Monocular Category-Level 3D Correspondence via Morphable Priors

Leonhard Sommer [1*] Artur Jesslen [1*] Basavaraj Sunagad [1,2] Adam Kortylewski [2]

[1]University of Freiburg [2]CISPA

Abstract

Understanding 3D objects from images is fundamental to robotics and AR/VR applications. While recent work has made progress in category-level pose estimation, current representations fail to capture the fine-grained semantics needed for reasoning about object parts, functions, and interactions. We identify the next frontier in 3D object understanding as ***monocular category-level 3D correspondence****—predicting, from a single image, 3D locations that remain consistent across instances within a category. To enable research in this direction, we introduce* ***HouseCorr3D****, the first large-scale benchmark for monocular category-level 3D correspondence with 178k images across 50 household object categories, 280 unique instances, and 3D keypoint annotations directly on CAD models. Crucially, HouseCorr3D provides amodal correspondence labels for occluded regions and explicit symmetry annotations, addressing key limitations of existing datasets. We further propose* ***Morpheus****, a framework that learns* ***morphable category-level shape priors*** *to establish semantically consistent 3D correspondences in camera space. By learning deformable canonical models, Morpheus moves beyond traditional pose-centric approaches to enable* ***fine-grained, correspondence-level object understanding****. Experiments demonstrate that Morpheus significantly outperforms baselines, establishing a new paradigm for 3D object understanding.*

1. Introduction

Understanding objects in 3D from images is a long-standing challenge in computer vision, with applications in robotics, augmented reality (AR), and virtual reality (VR). Traditional 3D object understanding has primarily focused on pose estimation, object detection, or 3D reconstruction. However, current approaches fail to capture the fine-grained semantics needed for reasoning about object parts, their functions, and how they can be manipulated or interacted with. A key step toward richer understanding is to establish semantic correspondences – estimating which points on different objects represent the same functional part. In 2D, this problem has driven extensive research [16, 26, 27, 29, 37], enabling applications like image matching, retrieval, and style transfer. Yet, 2D correspondences are inherently limited by viewpoint dependence, occlusion, and symmetry ambiguities. We therefore propose to move beyond 2D, and towards the prediction of semantically aligned 3D locations that remain consistent across all instances of a category (as illustrated in Fig. 1). Unlike prior work that maps pixels into normalized canonical spaces [19, 45], we suggest establishing correspondences directly in camera space, yielding an unambiguous representation for evaluation and downstream reasoning. Formally, we define this novel task as follows:

> **Monocular Category-level 3D correspondence:** Given two query and target RGB-D images I^q and I^t of objects from the same category, and a query 3D point $x^q \in \mathbb{R}^3$ in the camera space of I^q, the task is to predict the 3D point $x^t \in \mathbb{R}^3$ in I^t camera space that corresponds to the same semantic part.

We illustrate the monocular category-level 3D correspondence in camera space setup via 3D meshes in Fig. 3a. Unfortunately, existing benchmarks such as NOCS-Real275 [45], Wild6D [7], OmniNOCS [19], and Omni6DPose [61] only provide pose annotations, segmentation, and depth, but *lack category-level 3D correspondences*. To address this gap, we introduce **HouseCorr3D**, a large-scale benchmark for monocular category-level 3D correspondence in camera space. HouseCorr3D covers 50 everyday object categories with 178k images and 280 unique object instances, each annotated with semantic 3D keypoints directly on CAD models that project consistently across all views. Crucially, our annotations include *amodal correspondences*—correspondences for object parts that are occluded or not visible in the image. This capability is inspired by human reasoning [57], where we naturally infer the complete 3D structure of objects even under occlusion, and is essential for robotic manipulation where planning grasps and interactions requires understanding the full spatial extent of objects [53], not just visible surfaces. We also explicitly annotate object symmetries, ensuring symmetric

[*] Equal contribution

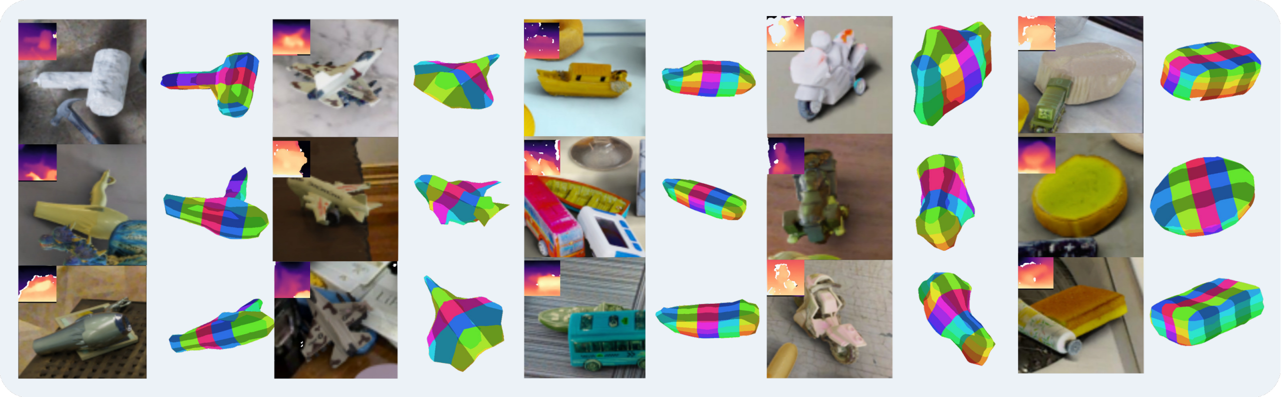


Figure 1. **Monocular Category-level 3D Correspondence.** We predict semantically consistent 3D keypoint locations across different instances of the same category from single RGB-D images. Our morphable priors enable establishing correspondences (shown with matching colors) that remain semantically aligned despite large shape variations, enabling fine-grained object understanding beyond traditional pose estimation and 2D semantic correspondence.

objects have multiple valid correspondences and avoiding unfair penalization of symmetry-equivalent predictions. Together, these properties address fundamental limitations of pose-focused datasets and, for the first time, enable quantitative evaluation of category-level 3D correspondence from single images.

Building on our benchmark, we propose **Morpheus**, a framework that learns *morphable category-level shape priors* to establish consistent 3D correspondences across instances directly in camera space. Instead of relying on a fixed canonical representation, Morpheus learns a deformable 3D template for each category that adapts to instance-specific shape variations while preserving correspondences. During training, our method jointly optimizes a 3D morphable prior, instance-specific shape deformations, and their 2D projection consistency. At inference, given a single RGB-D image, Morpheus predicts both the object's 3D shape in camera space and its semantically aligned keypoints, enabling direct correspondence evaluation without pose normalization.

In summary, our contributions are as follows:

(i) We identify **monocular category-level 3D correspondence in camera space** as a key next step beyond pose-centric representations toward semantically aligned 3D understanding.

(ii) We introduce **HouseCorr3D**, the first large-scale benchmark for category-level 3D correspondence, comprising 178k images across 50 household categories and 280 instances, with mesh-based keypoint annotations, amodal correspondences, and explicit symmetry labels.

(iii) We propose **Morpheus**, a framework that learns *morphable category-level shape priors* to establish semantically consistent 3D correspondences directly in camera space.

(iv) We demonstrate that Morpheus substantially outperforms existing baselines on HouseCorr3D, establishing a new paradigm for *fine-grained, correspondence-level 3D object understanding*.

2. Related work

2D Semantic Correspondence. 2D correspondence has advanced from local descriptors and dense flows (*e.g.*, SIFT [25], DAISY [42], SIFT Flow [22], DeepFlow [47]) to transformer-based self-supervised features [3, 32, 60, 63], which exhibit emergent semantic alignment and achieve strong results on benchmarks like SPair-71K, PF-PASCAL, and TSS [13, 21, 27]. Dedicated matchers such as LoFTR, COTR, DiffMatch [16, 29, 37], and spherical-map approaches [6, 26] further improve dense matching. While highly effective, these approaches remain limited to the image domain and do not predict 3D canonical coordinates or enforce semantic consistency across instances in 3D space.

3D Keypoint and Correspondence Methods. Prior work explored correspondence mapping in the 3D domain through keypoint detection and surface mapping. KeypointNet [58] introduced a large-scale dataset for learning category-consistent 3D keypoints, while others [15, 56] leverage keypoints for cage-based deformations and shape control. Canonical surface mapping [20] establishes correspondences by predicting UV coordinates on canonical templates, and Mesh R-CNN [8] jointly predicts mesh reconstructions with instance segmentation from 2D images. Recent semantic alignment methods [24, 44, 59] explore learning consistent correspondences across categories and human poses in 3D. DenseMatcher [64] extends matching to the mesh domain via functional maps, projecting multiview features onto 3D geometry. However, these approaches have fundamental limitations: Keypoint-

Net [58], Keypointdeformer [15], Yifan et al. [56], and DenseMatcher [64] require ground-truth 3D meshes as input; methods like [15, 24, 59] operate exclusively in 3D space without bridging to image-based features; and critically, none provide large-scale evaluation benchmarks with explicit handling of occlusion and symmetry. These limitations prevent their applicability to real-world scenarios where RGB(-D) images are predominantly available.

Morphable Models and Shape Priors. Morphable models achieve category-level understanding by capturing intra-class shape variability through deformable canonical templates. Classic work focused on faces and human bodies (*e.g.*, 3D Morphable Models [1], SMPL [23]), establishing the foundation for template-based shape modeling. Recent approaches [17, 30, 35, 36] extend these ideas to more diverse object classes using learned deformations or diffusion-guided generation. Deformation-based methods [11, 41, 46] map instances to template meshes using neural networks, while template-free approaches [31] learn canonical coordinate systems without relying on a single exemplar. More recent work leverages foundation models for semantic alignment across categories [30, 35], where semantically corresponding parts map to consistent representations. Domain-specific efforts have also addressed human bodies [12] and a range of animals [52]. Despite this progress, generalizing morphable models to diverse everyday objects with consistent 3D correspondences across instances remains an open challenge, especially for methods that operate only from image inputs.

Benchmarks for Category-Level 3D Understanding. To the best of our knowledge, there exists no dataset that enables category-level 3D correspondence evaluation from monocular images. Prior works [49] lift 2D images from domain-specific datasets [43, 48] to 3D using multi-view consistency but lack 3D evaluation benchmarks. Large-scale 3D shape collections such as ShapeNet [4] and ModelNet [50] provide CAD meshes, while ShapeNetPart [55] and PartNet [28] add part-level labels, but these lack consistent point-level correspondences across instances. Pose-focused datasets like Omni6DPose [61], CO3D [33], Pix3D [38], Pascal3D+ [51], and Omni3D [2] provide pose annotations in realistic scenes but do not supply semantic, amodal, or point-level correspondences across diverse instances. NOCS datasets [19, 45] introduced normalized coordinate spaces for pose estimation but are not designed for evaluating category-level correspondences. DenseCorr3D [64] takes a valuable step with part-level mesh annotations and functional-map evaluation, but operates exclusively in 3D with pre-reconstructed meshes. Thus, current 3D benchmarks do not bridge the gap between 2D-based and 3D correspondence methods.

In contrast, HouseCorr3D is explicitly designed for category-level 3D correspondence evaluation from monocular images, featuring 3D keypoints shared across all instances within 50 object categories, with amodal labels for occluded regions and explicit symmetry handling. This addresses a fundamental gap in current datasets and enables quantitative evaluation of correspondence-based 3D object understanding in camera space.

Table 1. **Comparison to existing correspondence datasets.** Prior benchmarks are limited in their evaluation to either 2D camera space or 3D object space. In contrast, **HouseCorr3D** focuses on 3D camera space across 50 classes. This allows us to evaluate the reasoning capabilities of monocular methods, including amodal correspondences without ambiguous object spaces.

Dataset	pairs	classes	input	eval. space
Pascal-Parts [5]	4k	20	2D	2D camera
PF-Pascal [13]	2k	20	2D	2D camera
Spair71k [27]	71k	18	2D	2D camera
KeyointNet [58]	N/A	16	3D	3D object
CorresPondenceNet [24]	N/A	25	3D	3D object
DenseCorr3D [64]	N/A	23	3D	3D object
HouseCorr3D	**178k**	**50**	**2.5D**	**3D camera**

3. Monocular Category-level 3D Correspondence Benchmark

Motivation. We introduce the first benchmark for category-level correspondences in 3D camera space. Unlike prior datasets that focus exclusively on correspondences in either 2D camera space [13, 27, 39, 43, 48] or 3D object space [64]. On the one hand, compared to reasoning in 3D object space, advancing monocular methods at estimating in 3D camera space, removes the need for ambiguous object-centric spaces, whereby neither the center nor the scale is well-defined. On the other hand, compared to estimation in 2D camera space the 3D camera space enables: a) the evaluation of amodal correspondences, b) modeling object symmetries explicitly, and c) enforcing methods to perform 3D over 2D reasoning.

Task definition. Given two RGB-D images I^q and I^t depicting objects from the same category, and a query 3D point $x^q \in \mathbb{R}^3$ in the camera space of I^q, the task is to predict the corresponding 3D point $x^t \in \mathbb{R}^3$ in the camera space of I^t that represents the same semantic part of the object. Formally, it can be expressed as a mapping $f : (x^q, \mathrm{I}^q, \mathrm{I}^t) \rightarrow x^t$. The evaluation is performed using the euclidean distance between the groundtruth target point x^t and the predicted target point $\hat{x}^t$, defined as

$$d(\hat{x}^t, x^t) = \left\|\hat{x}^t - x^t\right\|_2 . \tag{1}$$

The performance of a model is measured by computing the percentage of correctly predicted points within a given threshold on the euclidean distance (*e.g.*, PCK@0.1), using the largest of, width w, height h, and depth d of the object's

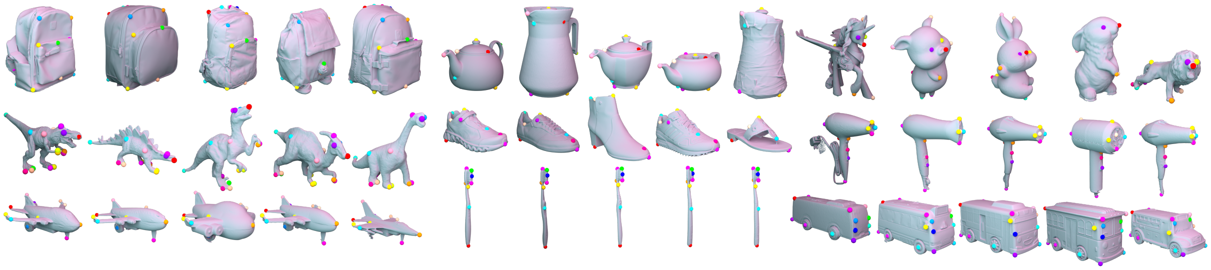


Figure 2. **Dataset Overview.** We annotate up to 19 3D keypoints directly on CAD meshes for 5–13 instances per category, covering 50 common household object classes. The keypoints are chosen to be semantically consistent and shared across all instances within each category We visualize a subset of these annotations across several categories to highlight their cross-instance and cross-shape consistency. Visualizations for the full dataset are provided in Sec. D.

3D bounding box, as: $d(\hat{x}^t, x^t) < 0.1 \cdot \max(h, w, d)$. This follows the conventions of other monocular 2D correspondence benchmarks [13, 27, 39, 43, 48], where the maximum width and height of the 2D bounding box are used to normalize the distance and compute PCK. Further discussion of correspondence evaluation, including the distinction between modal and amodal settings, is provided in Sec. G.

HouseCorr3D. We build our dataset on Omni6DPose [61], a large-scale synthetic dataset designed for category-level pose estimation in crowded scenes. We crop the images to obtain 178k test and 2.6M train images across 50 categories. We find 178k image pairs, by choosing a random image for each test image, which contains another instance. We specifically leverage Omni6DPose synthetic subset, which provides photorealistic renderings with high-quality CAD models of real object instances, natural lighting, cluttered scenes, and realistic occlusions. Unlike the real subset which contains limited instance diversity (typically 1–2 instances per category) and repetitive scene layouts due to video-frame extraction, the synthetic data provides greater scale and instance diversity, which is beneficial for learning robust category-level correspondences. We select 50 everyday object categories spanning household items (mugs, bottles, remotes), food items (fruits, vegetables), toys (cars, planes, animals), and accessories (backpacks, shoes, wallets), chosen to maximize shape diversity and practical relevance for robotic manipulation. For each category, between 2 and 16 semantic 3D keypoints are annotated directly on CAD meshes (see Fig. 2).

Keypoint Annotation Protocol. Keypoints must be shared across all instances of a category and are selected to be geometrically distinctive and semantically meaningful [40]—marking corners, edges, handle centers, or other salient structural features rather than arbitrary surface points. This ensures that annotations are both reliably localizable and transferable across instances. To ensure annotation quality and consistency, we employ a rigorous protocol (more details in Sec. D) involving two annotators[1] independently annotate the same set of meshes using an interactive 3D tool. Following this process, a two-stage merging process is applied including an initial *automatic merging* step which computes mutual nearest-neighbor matches between the two annotation sets across all instances based on distance (5%-threshold of object bounding-box diagonal) and consistency (pairs of keypoints are matched consistently), annotations are considered accepted or undecided. Then a second *manual merging* step is performed for undecided keypoints. Annotators use an interactive 3D viewer displaying multiple instances side-by-side to manually resolve ambiguities: accepting, rejecting, splitting, or merging annotations based on semantic and geometric consistency. The entire annotation process took approximately 65h across both annotators, yielding a total set of 2329 3D keypoint annotations on meshes by annotating between 2 and 19 keypoints per instance. Once keypoints are annotated on 3D meshes, we leverage ground-truth poses from Omni6DPose [61] to automatically project them into all rendered views, generating consistent 2D–3D correspondences across 178k pairs of images with minimal additional manual effort. This mesh-centric strategy offers three key advantages: (i) it enforces *semantic consistency* across all views and instances, (ii) it naturally provides *amodal* labels for occluded regions, and (iii) it efficiently scales a compact set of 3D annotations into a large-scale benchmark spanning 178k pairs across 50 categories and 280 instances. The resulting benchmark inherits the visual realism of Omni6DPose, featuring natural lighting, cluttered scenes, and partial occlusions.

Symmetry Handling. Many everyday objects exhibit geometric symmetries that introduce fundamental ambiguities in correspondence. For instance, a cylindrical mug body

[1] Annotators were trained on best practices for selecting geometrically distinctive and semantically meaningful keypoints that are reliably localizable and consistent across instances.

is rotationally symmetric—any point on the rim can rotate to any other without changing the object's shape. To the best of our knowledge, existing semantic correspondence benchmarks have not addressed symmetries, as they operate purely in 2D where such geometric constraints are difficult to define. By leveraging 3D annotations, HouseCorr3D explicitly handles *rotational* and *reflective* symmetries, ensuring that geometrically equivalent predictions are not unfairly penalized. Rotational symmetry is handled by treating all points on the orbit generated by rotations around the symmetry axis as valid correspondences, while reflective symmetry allows predictions to match either a keypoint or its mirrored counterpart. This yields a fair metric that respects the inherent geometric ambiguities in real-world objects and enables robust evaluation of category-level correspondence methods. Mathematical definitions and evaluation details are provided in Sec. G.

4. Method

Our goal is to predict the 3D shape of an object in camera space from a single RGB-D image, such that corresponding 3D keypoints across different object instances align consistently in the 3D camera coordinate frame. To achieve this, we introduce **Morpheus**, a framework that combines robust 6D pose diffusion with 3D morphable priors. We start by describing how to predict 3D correspondences in camera space in Sec. 4.1. Subsequently, we explain our architecture of morphable models in Sec. 4.2, and finally we elaborate on the objectives in Sec. 4.3.

Notation We denote a mesh as $\mathrm{M} = \{\mathrm{V}, \mathrm{E}\}$, with vertices $\mathrm{V} = \{\mathrm{v}_i \in \mathbb{R}^3\}_{i=1}^{|\mathrm{V}|}$ and edges $\mathrm{E} = \{(\mathrm{v}_i, \mathrm{v}_j)_e\}_{e=1}^{|\mathrm{E}|}$. For correspondence tasks, we distinguish query and target elements using superscripts $-^q$ and $-^t$ (*e.g.*, M^q and M^t). We denote a deformed mesh as M_{def} (defined in Sec. 4.2), and its transformation into camera space with pose π as $\mathrm{M}_{def}(\pi)$.

4.1. Mesh-based 3D Correspondence Prediction

Morpheus predicts for each RGB-D image a deformed template mesh M_{def} in canonical object-centric space. Using 6D pose diffusion [61], we can robustly estimate the 6D pose π for the query image and the target image, π^q and π^t respectively. Note, that this approach requires an object mask to sample 3D points from the object. Combined, we pose the deformed template mesh into the camera space as $\mathrm{M}_{def}(\pi)$. For each pair of images I^q and I^t, we first estimate their deformed meshes in 3D camera space, $\mathrm{M}^q_{def}(\pi^q)$ and $\mathrm{M}^t_{def}(\pi^t)$ respectively. Subsequently, we project each 3D query point x^q onto the surface of $\mathrm{M}^q_{def}(\pi^q)$, resulting in the surface point $\hat{x}^q$. Finally, we transform $\hat{x}^q$ to the target camera space using barycentric coordinates $\hat{x}^t$, illustrated in Fig. 3a.

4.2. 3D Morphable Priors

A central component of Morpheus is the *3D morphable prior*, a category-level canonical shape space that enables semantically consistent correspondences across instances. It consists of a canonical mesh capturing the common structure of a category, along with a learned deformation model that adapts it to individual instances. We call it a *prior* because all predictions are constrained to be deformations of this canonical representation. Since each vertex of the template retains its identity across deformations, semantic correspondences are preserved by design: pixels from different instances that map to the same canonical vertex correspond to the same semantic part. However, this approach raises several challenges that need to be adressed: (*i*) how to learn an appropriate canonical mesh that captures category structure, (*ii*) how to parametrize and learn deformations in a flexible yet stable manner, and (*iii*) how to ensure that learned deformations preserve semantic correspondence rather than collapsing to arbitrary mappings. We address these challenges through our hybrid volumetric mesh representation [34], vertex-wise affine deformation fields, and targeted regularization, as described below.

Canonical Shape Representation. Traditional mesh-only representations are often fragile and difficult to optimize directly, typically requiring manual interventions such as remeshing [9, 54]. To overcome this limitation, we employ a *hybrid volumetric mesh representation* [34]. This integrates the strengths of implicit and explicit 3D models. Concretely, the category-level shape is represented as a signed distance field ϕ_{sdf}, which provides the flexibility to model intricate geometries. Through Differentiable Marching Tetrahedra [34], the SDF is efficiently transformed into a mesh in a differentiable manner by evaluating SDF values on a tetrahedral grid. This formulation enables the use of mesh-based priors and regularizations, such as enforcing rigidity constraints during deformation learning.

Instance-Specific Deformations. To adapt this canonical mesh to specific instances, we apply an affine deformation field, following [62]. Unlike Zheng et al. [62], where deformations are applied directly to the signed distance field, we act on the template mesh vertices [49]. This avoids repeatedly extracting meshes for each instance and is thus more computationally efficient. Formally, we define an affine mapping $\phi_a : \mathbb{R}^3 \times \mathrm{L} \rightarrow \mathbb{R}^3$, which displaces each vertex $\boldsymbol{v}$ of the canonical mesh according to the instance-specific latent code l:

$$\phi_a(\boldsymbol{v}, \mathrm{l}) = \alpha(\boldsymbol{v}, \mathrm{l}) \odot \boldsymbol{v} + \delta(\boldsymbol{v}, \mathrm{l}), \tag{2}$$

where α, $\delta : \mathbb{R}^3 \times \mathrm{L} \rightarrow \mathbb{R}^3$ are produced by an MLP that takes both the vertex coordinate $\boldsymbol{v}$ and the latent code l as input. The latent code $\mathrm{l} = \psi_{\mathrm{l}}(\mathrm{I})$ itself is computed from the input image I by a deformation encoder ψ_{l} built from a DINOv2 backbone with a light convolutional head. This

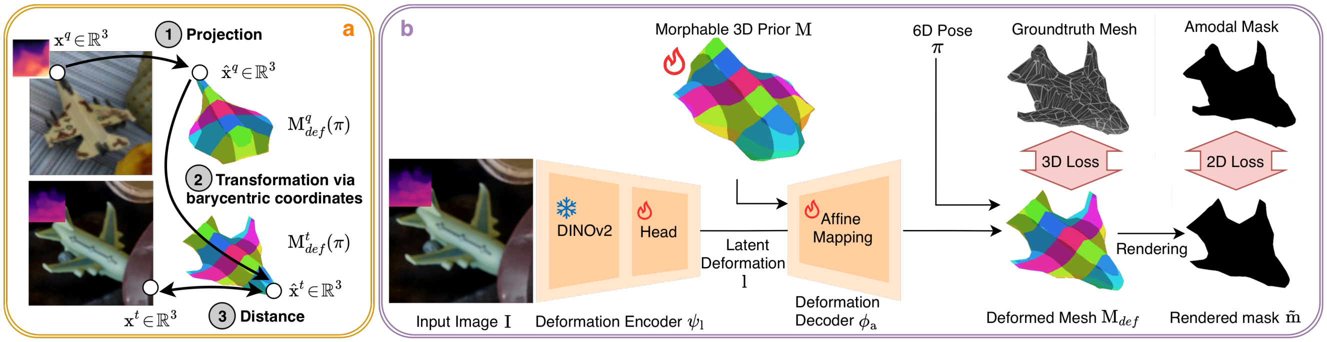


Figure 3. **(a) Monocular Category-Level 3D Correspondence.** Given a query point $x^q \in \mathbb{R}^3$ in one camera space and a target point $x^t \in \mathbb{R}^3$ in another camera space. The task is to predict the correspondence between these points. In our approach, we first project the query point onto the mesh M^q_{def} in the query camera space. Second we transform to the mesh M^t_{def} in the target camera space via barycentric coordinates. **(b) Pipeline Overview.** Given an RGB-D image, the deformation encoder ψ_l extracts an instance-specific latent deformation code l. This code drives a Deformation Decoder ϕ_a that morphs the learned 3D prior to fit the observed instance. The deformed prior is transformed to camera space using the 6D pose diffusion. During training, we supervise with amodal 2D and 3D objectives, to handle occlusions. Further, we provide 6D pose supervision to mitigate local minimas. Moreover, we regularize the deformation to encourage semantically consistent deformations.

code parametrizes vertex-wise displacements, enabling the mesh to morph into the observed instance while preserving semantic alignment. The resulting instance-adapted mesh is written as $M_{def}(I) = \{V_{def}(I), E\}$, where each deformed vertex is given by $V_{def}(I) = \{\phi_a(v_i, \psi_l(I))\}_{i=1}^{|V|}$. For simplicity, we simply rewrite it as $M_{def} = \phi_a(M, l)$. Through the deformation, vertices maintain consistent identities, enabling category-level correspondence prediction without requiring explicit supervision on keypoint locations.

4.3. Training Objectives

Morpheus is trained on object-centric images. We jointly optimize the template and deformation model using geometric objectives: 2D mask-based reconstruction, 3D mesh-based reconstruction, and mesh regularization. These reconstruction terms enforce consistency between rendered projections of the canonical mesh and the segmentation masks. Mesh regularization promotes rigidity in instance-specific deformations and hence maintains plausible canonicalization. Together, these objectives encourage the model to reconstruct category-consistent canonical meshes while preserving instance-specific details. In contrast to previous works [49], we provide 6D pose supervision to mitigate local minima. Moreover, we provide 2D amodal and 3D supervision to remain robust to occlusions.

2D Loss. We first supervise using amodal object masks. Given the predicted mask $\tilde{m}(M_{def}, I, \pi)$ rendered from the deformed mesh M_{def} under pose π, we compare against the ground-truth amodal mask m with a pixel-wise mean squared error:

$$\mathcal{L}_{m}(M_{def}, I, \pi, m) = \left\|\tilde{m}(M_{def}, I, \pi) - m\right\|^2. \quad (3)$$

Additionally, we encourage overlap with the distance transform m_{dt} of the ground-truth amodal mask:

$$\mathcal{L}_{mdt}(M_{def}, I, \pi, m_{dt}) = -\tilde{m}(M_{def}, I, \pi) m_{dt}, \quad (4)$$

with m_{dt} encoding the distance of each pixel inside the mask to the silhouette boundary, while pixels outside the mask are zero, which prevents disconnected parts from emerging when fitted across diverse instances.

3D Loss. To ensure accurate 3D instance reconstructions, we use a Chamfer distance between the deformed mesh vertices V_{def} and the ground-truth mesh vertices V_{gt}.

$$\mathcal{L}_{CD}(I, M_{def}, M_{gt}) = \frac{1}{|V_{def}|+|V_{gt}|} \left(\sum_{\boldsymbol{v}_i \in V_{def}} \|\boldsymbol{v}_i - \boldsymbol{v}'_{\chi(\boldsymbol{v}_i)}\| + \sum_{\boldsymbol{v}'_i \in V_{gt}} \|\boldsymbol{v}'_i - \boldsymbol{v}_{\chi(\boldsymbol{v}'_i)}\| \right), \quad (5)$$

where χ denotes the nearest neighbor operator.

Template and Deformation Regularization. Following [10], we enforce the SDF property with the Eikonal loss $\mathcal{L}_{sdf}$, penalize large deformation with an ℓ_2 term: $\mathcal{L}_{def}$, and encourage smoothness with an edge-based regularization $\mathcal{L}_{smooth}$ [62]. Their definitions are provided in Sec. E.

Full Training Loss. Our optimization proceeds in two stages. First, we refine the category-level template using only geometric terms:

$$\mathcal{L}_{geo} = \lambda_{CD}\mathcal{L}_{CD} + \lambda_{m}\mathcal{L}_{m} + \lambda_{mdt}\mathcal{L}_{mdt} + \lambda_{sdf}\mathcal{L}_{sdf}. \quad (6)$$

After convergence, we learn the instance deformations with the extended regularized loss:

$$\mathcal{L}_{geo\text{-}reg} = \mathcal{L}_{geo} + \lambda_{def}\mathcal{L}_{def} + \lambda_{smooth}\mathcal{L}_{smooth}. \quad (7)$$

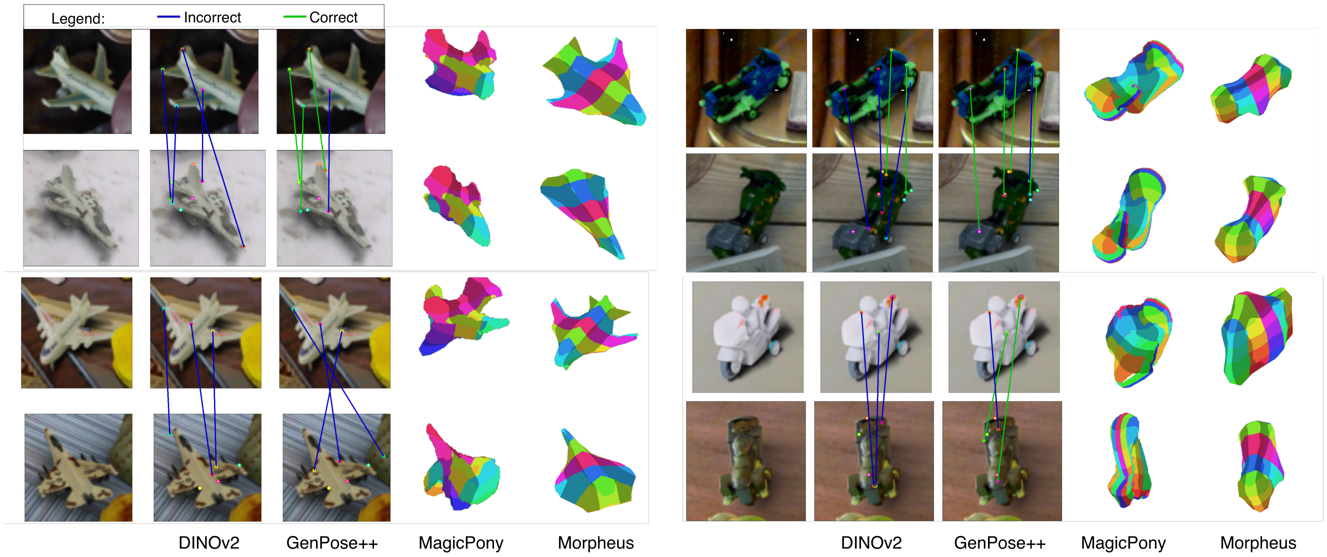


Figure 4. **Qualitative results.** We compare 2D feature matching method DINOv2, with 3D space matching methods GenPose++, MagicPony, and Morpheus. For DINOv2 and GenPose++ we visualize the 2D modal correspondences. For MagicPony and Morpheus, we visualize the predicted deformed meshes in camera space with a color scheme encoding semantic correspondences across instances (see Sec. F). Our occlusion-aware training allows Morpheus to be robust in case of occlusions in contrast to MagicPony, which is only trained on 2D supervision. Further, pose-aware training results in higher consistency for semantic parts across different viewpoints. Note that DINOv2 often confuses parts, and GenPose++ may predict points outside the object due to its lack of shape-deformation modeling.

5. Experiments

We evaluate Morpheus on the proposed HouseCorr3D benchmark, focusing on its ability to recover *category-level 3D correspondences*. We compare Morpheus with strong 2D correspondence baselines such as NOCS and DINOv2, as well as 3D space matching methods such as MagicPony and GenPose++. We first provide experimental details in Sec. 5.1, then elaborate on the evaluation metrics in Sec. 5.2. We describe all baselines in Sec. 5.3 and finally compare with prior work in Sec. 5.4.

5.1. Experimental Details

Morpheus uses a pretrained ViT-S DINOv2 image encoder [32] as backbone, and a pretrained 6D pose diffusion network [61]. From an input resolution of 448^2, the backbone maps to a 32^2 feature map. The deformation encoder is implemented as a ResNet head [14] that aggregates multi-scale feature maps with bottleneck blocks to produce refined latent deformation $\mathbb{l}$. The deformation decoder is a coordinate-conditioned MLP that fuses 3D point embeddings with latent deformation to predict deformations. To learn the initial template shape, we train each category-specific morphable model using the Adam optimizer [18] with a learning rate of 10^{-4} and a batch size of 30. Training proceeds in two stages: (i) 20 epochs optimizing the loss in Eq. (6), and (ii) 10 further epochs optimizing the extended loss in Eq. (7), which includes deformation regularizers. Training on a NVIDIA RTX 2080 takes about 12h.

5.2. 2D and 3D Metrics

For our benchmark, we use the percentage of correct keypoints (*i.e.*, PCK@0.1) as described in Sec. 3. We differentiate between 2D evaluation, where the distance is measured in pixel space and the threshold depends on the 2D bounding box, and 3D evaluation, where the distance is measured in camera space and the threshold depends on the 3D bounding box. In 3D, we further distinguish between modal correspondences (where the keypoint is visible in both images) and amodal correspondences (where one keypoint is occluded). Ambiguities due to object symmetries can lead to multiple valid correspondences, which we handle separately. We provide more details in Sec. G.

5.3. Baselines

Given our newly defined task, we made every effort to identify competitive baselines capable of processing RGB-D input data and producing predictions in both 2D and 3D domains. We first compare against 2D feature-matching baselines such as NOCS [45] and DINOv2 [32], where each pixel is represented by a feature vector in $\mathbb{R}^d$ and matched to its nearest neighbor in the target image. In this context, predicted NOCS coordinates are treated as features in $\mathbb{R}^3$. For MagicPony, we render its canonical-space coordinates and use the rendered results as a 2D feature-matching baseline (denoted as MagicPony$_{2D}$). Using the target image's depth map, the predicted 2D pixels can be reprojected into 3D, enabling 3D *modal* correspondences. However, since

Table 2. **PCK@0.1 results** for 2D, 3D modal, and 3D amodal correspondences on a subset of HouseCorr3D. Morpheus outperforms all 2D correspondence methods (DINOv2, NOCS, MagicPony_{2D}) and 3D methods (GenPose++ (GP), MagicPony). Note that MagicPony relies only on 2D supervision and is unaware of real-world scale, thus we use GenPose++ for translation and scale.

Method	🎒	✈	🚢	🚌	👕	🏍	mean (50)
2D							
DINOv2	7.0	15.2	17.1	13.3	14.0	10.6	22.9
MagicPony_{2D}	6.4	7.7	8.8	7.2	22.9	9.1	15.7
NOCS	27.2	20.7	14.0	42.6	23.7	16.6	26.7
GenPose++ (GP)	37.0	28.8	20.5	50.2	30.0	26.7	36.3
MagicPony+GP	4.8	7.2	4.1	4.2	22.1	8.1	10.7
Morpheus w/o Def.	39.9	32.0	22.5	51.8	34.2	29.5	39.1
Morpheus	**40.9**	**34.8**	**28.1**	**57.1**	**36.5**	**31.3**	**41.2**
3D (Modal + Amodal)							
GenPose++ (GP)	18.8	18.2	14.4	37.1	27.5	17.9	34.3
MagicPony+GP	1.2	2.1	0.9	0.9	10.8	2.2	7.1
Morpheus w/o Def.	22.7	21.6	16.5	41.3	35.0	19.8	38.4
Morpheus	**23.7**	**26.0**	**21.0**	**47.9**	**39.2**	**22.5**	**41.5**
3D Modal							
DINOv2+D	5.7	9.2	7.5	14.4	16.4	11.0	24.4
MagicPony_{2D}+D	3.9	3.1	2.7	4.7	22.7	10.1	14.0
NOCS+D	6.5	13.5	4.5	34.6	24.0	7.4	26.4
GenPose++ (GP)	22.9	14.9	12.9	38.5	27.9	27.5	37.0
MagicPony+GP	2.5	2.1	1.1	0.3	14.7	4.1	7.5
Morpheus w/o Def.	25.2	16.8	17.2	44.5	35.2	27.8	40.2
Morpheus	**26.0**	**23.6**	**19.9**	**49.2**	**38.8**	**33.8**	**43.7**
3D Amodal							
GenPose++ (GP)	17.1	19.2	14.8	36.7	27.3	15.0	32.9
MagicPony+GP	0.7	2.1	0.9	1.1	9.1	1.6	7.1
Morpheus w/o Def.	21.6	23.1	16.3	40.3	34.9	17.3	37.8
Morpheus	**22.8**	**26.7**	**21.3**	**47.5**	**39.4**	**19.0**	**40.8**

occluded regions are not visible, the 3D *amodal* correspondence task cannot be solved using any 2D baseline.

On the other hand, we compare with 3D space matching baselines such as GenPose++ and MagicPony. MagicPony also uses a 3D morphable prior; thus, the template mesh can be used to match points in 3D as explained in Sec. 4.1. In contrast, GenPose++ does not predict a 3D shape, but only a 6D pose. However, we can transform the query points from camera space into the normalized object-centric space using the inverse query camera pose, and further to the target camera space using the target camera pose. As MagicPony is unaware of the real scale, we apply the translation and object scale from GenPose++ for 3D space matching. The rotation from GenPose++ is not applicable to MagicPony, as it has an arbitrary canonical template pose.

5.4. Comparison with Prior Work

Overall, Tab. 2 shows that Morpheus sets a new state-of-the-art result on all 2D and 3D correspondence metrics. Fig. 4 illustrates qualitative predictions of Morpheus.

Occlusions. 2D feature matching methods such as NOCS or DINOv2 cannot handle occlusions by design, and cannot evaluate them in the 3D amodal setting. In Fig. 4, we also observe how DINOv2 confuses the back of the airplane with the front of another one, which is occluded in the query image. Furthermore, we observe qualitatively that MagicPony suffers from lack of occlusion-aware training, such that the query airplane is cut off in the estimated mesh. In contrast, Morpheus successfully reconstructs occluded parts. We can see that the PCK@0.1 drops from 43.7% for modal 3D correspondences to 40.8% for amodal correspondences.

Normalized Object Space. Finding correspondences using a normalized object space alone is insufficient. We can see this from the NOCS baseline, which Morpheus outperforms for both 2D and 3D modal correspondences, and from the fact that Morpheus improves over GenPose++, which uses the NOCS space to match query to target points. Qualitatively, in Fig. 4, we observe how GenPose++ incorrectly matches the right wing of an airplane to a location outside the target airplane's smaller wing range. Similarly, the right handle of the motorcycle is matched inside the target motorcycle, even though it should lie further outward.

MagicPony. Beyond occlusions, MagicPony struggles to recover consistent rotations across images, essential for reliable correspondences. Qualitatively, we observe that objects (e.g., airplanes) barely rotate; instead, the model compensates through deformation to fit the image, leading to incorrect correspondences. In Tab. 2, MagicPony_{2D} outperforms the 3D space matching variant, likely due to its 2D-only supervision and noisy pose predictions. Morpheus outperforms MagicPony through pose- and occlusion-aware training, achieving better shape prediction and higher consistency across viewpoints. We encourage the community to build upon MagicPony's impressive approach to learning 3D morphable models with only 2D supervision.

SPair71k. Thanks to its broader category diversity, our benchmark is more challenging than SPair71k [27], as seen in the performance gap: DINOv2 achieves 52.7% on SPair71k but drops to only 22.9% on HouseCorr3D.

6. Conclusion

This paper introduces a paradigm shift from correspondence evaluation in 2D camera space or 3D object space toward *monocular category-level 3D correspondences in camera space*. **HouseCorr3D** provides 50 everyday categories in crowded scenes with mesh-based annotations, establishing a solid foundation for comparing monocular 3D correspondence methods with explicit handling of symmetries, occlusions, and challenging amodal correspondences. We demonstrate that solving this task requires moving beyond 2D feature matching. **Morpheus** leverages morphable priors to achieve state-of-the-art performance through pose- and occlusion-aware supervision, successfully morphing objects while maintaining consistent correspondences across instances with varying shapes and poses. We also show that approaches relying only on 2D supervision remain insufficient. Our benchmark provides a foundation for expanding correspondence learning toward embodied robotics applications, where reasoning about full 3D object geometry, including occluded parts, is essential.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 3

[2] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3D: A large benchmark and model for 3D object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, 2023. IEEE. 3

[3] 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. 2

[4] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 3

[5] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1971–1978, 2014. 3

[6] 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. 2

[7] Yang Fu and Xiaolong Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1

[8] Georgia Gkioxari, Justin Johnson, and Jitendra Malik. Mesh r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9784–9794, 2019. 2

[9] Shubham Goel, Georgia Gkioxari, and Jitendra Malik. Differentiable stereopsis: Meshes from multiple views using differentiable rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8635–8644, 2022. 5

[10] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 6, 15

[11] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. 3d-coded : 3d correspondences by deep deformation. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[12] 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)*, pages 7297–7306, 2018. 3

[13] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2016. 2, 3, 4

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 7, 13

[15] Tomas Jakab, Richard Tucker, Ameesh Makadia, Jiajun Wu, Noah Snavely, and Angjoo Kanazawa. Keypointdeformer: Unsupervised 3d keypoint discovery for shape control. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12778–12787, 2021. 2, 3

[16] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. COTR: Correspondence Transformer for Matching Across Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[17] Hyunwoo Kim, Itai Lang, Noam Aigerman, Thibault Groueix, Vladimir G. Kim, and Rana Hanocka. Meshup: Multi-target mesh deformation via blended score distillation. In *International Conference on 3D Vision (3DV)*, 2025. 3, 16

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 7

[19] Akshay Krishnan, Abhijit Kundu, Kevis-Kokitsi Maninis, James Hays, and Matthew Brown. Omninocs: A unified nocs dataset and model for 3d lifting of 2d objects. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 3

[20] 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)*, pages 2202–2211, 2019. 2

[21] Xinghui Li, Kai Han, Xingchen Wan, and Victor Adrian Prisacariu. Simsc: A simple framework for semantic correspondence with temperature learning. *arXiv preprint arXiv:2305.02385*, 2023. 2

[22] 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. 2

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 3

[24] Yujing Lou, Yang You, Chengkun Li, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Weiming Wang, and Cewu Lu. Human correspondence consensus for 3d object semantic understanding. In *European Conference on Computer Vision (ECCV)*, page 496–512, Berlin, Heidelberg, 2020. Springer-Verlag. 2, 3

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

[26] 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)*, pages 19521–19530, 2024. 1, 2

[27] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence, 2019. 1, 2, 3, 4, 8

[28] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[29] Jisu Nam, Gyuseong Lee, Sunwoo Kim, Hyeonsu Kim, Hyoungwon Cho, Seyeon Kim, and Seungryong Kim. Diffusion model for dense matching. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024. 1, 2

[30] 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. 3, 16

[31] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[32] 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. 2, 7

[33] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[34] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Conference on Neural Information Processing Systems (NeurIPS)*, 34: 6087–6101, 2021. 5

[35] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. Shic: Shape-image correspondences with no keypoint supervision. In *European Conference on Computer Vision (ECCV)*, 2024. 3, 16, 17

[36] 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)*, pages 6468–6479, 2025. 3, 12

[37] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[38] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[39] Yixuan Sun, Yiwen Huang, Haijing Guo, Yuzhou Zhao, Runmin Wu, Yizhou Yu, Weifeng Ge, and Wenqiang Zhang. Misc210k: A large-scale dataset for multi-instance semantic correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7121–7130, 2023. 3, 4

[40] Supasorn Suwajanakorn, Noah Snavely, Jonathan Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 2063–2074, 2018. 4

[41] Meng Tian, Marcelo H. Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision (ECCV)*, page 530–546, Berlin, Heidelberg, 2020. Springer-Verlag. 3

[42] Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32:815–30, 2010. 2

[43] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. Cub-200-2011, 2022. 3, 4

[44] 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)*, pages 1138–1147, 2025. 2

[45] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651, 2019. 1, 3, 7, 13, 16

[46] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[47] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow: Large displacement optical flow with deep matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1385–1392, 2013. 2

[48] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. DOVE: Learning deformable 3d objects by

watching videos. *International Journal of Computer Vision (IJCV)*, 2023. 3, 4

[49] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. MagicPony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 5, 6, 12, 13

[50] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[51] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 75–82. IEEE, 2014. 3

[52] 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. 3

[53] Zhenjia Xu, Zhanpeng He, Jiajun Wu, and Shuran Song. Learning 3d dynamic scene representations for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2020. 1

[54] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15980–15989, 2021. 5

[55] 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 *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2016. 3

[56] Wang Yifan, Noam Aigerman, Vladimir G. Kim, Siddhartha Chaudhuri, and Olga Sorkine-Hornung. Neural cages for detail-preserving 3d deformations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 72–80, 2020. 2, 3

[57] Ilker Yildirim, Max H Siegel, Amir A Soltani, Shraman Ray Chaudhuri, and Joshua B Tenenbaum. Perception of 3D shape integrates intuitive physics and analysis-by-synthesis. *Nature Human Behaviour*, 8(2):320–335, 2024. 1

[58] Yang You, Yujing Lou, Chengkun Li, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Cewu Lu, and Weiming Wang. Keypointnet: A large-scale 3d keypoint dataset aggregated from numerous human annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13644–13653, 2020. 2, 3

[59] Yang You, Chengkun Li, Yujing Lou, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Weiming Wang, and Cewu Lu. Understanding pixel-level 2d image semantics with 3d keypoint knowledge engine. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 44(9):5780–5795, 2022. 2, 3

[60] 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. 2, 13

[61] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: Large-scale multi-object 6d pose estimation with realistic rendering. In *European Conference on Computer Vision (ECCV)*, pages 3110–3120, 2024. 1, 3, 4, 5, 7, 12, 13, 14

[62] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3d shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1429–1439, 2021. 5, 6, 15

[63] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022. 2

[64] 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. 2, 3, 16

Monocular Category-Level 3D Correspondence via Morphable Priors

Supplementary Material

A. Additional results

In addition to the results reported in the main paper, we provide in Tab. A3 the complete set of quantitative results for our method, covering all categories of HouseCorr3D. These extended results complement the main text by offering a more fine-grained view of per-class performance. Importantly, we observe the same overall trends as in the main paper. This consistency arises because the categories highlighted in the main figures were chosen at random, rather than being selected to favor particular outcomes. Thus, the additional results confirm that our observations hold uniformly across the entire benchmark and are not biased by the choice of examples shown in the main paper.

Despite the overall robustness of our method, some limitations can be observed in challenging scenarios. A first source of error arises from inaccurate pose estimation from [61]. Since canonical alignment is a prerequisite for predicting consistent correspondences, pose misalignment can propagate through the pipeline and lead to incorrect predictions. A second limitation concerns the deformation decoder. The learned deformations are constrained by both the template representation and the distribution of training data. As a result, objects that exhibit high intra-class variability, or that contain fine-scale structures not well captured in the template, often cannot be deformed adequately. This is especially evident for thin or elongated extremities such as airplane wings, bottle tips, or animal legs, where the predicted deformation either underestimates the required displacement or, in extremely rare cases, collapses the geometry entirely. Finally, the model may fail in cases where very large non-linear deformations are required. Since the decoder is trained to interpolate within the observed shape distribution, extrapolations to unseen structural variations remain difficult. Consequently, regions that extend far beyond the canonical template tend to remain under-deformed, leading to visible artifacts such as truncated parts or floating geometry. While these errors are relatively rare, they underscore the inherent trade-off between enforcing a shared canonical prior and maintaining sufficient flexibility to capture extreme shape variations across object instances. We also provide additional qualitative limitations in Fig. A1.

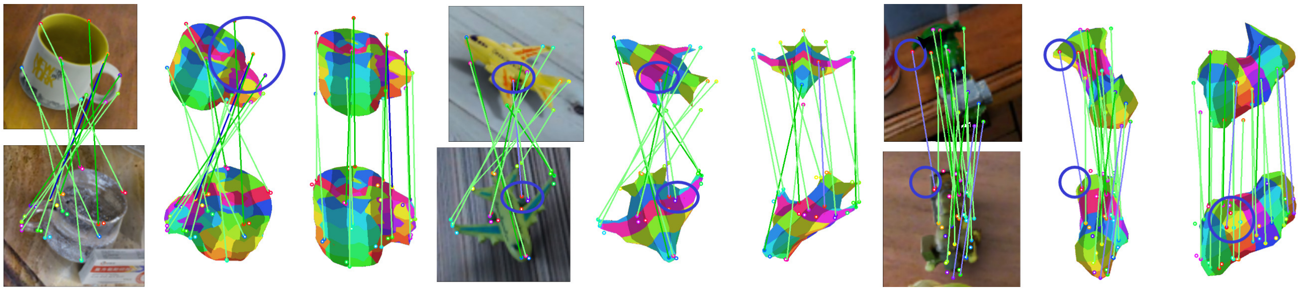


Figure A1. **Qualitative Results.** We illustrate some limitations qualitatively. In the first example, the pose estimation for the query object is slightly off, resulting in wrong projections on the estimated mesh. Second, coarse estimation of the mesh results in wrong correspondence. Third, wrong depth estimation, leads to wrong 3D correspondence estimation, despite the 2D projection is accurate.

B. Experimental details

B.1. Hyperparameters

Training Morpheus involves multiple components and multiple losses, so we draw inspiration from [36, 49, 61] for our hyperparameter settings. Table A1 summarizes the overall training setup, loss weights, and model architectures used across our experiments.

B.2. DINOv2

For the DINOv2 baseline, we use the ViT-S backbone initialized from the public weights. Images are resized to 448^2, yielding a 32^2 patch grid, and we L2-normalize the resulting feature map before computing correspondences.

Training Hyperparameters	
Optimizer	Adam
Batch Size	30
Batch Accumulation	2
Learning Rate	1.0×10^{-3}
Epsilon	1.0×10^{-8}
Beta 1	0.9
Beta 2	0.999
Weight Decay	0
Learning Rate Scheduler	Exponential LR
Warmup	100
Gamma	0.98
LR Min.	1.0×10^{-4}

Deformation Architecture	
Backbone	DINOv2 ViT-S
Deformation Encoder	ResNet Blocks
ResNet Blocks	4
ResNet Block Type	bottleneck
Out Dimensions	[256, 256, 256, 256]
Strides	[2, 2, 2, 2]
Pre-Upsampling	[1, 1, 1]
Deformation Decoder	Coordinate MLP
Layers	5
Hidden Dimension	256
Out Dimension	6

Loss Weights	
Mesh Chamfer Distance (λ_{CD})	0.1
Mask Mean Square Error (λ_{m})	2
Mask Distance Transform (λ_{mdt})	200
SDF Regularization (λ_{sdf})	0.01
Deformation Regularization (λ_{def})	0.075
Smoothness Regularization (λ_{smooth})	0.0075

Template Architecture	
Type	Coordinate MLP
Layers	5
Hidden Dimension	256
Out Dimension	1
DMTet Resolution	16

Table A1. Full-width hyperparameter overview including training setup, loss weights, and model architectures.

B.3. NOCS

We closely follow the procedure introduced by Wang et al. [45] to evaluate the NOCS baseline on HouseCorr3D. We use the same ResNet50 [14] backbone together with Feature Pyramid Network (FPN). For every training image we generate ground-truth NOCS targets by normalizing each object mesh to the unit cube and encoding the resulting XYZ coordinates directly as RGB values. Using the camera poses provided in Omni6DPose[61], we then render these NOCS maps so that every pixel stores its canonical 3D coordinate. Training uses the official ground-truth instance masks, category labels, and depth maps from Omni6DPose to supervise the model and to restrict supervision to the visible object regions. At inference time we predict a dense NOCS map for each input image. For 2D correspondence queries, we read the predicted canonical coordinate at the query pixel and find the nearest neighbor in NOCS space among all image pixels in the target image; the location of that neighbor serves as the correspondence prediction. For 3D correspondence queries, given a 3D query point x^q in the source image, we first find the corresponding canonical coordinate by projecting x^q into the source image and reading the predicted NOCS value at that pixel. We then find the nearest neighbor in NOCS space among all pixels in the target image; we back-project that pixel using the depth map to obtain the predicted 3D correspondence x^t.

B.4. MagicPony

Following MagicPony [49], we sample 5K images, extract object features using the provided modal masks, and apply PCA to reduce the feature dimension to 16. We replace the original DINOv1 encoder with DINOv2, which improves category-level 2D correspondence estimation [60]. Due to memory constraints, each category-level model is trained for 120 epochs with a grid resolution of 128, whereas the original implementation switches to resolution 256 for the final 30 epochs. Because our evaluation emphasizes correspondence accuracy within a 10% object-size tolerance rather than fine-grained reconstruction, sub-percent shifts (e.g., $< 0.5\%$) are negligible.

C. Additional dataset statistics

We rely exclusively on the realistic synthetic subset of Omni6DPose [61]. Preliminary experiments showed that the real captures provide limited diversity: most categories contain at most two unique object instances, scenes are often repeated across long video sequences, and overall variation in layout is low. As a result, the number of reliable correspondences that can be established from the real subset is severely restricted.

In contrast, the synthetic pipeline offers large-scale variation in both object instances and scene composition. This diversity is crucial for learning robust 2D–3D semantic correspondences across categories. Moreover, the synthetic subset has been designed to closely mimic real-world conditions, with natural lighting, cluttered environments, and realistic occlusions. This ensures that models trained on our benchmark generalize well beyond simplified synthetic settings. Therefore, our benchmark focuses on the high-quality synthetic subset, which provides both realism and sufficient coverage for large-scale correspondence evaluation. In total, HouseCorr3D contains 178k images across 280 unique object instances from 50 categories, making it the first large-scale dataset with dense, semantically consistent 2D–3D correspondences for everyday objects. To better illustrate the scope of the annotations, Tab. A4 reports the number of annotated keypoints for each category, highlighting differences in semantic coverage across classes. In Fig. A2, we further visualize the total number of keypoints annotated per class and indicate, through color coding, how many object instances were annotated. Together, these results offer a clear overview of the dataset's scale and diversity and underscore its suitability as a benchmark for category-level 3D correspondence.

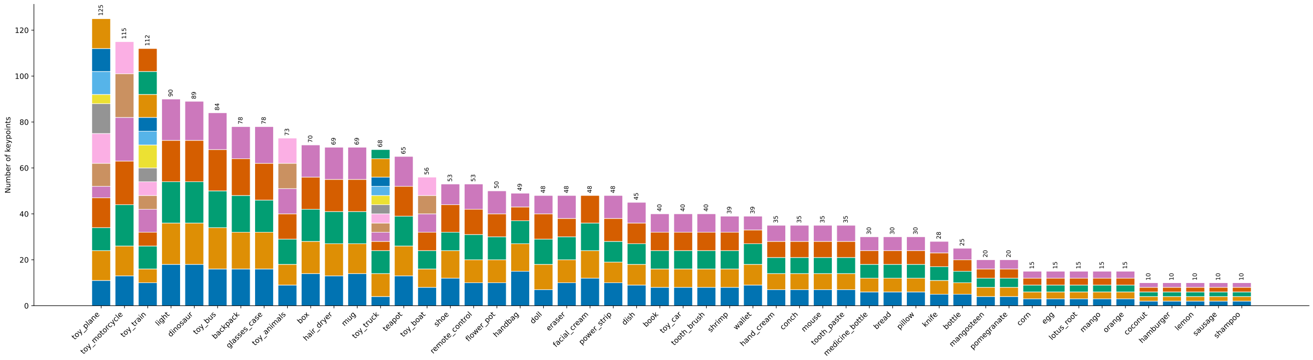


Figure A2. **Total number of annotated keypoints per class.** Different object instances are shown in different colors. *Note.* The number of keypoints per instance can vary within a class because instances often differ in shape and semantics. For example, two toy_plane instances have fewer keypoints because they are helicopters, and roughly half of the toy_train instances are high-speed bullet trains while the others are conventional locomotives.

D. Mesh annotation process

For mesh annotation, we convert each CAD mesh into a point cloud to facilitate visual inspection and interaction. Annotators are then provided with up to 20 3D keypoints per category that must be placed consistently across all instances. These keypoints are chosen to be semantically meaningful and geometrically well-defined: rather than marking the center of a continuous surface, annotators focus on distinctive structures such as corners, edges, wheel centers, handles, or wing tips. This strategy ensures that annotated points are both discriminative and reliably transferable across different instances of a category. To guarantee annotation quality, each instance was independently annotated by two annotators. The two annotation sets are then automatically merged using a correspondence-based algorithm. First, keypoints from both annotators are transformed to an object-centric coordinate frame and mutual nearest-neighbor correspondences are computed across all instances of a category. Matched keypoints are either classified as close (within 5% of the object's bounding-box diagonal) or distant. Based on matching patterns across instances, keypoints are automatically accepted (pairs are always matched and close, AUTO_ACCEPT), split into separate entries (pairs are always matched but distant, indicating semantic disagreement between annotators, AUTO_SPLIT), or kept as-is (never matched, AUTO_UNMATCHED). Ambiguous cases (*i.e.*, all remaining keypoints not falling in any previous categories), which includes keypoints with inconsistent matching behavior or mixed proximity patterns, are resolved through an interactive post-merging step, where both annotators visualize correspondences across multiple instances and manually and mutually decide whether to accept a single keypoint (*e.g.*, MANUAL_ACCEPT + SET1 & REJECT for the second keypoint), merge keypoints (*i.e.*, use the mean of both keypoints, MANUAL_ACCEPT + MEAN), split keypoints (*i.e.*, create separate keypoints and keep both when they refer to different semantic concepts, MANUAL_ACCEPT + SET1&SET2), or reject both keypoint (REJECT). We summarize the final merged status and manual decision distributions in Tab. A2. This systematic merging procedure reduces noise and ensures high-quality annotations consistent across the dataset. In addition, the reference mesh for each category was annotated first, and subsequent instances were aligned to this reference using a 3D interface. This alignment step further reduced ambiguities and ensured that annotations across different instances adhered to the same semantic standard. Overall, this process yields a compact yet semantically robust set of 3D keypoints that serve as the foundation for our correspondence benchmark.

Table A2. Final merged status (left) and manually accepted decision (right) distributions over all categories.

Status	Percentage
AUTO_ACCEPT	23.1%
AUTO_SPLIT	16.6%
AUTO_UNMATCHED	24.9%
MANUAL_ACCEPT	21.9%
REJECT	13.4%

Decision	Percentage
MEAN	48.2%
SET1	24.5%
SET2	27.3%

Note. Most rejected keypoints occur when only one annotator set (SET1 or SET2) is retained during manual validation. This happens when both sets target the same semantic zones, but one of them is judged to be of relative higher quality and the other is therefore discarded.

E. Additional Losses

Learning accurate correspondences requires not only supervision on visible matches but also strong geometric regularization to stabilize training and enforce plausible shapes. To this end, we use additional loss terms (Eqs. (A1) to (A3)) that impose additional constraints to the learned deformation and shape representation.

Eikonal loss. To enforce the signed distance function (SDF) property, we adopt the Eikonal regularizer [10], which encourages unit-norm gradients of the implicit function. Because gradients are only reliable near the extracted surface, we additionally sample auxiliary points $\mathcal{P}_{sdf}$ throughout the canonical space:

$$\mathcal{L}_{sdf}(\mathrm{M}, x) = \big(\|\nabla\phi_{sdf}(x)\|_2 - 1\big)^2, \quad x \in \mathcal{P}_{sdf}. \tag{A1}$$

This prevents degenerate fields and stabilizes the geometry across unseen regions.

Deformation regularizer. To avoid arbitrary or excessive deformations, we penalize ℓ_2 deviations of vertices from the category template:

$$\mathcal{L}_{def}(\mathrm{M}, \mathrm{M}_{def}, \mathrm{I}) = \frac{1}{|\mathrm{V}|}\sum_{\boldsymbol{v}\in\mathrm{V}} \big\|\boldsymbol{v} - \phi_a(\boldsymbol{v}, \mathrm{l})\big\|^2, \text{ with } \mathrm{l} = \psi_{\mathrm{l}}(\mathrm{I}) \tag{A2}$$

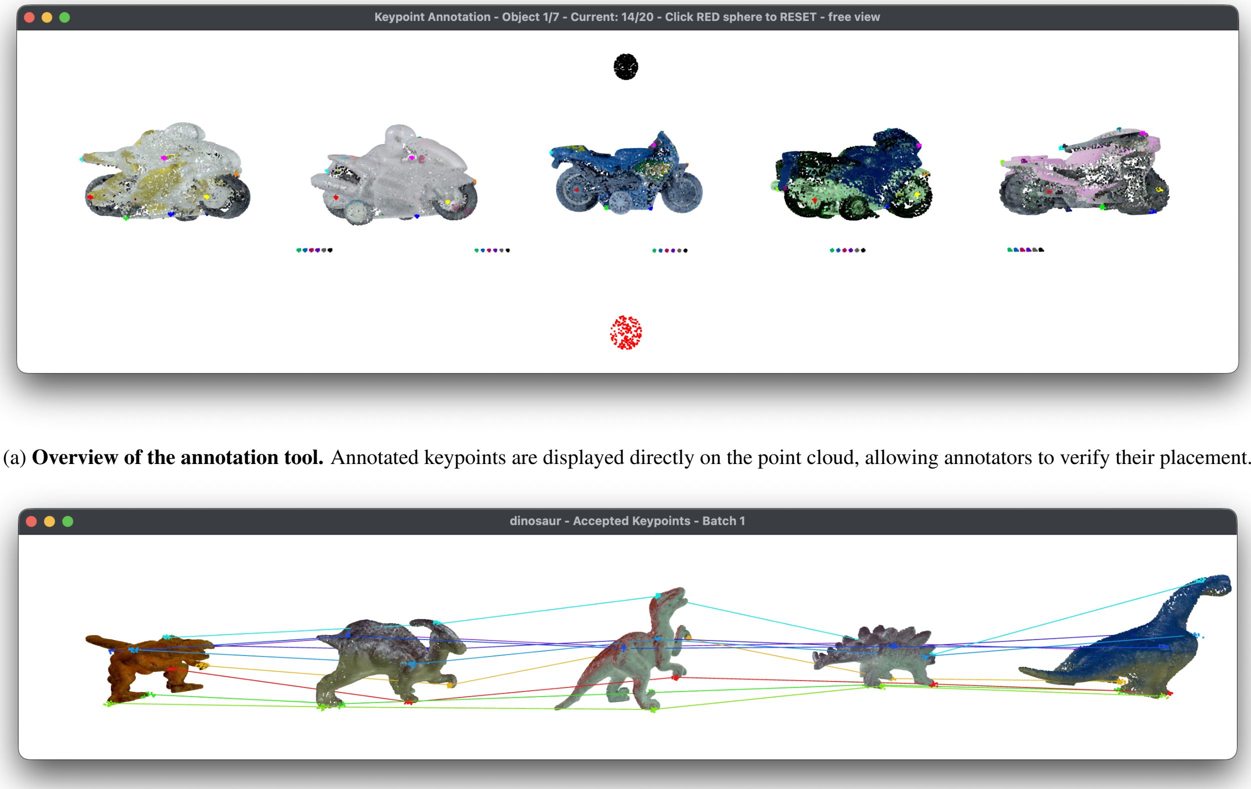
This term encourages learned shapes to remain close to the canonical prototype while still allowing instance-specific variation.

Smoothness regularizer. Finally, we promote locally coherent deformations by enforcing smooth displacements across neighboring vertices, following [62]:

$$\mathcal{L}_{smooth}(\mathrm{M}, \mathrm{M}_{def}, \mathrm{I}) = \frac{1}{|\mathrm{E}|}\sum_{\boldsymbol{i},\boldsymbol{j}\in\mathrm{E}} \frac{\big\|[\boldsymbol{i} - \phi_a(\boldsymbol{i}, \psi_{\mathrm{l}}(\mathrm{I}))] - [\boldsymbol{j} - \phi_a(\boldsymbol{j}, \psi_{\mathrm{l}}(\mathrm{I}))]\big\|_2}{\|\boldsymbol{i} - \boldsymbol{j}\|_2}. \tag{A3}$$

This regularizer suppresses spurious local distortions while still allowing non-rigid articulation.

Together, these terms ensure that the learned representation respects the SDF property, stays anchored to a canonical template, and maintains smooth, realistic deformations.


Keypoint Annotation - Object 1/7 - Current: 14/20 - Click RED sphere to RESET - free view

(a) **Overview of the annotation tool.** Annotated keypoints are displayed directly on the point cloud, allowing annotators to verify their placement.

dinosaur - Accepted Keypoints - Batch 1

(b) All annotated keypoints and their correspondences for the dinosaur category are visualized, enabling inspection of annotation quality.

Figure A3. **Annotation process illustration.** Using our interactive 3D interface, annotators align 5 instances per category and assign 3D keypoints to their respective meshes. We also visualize the resulting correspondences to assess their quality and consistency.

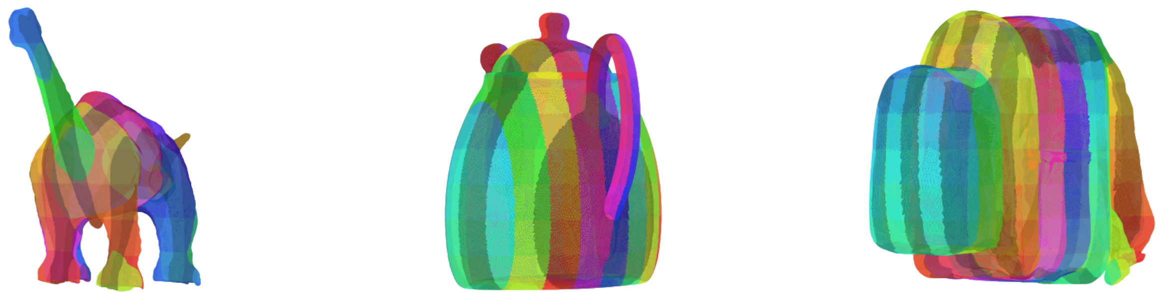


Figure A4. **HueGrid visualization.** We integrate 3D-based color encoding with a structured checkerboard pattern which allows to jointly highlight absolute correspondences and local deformations. We show the HueGrid projection for three example objects.

F. HueGrid Visualization

To visualize dense correspondences, we introduce the *HueGrid* representation. Classical 3D-aware coloring schemes such as NOCS [45] (widely adopted in [17, 30, 35, 64]) encode XYZ coordinates directly as RGB values, but this makes local distortions hard to perceive given the continuous nature of the color mapping. Conversely, Shtedritski et al. [35] texture

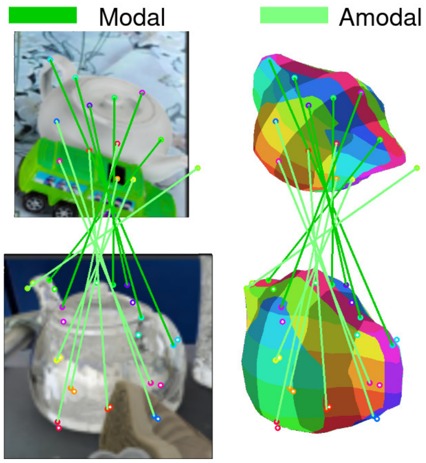


Figure A5. **Modal vs. Amodal Correspondences.** Choosing the 3D camera space as evaluation space, means we can also evaluate amodal correspondences. Here we show the three types of amodal correspondences in lightgreen, a) self-occlusion, b) occlusion from another object, and c) outside of the camera frustum. Not that it is sufficient if a point is occluded in either the query or the target space.

meshes with a colored checkerboard pattern, which clearly reveals local stretching because square cells deform into visible shapes once projected into the image.

HueGrid combines the best of both ideas: we keep the informative 3D-based color coding of NOCS while superimposing the structured checkerboard cues from Shtedritski et al. [35]. The resulting visualization simultaneously conveys absolute correspondence information and local geometric deformation. The visualization is illustrated in Fig. A4 for three representative mesh examples. We will also provide the code to generate HueGrid visualizations for all meshes and point clouds.

G. Discussion about correspondence evaluation

Modal vs. Amodal masks. We distinguish between *modal* and *amodal* correspondences in 3D, see Fig. A5. Modal correspondences are defined only on the subset of surface points that are visible from a given viewpoint, mapping observed 2D pixels to their canonical surface counterparts. In contrast, amodal correspondences extend this mapping to the full object surface, including parts that may be (self-)occluded. Modal evaluation reflects how well a method can align observed geometry with a canonical template and is directly comparable to tasks such as 2D keypoint transfer. Amodal evaluation goes further: it measures whether a model has learned a complete category-level shape prior that can predict correspondences even for unobserved surfaces. This distinction is critical for downstream tasks that require holistic understanding, such as shape completion, scene reasoning, or part-level manipulation. In 2D, we are restricted to image pixels, which by definition correspond only to visible regions; there is no ground-truth notion of a pixel for an occluded surface. In 3D, however, we can explicitly represent the canonical surface $\mathcal{C}$ and predict both visible and occluded points across poses. This makes it possible to evaluate amodal correspondences, providing a stronger test of a model's ability to infer complete, semantically consistent shapes across instances.

Evaluation under symmetry. Many everyday objects exhibit geometric symmetries that introduce fundamental ambiguities in correspondence. To the best of our knowledge, existing semantic correspondence benchmarks have not addressed symmetries, as they operate purely in 2D where such geometric constraints are difficult to define. By leveraging 3D annotations, HouseCorr3D explicitly handles *reflective* and *rotational* symmetries, ensuring that geometrically equivalent predictions are not unfairly penalized, see Fig. A6.

Reflective symmetry is invariance under mirror reflection across a plane. For each target point $\boldsymbol{x}$ we obtain multiple correct mirrored solutions $\boldsymbol{x}'$ and treat them equivalent by choosing the one that is closest to the predicted point $\hat{\boldsymbol{x}}$

$$e_{\text{refl-sym}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \min_{\boldsymbol{x}'} \{\|\boldsymbol{x}' - \hat{\boldsymbol{x}}\|\}.$$

Rotational symmetry is defined as invariance under rotations about a fixed axis. Given a predicted point $\hat{\boldsymbol{x}}$ and rotation $R_{\mathbf{a}}(\theta)$ about axis $\mathbf{a}$, the correspondence error is

$$e_{\text{rot-sym}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \min_{\theta} \|R_{\mathbf{a}}(\theta)\, \boldsymbol{x} - \hat{\boldsymbol{x}}\|,$$

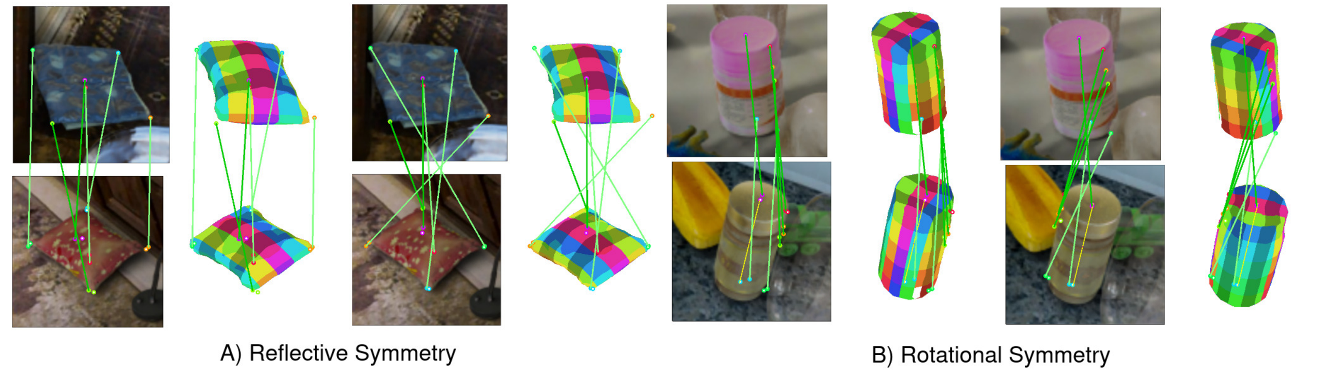


Figure A6. **Evaluation under symmetry.** Illustration of two correct correspondence estimations with explicit symmetries. First, in A), we show two possible predictions under reflective symmetry. Despite flipping the pillow, the correspondences are correct. Second, in B), we show two possible predictions under rotational symmetry. We visualize the rotation axis in yellow.

with $\theta \in [0, 2\pi)$. Geometrically, this equals the distance from $\hat{\boldsymbol{x}}$ to the circular orbit of $\boldsymbol{x}$ around the symmetry axis.

With these symmetry-aware definitions, predictions are correct if they align with any symmetric equivalent: rotationally symmetric points are judged by distance to their orbit, and mirror-symmetric points by distance to the closer counterpart. This yields a fair metric that respects the inherent geometric ambiguities in real-world objects and enables robust evaluation of category-level correspondence methods.

Reproducibility and LLM assistance

To ensure full reproducibility of our work, we will release all code and data used in this paper. The complete processing pipeline, including scripts for dataset preparation and annotation generation, will be made publicly available on GitHub. Our training and inference code for the proposed model will be provided in the same repository, together with configuration files and instructions for reproducing all experiments reported in the paper. The dataset itself, including the annotated 3D meshes and projected 2D keypoints, will be released on Hugging Face for easy access and long-term hosting. In addition, we will provide helper functions to compute the 3D correspondence metrics introduced in this paper, ensuring that results can be evaluated in a consistent and standardized manner.

We used large language models (LLMs) in a limited capacity to assist with the writing of this paper. Specifically, LLMs were employed only to (i) improve sentence clarity and conciseness, and (ii) condense overly lengthy paragraphs. All technical contributions — including the method design, experimental setup, results, and analyses — are entirely our own work.

Table A3. Category-level comparison across all 50 classes for Morpheus and the baselines. Beyond showcasing which categories are almost solved versus still challenging, the table reveals how object variability drives performance: classes with low deformation and consistent shapes (*e.g.*, *shampoo*, *corn*) are nearly saturated, whereas highly diverse toy categories (*e.g.*, *toy car*, *toy animal*) remain difficult.

	mean	backpack	book	bottle	box	bread	coconut	conch	corn	dinosaur	dish	doll	egg	eraser	facial cream	flower pot	glasses case
2D																	
DINOv2	22.9	7.0	14.9	25.9	41.1	14.0	30.9	15.8	30.8	16.4	8.6	5.0	12.7	64.4	11.7	9.5	44.2
MagicPony$_{2D}$	15.7	6.4	7.1	36.9	41.8	22.9	22.1	5.0	21.8	7.4	10.2	5.6	11.2	36.9	14.4	9.7	38.1
NOCS	26.7	27.2	28.4	35.8	49.7	23.7	59.3	2.4	57.0	8.0	3.0	6.7	1.5	35.1	12.1	4.5	55.2
GenPose++ (GP)	36.3	37.0	43.2	42.8	31.1	30.0	70.4	12.2	73.6	13.3	13.2	4.3	11.6	40.8	27.3	16.0	64.3
MagicPony+GP	10.7	4.8	4.5	20.2	21.8	22.1	32.0	1.9	16.8	8.1	9.1	6.9	8.2	21.2	12.0	8.1	13.6
Morpheus	41.2	40.9	46.2	47.2	61.8	36.5	73.3	16.0	75.2	15.1	15.6	8.5	15.4	46.5	26.7	18.9	73.1
Morpheus w/o Def.	39.1	39.9	46.9	46.3	34.6	34.2	73.3	13.2	74.7	14.8	15.2	5.0	16.1	47.4	26.6	19.1	72.5
3D																	
GenPose++ (GP)	34.3	18.8	36.9	62.8	11.1	27.5	77.0	19.6	89.8	7.8	31.8	2.8	19.5	28.0	27.4	28.0	57.2
MagicPony+GP	7.1	1.2	1.0	30.7	4.6	10.8	29.8	1.3	18.4	1.8	16.3	2.8	6.7	5.8	14.8	13.0	4.2
Morpheus	41.5	23.7	42.5	73.2	42.9	39.2	85.3	26.3	91.2	6.9	46.7	5.0	22.1	34.1	46.2	39.7	68.5
Morpheus w/o Def.	38.4	22.7	42.6	71.0	10.4	35.0	85.1	26.0	91.6	8.4	42.4	3.5	22.1	35.1	28.0	38.9	67.8
3D Modal																	
DINOv2+D	24.4	5.7	18.0	29.0	27.0	16.4	52.9	16.7	40.8	8.1	15.3	2.6	10.4	39.6	31.3	28.7	37.4
MagicPony$_{2D}$+D	14.0	3.9	4.8	32.2	23.1	22.7	27.2	10.6	24.0	4.7	10.9	0.0	13.9	20.8	26.3	21.3	26.6
NOCS+D	26.4	6.5	31.9	58.7	42.3	24.0	69.9	2.7	75.2	6.5	34.9	1.9	6.6	18.9	26.4	22.5	51.2
GenPose++ (GP)	37.0	22.9	43.2	60.0	14.6	27.9	84.6	14.3	92.5	11.6	30.4	2.6	19.8	30.2	39.3	33.9	59.4
MagicPony+GP	7.5	2.5	2.3	35.8	5.3	14.7	1.5	3.6	20.4	1.8	10.6	2.2	9.6	5.5	23.8	17.5	3.9
Morpheus	43.7	26.0	48.1	71.3	59.6	38.8	91.2	24.3	94.1	12.2	51.9	2.6	23.6	37.8	45.1	40.0	71.0
Morpheus w/o Def.	40.2	25.2	48.5	70.2	14.6	35.2	91.2	24.3	94.4	14.5	43.7	2.6	23.6	37.2	38.8	35.9	70.1
3D Amodal																	
GenPose++ (GP)	32.9	17.1	35.2	64.4	9.4	27.3	69.5	21.9	88.2	6.8	32.1	2.9	19.3	27.3	22.0	26.5	56.7
MagicPony+GP	7.1	0.7	0.6	27.3	4.3	9.1	58.1	0.4	17.1	1.9	17.6	3.0	4.9	5.9	10.8	11.8	4.2
Morpheus	40.8	22.8	41.1	74.2	35.1	39.4	79.4	27.2	89.5	5.4	45.5	5.8	21.1	33.0	46.7	39.6	67.8
Morpheus w/o Def.	37.8	21.6	41.1	71.5	8.4	34.9	79.0	26.8	89.9	6.8	42.2	3.9	21.1	34.4	23.0	39.7	67.2

	mean	hair dryer	hamburger	hand cream	handbag	knife	lemon	light	lotus root	mango	mangosteen	medicine bottle	mouse	mug	orange	pillow	pomegranate	power strip
2D																		
DINOv2	22.9	19.2	18.0	30.0	16.9	15.6	31.7	16.0	28.5	23.9	26.3	17.6	17.9	13.2	33.3	39.0	21.8	22.7
MagicPony$_{2D}$	15.7	6.8	36.8	19.7	8.2	13.0	11.1	17.4	21.1	11.8	12.8	13.5	6.4	7.0	11.1	22.2	12.7	13.1
NOCS	26.7	21.9	67.3	43.2	17.7	11.6	20.3	19.3	28.0	28.0	3.0	24.9	53.4	26.5	24.6	47.7	9.3	27.8
GenPose++ (GP)	36.3	27.0	76.4	48.5	34.0	26.8	29.4	37.0	39.9	35.4	27.9	31.7	58.9	37.0	30.6	59.3	30.2	37.0
MagicPony+GP	10.7	4.7	55.7	11.2	7.5	9.0	3.3	12.1	8.6	5.1	15.4	9.6	5.2	5.1	15.8	22.7	7.5	8.6
Morpheus	41.2	35.1	79.3	55.8	31.0	30.7	33.9	37.6	39.7	43.4	21.8	39.1	63.6	43.3	38.9	65.2	26.6	44.5
Morpheus w/o Def.	39.1	28.8	80.5	54.3	36.7	29.3	34.4	40.5	40.3	43.8	21.0	36.2	61.2	39.6	35.5	63.8	24.2	41.2
3D																		
GenPose++ (GP)	34.3	19.6	74.8	44.0	18.0	28.0	23.3	36.5	52.7	21.5	42.6	35.9	51.4	20.2	40.8	41.5	38.1	31.6
MagicPony+GP	7.1	0.3	50.5	2.7	1.2	5.0	0.6	8.4	9.3	0.7	19.7	8.3	2.0	0.7	14.7	7.9	4.8	2.6
Morpheus	41.5	27.4	81.1	50.2	16.3	32.4	32.8	45.7	53.5	30.3	38.0	56.1	59.0	24.4	54.3	51.5	36.9	41.5
Morpheus w/o Def.	38.4	19.5	82.3	48.8	21.1	30.4	33.3	45.7	52.8	27.6	39.6	48.4	56.1	21.5	50.6	48.5	34.1	37.3
3D Modal																		
DINOv2+D	24.4	18.4	36.4	24.5	11.2	16.9	25.6	34.3	28.0	23.4	24.7	45.8	17.7	4.9	57.6	24.3	33.1	24.7
MagicPony$_{2D}$+D	14.0	5.8	40.9	13.1	4.1	13.5	7.8	24.9	19.3	9.3	12.3	21.5	6.1	1.4	11.1	16.5	23.1	8.6
NOCS+D	26.4	14.4	60.1	37.2	2.5	15.9	19.0	43.4	40.4	17.6	18.2	50.7	48.8	12.0	27.4	42.7	9.2	28.4
GenPose++ (GP)	37.0	21.0	83.2	45.6	14.9	32.6	24.4	42.4	57.0	19.6	38.3	40.2	59.1	30.4	34.6	53.5	37.3	36.7
MagicPony+GP	7.5	0.2	54.1	3.2	1.6	6.8	1.1	12.7	14.0	1.8	27.3	4.3	3.0	0.8	5.0	8.9	6.3	3.4
Morpheus	43.7	29.1	81.4	53.4	14.0	33.4	36.7	62.3	57.0	34.6	33.3	63.6	64.8	33.1	41.5	59.7	26.3	46.6
Morpheus w/o Def.	40.2	23.4	83.2	51.3	16.1	31.8	37.8	55.1	56.4	29.9	32.7	55.1	61.3	30.1	42.4	55.0	24.6	41.8
3D Amodal																		
GenPose++ (GP)	32.9	19.4	66.4	43.4	19.3	23.3	22.2	34.6	49.8	22.6	45.8	33.7	42.1	16.6	45.6	37.3	38.8	29.4
MagicPony+GP	7.1	0.4	46.8	2.6	1.0	3.1	0.0	7.1	6.1	0.0	14.4	10.0	0.8	0.7	22.1	7.6	3.2	2.3
Morpheus	40.8	27.1	80.9	49.1	17.3	31.3	28.9	40.4	51.2	27.9	41.6	52.2	52.0	21.4	64.3	48.7	46.3	39.4
Morpheus w/o Def.	37.8	18.8	81.4	47.9	23.1	29.0	28.9	42.7	50.4	26.3	44.9	44.9	49.8	18.5	57.0	46.3	42.5	35.3

	mean	remote	sausage	shampoo	shoe	shrimp	teapot	tooth brush	tooth paste	toy animal	toy boat	toy bus	toy car	toy m'bike	toy plane	toy train	toy truck	wallet
2D																		
DINOv2	22.9	19.1	35.2	56.5	30.0	23.0	13.1	37.7	41.1	14.9	17.1	13.3	9.6	10.6	15.2	14.5	13.0	25.4
MagicPony$_{2D}$	15.7	15.2	20.4	49.0	13.8	9.5		8.7	29.6	8.1	8.8	7.2	5.1	9.1	7.7	11.7	8.5	12.5
NOCS	26.7	27.7	15.0	69.7	45.2	9.5	21.7	52.2	55.9	0.9	14.0	42.6	5.5	16.3	20.7	27.3	28.5	28.2
GenPose++ (GP)	36.3	41.2	25.0	90.8	62.6	17.5	23.1	62.2	65.4	6.5	20.5	50.2	7.1	26.7	28.8	34.4	41.5	35.8
MagicPony+GP	10.7	4.5	2.3	22.6	11.9	4.3	3.9	0.4	11.9	4.7	4.1	4.2	3.8	8.1	7.2	5.7	7.7	11.5
Morpheus	41.2	45.6	31.0	89.7	62.9	21.4	37.3	68.3	68.9	8.9	28.1	57.1	9.2	31.3	34.8	43.3	47.1	39.6
Morpheus w/o Def.	39.1	44.9	32.4	91.8	59.1	18.2	25.0	68.3	68.5	7.9	22.5	51.8	6.7	29.5	32.0	41.8	44.9	38.1
3D																		
GenPose++ (GP)	34.3	40.9	24.1	90.1	62.9	12.3	16.0	67.2	65.2	1.6	14.4	37.1	2.2	17.9	18.2	23.6	30.7	24.9
MagicPony+GP	7.1	1.7	1.9	19.5	7.2	1.4	1.0	0.4	4.0	2.2	0.9	0.9	0.7	2.2	2.1	1.5	3.2	2.8
Morpheus	41.5	44.7	31.0	92.8	62.9	16.4	26.4	72.8	69.4	5.4	21.0	47.9	3.3	22.5	26.0	31.5	37.3	31.1
Morpheus w/o Def.	38.4	44.2	31.9	91.4	57.6	15.5	15.3	73.0	68.5	1.6	16.5	41.3	1.9	19.8	21.6	30.2	33.5	30.5
3D Modal																		
DINOv2+D	24.4	15.4	30.6	52.1	30.0	14.0	11.4	51.0	41.4	14.8	7.5	14.4	11.6	11.0	9.2	11.6	20.2	20.6
MagicPony$_{2D}$+D	14.0	12.0	22.2	31.5	11.5	7.4		12.9	29.2	4.9	2.7	4.7	2.7	10.1	3.1	5.0	8.8	6.9
NOCS+D	26.4	28.7	13.0	71.4	39.0	5.5	12.0	66.8	60.6	1.8	4.5	34.6	0.3	7.4	13.5	25.2	14.7	16.8
GenPose++ (GP)	37.0	45.0	28.7	92.5	64.6	11.7	21.5	80.7	67.4	2.5	12.9	38.5	2.4	27.5	14.9	32.9	22.8	25.2
MagicPony+GP	7.5	2.0	0.9	21.9	11.9	0.8	1.0	0.4	3.0	1.5	1.1	0.3	1.2	4.1	2.1	1.1	3.4	4.9
Morpheus	43.7	49.7	33.3	87.0	58.5	17.5	31.5	82.3	72.9	3.7	19.9	49.2	3.1	33.8	23.6	41.2	29.8	30.5
Morpheus w/o Def.	40.2	48.2	33.3	87.7	50.8	14.4	16.8	82.8	71.9	1.2	17.2	44.5	1.0	27.8	16.8	40.9	25.4	30.5
3D Amodal																		
GenPose++ (GP)	32.9	39.2	19.4	87.7	61.9	12.5	13.7	57.9	64.4	1.4	14.8	36.7	2.1	15.0	19.2	20.1	34.0	24.8
MagicPony+GP	7.1	1.6	2.8	17.1	3.8	1.7	1.0	0.5	4.4	2.4	0.9	1.1	0.6	1.6	2.1	1.6	3.1	1.6
Morpheus	40.8	42.6	28.7	98.6	65.7	16.0	24.3	66.3	68.1	5.9	21.3	47.5	3.3	19.0	26.7	27.8	40.5	31.4
Morpheus w/o Def.	37.8	42.5	30.6	95.2	61.9	15.9	14.7	66.3	67.3	1.7	16.3	40.3	2.1	17.3	23.1	26.2	36.9	30.6

Category	Number of keypoints
backpack	16
book	8
bottle	5
box	14
bread	6
coconut	2
conch	7
corn	3
dinosaur	18
dish	9
doll	11
egg	3
eraser	10
facial_cream	12
flower_pot	10
glasses_case	16
hair_dryer	14
hamburger	2
hand_cream	7
handbag	15
knife	6
lemon	2
light	18
lotus_root	3
mango	3
mangosteen	4
medicine_bottle	6
mouse	7
mug	14
orange	3
pillow	6
pomegranate	4
power_strip	10
remote_control	11
sausage	2
shampoo	2
shoe	12
shrimp	8
teapot	13
tooth_brush	8
tooth_paste	7
toy_animals	11
toy_boat	8
toy_bus	18
toy_car	8
toy_motorcycle	19
toy_plane	13
toy_train	10
toy_truck	10
wallet	9

Table A4. Maximum number of annotated keypoints observed for each category.

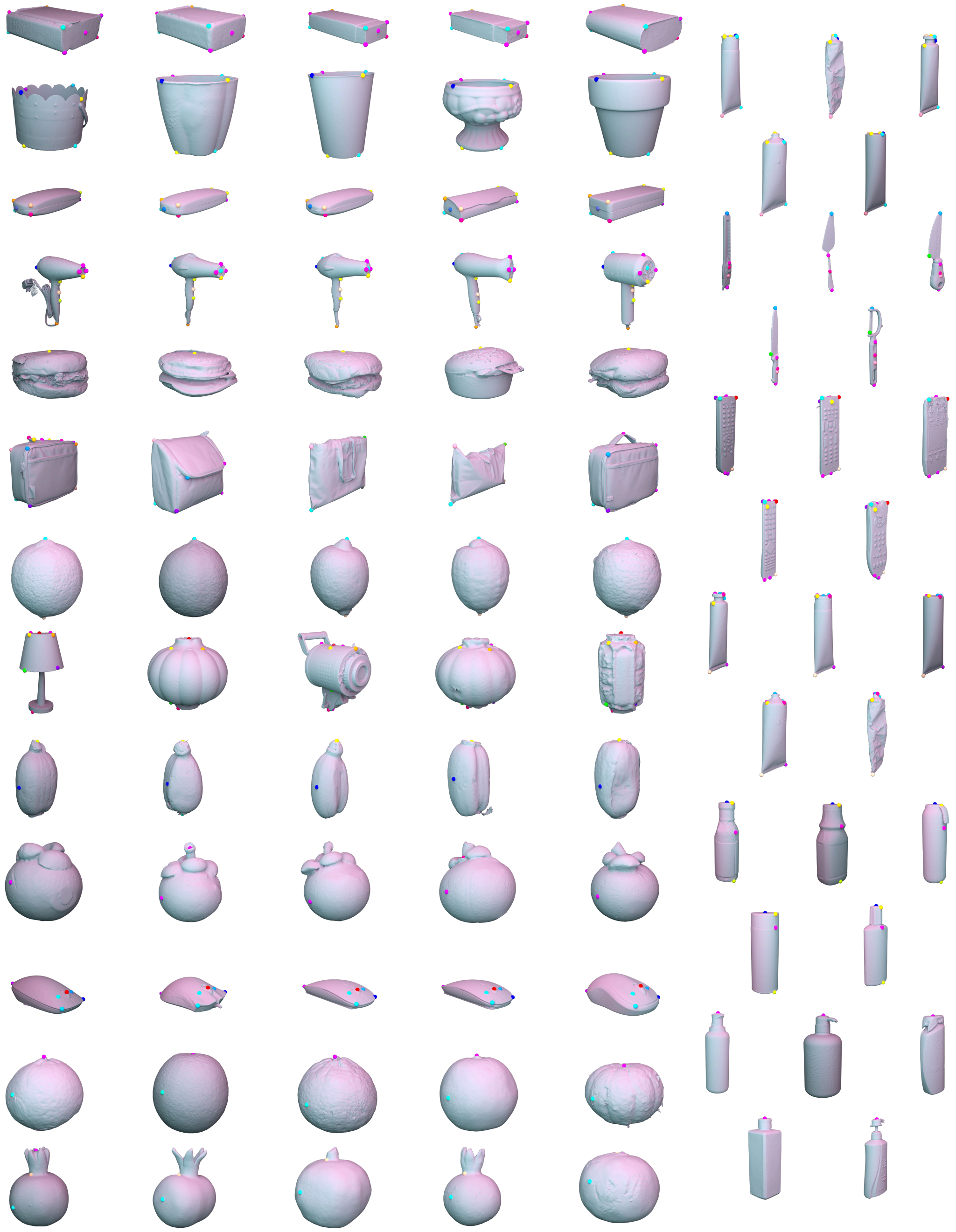


Figure A7. Full keypoints annotation overview of HouseCorr3D (part 1 of 3)

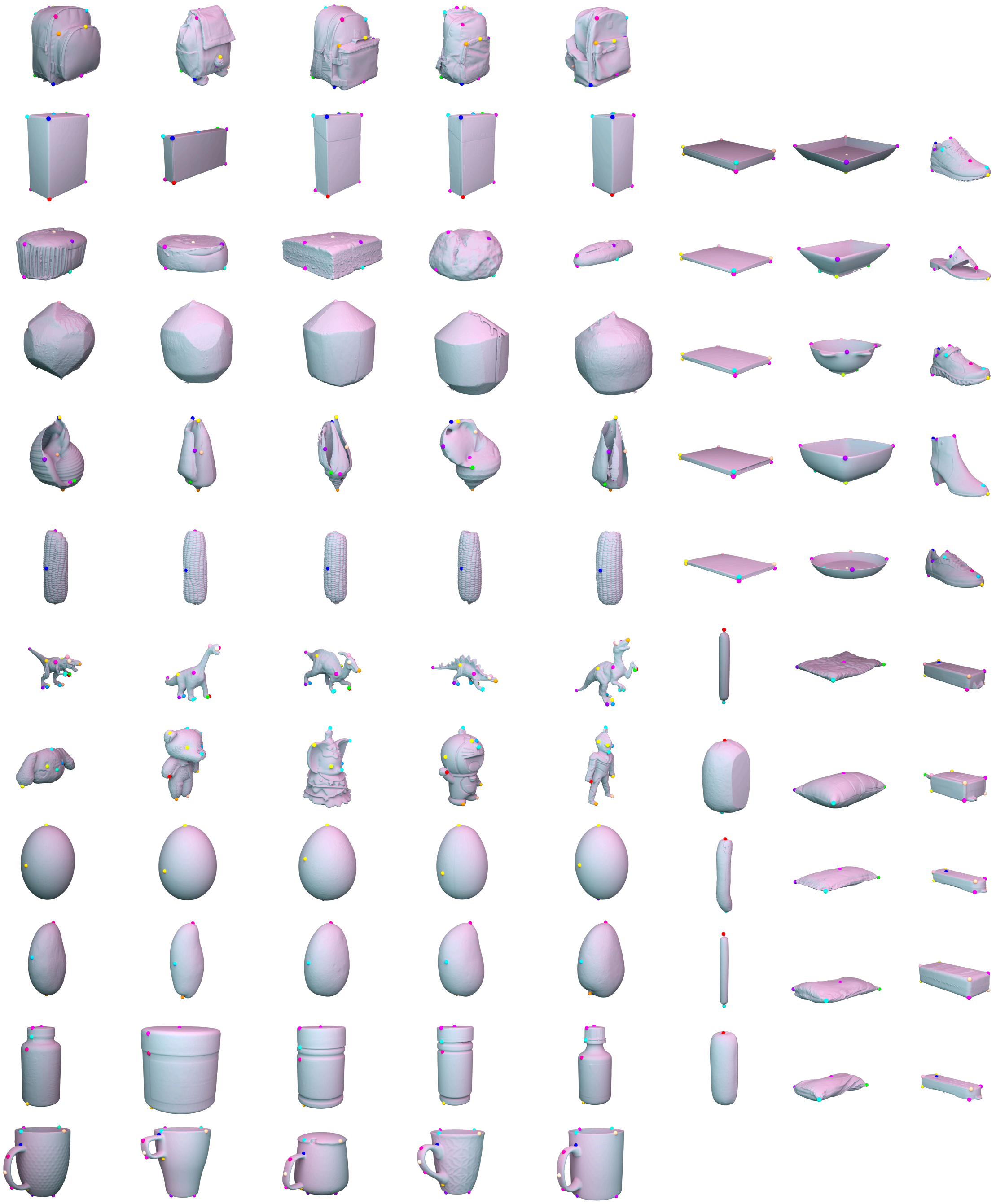


Figure A7. Full keypoints annotation overview of HouseCorr3D (part 2 of 3)

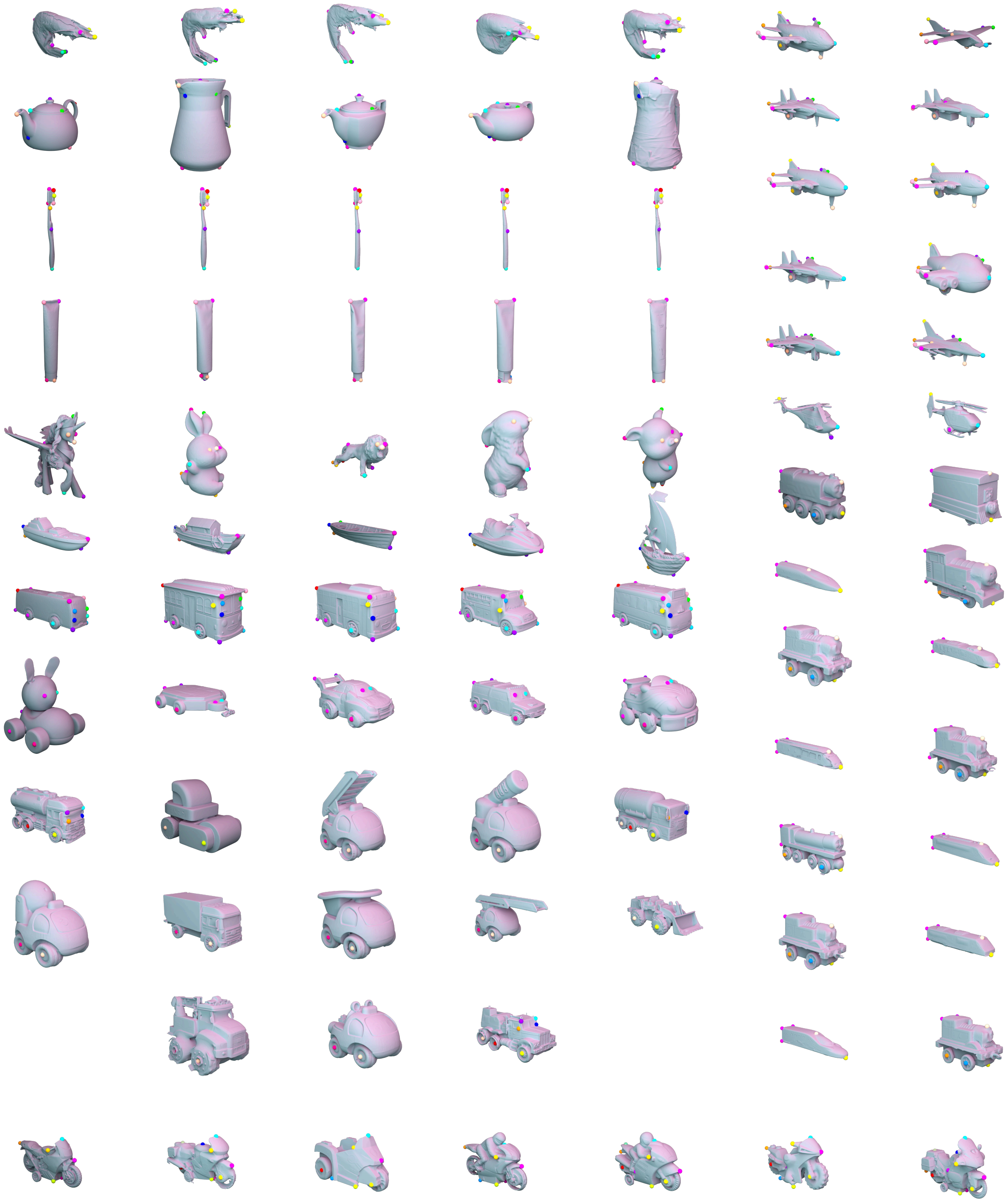


Figure A7. Full keypoints annotation overview of HouseCorr3D (part 3 of 3)