
DIY-SC	wS	53k†	70.9	65.0	88.5	80.0	69.7	64.6	86.8	69.0	58.9	70.7	44.0
GECO	S	76k‡	74.2	57.4	91.6	78.9	69.1	65.3	83.8	82.9	62.4	70.5	47.6
GeoAware-SC	S	53k†	84.9	56.1	86.9	75.9	73.1	69.5	82.3	76.2	63.9	74.1	55.0
MARCO	S	53k†	87.2	67.8	89.9	80.6	78.6	75.9	87.5	76.9	65.9	76.5	56.8
$K2D_{SPAIR}$	S	53k†	85.7	71.6	92.6	83.0	81.6	80.2	88.0	80.8	70.0	80.6	61.2
$K2D_{DIVERSE}$	S	54k	88.5	**74.4**	**95.8**	83.3	90.5	88.1	96.5	87.4	75.0	83.8	**72.7**
$K2D_{FULL}$	S	155k	**88.7**	**74.4**	**95.8**	**83.9**	**91.2**	**89.2**	**96.6**	**89.5**	**76.4**	**84.0**	**72.7**
$K2D_{ORACLE}$	S	413k	99.2	94.8	95.8	95.8	98.5	95.4	99.1	97.7	85.0	89.5	74.5

Sensitivity to settings. C3D has three settings: the number of sampled queries N, the kernel and scale σ of Eq. (2), and the combination weight of Eq. (4). The method ranking is identical for every number of queries we test. Around the default scale, each of the three kernels and each neighbouring combination weight swaps at most a few pairs of methods, and every kernel converges to the unweighted ranking as it widens; only a kernel much narrower than the default, which limits the error to the immediate neighbourhood of each query, reorders the methods. The conclusions therefore do not hinge on a tuned setting; App. I reports the numbers (Tables A10 and A11).

5.3 Evaluation with C3D

For every method, Table 1 compares keypoint accuracy with C3D on SPair-71k; Tables A14 to A16 give more details. The three scores order the methods almost identically (Table A11, bottom), so the annotation-free $C3D_{Free}$ alone can also be used to rank methods on non-annotated datasets.

Where C3D and PCK disagree. PCK and C3D agree on the top of Table 1 but disagree on two supervised baselines. GeoAware-SC is 5th by PCK on SPair-71k but 10th of eleven by C3D, and GECO is 6th by PCK but last by C3D. SPair-U sides with C3D: on unseen landmarks the two methods fall to 10th and 9th, as the DensePose surface points already suggested for GeoAware-SC (Sec. 5.2). Both fit the training landmarks, but not the field between them, and only the structural score reveals it from in-domain data. The same holds over the whole panel: ranked on SPair-71k, C3D orders the methods closer to SPair-U than SPair-71k PCK does (Kendall τ_b with SPair-U of 0.60 for PCK, against 0.75 for C3D and 0.82 for $C3D_{Free}$; Table A13), and the gap widens on the eight released baselines alone. A structural score on the in-domain test pairs thus anticipates how methods rank on landmarks none of them was trained on.

K2D under C3D. Our rows are the best entries in every column, including the two annotation-free ones: $K2D_{FULL}$ has the lowest error in all four structural columns, followed by $K2D_{DIVERSE}$, and both lead MARCO by a wide margin. We also observe that $K2D_{ORACLE}$ has the best score, which shows it has learn the most coherent latent representation. Because the labeler shares the geometry prior of the metric, we do not read these values as an independent confirmation of K2D's performances; the independent evidence is Table 2. What the structural columns add is that the gain on the annotated keypoints is not obtained by fitting the landmarks at the expense of the field.

5.4 Sparse Keypoint Accuracy

Table 2 compares all methods at PCK@0.10. K2D, trained on 155k pairs from eight datasets with every test split held out ($K2D_{FULL}$), is the best or tied-best method on all eleven benchmarks; $K2D_{DIVERSE}$, trained under the same recipe on around $6k$ samples per benchmark, ties it on SPair-U, PF-Pascal and COCO-DP and is second everywhere else. Against MARCO, the strongest supervised baseline, K2D gains $+1.5$ on SPair-71k and $+6.6$ on SPair-U, the two benchmarks on which MARCO trains and which neither model has seen the landmarks of, respectively. The larger margins on SOCO ($+12.6$ and $+13.3$), AP-10K ($+12.6$), CUB-200 ($+9.1$) and COCO-DensePose ($+15.9$) are due to our training set containing the train splits of these datasets and MARCO's does not, so they measure

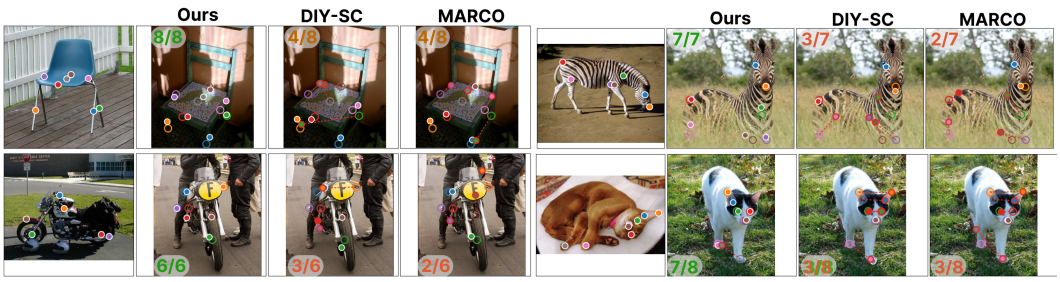

Figure 3: **Qualitative results.** Keypoint transfer of K2D (Ours), DIY-SC and MARCO on four challenging pairs in terms of shape (chair), viewpoint (motorbike) or pose (zebra/cat). The source image shows the annotated keypoints; in each target, filled dots are the predictions and rings of the same color the ground truth. We illustrate the dense features via PCA for these pairs in Fig. A9.

what more diverse data buys rather than transfer. The margins that measure transfer are SPair-U ($+6.6$), PF-Willow ($+3.3$), which no method trains on, and MP-100 ($+10.5$), whose test categories are absent from every training split.

Generalization beyond the training landmarks. The SPair-U column separates methods that learned the SPair-71k landmarks from methods that learned a correspondence field. The accuracy of the three supervised baselines drops sharply from SPair-71k to SPair-U (GeoAware-SC -28.8, MARCO -19.4, GECO -16.8); GeoAware-SC, second among the baselines on SPair-71k, falls below the self-supervised SD+DINO on SPair-U. K2D drop the least (-14.3) among supervised methods; the weakly supervised DIY-SC loses only 5.9 but from a 17.8-point lower start. Dense structural supervision therefore does not trade landmark accuracy for generality: it improves both.

Training data. The row $\text{K2D}_{\text{SPAIR}}$ is our recipe on the SPair-71k training pairs alone, the same data as most baselines. It already outperforms MARCO on ten of the eleven benchmarks, by $+3.8$ on SPair-U, and trails it only on SPair-71k itself. Training on the diverse pool instead, which has the same number of pairs but 27 times as many categories (App. G), consistently improves: as observed for self-supervised features (Sunagad et al., 2026), a more diverse image collection yields a more general representation at equal budget. Lifting the cap to every labelled train/val pair helps only if the pool stays balanced: without a cap on CUB-200 the 86,778 bird pairs make up 37% of the pool, and the resulting model matches the full model's mean PCK@0.10 but loses on SPair-71k and SPair-U (details in App. G). For the 155k pool of K2D_{FULL}, gains over $\text{K2D}_{\text{DIVERSE}}$ fall where the training data grew: AP-10K $+2.1$, MP-100 $+1.4$, X-SOC $+1.1$, with the remaining columns tied.

Limitations and Future Work. Structural consistency complements sparse semantic metrics: C3D_{Free} and C3D_{GT} show consistent trends and agree with rankings across benchmarks. Broader studies can extend that to further latent representation, or additionnal object types. Off-the-shelf masks, monocular geometry, and finite sampling keep dense evaluation practical without dense annotations but further extensions include confidence-aware matching, no-match predictions, adaptive sampling, and alternative geometry models.

6 Conclusion

Semantic correspondence is meant to produce a field that can be queried anywhere on an object, yet it is evaluated at a handful of annotated keypoints. We introduced C3D to measure the field where no keypoint is annotated: it lifts sampled queries and their matches into monocular point maps and scores how well the match preserves the object's 3D structure, with keypoints as anchors or without any annotation. Computed on SPair-71k, it predicts how methods rank on the unseen landmarks of SPair-U, and it exposes supervised models whose keypoint accuracy does not transfer. The same structural residual also works as supervision. K2D densifies annotated keypoints by anchored semantic–structural transport and is the best or tied-best method on all eleven keypoint benchmarks, with the smallest drop from SPair-71k to SPair-U, among supervised methods. At equal size, a more diverse training pool improves every benchmark, whereas nearly tripling it helps only where the new data lies. We hope C3D will be reported alongside PCK, so that progress in semantic correspondence is measured on the whole object and not only at its landmarks.

AI USE STATEMENT

In this work, we used generative AI tools for designing or providing feedback on research methodology or experiments, help in implementing methods, discover research topics or identify missed relevant literature, brainstorming, edit the manuscript to improve readability. 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. We have reviewed all AI-assisted work. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The metric is described in Sec. 3.2, with the verification of what it measures in Sec. 5 with additional details introduced in App. I. Our method is described in Sec. 4, with additional description and hyperparameters detailed in App. F. We did the utmost effort to add as many details as possible to be able to reproduce the results. Additionally, the complete code, including data processing, metrics implementation and model weights, will be made available upon acceptance on Github.

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 (ECCV Workshops)*, 2022.

Amir Beck and Shoham Sabach. Weiszfeld's method: Old and new results. *Journal of Optimization Theory and Applications*, 164(1):1–40, 2015.

Jelena Bratulić, Sudhanshu Mittal, Thomas Brox, and Christian Rupprecht. On geometric understanding and learned priors in feed-forward 3d reconstruction models. In *European Conference on Computer Vision (ECCV)*, 2026.

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.

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

Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 43(10):3632–3647, 2020.

Xinle Cheng, Congyue Deng, Adam W Harley, Yixin Zhu, and Leonidas Guibas. Zero-shot image feature consensus with deep functional maps. In *European Conference on Computer Vision (ECCV)*, 2024.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018.

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.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Conference on Neural Information Processing Systems (NeurIPS)*, 2013.

Norman Di Palo and Edward Johns. Dinobot: Robot manipulation via retrieval and alignment with vision foundation models. In *International Conference on Robotics and Automation (ICRA)*, 2024.

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.

Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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.

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, Olaf Dünkel, and Adam Kortylewski. Geometry matters: 3d foundation priors for learning semantic correspondence. arXiv:2605.30093, 2026.

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.

Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision (ECCV)*, 2024.

Or Litany, Tal Remez, Emanuele Rodolà, Alex M. Bronstein, and Michael M. Bronstein. Deep functional maps: Structured prediction for dense shape correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2017.

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, 2010.

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.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

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

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 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)*, 31(4):1–11, 2012.

Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

M Yunus Seker, Shobhit Aggarwal, and Oliver Kroemer. Grounded task axes: Zero-shot semantic skill generalization via task-axis controllers and visual foundation models. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 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.

Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. Cleandift: Diffusion features without noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Basavaraj Sunagad, Artur Jesslen, and Adam Kortylewski. Two global crops suffice: Locating semantic emergence in dino-style self-supervised learning. arXiv:2609.28187, 2026.

Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Probabilistic warp consistency for weakly-supervised semantic correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Endre Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386, 1937.

Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, Wei Ji, Chen Wang, Xiaoding Yuan, Prakhar Kaushik, Guofeng Zhang, Jie Liu, Yushan Xie, Yawen Cui, Alan Yuille, and Adam Kortylewski. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Lumin Xu, Sheng Jin, Wang Zeng, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Pose for everything: Towards category-agnostic pose estimation. In *European Conference on Computer Vision (ECCV)*, 2022.

Songlin Yang, Tianyi Wei, Yushi Lan, Zeqi Xiao, Anyi Rao, and Xingang Pan. Dense semantic matching with VGGT prior. In *European Conference on Computer Vision (ECCV)*, 2026.

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 Polania Cabrera, 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.

Kaifeng Zhang, Yang Fu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. Self-supervised geometric correspondence for category-level 6d object pose estimation in the wild. In *International Conference on Learning Representations (ICLR)*, 2022.

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.