Category-Level 3D Correspondence in Camera Space via Morphable Object Priors

Leonhard Sommer[1⋆✉], Artur Jesslen[1⋆✉], Basavaraj Sunagad[2], and Adam Kortylewski[2]

[1] University of Freiburg, Germany
[2] CISPA Helmholtz Center for Information Security, Germany

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. In this work, we study **category-level 3D correspondence in camera space**—predicting, from a single image, 3D locations that remain consistent across instances within a category—and show that it can emerge without explicit correspondence supervision by learning a shared morphable object prior. To enable research in this direction, we introduce **HouseCorr3D**, the first large-scale benchmark for single-view 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 method that learns **morphable category-level shape priors** by disentangling canonical shape, deformation, and object pose. Through this shared canonical grounding, semantically meaningful 3D correspondences **in camera space** emerge implicitly. These emerging 3D correspondences set a new state of the art on HouseCorr3D, demonstrating that semantic 3D object understanding can arise without direct correspondence supervision. Data and code: /GenIntel/HouseCorr3D.

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

⋆ Equal contribution.
✉ Corresponding authors: {sommerl,jesslen}@cs.uni-freiburg.de.

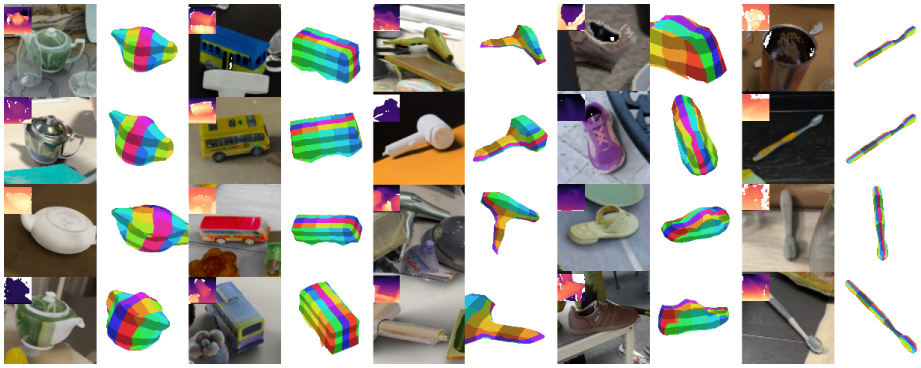

Fig. 1: Single-view Category-level 3D Correspondence. We predict semantically consistent 3D keypoint locations across different instances of the same category from single RGB-D images. Our morphable priors enable establishing correspondences (shown with matching colors) that remain semantically aligned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

problem has driven extensive research [21, 31, 32, 34, 44], 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 [24, 52], we propose to establish correspondences directly in 3D camera space, resolving fundamental ambiguities that arise in image-space matching due to occlusion, viewpoint change, and scale variation. Formally, we define this novel task as follows:

Single-view 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 point.

Intuitively, the task asks: *if we select a semantic part on one object, where does the same part lie on another instance of the category?* Our approach answers this question by mediating correspondence through a shared deformable template. An overview of this camera-space correspondence setup is illustrated in Fig. 3a. Unfortunately, existing benchmarks such as NOCS-Real275 [52], Wild6D [8], OmniNOCS [24], and Omni6DPose [68] 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 single-view 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 [65], 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 [61], not just visible surfaces. We also explicitly support object symmetries, ensuring symmetric 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.

On HouseCorr3D, we show that single-view category-level 3D correspondence can emerge **without explicit correspondence supervision** by constraining object instances through a shared deformable representation. To this end, we propose **Morpheus**, a framework that learns **morphable category-level shape priors** to produce semantically consistent 3D correspondences directly in camera space. Instead of relying on a fixed 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 correspondence evaluation without pose normalization.

In summary, our contributions are as follows:

(i) We identify **single-view 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 **correspondence-level 3D object understanding**.

2 Related work

2D Semantic Correspondence. 2D correspondence has advanced from local descriptors and dense flows [27, 30, 49, 54] to transformer-based self-supervised features [3, 16, 17, 37, 69, 71], which exhibit emergent semantic alignment and achieve strong results on benchmarks like SPair-71K, PF-PASCAL, and TSS [14, 26, 32].

Dedicated matchers such as LoFTR, COTR, DiffMatch [21,34,44], spherical-map-based approaches [7,31], and geometry-guided approaches [19] further improve correspondence matching. While highly effective, these approaches remain limited to images 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 [67] introduced a large-scale dataset for learning category-consistent 3D keypoints, while others [18,64] leverage keypoints for cage-based deformations and shape control. Canonical surface mapping [25] establishes correspondences by predicting UV coordinates on canonical templates, and Mesh R-CNN [9] jointly predicts mesh reconstructions with instance segmentation from 2D images. Some methods [20, 38, 43] learn implicit correspondences anchored in 3D geometry through contrastive learning at the vertex level, but rely on non-deformable templates. Recent semantic alignment methods [29,51,66] explore learning consistent correspondences across categories and human poses in 3D. DenseMatcher [72] extends matching to the mesh domain, projecting multiview features onto 3D geometry. Crucially, these methods match *pre-reconstructed* meshes rather than RGB-D images; adapting them would demand a brittle pipeline with compounding errors: shape reconstruction [6, 58], pose estimation [68], canonicalization, and functional-map matching [72]. Similarly, many explicit 3D correspondence approaches have fundamental limitations and require ground-truth 3D meshes as input [18,64,67,72]; or operate exclusively in 3D space without bridging to image-based features [18,29,66]; and critically, none provide large-scale evaluation benchmarks with explicit handling of occlusion and symmetry. These limitations prevent their applicability to real-world scenarios.

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 [28]), establishing the foundation for template-based shape modeling. Recent approaches [22,35,41,42] extend these ideas to more diverse object classes using learned deformations or diffusion-guided generation. Among these, Common3D [42] is the most closely related category-level morphable-prior approach; it learns from object-centric RGB videos, but assumes a known object distance, limited occlusion, and mostly upright objects at inference. Deformation-based methods [12,48,53] map instances to template meshes using neural networks, while template-free approaches [36] learn canonical coordinate systems without relying on a single exemplar. More recent work leverages foundation models for semantic alignment across categories [35,41], where semantically corresponding parts map to consistent representations. Domain-specific efforts have also addressed human bodies [13] and a range of animals [60]. 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.

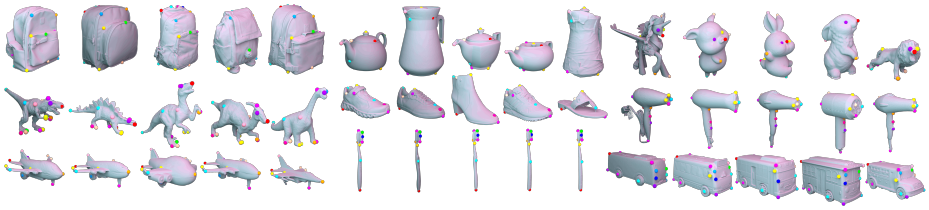

Fig. 2: Dataset Overview. We annotate up to 19 3D keypoints directly on CAD meshes for 5–13 instances per category, covering 50 common household object classes. The keypoints are chosen to be semantically consistent and shared across all instances within each category We visualize a subset of these annotations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

Benchmarks for Category-Level 3D Understanding. To the best of our knowledge, there exists no dataset that enables single-view category-level 3D correspondence evaluation. Prior works [56] lift 2D images from domain-specific datasets [50,55] to 3D using multi-view consistency but lack 3D evaluation benchmarks. Large-scale 3D shape collections such as ShapeNet [4] and ModelNet [57] provide CAD meshes, while ShapeNetPart [63] and PartNet [33] add part-level labels, but these lack consistent point-level correspondences across instances. Pose-focused datasets like Omni6DPose [68], CO3D [39], Pix3D [45], Pascal3D+ [59], and Omni3D [2] provide pose annotations in realistic scenes but do not supply semantic, amodal, or point-level correspondences across diverse instances. NOCS datasets [24,52] introduced normalized coordinate spaces for pose estimation but are not designed for evaluating category-level correspondences, as described in **??**. DenseCorr3D [72] 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 single-view category-level 3D correspondence evaluation from 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.

3 The HouseCorr3D Benchmark

Motivation. We introduce the first benchmark for category-level correspondences in 3D camera space (Tab. 1), unlike prior datasets that focus exclusively on correspondences in either 2D camera space [14,32,46,50,55] or 3D object space [72]. 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.

Table 1: Comparison to existing correspondence datasets. Prior benchmarks evaluate in either 2D camera or 3D object space. **HouseCorr3D** is the first to target 3D camera space, enabling amodal evaluation across 50 classes.

Dataset	pairs	classes	input	eval. space	symmetry	occlusion
Pascal-Parts [5]	4k	20	2D	2D camera	✗	✗
PF-Pascal [14]	2k	20	2D	2D camera	✗	✗
Spair71k [32]	71k	18	2D	2D camera	✗	✓
KeypointNet [67]	N/A	16	3D	3D object	✗	✗
CPNet [29]	N/A	25	3D	3D object	✓	✗
DenseCorr3D [72]	N/A	23	3D	3D object	✓	✗
HouseCorr3D	**178k**	**50**	**2.5D**	**3D camera**	✓	✓

Moreover, compared to estimation in 2D camera space the **3D camera space has several critical advantages**: a) the evaluation of amodal correspondences, b) modeling object symmetries explicitly, and c) enforcing methods to perform 3D over 2D reasoning. Importantly, HouseCorr3D is designed as a *test-only benchmark*: keypoints annotations are used exclusively for evaluation.

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 $d(\hat{x}^t, x^t) = \|\hat{x}^t - x^t\|_2$. 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 3D bounding box, as: $d(\hat{x}^t, x^t) < 0.1 \cdot \max(h, w, d)$. This follows the conventions of other 2D correspondence benchmarks [14, 32, 46, 50, 55], 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 **??**.

HouseCorr3D. We build our dataset on Omni6DPose [68], 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 photo-realistic 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. To nonetheless assess real-world generalization, we additionally annotate a small evaluation-only subset of the real ROPE data, comprising 5 representative classes (24 instances and 134 keypoints); we detail its selection and evaluation in **??**. We additionally define

two topology-based evaluation subsets to probe robustness to structural variation. Because a keypoint is annotated only where its corresponding semantic part exists, the number of keypoints on an instance reflects its topology: instances that differ topologically through missing parts or markedly different shapes (*e.g.*, a helicopter versus an airplane), have different numbers of keypoints annotated. We therefore label an image pair as *different-topology* when its two instances differ by two or more annotated keypoints, and as *similar-topology* otherwise. 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 19 semantic 3D keypoints are annotated directly on CAD meshes (see Fig. 2).

Keypoint Annotation Protocol. Keypoints are shared across all instances of a category and chosen to be geometrically distinctive and semantically meaningful [47]—marking corners, edges, or handle centers rather than arbitrary surface points—so that they are reliably localizable and transferable across instances. To ensure quality, two annotators independently label the same meshes with an interactive 3D tool, and the two annotation sets are merged via an automatic step (mutual nearest-neighbor matching) followed by manual resolution of ambiguous cases (details in **??**). This yields 2329 3D keypoints (2–19 per instance) over roughly 65h of annotation. Using Omni6DPose ground-truth poses [68], we project them into all rendered views, scaling this compact annotation set into 178k image pairs with consistent, *amodal* 2D–3D correspondences across 50 categories and 280 instances. Our benchmark inherits the visual realism of Omni6DPose, featuring natural lighting, cluttered scenes, and partial occlusions.

Symmetry. Many everyday objects exhibit geometric symmetries that introduce fundamental ambiguities in correspondence. For instance, a cylindrical mug body 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 *discrete* and *continuous* symmetries, ensuring that geometrically equivalent predictions are not unfairly penalized. Symmetry is handled by treating all points on the orbit generated by rotations around the symmetry axis as valid correspondences. This yields a fair metric that respects the inherent geometric ambiguities in real-world objects and enables robust evaluation of category-level correspondence methods. More details are provided in **??**.

4 Method

Our goal is to recover category-level 3D correspondences directly in camera space from single-view RGB-D observations. To achieve this, we introduce **Morpheus**, a model that, from a single image, predicts a 6D object pose and a deformable 3D shape whose semantic structure remains consistent across object instances. The

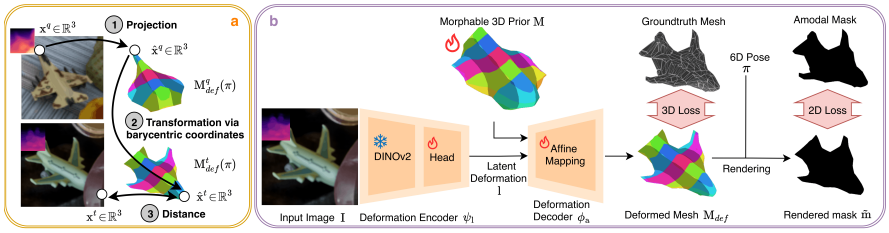

Fig. 3: **(a) Single-view category-level 3D correspondence.** Given a query point $x^q \in \mathbb{R}^3$, we project it onto the deformed query mesh M^q_{def} and encode its location as barycentric coordinates. Since query and target instances share the same mesh topology, these coordinates transfer directly to M^t_{def}, yielding the corresponding point $x^t \in \mathbb{R}^3$. **(b) Pipeline.** Given an RGB-D image, the deformation encoder ψ_{l} predicts a latent code l that drives the decoder ϕ_a to adapt the category shape prior to the observed instance. The deformed mesh is placed in camera space using the predicted 6D pose. Training uses amodal 2D and 3D losses together with pose supervision.

central idea of Morpheus is to represent all objects within a category as *identity-preserving deformations* of a shared template mesh. Because template vertices maintain persistent identities during deformation, semantic correspondences arise naturally. We start by describing how to predict 3D correspondences in camera space in Sec. 4.1. Subsequently, we explain our architecture 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, $-^q$ and $-^t$ distinguish query and target elements (*e.g.*, M^q and M^t). We denote a deformed mesh as M_{def}, and its transformation into camera space with pose π as $\mathrm{M}_{def}(\pi)$.

4.1 Mesh-based 3D Correspondence Prediction

Morpheus establishes correspondences by mediating all predictions through a shared deformable template. For each RGB-D image I, the model predicts: (i) an instance-specific deformation of the template mesh, and (ii) a 6D pose ω estimated from pretrained pose diffusion [68]. The deformed mesh is then transformed into camera space as $M_{\mathrm{def}}(\omega)$. Given a query–target image pair (I^q, I^t), we obtain their posed meshes as $M^q_{\mathrm{def}}(\omega^q)$ and $M^t_{\mathrm{def}}(\omega^t)$.

3D Correspondence Prediction via Mesh Transfer. A query 3D point $\mathbf{x}_q$ is first projected onto the surface of the query mesh, producing a surface point $\hat{\mathbf{x}}_q$. We represent this point using barycentric coordinates with respect to the underlying mesh face. Because both instances share identical mesh topology, these barycentric coordinates define a category-level surface identifier. The same identifier is then transferred onto the target mesh, yielding the predicted correspondence $\hat{\mathbf{x}}_t$ in the target camera space (Fig. 3a). Thus, single-view 3D correspondence estimation reduces to predicting the pose and deformations of a shared template mesh rather than directly matching points between images.

4.2 3D Morphable Priors

A central component of Morpheus is the *3D morphable prior*, which models all instances of a category as deformations of a shared canonical template, hence enabling semantically consistent correspondences across instances. It consists of a canonical mesh capturing the common topological structure of a category, along with a learned deformation model that adapts it to individual instances. We refer to the model as 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: observations mapped to the same template vertex correspond to the same semantic part across instances. This converts 3D correspondence estimation into a pose and deformation estimation problem. Unlike prior morphable-model approaches designed for reconstruction, our formulation leverages persistent template vertex identity as the fundamental mechanism enabling single-view category-level 3D correspondence.

Canonical Shape Representation. Traditional mesh-only representations are often fragile and difficult to optimize directly, typically requiring manual interventions such as remeshing [10,62]. To overcome this limitation, we employ a hybrid volumetric mesh representation [40]. This integrates the strengths of implicit and explicit 3D models. Concretely, the category-level shape is represented as a signed distance field ϕ_{sdf}, providing flexibility to model intricate geometries. Through Differentiable Marching Tetrahedra [40], 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 learn an affine deformation field, following [70]. Unlike [70], where deformations are applied directly to the signed distance field, we act on the template mesh vertices [56]. This avoids repeatedly extracting meshes for each instance and is thus more computationally efficient. Formally, we define an affine mapping $\phi_a : \mathbb{R}^3 \times \mathrm{L} \to \mathbb{R}^3$, which displaces each vertex $\boldsymbol{v}$ individually according to the instance-specific latent code l:

$$\phi_a(\boldsymbol{v}, \mathrm{l}) = \alpha(\boldsymbol{v}, \mathrm{l}) \odot \boldsymbol{v} + \delta(\boldsymbol{v}, \mathrm{l}), \tag{1}$$

where α, $\delta : \mathbb{R}^3 \times \mathrm{L} \to \mathbb{R}^3$ are produced by an MLP that takes both the vertex coordinate $\boldsymbol{v}$ and 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 code parametrizes vertex-wise displacements, enabling the mesh to morph into the observed instance while preserving semantic alignment. The resulting instance mesh is $\mathrm{M}_{def}(\mathrm{I}) = \{\mathrm{V}_{def}(\mathrm{I}), \mathrm{E}\}$, where each deformed vertex is given by $\mathrm{V}_{def}(\mathrm{I}) = \{\phi_a(\mathrm{v}_i, \psi_{\mathrm{l}}(\mathrm{I}))\}_{i=1}^{|\mathrm{V}|}$. For simplicity, we simply rewrite it as $\mathrm{M}_{def} = \phi_a(\mathrm{M}, l)$. Through the deformation, vertices maintain consistent identities, enabling category-level correspondence prediction without explicit correspondence supervision.

4.3 Training Objectives

Morpheus is trained using geometric supervision that encourages all object instances to explain observations through a shared morphable category-level prior. Importantly, no explicit correspondence supervision is used during training. Instead, semantic alignment emerges implicitly because every instance must deform the same canonical template while remaining consistent with observed data. We jointly optimize the encoder, decoder and morphable prior using reconstruction and regularization objectives. Specifically, training enforces consistency between the deformed mesh and the input observations (Fig. 3b) through (i) 2D mask-based reconstruction, (ii) 3D geometry alignment, and (iii) deformation regularization. Together, these objectives encourage the model to learn category-consistent canonical structure while preserving instance-specific shape variation. In contrast to prior work [56], we additionally leverage a 6D pose, used as a known transform into camera space to stabilize optimization and overcome local minima. Furthermore, amodal 2D supervision and 3D geometric losses improve robustness under occlusion by encouraging reasoning about both visible and occluded object regions.

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

$$\mathcal{L}_{\mathrm{m}}(\mathrm{M}_{def}, \mathrm{I}, \pi, \mathrm{m}) = \left\| \tilde{\mathrm{m}}(\mathrm{M}_{def}, \mathrm{I}, \pi) - \mathrm{m} \right\|^2. \tag{2}$$

Additionally, we encourage overlap with the distance transform of the ground-truth amodal mask $\mathtt{dt}(\mathrm{m})$:

$$\mathcal{L}_{\mathrm{mdt}}(\mathrm{M}_{def}, \mathrm{I}, \pi, \mathrm{m}) = -\tilde{\mathrm{m}}(\mathrm{M}_{def}, \mathrm{I}, \pi) \odot \mathtt{dt}(\mathrm{m}), \tag{3}$$

with $\mathtt{dt}(\mathrm{m})$ 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. For accurate 3D instance reconstruction, we use a Chamfer distance between the deformed mesh vertices V_{def} and the GT mesh vertices V_{gt}.

$$\mathcal{L}_{CD}(\mathrm{I}, \mathrm{M}_{def}, \mathrm{M}_{gt}) = \frac{1}{|\mathrm{V}_{def}|+|\mathrm{V}_{gt}|} \Big(\sum_{\boldsymbol{v}_i \in \mathrm{V}_{def}} \|\boldsymbol{v}_i - \boldsymbol{v}'_{\chi(\boldsymbol{v}_i)}\| + \sum_{\boldsymbol{v}'_i \in \mathrm{V}_{gt}} \|\boldsymbol{v}'_i - \boldsymbol{v}_{\chi(\boldsymbol{v}'_i)}\| \Big), \tag{4}$$

where χ denotes the nearest neighbor operator.

Template and Deformation Regularization. Following [11], 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 $\mathcal{L}_{smooth}$ [70]. Their definitions are provided in **??**.

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

$$\mathcal{L}_{\mathrm{geo}} = \lambda_{CD}\mathcal{L}_{CD} + \lambda_{\mathrm{m}}\mathcal{L}_{\mathrm{m}} + \lambda_{\mathrm{mdt}}\mathcal{L}_{\mathrm{mdt}} + \lambda_{sdf}\mathcal{L}_{sdf}. \tag{5}$$

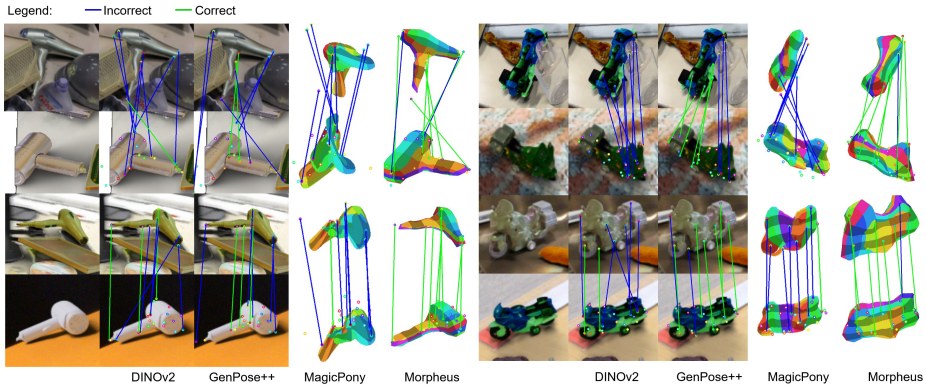

Fig. 4: Qualitative results. We compare 2D feature matching method DINOv2 [37], with 3D space matching methods GenPose++ [68], MagicPony [56], and Morpheus. For DINOv2 and GenPose++ we visualize the 2D correspondences. For MagicPony and Morpheus, we visualize the predicted deformed meshes in camera space, along with overlaid correspondence lines. MagicPony's predictions may appear visually plausible when projected in 2D but are often incorrect in 3D. Additionally, pose-aware training results in higher consistency for semantic parts across different viewpoints.

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

$$\mathcal{L}_{\text{geo-reg}} = \mathcal{L}_{\text{geo}} + \lambda_{def}\mathcal{L}_{def} + \lambda_{smooth}\mathcal{L}_{smooth}. \tag{6}$$

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 [52] and DINOv2 [37], as well as 3D space matching methods such as MagicPony [56] and GenPose++ [68]. We first provide experimental details in Sec. 5.1. We describe all baselines in Sec. 5.2 and finally compare with prior work in Sec. 5.3.

5.1 Experimental Details

Morpheus uses a pretrained ViT-S DINOv2 image encoder [37] as backbone, and a pretrained 6D pose diffusion network [68]. 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 [15] that aggregates multi-scale feature maps with bottleneck blocks to produce refined latent deformation 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 [23] with a learning rate of 10^{-4} and a batch size of 30. Training proceeds in two

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 [37], MagicPony$_{2D}$ [56], NOCS [52] and 3D methods (GenPose++ (GP++) [68], MagicPony [56], Common3D [42], and Morpheus). ⋆2D predictions lifted to 3D via depth; amodal evaluation is not applicable (occluded points have no depth).

Method	🚘	✈	🚘	🚌	☎	🏍	mean	🚘	✈	🚘	🚌	☎	🏍	mean
	2D							**3D Modal**						
DINOv2⋆	7.0	15.2	17.1	13.3	14.0	10.6	22.9	5.7	9.2	7.5	14.4	16.4	11.0	24.4
MagicPony$_{2D}$⋆	6.4	7.7	8.8	7.2	22.9	9.1	15.7	3.9	3.1	2.7	4.7	22.7	10.1	14.0
NOCS⋆	27.2	20.7	14.0	42.6	23.7	16.6	26.7	6.5	13.5	4.5	34.6	24.0	7.4	26.4
GP++	37.0	28.8	20.5	50.2	30.0	26.7	36.3	22.9	14.9	12.9	38.5	27.9	27.5	37.0
MagicPony+GP++	4.8	7.2	4.1	4.2	22.1	8.1	10.7	2.5	2.1	1.1	0.3	14.7	4.1	7.5
Common3D	24.8	10.6	3.6	51.3	14.1	19.2	21.2	19.8	2.8	0.0	39.7	7.8	12.9	16.0
Morpheus w/o Def.	39.9	32.0	22.5	51.8	34.2	29.5	39.1	25.2	16.8	17.2	44.5	35.2	27.8	40.2
Morpheus (Ours)	**40.9**	**34.8**	**28.1**	**57.1**	**36.5**	**31.3**	**41.2**	**26.0**	**23.6**	**19.9**	**49.2**	**38.8**	**33.8**	**43.7**
	3D Amodal							**3D (Modal + Amodal)**						
GP++	17.1	19.2	14.8	36.7	27.3	15.0	32.9	18.8	18.2	14.4	37.1	27.5	17.9	34.3
MagicPony+GP++	0.7	2.1	0.9	1.1	9.1	1.6	7.1	1.2	2.1	0.9	0.9	10.8	2.2	7.1
Common3D	9.0	3.1	2.1	31.8	8.0	7.3	11.2	9.8	2.9	1.8	32.3	8.0	7.7	12.1
Morpheus w/o Def.	21.6	23.1	16.3	40.3	34.9	17.3	37.8	22.7	21.6	16.5	41.3	35.0	19.8	38.4
Morpheus (Ours)	**22.8**	**26.7**	**21.3**	**47.5**	**39.4**	**19.0**	**40.8**	**23.7**	**26.0**	**21.0**	**47.9**	**39.2**	**22.5**	**41.5**

stages: (i) 20 epochs optimizing the loss in Eq. (5), and (ii) 10 further epochs optimizing the extended loss in Eq. (6), which includes deformation regularizers. Training on a NVIDIA RTX 2080 takes 12h and is trained per category following standard practice for morphable models [28, 56]. The lightweight deformation head sits on top of a shared ViT-S DINOv2 backbone dominating inference cost. This factorization scales practically: adding a category requires training only on the ∼50k images of that class rather than retraining on all 2.6M images.

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 **??**.

5.2 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 [52] and DINOv2 [37], 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. 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

Table 3: Real-world generalization and topology robustness (PCK@0.1). **(a)** On a filtered real subset of ROPE (5 classes, 24 instances, 134 keypoints), Morpheus outperforms all baselines, confirming synthetic-to-real transfer. **(b)** Split by topology, Morpheus drops by 19pp in 3D from similar- to different-topology pairs, isolating topological mismatch as a key limitation.

Real subset	2D	3D
$\text{MagicPony}_{\text{2D}}$	16.8	N/A
GP++	37.0	25.1
MagicPony+GP++	12.6	7.3
Morpheus	**44.7**	**34.8**

(a) Real-world subset

Morpheus	2D	3D
All	41.2	41.5
Similar topology	44.6	46.7
Different topology	36.3	27.7

(b) Topology subsets

3D *modal* correspondences. However, since occluded regions are not visible, the 3D *amodal* correspondence task cannot be solved using any 2D baseline. Additionally, 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 learns canonical space and camera space entangled together, an external orientation estimator cannot be applied as its canonical space would not match the learned canonical orientation. However, we still require the translation and object scale from GenPose++ to obtain 3D correspondence predictions. Finally, we adapt Common3D [42], the closest category-level morphable-prior baseline. As it is designed to learn from object-centric videos and assumes a known object distance, limited occlusion, and upright objects at inference, we train it on HouseCorr3D and extend it with a median depth-based scaling to recover metric placement.

5.3 Comparison with Prior Work

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

Occlusions. 2D feature matching methods (*e.g.*, NOCS, DINOv2) cannot handle occlusions by design, and cannot evaluate them in the 3D amodal setting. In Fig. 4, we also observe how DINOv2 matches the back of the hairdryer with the front of another one, which is truncated in the target image. Furthermore, we observe qualitatively that MagicPony fails to correctly reconstruct occluded parts, as seen for the hairdryer in Fig. 4. In contrast, Morpheus successfully reconstructs occluded parts. As shown in Tab. 2, Morpheus experiences an average drop of 2.9pp PCK@0.1 between modal and amodal correspondences, confirming that occlusions pose a greater challenge, yet performance remains competitive.

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 front of a hairdryer to a location outside the target hairdryer's due to the smaller size.

MagicPony predict deformations that can fit 2D images well in most cases. However, since it is designed for 2D alignment, it struggles to recover consistent 3D rotations across images resulting in the model compensating through deformation rather than rotation, which, while plausible in 2D, leads to unreliable 3D correspondences (see Fig. 4). This is further shown in Tab. 2, where $\text{MagicPony}_{\text{2D}}$ outperforms MagicPony+GP++, confirming that learning entangled canonical and camera space together does not lead to reliable correspondences in 3D. Morpheus addresses this through explicit disentanglement of pose, shape, and canonicalization during training, yielding more accurate 3D structure and semantic consistency across viewpoints.

Common3D. Even after training on HouseCorr3D and applying the depth-based scaling, Morpheus outperforms Common3D [42] by more than 20pp across all settings. This substantial gap highlights two fundamental limitations of Common3D. First, it does not effectively integrate RGB and depth information, instead relying primarily on RGB features with depth used only for post-hoc scaling. Second, it introduces inconsistencies between the implicit deformation learned through 2D-3D correspondences and the explicit mesh deformation, which can negatively affect pose estimation accuracy.

Performance gains. Using the predicted poses [68], the improvement of Morpheus decomposes into two contributions: replacing rigid object-space transfer with a learned category template improves performance by 4.1 pp, and adding instance-specific deformations brings a further 3.1pp gain (Morpheus vs. Morpheus w/o Def. in Tab. 2). While the category template explains a large part of the improvement, deformations add clear gains, especially for high-variation categories (up to 14.7pp), confirming their importance.

Real-world transfer. Although HouseCorr3D is synthetic, Morpheus transfers to real data. On a filtered subset of ROPE (5 classes, 24 instances, 134 keypoints), it reaches 44.7 and 34.8 PCK@0.1 in 2D and 3D (Tab. 3a), on par with its synthetic 2D performance and with only an approximately 7pp drop in 3D, which we attribute partly to noisier depth and pose annotation rather than a failure of generalization. We provide the full protocol and results in **??**.

Topology. Topological variation remains a major challenge (Tab. 3b). On similar-topology pairs, Morpheus achieves 46.7 PCK@0.1 in 3D (A+M), whereas performance decreases to 27.7 on different-topology pairs, a 19-point reduction. This gap persists despite excluding keypoints that appear on only one of the paired objects. The results suggest that the fixed-connectivity template mesh imposes a structural bias that limits robustness to topological discrepancies.

5.4 Discussion

While Morpheus enables semantically consistent correspondences across diverse object instances, several limitations remain. *(i) Pose sensitivity:* since correspondences live in camera space, they inherit pose-estimation errors that misalign the whole object even when its shape is recovered correctly; supplying ground-truth pose then recovers up to 14pp (**??**). We also found it preferable to disentangle pose and shape rather than optimize them jointly, though coupling the two more tightly is an appealing direction. *(ii) Topology:* because the template has fixed connectivity, it cannot absorb large topological variation such as missing parts, and accuracy degrades the further an instance's structure departs from it (**??**). *(iii) Fine-grained details:* the same smoothness regularization that stabilizes training can also oversmooth thin structures. Despite these limitations, identity-preserving morphable priors are a highly effective mechanism for establishing single-view category-level 3D correspondences in camera space.

6 Conclusion

This paper introduces a paradigm shift from correspondence evaluation in 2D camera space or 3D object space toward **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 single-view 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.

Acknowledgments

AK acknowledges support via his Emmy Noether Research Group funded by the German Research Foundation (DFG) under grant number 468670075. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant number 539134284, through EFRE (FEIH_2698644) and the state of Baden-Württemberg.

References

1. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. p. 187–194. SIGGRAPH '99, ACM Press/Addison-Wesley Publishing Co., USA (1999). https://doi.org/10.1145/311535.311556, https://doi.org/10.1145/311535.311556 4
2. Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., Gkioxari, G.: 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). IEEE, Vancouver, Canada (June 2023) 5
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 3
4. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015) 5
5. Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A.: 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). pp. 1971–1978 (2014) 6
6. Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., et al.: Sam 3d: 3dfy anything in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7220–7232 (2026) 4
7. Dünkel, O., Wimmer, T., Theobalt, C., Rupprecht, C., Kortylewski, A.: Do it yourself: Learning semantic correspondence from pseudo-labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025) 4
8. Fu, Y., Wang, X.: 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) 2
9. Gkioxari, G., Johnson, J., Malik, J.: Mesh r-cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9784–9794 (2019). https://doi.org/10.1109/ICCV.2019.00988 4
10. Goel, S., Gkioxari, G., Malik, J.: Differentiable stereopsis: Meshes from multiple views using differentiable rendering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8635–8644 (2022) 9
11. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. In: Proceedings of the International Conference on Machine Learning (ICML) (2020) 10
12. Groueix, T., Fisher, M., Kim, V.G., Russell, B., Aubry, M.: 3d-coded : 3d correspondences by deep deformation. In: European Conference on Computer Vision (ECCV) (2018) 4
13. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7297–7306 (2018) 4
14. Ham, B., Cho, M., Schmid, C., Ponce, J.: Proposal flow: Semantic correspondences from object proposals. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2016) 3, 5, 6

15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90 11
16. Hedlin, E., Sharma, G., Mahajan, S., He, X., Isack, H., Kar, A., Rhodin, H., Tagliasacchi, A., Yi, K.M.: Unsupervised keypoints from pretrained diffusion models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22820–22830 (2024) 3
17. Hedlin, E., Sharma, G., Mahajan, S., Isack, H., Kar, A., Tagliasacchi, A., Yi, K.M.: Unsupervised semantic correspondence using stable diffusion. Conference on Neural Information Processing Systems (NeurIPS) **36**, 8266–8279 (2023) 3
18. Jakab, T., Tucker, R., Makadia, A., Wu, J., Snavely, N., Kanazawa, A.: Keypointdeformer: Unsupervised 3d keypoint discovery for shape control. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12778–12787 (2021). https://doi.org/10.1109/CVPR46437.2021.01259 4
19. Jesslen, A., Dünkel, O., Kortylewski, A.: Geometry matters: 3d foundation priors for learning semantic correspondence (2026), https://arxiv.org/abs/2605.30093 4
20. Jesslen, A., Zhang, G., Wang, A., Ma, W., Yuille, A., Kortylewski, A.: Novum: Neural object volumes for robust object classification. In: European Conference on Computer Vision (ECCV) (2024), https://arxiv.org/abs/2305.14668 4
21. Jiang, W., Trulls, E., Hosang, J., Tagliasacchi, A., Yi, K.M.: COTR: Correspondence Transformer for Matching Across Images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 2, 4
22. Kim, H., Lang, I., Aigerman, N., Groueix, T., Kim, V.G., Hanocka, R.: Meshup: Multi-target mesh deformation via blended score distillation. In: International Conference on 3D Vision (3DV) (2025) 4
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015), https://arxiv.org/abs/1412.6980 11
24. Krishnan, A., Kundu, A., Maninis, K.K., Hays, J., Brown, M.: Omninocs: A unified nocs dataset and model for 3d lifting of 2d objects. In: European Conference on Computer Vision (ECCV) (2024) 2, 5
25. Kulkarni, N., Tulsiani, S., Gupta, A.: Canonical surface mapping via geometric cycle consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2202–2211 (2019). https://doi.org/10.1109/ICCV.2019.00229 4
26. Li, X., Han, K., Wan, X., Prisacariu, V.A.: Simsc: A simple framework for semantic correspondence with temperature learning. arXiv preprint arXiv:2305.02385 (2023) 3
27. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **33**(5), 978–994 (2011) 3
28. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015) 4, 12
29. Lou, Y., You, Y., Li, C., Cheng, Z., Li, L., Ma, L., Wang, W., Lu, C.: Human correspondence consensus for 3d object semantic understanding. In: European Conference on Computer Vision (ECCV). p. 496–512. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58542-6_30, https://doi.org/10.1007/978-3-030-58542-6_30 4, 6

30. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (IJCV) **60**(2), 91–110 (2004) 3
31. Mariotti, O., Mac Aodha, O., Bilen, H.: Improving semantic correspondence with viewpoint-guided spherical maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19521–19530 (2024) 2, 4
32. Min, J., Lee, J., Ponce, J., Cho, M.: Spair-71k: A large-scale benchmark for semantic correspondence (2019), https://arxiv.org/abs/1908.10543 2, 3, 5, 6
33. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L., Su, H.: 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) 5
34. Nam, J., Lee, G., Kim, S., Kim, H., Cho, H., Kim, S., Kim, S.: Diffusion model for dense matching. In: International Conference on Learning Representations (ICLR). OpenReview.net (2024), https://openreview.net/forum?id=Zsfiqpft6K 2, 4
35. Neverova, N., Novotny, D., Khalidov, V., Szafraniec, M., Labatut, P., Vedaldi, A.: Continuous surface embeddings for deformable shape correspondence. Conference on Neural Information Processing Systems (NeurIPS) (2020) 4
36. Novotny, D., Ravi, N., Graham, B., Neverova, N., Vedaldi, A.: C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 4
37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023) 3, 11, 12
38. Pham, N., Jesslen, A., Schiele, B., Kortylewski, A., Fischer, J.: Interpretable 3d neural object volumes for robust conceptual reasoning. In: International Conference on Learning Representations (ICLR) (2026), https://openreview.net/forum?id=VSPLa2Sito 4
39. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: 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) 5
40. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. Conference on Neural Information Processing Systems (NeurIPS) **34**, 6087–6101 (2021) 9
41. Shtedritski, A., Rupprecht, C., Vedaldi, A.: Shic: Shape-image correspondences with no keypoint supervision. In: European Conference on Computer Vision (ECCV) (2024) 4
42. Sommer, L., Dünkel, O., Theobalt, C., Kortylewski, A.: 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). pp. 6468–6479 (June 2025) 4, 12, 13, 14
43. Sommer, L., Jesslen, A., Ilg, E., Kortylewski, A.: Unsupervised learning of category-level 3d pose from object-centric videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22787–22796 (2024) 4
44. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2, 4

45. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3d: Dataset and methods for single-image 3d shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 5
46. Sun, Y., Huang, Y., Guo, H., Zhao, Y., Wu, R., Yu, Y., Ge, W., Zhang, W.: Misc210k: A large-scale dataset for multi-instance semantic correspondence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7121–7130 (2023) 5, 6
47. Suwajanakorn, S., Snavely, N., Tompson, J., Norouzi, M.: Discovery of latent 3d keypoints via end-to-end geometric reasoning. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Conference on Neural Information Processing Systems (NeurIPS). pp. 2063–2074 (2018), https://proceedings.neurips.cc/paper/2018/hash/24146db4eb48c718b84cae0a0799dcfc-Abstract.html 7
48. Tian, M., Ang, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: European Conference on Computer Vision (ECCV). p. 530–546. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58589-1_32, https://doi.org/10.1007/978-3-030-58589-1_32 4
49. Tola, E., Lepetit, V., Fua, P.: Daisy: An efficient dense descriptor applied to wide baseline stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **32**, 815–30 (05 2010). https://doi.org/10.1109/TPAMI.2009.77 3
50. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: Cub-200-2011 (Apr 2022). https://doi.org/10.22002/D1.20098 5, 6
51. Wandel, K., Wang, H.: 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). pp. 1138–1147 (2025). https://doi.org/10.1109/CVPR52734.2025.00114 4
52. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: 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). pp. 2642–2651 (2019) 2, 5, 11, 12
53. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: European Conference on Computer Vision (ECCV) (2018) 4
54. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: Deepflow: Large displacement optical flow with deep matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1385–1392 (2013) 3
55. Wu, S., Jakab, T., Rupprecht, C., Vedaldi, A.: DOVE: Learning deformable 3d objects by watching videos. International Journal of Computer Vision (IJCV) (2023) 5, 6
56. Wu, S., Li, R., Jakab, T., Rupprecht, C., Vedaldi, A.: MagicPony: Learning articulated 3d animals in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 5, 9, 10, 11, 12
57. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 5
58. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21469–21480 (2025) 4

59. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). pp. 75–82. IEEE (2014) 5
60. Xu, J., Zhang, Y., Peng, J., Ma, W., Jesslen, A., Ji, P., Hu, Q., Zhang, J., Liu, Q., Wang, J., Ji, W., Wang, C., Yuan, X., Kaushik, P., Zhang, G., Liu, J., Xie, Y., Cui, Y., Yuille, A., Kortylewski, A.: Animal3d: A comprehensive dataset of 3d animal pose and shape. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023) 4
61. Xu, Z., He, Z., Wu, J., Song, S.: Learning 3d dynamic scene representations for robot manipulation. In: Conference on Robot Learning (CoRL) (2020) 3
62. Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Chang, H., Ramanan, D., Freeman, W.T., Liu, C.: Lasr: Learning articulated shape reconstruction from a monocular video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15980–15989 (2021) 9
63. Yi, L., Kim, V.G., Ceylan, D., Shen, W., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L.: A scalable active framework for region annotation in 3d shape collections. In: ACM Trans. Graphics (Proc. SIGGRAPH Asia) (2016) 5
64. Yifan, W., Aigerman, N., Kim, V.G., Chaudhuri, S., Sorkine-Hornung, O.: Neural cages for detail-preserving 3d deformations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 72–80 (2020). https://doi.org/10.1109/CVPR42600.2020.00015 4
65. Yildirim, I., Siegel, M.H., Soltani, A.A., Ray Chaudhuri, S., Tenenbaum, J.B.: Perception of 3D shape integrates intuitive physics and analysis-by-synthesis. Nature Human Behaviour **8**(2), 320–335 (Feb 2024) 3
66. You, Y., Li, C., Lou, Y., Cheng, Z., Li, L., Ma, L., Wang, W., Lu, C.: 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). https://doi.org/10.1109/TPAMI.2021.3072659 4
67. You, Y., Lou, Y., Li, C., Cheng, Z., Li, L., Ma, L., Lu, C., Wang, W.: 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). pp. 13644–13653 (2020). https://doi.org/10.1109/CVPR42600.2020.01366 4, 6
68. Zhang, J., Huang, W., Peng, B., Wu, M., Hu, F., Chen, Z., Zhao, B., Dong, H.: Omni6dpose: Large-scale multi-object 6d pose estimation with realistic rendering. In: European Conference on Computer Vision (ECCV). pp. 3110–3120 (2024) 2, 4, 5, 6, 7, 8, 11, 12, 14
69. Zhang, J., Herrmann, C., Hur, J., Cabrera, L.P., Jampani, V., Sun, D., Yang, M.H.: A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. In: Conference on Neural Information Processing Systems (NeurIPS) (2023) 3
70. Zheng, Z., Yu, T., Dai, Q., Liu, Y.: Deep implicit templates for 3d shape representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1429–1439 (2021) 9, 10
71. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. International Conference on Learning Representations (ICLR) (2022) 3
72. Zhu, J., Ju, Y., Zhang, J., Wang, M., Yuan, Z., Hu, K., Xu, H.: Densematcher: Learning 3d semantic correspondence for category-level manipulation from a single demo. International Conference on Learning Representations (ICLR) (2025) 4, 5, 6