

3D Congealing: 3D-Aware Image Alignment in the Wild

Yunzhi Zhang¹, Zizhang Li¹, Amit Raj², Andreas Engelhardt³, Yuanzhen Li²,
Tingbo Hou², Jiajun Wu¹, and Varun Jampani⁴

¹ Stanford University

² Google Research

³ University of Tübingen

⁴ Stability AI

Abstract. We propose 3D Congealing, a novel problem of 3D-aware alignment for 2D images capturing semantically similar objects. Given a collection of unlabeled Internet images, our goal is to associate the shared semantic parts from the inputs and aggregate the knowledge from 2D images to a shared 3D canonical space. We introduce a general framework that tackles the task without assuming shape templates, poses, or any camera parameters. At its core is a canonical 3D representation that encapsulates geometric and semantic information. The framework optimizes for the canonical representation together with the pose for each input image, and a per-image coordinate map that warps 2D pixel coordinates to the 3D canonical frame to account for the shape matching. The optimization procedure fuses prior knowledge from a pre-trained image generative model and semantic information from input images. The former provides strong knowledge guidance for this under-constraint task, while the latter provides the necessary information to mitigate the training data bias from the pre-trained model. Our framework can be used for various tasks such as pose estimation and image editing, achieving strong results on real-world image datasets under challenging illumination conditions and on in-the-wild online image collections. Project page at <https://ai.stanford.edu/~yzzhang/projects/3d-congealing/>.

1 Introduction

We propose the task of *3D Congealing*, where the goal is to align a collection of images containing semantically similar objects into a shared 3D space. Specifically, we aim to obtain a canonical 3D representation together with the pose and a dense map of 2D-3D correspondence for each image in the collection. The input images may contain object instances belonging to a similar category with varying shapes and textures, and are captured under distinct camera viewpoints and illumination conditions, which all contribute to the pixel-level difference as shown in Figure 1. Despite such inter-image differences, humans excel at aligning such



Fig. 1: Objects with different shapes and appearances, such as these sculptures, may share similar semantic parts and a similar geometric structure. We study 3D Congealing, inferring and aligning such a shared structure from an unlabeled image collection. Such alignment can be used for tasks such as pose estimation and image editing. See Appendix A for full results.

images with one another in a geometrically and semantically consistent manner based on their 3D-aware understanding.

Obtaining a canonical 3D representation and grounding input images to the 3D canonical space enable several downstream tasks, such as 6-DoF object pose estimation, pose-aware image filtering, and image editing. Unlike the task of 2D congealing [11, 29, 31], where the aim is to align the 2D pixels across the images, 3D Congealing requires aggregating the information from the image collection altogether and forming the association among images in 3D. The task is also closely related to 3D reconstruction from multiview images, with a key distinction in the problem setting, as inputs here do not necessarily contain identical objects but rather semantically similar ones. Such a difference opens up the possibility of image alignment from readily available image collections on the Internet, *e.g.*, online search results, landmark images, and personal photo collections.

3D Congealing represents a challenging problem, particularly for arbitrary images without camera pose or lighting annotations, even when the input images contain identical objects [1, 4, 20, 44], because the solutions for pose and shape are generally entangled. On the one hand, the definition of poses is specific to the coordinate frame of the shape; on the other hand, the shape optimization is typically guided by the pixel-wise supervision of images under the estimated poses. To overcome the ambiguity in jointly estimating poses and shapes, prior works mostly start from noisy pose initializations [20], data-specific initial pose distributions [25, 44], or rough pose annotations such as pose quadrants [1]. They then perform joint optimization for a 3D representation using an objective of reconstructing input image pixels [1, 20, 44] or distribution matching [25].

In this work, instead of relying on initial poses as starting points for shape reconstruction, we propose to tackle the joint optimization problem from a different perspective. We first obtain a plausible 3D shape that is compliant with the input image observations using pre-trained generative models, and then use semantic-aware visual features, *e.g.*, pre-trained features from DINO [2, 30] and Stable-Diffusion [36], to register input images to the 3D shape. Compared to

photometric reconstruction losses, these features are more tolerant of variance in object identities among image inputs.

We make deliberate design choices to instantiate such a framework that fuses the knowledge from pre-trained text-to-image (T2I) generative models with real image inputs. First, to utilize the prior knowledge from generative models, we opt to apply a T2I personalization method, Textual Inversion [7], which aims to find the most suitable text embedding to reconstruct the input images via the pre-trained model. Furthermore, a semantic-aware distance is proposed to mitigate the appearance discrepancy between the rendered image and the input photo collection. Finally, a canonical coordinate mapping is learned to find the correspondence between 3D canonical representation and 2D input images.

To prove the effectiveness of the proposed framework, we compare the proposed method against several baselines on the task of pose estimation on a dataset with varying illuminations and show that our method surpasses all the baselines significantly. We also demonstrate several applications of the proposed method, including image editing and object alignment on web image data.

In summary, our contributions are:

1. We propose a novel task of 3D Congealing that involves aligning images of semantically similar objects in a shared 3D space.
2. We develop a framework tackling the proposed task and demonstrate several applications using the obtained 2D-3D correspondence, such as pose estimation and image editing.
3. We show the effectiveness and applicability of the proposed method on a diverse range of in-the-wild Internet images.

2 Related Works

Image Congealing. Image congealing methods [12, 13, 18, 27] aim to align a collection of input images based on the semantic similarity of parts. To tackle this task, Neural Congealing [29] proposes to use neural atlases, which are 2D feature grids, to capture common semantic features from input images and recover a dense mapping between each input image and the learned neural atlases. GANgealing [31] proposes to use a spatial transformer to map a randomly generated image from a GAN [8] to a jointly aligned space. These 2D-warping-based methods are typically applied to source and target image pairs with no or small camera rotation, and work best on in-plane transformation, while our proposed framework handles a *larger variation of viewpoints* due to 3D reasoning.

Object Pose Estimation. Object pose estimation aims to estimate the pose of an object instance with respect to the coordinate frame of its 3D shape. Classical methods for pose estimation recover poses from multi-view images using pixel- or feature-level matching to find the alignment between different images [38]. These methods are less suitable in the in-the-wild setting due to the increasing variance among images compared to multi-view captures. Recent methods proposed to tackle this task by training a network with pose annotations as supervision [19, 42, 48], but it remains challenging for these methods to generalize beyond the

training distribution. Another class of methods propose to use an analysis-by-synthesis framework to estimate pose given category-specific templates [3] or a pre-trained 3D representation [46]; these assumptions make it challenging to apply these methods to generic objects in the real world. ID-Pose [5] leverages Zero-1-to-3 [21], a view synthesis model, and optimizes for the relative pose given a source and a target image. Goodwin *et al.* [9] use pre-trained self-supervised features for matching, instead of doing it at the pixel level, but require both RGB and depth inputs; in contrast, we assume access to only RGB images.

Shape Reconstruction from Image Collections. Neural rendering approaches [26, 43, 45] use images with known poses to reconstruct the 3D shape and appearance from a collection of multiview images. The assumptions of known poses and consistent illumination prevent these methods from being applied in the wild. Several works have extended these approaches to relax the pose assumption, proposing to handle noisy or unknown camera poses of input images through joint optimization of poses and 3D representation [4, 20, 44]. SAMURAI [1] also proposes to extend the NeRF representation to handle scenes under various illuminations, provided with coarse initial poses in the form of pose quadrant annotations. Unlike these methods, we do not have any assumption on the camera pose of input images, and handle image inputs with variations in illumination conditions. Moreover, our framework allows for aligning *multiple objects* of the same category with geometric variations into a common coordinate frame.

3D Distillation from 2D Diffusion Models. Recently, text-to-image diffusion models have shown great advancement in 2D image generation, and DreamFusion [32] has proposed to distill the prior on 2D images from pre-trained text-to-image models to obtain text-conditioned 3D representations. Other methods have extended this idea to optimize for 3D assets conditioned on image collections [33]. DreamFusion uses a full 3D representation, while Zero-1-to-3 [21] extracts the 3D knowledge using view synthesis tasks. Specifically, it finetunes Stable-Diffusion [36] on synthetic data from Objaverse [6] to generate a novel view conditioned on an image and a relative pose. The fine-tuned model can be further combined with gradients proposed in DreamFusion to obtain full 3D assets. MVDream [39] extends the idea of fine-tuning with view synthesis tasks, but proposes to output 4 views at once and has shown better view consistency in the final 3D asset outputs. DreamBooth3D [33] proposed to utilize fine-tuned diffusion model [37] for the image-conditioned 3D reconstruction task. Their goal is to faithfully reconstruct the 3D shape; therefore, the model requires a multi-stage training that involves applying diffusion model guidance and photometric-reconstruction-based updates. These works provide a viable solution for 3D reconstruction from image collections, but they do not *ground the input images* to the 3D space as in ours.

3 Method

Problem Formulation. We formulate the problem of 3D Congealing as follows. Given a set of N object-centric images $\mathcal{D} = \{x_n\}_{n=1}^N$ that captures objects sharing

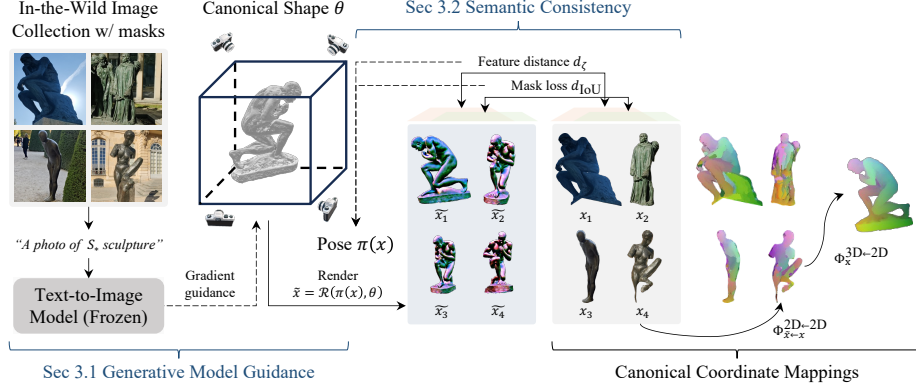


Fig. 2: Pipeline. Given a collection of in-the-wild images capturing similar objects as inputs, we develop a framework that “congeals” these images in 3D. The core representation consists of a canonical 3D shape that captures the geometric structure shared among the inputs, together with a set of coordinate mappings that register the input images to the canonical shape. The framework utilizes the prior knowledge of plausible 3D shapes from a generative model, and aligns images in the semantic space using pre-trained semantic feature extractors.

semantic components, *e.g.*, objects from one category, we seek to align the object instances in these images into a canonical 3D representation, *e.g.*, NeRF [26], parameterized by θ . We refer to the coordinate frame of this 3D representation as the canonical frame. We also recover the camera pose of each observation $x \in \mathcal{D}$ in the canonical frame, denoted using a pose function $\pi : x \mapsto (\xi, \kappa)$ where ξ represents the object pose in $SE(3)$ and κ is the camera intrinsic parameters. We assume access to instance masks, which can be easily obtained using an off-the-shelf segmentation method [16].

The 3D representation should be consistent with the physical prior of objects in the natural world, and with input observations both geometrically and semantically. These constraints can be translated into an optimization problem:

$$\max_{\pi, \theta} p_{\Theta}(\theta), \text{ s.t. } x = \mathcal{R}(\pi(x), \theta), \forall x \in \mathcal{D}, \quad (1)$$

where p_{Θ} is a prior distribution for the 3D representation parameter θ that encourages physically plausible solutions, \mathcal{R} is a predefined rendering function that enforces geometric consistency, and the equality constraint on image reconstruction enforces compliance with input observations.

In the following sections, we will describe an instantiation of the 3D prior p_{Θ} (Sec. 3.1), an image distance function that helps enforce the equality constraint (Sec. 3.2), followed by the 3D Congealing optimization (Sec. 3.3).

3.1 3D Guidance from Generative Models

As illustrated in the left part of Figure 2, we extract the prior knowledge for 3D representations $p_{\Theta}(\cdot)$ from a pre-trained text-to-image (T2I) model such as

Stable-Diffusion [36]. DreamFusion [32] proposes to turn a text prompt y into a 3D representation θ using the following Score Distillation Sampling (SDS) objective, leveraging a T2I diffusion model with frozen parameters ϕ ,

$$\min_{\theta} \mathbb{E}_{x \in \mathcal{D}(\theta)} \mathcal{L}_{\text{diff}}^{\phi}(x, y). \quad (2)$$

Here $\mathcal{D}(\theta) := \{\mathcal{R}(\pi, \theta) \mid \pi \sim p_{\Pi}(\cdot)\}$ contains images rendered from the 3D representation θ under a prior camera distribution $p_{\Pi}(\cdot)$, and $\mathcal{L}_{\text{diff}}^{\phi}$ is the training objective of image diffusion models specified as follows:

$$\mathcal{L}_{\text{diff}}^{\phi}(x, y) := \mathbb{E}_{t \sim \mathcal{U}([0,1]), \epsilon \sim \mathcal{N}(\mathbf{0}, I)} [\omega(t) \|\epsilon_{\phi}(\alpha_t x + \sigma_t \epsilon, y, t) - \epsilon\|_2^2], \quad (3)$$

where ϵ_{ϕ} is the pre-trained denoising network, $\omega(\cdot)$ is the timestep-dependent weighting function, t is the diffusion timestep and α_t, σ_t are timestep-dependent coefficients from the diffusion model schedule.

The above loss can be used to guide the optimization of a 3D representation θ , whose gradient is approximated by

$$\nabla_{\theta} \mathcal{L}_{\text{diff}}^{\phi}(x = \mathcal{R}(\xi, \kappa, \theta), y) \approx \mathbb{E}_{t, \epsilon} \left[\omega(t) (\epsilon_{\phi}(\alpha_t x + \sigma_t \epsilon, y, t) - \epsilon) \frac{\partial x}{\partial \theta} \right], \quad (4)$$

where ξ and κ are the extrinsic and intrinsic camera parameters, respectively. The derived gradient approximation is adopted by later works such as MVDream [39], which we use as the backbone.

The original SDS objective is optimizing for a text-conditioned 3D shape with a user-specified text prompt y and does not consider image inputs. Here, we use the technique from Textual Inversion [7] to recover the most suitable text prompt y^* that explains input images, defined as follows:

$$y^* = \arg \min_y \mathbb{E}_{x \in \mathcal{D}} \mathcal{L}_{\text{diff}}^{\phi}(x, y). \quad (5)$$

Eq. (2) and Eq. (5) differ in that both the sources of the observations x (an infinite dataset of rendered images $\mathcal{D}(\theta)$ for the former, and real data \mathcal{D} for the latter) and the parameters being optimized over (θ and y , respectively). In our framework, we incorporate the real image information to the SDS guidance via first solving for y^* (Eq. (5)) and keep it frozen when optimizing for θ (Eq. (2)). The diffusion model parameter ϕ is frozen throughout the process, requiring significantly less memory compared to the alternative of integrating input image information via finetuning ϕ as in DreamBooth3D [33].

3.2 Semantic Consistency from Deep Features

The generative model prior from Sec. 3.1 effectively constrains the search space for the solutions. However, the objectives from Eqs. (2) and (5) use the input image information only indirectly, via a text embedding y^* . To explain the relative geometric relation among input images, we explicitly recover the pose of each input image w.r.t. θ , as illustrated in Figure 2 (middle) and as explained below.

To align input images, we use an image distance metric defined by semantic feature dissimilarity. In particular, pre-trained deep models such as DINO [2, 30] have been shown to be effective semantic feature extractors. Denote such a model as f parameterized by ζ . The similarity of two pixel locations u_1 and u_2 from two images x_1 and x_2 , respectively, can be measured with

$$d_{\zeta}^{u_1, u_2}(x_1, x_2) := 1 - \frac{\langle [f_{\zeta}(x_1)]_{u_1}, [f_{\zeta}(x_2)]_{u_2} \rangle}{\| [f_{\zeta}(x_1)]_{u_1} \|_2 \| [f_{\zeta}(x_2)]_{u_2} \|_2}, \quad (6)$$

where $[\cdot]$ is an indexing operator. It thereafter defines an image distance function

$$\|x_1 - x_2\|_{d_{\zeta}} := \frac{1}{HW} \sum_u d_{\zeta}^{u, u}(x_1, x_2), \quad (7)$$

where x_1 and x_2 have resolution $H \times W$, and the sum is over all image coordinates.

The choice of semantic-aware image distance, instead of photometric differences as in the classical problem setting of multiview 3D reconstruction [38, 43, 45], leads to solutions that maximally align input images to the 3D representation with more tolerance towards variance in object shape, texture, and environmental illuminations among input images, which is crucial in our problem setting.

3.3 Optimization

The Canonical Shape and Image Poses. Combining Secs. 3.1 and 3.2, we convert the original problem in Eq. (1) into

$$\min_{\pi, \theta} \underbrace{\mathbb{E}_{x \in \mathcal{D}(\theta)} \mathcal{L}_{\text{diff}}^{\phi}(x, y^*)}_{\text{generative model guidance}} + \lambda \underbrace{\mathbb{E}_{x \in \mathcal{D}} \|\mathcal{R}(\pi(x), \theta) - x\|_d}_{\text{data reconstruction}}, \quad (8)$$

where y^* come from Eq. (5) and λ is a loss weight. Compared to Eq. (5), here the first term instantiates the generative modeling prior and the second term is a soft constraint of reconstructing input observations. Specifically, $d = \lambda_{\zeta} d_{\zeta} + \lambda_{\text{IoU}} d_{\text{IoU}}$, where d_{ζ} is the semantic-space distance metric from Sec. 3.2, and d_{IoU} is the Intersection-over-Union (IoU) loss for masks, $\|m_1 - m_2\|_{d_{\text{IoU}}} := 1 - (\|m_1 \odot m_2\|_1) / (\|m_1\|_1 + \|m_2\|_1 - \|m_1 \odot m_2\|_1)$, where m_1 and m_2 are image masks, which in Eq. (8) are set to be the mask rendering and the instance mask for x . The use of both d_{ζ} and d_{IoU} tolerates shape variance among input instances.

For the shape representation, we follow NeRF [26] and use neural networks $\sigma_{\theta} : \mathbb{R}^3 \rightarrow \mathbb{R}$ and $c_{\theta} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ to map a 3D spatial coordinate to a density and an RGB value, respectively. The rendering operation \mathcal{R} is the volumetric rendering operation specified as follows:

$$\mathcal{R}(r, \xi, \theta; c_{\theta}) = \int T(t) \sigma_{\theta}(\xi r(t)) c_{\theta}(\xi r(t)) dt, \quad (9)$$

where $T(t) = \exp(-\int \sigma_{\theta}(r(t')) dt')$, $r : \mathbb{R} \rightarrow \mathbb{R}^3$ is a ray shooting from the camera center to the image plane, parameterized by the camera location and the ray's direction, and ξ is the relative pose that transforms the ray from the camera frame to the canonical frame.

Forward Canonical Coordinate Mappings. After the above optimization, each image x from the input image collection can be “congealed” to the shape θ via a *canonical coordinate mapping*, i.e., a forward warping operation $\Phi_x^{\text{fwd}} : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ that maps a 2D image coordinate to a 3D coordinate in the canonical frame of reference as illustrated in Figure 2. Φ_x^{fwd} consists of the following two operations.

First, we warp a coordinate u from the real image x to the rendering of the canonical shape under its pose $\pi(x)$, denoted as $\tilde{x} := \mathcal{R}(\pi(x), \theta)$. Specifically,

$$\Phi_{\tilde{x} \leftarrow x}^{2\text{D} \leftarrow 2\text{D}}(u) := \arg \min_{\tilde{u}} d_{\zeta}^{\tilde{u}, u}(\tilde{x}, x) + \lambda_{\ell_2} \|\tilde{u} - u\|_2^2 + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}(\tilde{u}, u), \quad (10)$$

where d_{ζ} follows Eq. (6), the 2D coordinates u and \tilde{u} are normalized into range $[0, 1]$ before computing the ℓ_2 norm, the smoothness term $\mathcal{L}_{\text{smooth}}$ is specified in Appendix B, and λ_{ℓ_2} and λ_{smooth} are scalar weights. This objective searches for a new image coordinate \tilde{u} (from the rendering \tilde{x}) that shares a semantic feature similar to u (from the real image x), and ensures that \tilde{u} stays in the local neighborhood of u via a soft constraint of the coordinate distance. Afterward, a 2D-to-3D operation takes in the warped coordinate from above and outputs its 3D location in the normalized object coordinate space (NOCS) [41] of θ :

$$\Phi_x^{3\text{D} \leftarrow 2\text{D}}(\tilde{u}) := [\mathcal{R}_{\text{NOCS}}(\pi(x), \theta)]_{\tilde{u}}, \quad (11)$$

where $\mathcal{R}_{\text{NOCS}}$ is identical to \mathcal{R} from Eq. (9), but replacing the color field c_{θ} with a canonical object coordinate field, $c_{\text{NOCS}} : \mathbb{R}^3 \rightarrow \mathbb{R}^3, p \mapsto (p - p_{\min}) / (p_{\max} - p_{\min})$, where p_{\min} and p_{\max} are the two opposite corners of the canonical shape’s bounding box. These bounding boxes are determined by the mesh extracted from the density neural field σ_{θ} using the Marching Cube [22] algorithm.

Combining the above, given an input image coordinate u , $\Phi_x^{\text{fwd}}(u) := \Phi_x^{3\text{D} \leftarrow 2\text{D}} \circ \Phi_{\tilde{x} \leftarrow x}^{2\text{D} \leftarrow 2\text{D}}(u)$ identifies a 3D location in the canonical frame corresponding to u .

Reverse Canonical Coordinate Mappings. Each image can be “uncongealed” from the canonical shape using $\Phi_x^{\text{rev}} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$, which is the reverse operation of $\Phi_x^{\text{fwd}}(u)$ and is approximately computed via nearest-neighbor inversion as explained below.

Given a 3D location within a unit cube, $p \in [0, 1]^3$, $\Phi_x^{\text{rev}}(p) := \Phi_{\tilde{x} \leftarrow x}^{2\text{D} \leftarrow 2\text{D}} \circ \Phi_x^{2\text{D} \leftarrow 3\text{D}}(p)$. In particular,

$$\Phi_x^{2\text{D} \leftarrow 3\text{D}}(p) := \arg \min_{\tilde{u}} \|p - \Phi_x^{3\text{D} \leftarrow 2\text{D}}(\tilde{u})\|_2 \quad (12)$$

is an operation that takes in a 3D coordinate p in the canonical frame and searches for a 2D image coordinate whose NOCS value is the closest to p , and $\Phi_{\tilde{x} \leftarrow x}^{2\text{D} \leftarrow 2\text{D}}$ is computed via inverting $\Phi_{\tilde{x} \leftarrow x}^{2\text{D} \leftarrow 2\text{D}}$ from Eq. (10),

$$\Phi_{\tilde{x} \leftarrow x}^{2\text{D} \leftarrow 2\text{D}}(\tilde{u}) := \arg \min_u \|\tilde{u} - \Phi_{\tilde{x} \leftarrow x}^{2\text{D} \leftarrow 2\text{D}}(u)\|_2. \quad (13)$$

In summary, the above procedure establishes the 2D-3D correspondence between an input image x and the canonical shape via Φ_x^{fwd} , and defines the dense 2D-2D correspondences between two images x_1, x_2 via $\Phi_{x_2}^{\text{rev}} \circ \Phi_{x_1}^{\text{fwd}}$ which enables image editing (Figure 8). The full framework is described in Algorithm 1.

3.4 Implementation Details

Input images are cropped with the tightest bounding box around the foreground masks. The masks come from dataset annotations, if available, or from Grounded-SAM [16, 35], an off-the-shelf segmentation model, for all Internet images.

Across all experiments, we optimize for y^* (Algorithm 1, line 2) for 1,000 iterations using an AdamW [23] optimizer with learning rate 0.02 and weight decay 0.01. We optimize for θ (line 3) with $\lambda = 0$ for 10,000 iterations, with AdamW and learning rate 0.001. The NeRF model θ has 12.6M parameters. It is frozen afterwards and defines the coordinate frame for poses.

Since directly optimizing poses and camera parameters with gradient descents easily falls into local minima [20], we initialize π using an analysis-by-synthesis approach (Algorithm 1, line 5-7). Specifically, we parameterize the camera intrinsics using a pinhole camera model with a scalar Field-of-View (FoV) value, and sample the camera parameter (ξ, κ) from a set of candidates determined by an exhaustive combination of 3 FoV values, 16 azimuth values, and 16 elevation values uniformly sampled from $[15^\circ, 60^\circ]$, $[-180^\circ, 180^\circ]$, and $[-90^\circ, 90^\circ]$, respectively. In this pose initialization stage, all renderings use a fixed camera radius and are cropped with the tightest bounding boxes, computed using the rendered masks, before being compared with the real image inputs. We then select the candidate with the lowest error measured by the image distance function d from Sec. 3.2, with $\lambda_\zeta = 1$ and $\lambda_{\text{IoU}} = 0$.

After pose initialization, we use the $\mathfrak{se}(3)$ Lie algebra for camera extrinsics parameterization following BARF [20], and optimize for the extrinsics and intrinsics of each input image (Algorithm 1, line 8), with $\lambda_\zeta = 0$ and $\lambda_{\text{IoU}} = 1$, for 1,000 iterations with the Adam [15] optimizer and learning rate 0.001. Since θ is frozen, the optimization effectively only considers the second term from Eq. (8). Finally, to optimize for the canonical coordinate mappings (Algorithm 1, line 9), for each input image, we run 4,000 iterations for Eq. (10) with AdamW and learning rate 0.01. All experiments are run on a single 24GB A5000 GPU.

4 Experiments

In this section, we first benchmark the pose estimation performance of our method on in-the-wild image captures (Sec. 4.1), and then show qualitative results on diverse input data and demonstrate applications such as image editing (Sec. 4.2).

4.1 Pose Estimation

Dataset. While our method does not require input object instances to be identical, we use a dataset of multi-illumination object-centric image captures

```

1: procedure RUN( $\mathcal{D} = \{x_n\}_{n=1}^N$ )
2:    $y^* \leftarrow$  Solution to Eq. (5)
3:   Optimize  $\theta$  with Eq. (8)
4:   Sample pose candidates  $\{\xi_i\}_i$ 
5:   for  $n \leftarrow 1$  to  $N$  do  $\triangleright$  Pose initialization
6:      $\pi(x_n) \leftarrow \arg \min_{\xi_i} \|\mathcal{R}(\xi, \theta) - x_n\|_{d_\zeta}$ 
7:   end for
8:   Optimize  $\pi(x_n)$  with Eq. (8) for all  $n$ 
9:   Determine  $\Phi_{x_n}^{\text{fwd}}$  and  $\Phi_{x_n}^{\text{rev}}$  for all  $n$ 
10:  return  $\theta, \pi, \{\Phi_{x_n}^{\text{fwd}}\}_{n=1}^N, \{\Phi_{x_n}^{\text{rev}}\}_{n=1}^N$ 
11: end procedure

```

Algorithm 1: Overview.

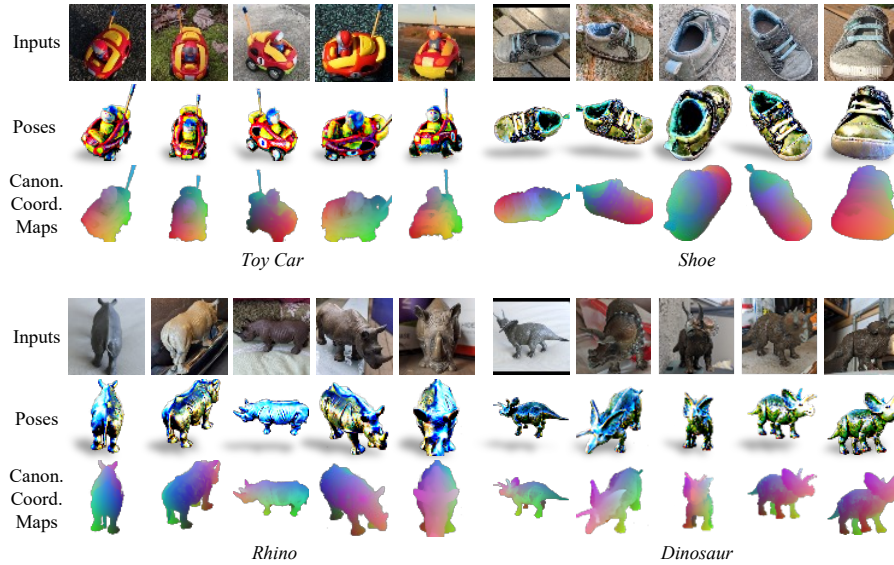


Fig. 3: Pose Estimation from Multi-Illumination Captures. The figure shows 4 example scenes from the NAVI dataset, displaying the real image inputs, canonical shapes under estimated poses, and the canonical coordinate maps.

with ground truth camera poses for evaluation. Specifically, we use the in-the-wild split of the NAVI [14] dataset, which contains 35 object image collections in its official release. Each image collection contains an average of around 60 casual image captures of an object instance placed under different illumination conditions, backgrounds and cameras with ground truth poses.

We use identical hyperparameters for all scenes. We do not introduce additional semantic knowledge for objects contained in the scene and use a generic text prompt, “a photo of sks object”, for initialization for all scenes. The text embeddings corresponding to the tokens for “sks object” are being optimized using Eq. (5) with the embeddings for others being frozen. For each scene, it takes around 1 hr to optimize for the NeRF model, 15 min for pose initialization, and 45 min for pose optimization.

Baselines. We compare with several multiview reconstruction baselines. In particular, NeROIC [17] uses the poses from COLMAP, and NeRS [47] and SAMURAI [1] require initial camera directions. GNeRF [25] is a pose-free multiview 3D reconstruction method that is originally designed for single-illumination scenes, and is adapted as a baseline using the same input assumption as ours. PoseDiffusion [42] is a learning-based framework that predicts relative object poses, using ground truth pose annotations as training supervision. The original paper takes a model pre-trained on CO3D [34] and evaluates the pose prediction performance in the wild, and we use the same checkpoint for evaluation.

Metrics. The varying illuminations pose challenges to classical pose estimation methods such as COLMAP [38]. We use the official split of the data which

Labels	Methods	Rotation $^{\circ}\downarrow$		Translation \downarrow	
		S_C	$\sim S_C$	S_C	$\sim S_C$
Pose	NeROIC [17]	42.11	-	0.09	-
	NeRS [47]	122.41	123.63	0.49	0.52
	SAMURAI [1]	26.16	36.59	0.24	0.35
None	GNeRF [25]	93.15	80.22	1.02	1.04
	PoseDiffusion [42]	46.79	46.34	0.81	0.90
	Ours (3 seeds)	26.97\pm2.24	32.56\pm2.90	0.40\pm0.01	0.41\pm0.04
	Ours (No Pose Init)	53.45	57.87	0.97	0.96
	Ours (No IoU Loss)	31.29	31.15	0.87	0.85

Table 1: Pose Estimation from Multi-Illumination Image Captures. Our method performs better than both GNeRF and PoseDiffusion with the same input information, and on par with SAMURAI which additionally assumes camera pose direction as inputs. Different random seeds lead to different canonical shapes, but our method is robust to such variations. \pm denotes means followed by standard deviations.

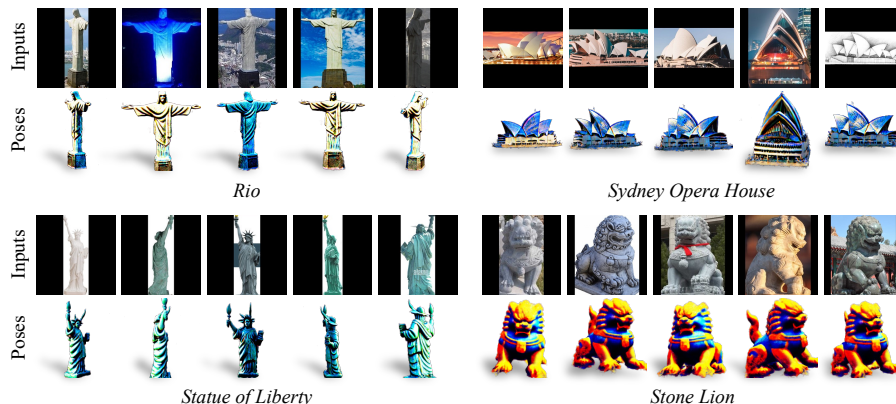


Fig. 4: Pose Estimation for Tourist Landmarks. This is a challenging problem setting due to the varying viewpoints and lighting conditions, and the proposed method can successfully align online tourist photos taken at different times and possibly at different geographical locations, into one canonical representation.

partitions the 35 scenes into 19 scenes where COLMAP converges (S_C in Table 1), and 16 scenes where COLMAP fails to converge ($\sim S_C$). Following [14], we report the absolute rotation and translation errors using Procrustes analysis [10], where for each scene, the predicted camera poses are aligned with the ground truth pose annotations using a global transformation before computing the pose metrics.

Results. Handling different illumination conditions is challenging for all baselines using photometric-reconstruction-based optimization [1, 17, 47] even with additional information for pose initialization. As shown in Table 1, our approach significantly outperforms both GNeRF and PoseDiffusion and works on par with SAMURAI which requires additional pose initialization. We run our full pipeline with 3 random seeds and observe a consistent performance across seeds. As shown in Figure 3 results, across a wide range of objects captured in this dataset, our

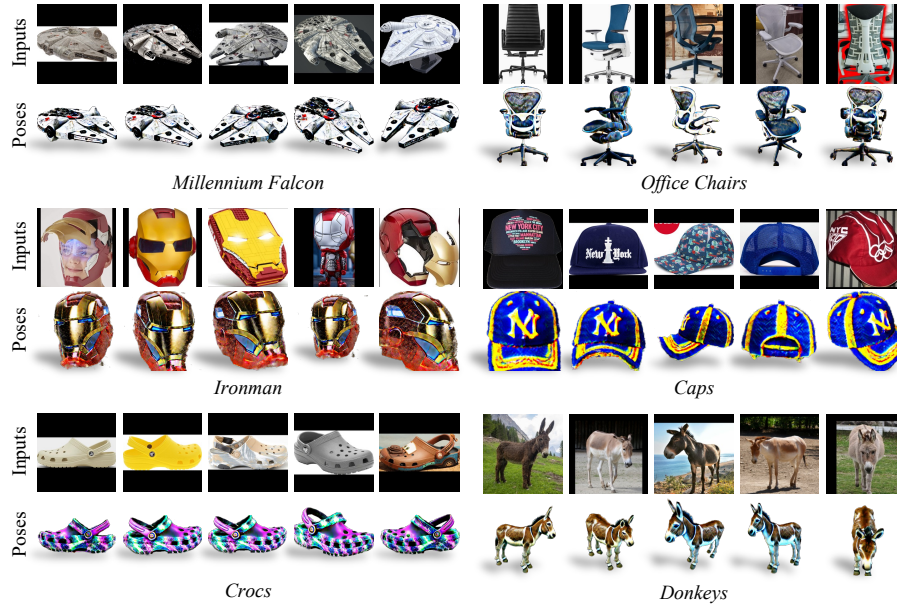


Fig. 5: Object Alignment from Internet Images. Results of an online image search may contain various appearances, identities, and articulated poses of the object. Our method can successfully associate these in-the-wild images with one shared 3D space.

method accurately estimates the poses for the input images and associates all inputs together in 3D via the canonical coordinate maps.

Ablations. Table 1 also shows ablation for the pose fitting objectives. The initialization is critical (“No Pose Init”), which is expected as pose optimization is susceptible to local optima [20]. “No IoU Loss”, which is equivalent to using the initialized poses as final predictions, also negatively affects the performance.

4.2 Applications

We show qualitative results on various in-the-wild image data. Inputs for Figures 4 and 5 are crawled with standard online image search engines and are CC-licensed, each consisting of 50 to 100 images. Inputs for Figures 6 and 7 come from the SPair-71k dataset [28]. We use identical hyperparameters for all datasets, except for text prompt initialization where we use a generic description of the object, *e.g.*, “a photo of sks sculpture”, or “a photo of cats plus dogs” for Figure 6.

Single-Instance. Figure 4 shows the result on Internet photos of tourist landmarks, which may contain a large diversity in illuminations (*e.g.*, the Rio) and styles (*e.g.*, the sketch of the Sydney Opera House). The proposed method can handle such variances and align these photos or art pieces, which are abundant from the Internet image database, to the same canonical 3D space and recover the relative camera poses.



Fig. 6: Cross-Category Results. The method can associate images from different categories, such as cats and dogs, by leveraging a learned average shape.



Fig. 7: Results on Deformable Objects. The method can be applied to images with highly diverse articulated poses and shapes as shown in the examples above.

Cross-Instance, Single-Category. Internet images from generic objects may contain more shape and texture variations compared to the landmarks. Figure 5 shows results for various objects. Our framework consistently infers a canonical shape from the input images to capture the shared semantic components being observed. For example, the corresponding semantic parts of faces of humans and the Ironman are identified as similar and are aligned with each other.

Cross-Category. The method does not make an assumption on the category of inputs. Given cross-category inputs, such as a mixture of cats and dogs as shown in Figure 6, the method effectively infers an average shape as an anchor to further reason about the relative relation among images from different categories.

Inputs with Deformable Shapes. To test the robustness of the method, we run the pipeline on images of humans with highly diverse poses. Figures 1 and 7 show that the method assigns plausible poses to the inputs despite the large diversity of shapes and articulated poses contained in the inputs.

Image Editing. The proposed method finds image correspondence and can be applied to image editing, as shown in Figure 8 (a-b). Figure 8 (c) shows that our method obtains more visually plausible results compared to the Nearest-Neighbor (NN) baseline using the same DINO features. The baseline matches features in 2D for each pixel individually and produce noisy results, as discussed in Appendix C. Quantitative evaluation of correspondence matching is included in Appendix D.

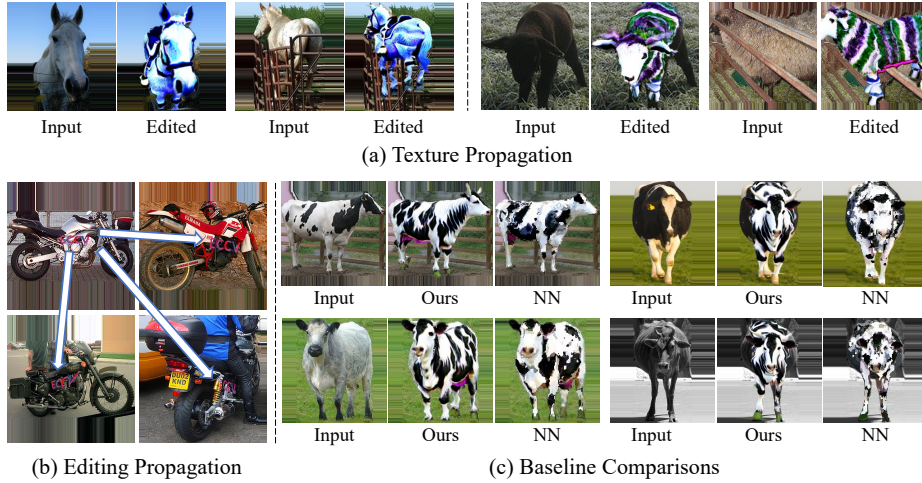


Fig. 8: Image Editing. Our method propagates (a) texture and (b) regional editing to real images, (c) achieving smoother results compared to the nearest-neighbor (NN) baseline thanks to the 3D geometric reasoning.

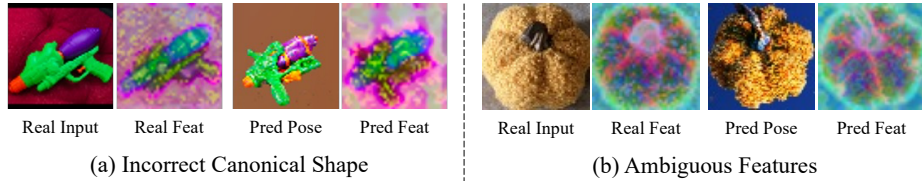


Fig. 9: Failure Modes. Our method inherits the failure from (a) canonical shape optimization and (b) pre-trained feature extractors.

4.3 Failure Modes

We have identified two failure modes of the proposed method: (1) incorrect shapes from the generative model distillation process, *e.g.*, the incorrect placement of the water gun handle from Figure 9 (a), and (2) incorrect poses due to feature ambiguity, *e.g.*, the pumpkin is symmetric and DINO features cannot disambiguate sides from Figure 9 (b).

5 Conclusion

We have introduced 3D Congealing, 3D-aware alignment for 2D images capturing semantically similar objects. Our proposed framework leverages a canonical 3D representation that encapsulates geometric and semantic information and, through optimization, fuses prior knowledge from a pre-trained image generative model and semantic information from input images. We show that our model achieves strong results on real-world image datasets under challenging identity, illumination, and background conditions.

Acknowledgments. We thank Chen Geng and Sharon Lee for their help in reviewing the manuscript. This work is in part supported by NSF RI #2211258, #2338203, and ONR MURI N00014-22-1-2740.

References

1. Boss, M., Engelhardt, A., Kar, A., Li, Y., Sun, D., Barron, J., Lensch, H., Jampani, V.: Samurai: Shape and material from unconstrained real-world arbitrary image collections. *Advances in Neural Information Processing Systems* **35**, 26389–26403 (2022) [2](#), [4](#), [10](#), [11](#)
2. 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*. pp. 9650–9660 (2021) [2](#), [7](#)
3. Chen, X., Dong, Z., Song, J., Geiger, A., Hilliges, O.: Category level object pose estimation via neural analysis-by-synthesis. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI* 16. pp. 139–156. Springer (2020) [4](#)
4. Chen, Y., Chen, X., Wang, X., Zhang, Q., Guo, Y., Shan, Y., Wang, F.: Local-to-global registration for bundle-adjusting neural radiance fields. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8264–8273 (2023) [2](#), [4](#)
5. Cheng, W., Cao, Y.P., Shan, Y.: Id-pose: Sparse-view camera pose estimation by inverting diffusion models. *arXiv preprint arXiv:2306.17140* (2023) [4](#)
6. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., Vanderbilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13142–13153 (2023) [4](#)
7. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022) [3](#), [6](#)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *Communications of the ACM* **63**(11), 139–144 (2020) [3](#)
9. Goodwin, W., Vaze, S., Havoutis, I., Posner, I.: Zero-shot category-level object pose estimation. In: *European Conference on Computer Vision*. pp. 516–532. Springer (2022) [4](#)
10. Gower, J.C., Dijksterhuis, G.B.: *Procrustes problems*, vol. 30. OUP Oxford (2004) [11](#)
11. Gupta, K., Jampani, V., Esteves, C., Shrivastava, A., Makadia, A., Snavely, N., Kar, A.: Asic: Aligning sparse in-the-wild image collections. *arXiv preprint arXiv:2303.16201* (2023) [2](#), [19](#), [20](#)
12. Huang, G., Mattar, M., Lee, H., Learned-Miller, E.: Learning to align from scratch. *Advances in neural information processing systems* **25** (2012) [3](#)
13. Huang, G.B., Jain, V., Learned-Miller, E.: Unsupervised joint alignment of complex images. In: *ICCV*. pp. 1–8. IEEE (2007) [3](#)
14. Jampani, V., Maninis, K.K., Engelhardt, A., Karpur, A., Truong, K., Sargent, K., Popov, S., Araujo, A., Martin-Brualla, R., Patel, K., et al.: Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. *arXiv preprint arXiv:2306.09109* (2023) [10](#), [11](#)

15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015) [9](#)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) [5](#), [9](#)
17. Kuang, Z., Olszewski, K., Chai, M., Huang, Z., Achlioptas, P., Tulyakov, S.: Neroic: Neural rendering of objects from online image collections. *ACM Transactions on Graphics (TOG)* **41**(4), 1–12 (2022) [10](#), [11](#)
18. Learned-Miller, E.G.: Data driven image models through continuous joint alignment. *IEEE TPAMI* **28**(2), 236–250 (2005) [3](#)
19. Lin, A., Zhang, J.Y., Ramanan, D., Tulsiani, S.: Relpose++: Recovering 6d poses from sparse-view observations. arXiv preprint arXiv:2305.04926 (2023) [3](#)
20. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5741–5751 (2021) [2](#), [4](#), [9](#), [12](#)
21. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023) [4](#)
22. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. *ACM SIGGRAPH Computer Graphics* **21**(4), 163–169 (1987) [8](#)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018) [9](#)
24. Luo, G., Dunlap, L., Park, D.H., Holynski, A., Darrell, T.: Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *NeurIPS* **36** (2024) [18](#)
25. Meng, Q., Chen, A., Luo, H., Wu, M., Su, H., Xu, L., He, X., Yu, J.: Gnerf: Gan-based neural radiance field without posed camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6351–6361 (2021) [2](#), [10](#), [11](#)
26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **65**(1), 99–106 (2021) [4](#), [5](#), [7](#)
27. Miller, E.G., Matsakis, N.E., Viola, P.A.: Learning from one example through shared densities on transforms. In: CVPR. vol. 1, pp. 464–471. IEEE (2000) [3](#)
28. Min, J., Lee, J., Ponce, J., Cho, M.: Spair-71k: A large-scale benchmark for semantic correspondence. arXiv preprint arXiv:1908.10543 (2019) [12](#), [19](#), [22](#)
29. Ofri-Amar, D., Geyer, M., Kasten, Y., Dekel, T.: Neural congealing: Aligning images to a joint semantic atlas. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19403–19412 (2023) [2](#), [3](#), [18](#), [19](#), [20](#)
30. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) [2](#), [7](#), [18](#), [19](#)
31. Peebles, W., Zhu, J.Y., Zhang, R., Torralba, A., Efros, A.A., Shechtman, E.: Gan-supervised dense visual alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13470–13481 (2022) [2](#), [3](#), [18](#), [19](#), [20](#)
32. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) [4](#), [6](#)

33. Raj, A., Kaza, S., Poole, B., Niemeyer, M., Ruiz, N., Mildenhall, B., Zada, S., Aberman, K., Rubinstein, M., Barron, J., et al.: Dreambooth3d: Subject-driven text-to-3d generation. arXiv preprint arXiv:2303.13508 (2023) [4](#), [6](#)
34. 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. pp. 10901–10911 (2021) [10](#)
35. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024) [9](#)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) [2](#), [4](#), [6](#)
37. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023) [4](#)
38. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016) [3](#), [7](#), [10](#)
39. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023) [4](#), [6](#)
40. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. arXiv preprint arXiv:2306.03881 (2023) [18](#)
41. 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/CVF Conference on Computer Vision and Pattern Recognition. pp. 2642–2651 (2019) [8](#)
42. Wang, J., Ruppel, C., Novotny, D.: Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9773–9783 (2023) [3](#), [10](#), [11](#)
43. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021) [4](#), [7](#)
44. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: Nerf-: Neural radiance fields without known camera parameters. arXiv preprint arXiv:2102.07064 (2021) [2](#), [4](#)
45. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems* **34**, 4805–4815 (2021) [4](#), [7](#)
46. Yen-Chen, L., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.Y.: inerf: Inverting neural radiance fields for pose estimation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1323–1330. IEEE (2021) [4](#)
47. Zhang, J., Yang, G., Tulsiani, S., Ramanan, D.: Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. In: *Advances in Neural Information Processing Systems*. vol. 34, pp. 29835–29847 (2021) [10](#), [11](#)
48. Zhang, J.Y., Ramanan, D., Tulsiani, S.: Relpose: Predicting probabilistic relative rotation for single objects in the wild. In: *European Conference on Computer Vision*. pp. 592–611. Springer (2022) [3](#)

49. 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. arXiv preprint arXiv:2305.15347 (2023) 18

A Additional Qualitative Results

The complete set of input images used for the sculpture dataset from Figure 1 and the corresponding results are shown in Figure 10. Images come from a personal photo collection.

B Implementation Details

Feature Extractors. We use the ViT-G/14 variant of DINO-V2 [30] as the feature extractor f_ζ from Sec. 3.2 and extract tokens from its final layer for all quantitative experiments. For qualitative results from Figure 1, following [40], we use features from the first upsampling block of the UNet from Stable Diffusion 2.1 with diffusion timestep 261 as f_ζ , as these features are similar for semantically-similar regions [24, 40, 49] but are locally smoother compared to DINO, which is consistent with the observations from [49].

Smoothness Loss. The smoothness loss from Eq. (10) is specified as follows. Following [29], we define

$$\mathcal{L}_{\text{rigidity}, \nabla}(T) := \|J_\nabla(T)^T J_\nabla(T)\|_F + \|(J_\nabla(T)^T J_\nabla(T))^{-1}\|_F, \quad (14)$$

where T jointly considers all neighboring coordinates u , instead of only one coordinate u at once. We define $[T]_u = \tilde{u} - u$, where \tilde{u} is the optimization variable for input u from Eq. (10), and $J_\nabla(T)$ computes the Jacobian matrix of T approximated with finite differences with pixel offset ∇ . Following [31], we denote huber loss with $\mathcal{L}_{\text{huber}}$ and define the total variation loss as

$$\mathcal{L}_{\text{TV}}(T) = \mathcal{L}_{\text{huber}}(\nabla_x T) + \mathcal{L}_{\text{huber}}(\nabla_y T), \quad (15)$$

where ∇_x and ∇_y are partial derivatives *w.r.t.* x and y coordinates, approximated with finite differences. The final smoothness loss is defined as

$$\mathcal{L}_{\text{smooth}} = \lambda_{\text{rigidity}, \nabla=10} + 0.1\mathcal{L}_{\text{rigidity}, \nabla=1} + 10\mathcal{L}_{\text{TV}}. \quad (16)$$

C Feature Visualizations

Matching features independently for each pixel gives noisy similarity heatmaps (Figure 12 (b)), due to the noise of feature maps (Figure 12 (d-e)), and the lack of geometric reasoning in the matching process. Our method is robust to such noises as it seeks to align the input with a posed rendering considering all pixel locations in the input altogether.

	Aero	Bike	Boat	Bus	Car	Chair	Motor	Train	TV	Mean
GANgealing. [31]	-	37.5	-	-	-	-	-	-	-	-
Neural Congealing [29]	-	29.1	-	-	-	-	-	-	-	-
ASIC [11]	57.9	25.2	24.7	28.4	30.9	21.6	26.2	49.0	24.6	32.1
DINOv2-ViT-G/14 [30]	72.5	67.0	45.5	54.6	53.5	40.7	71.8	53.5	36.3	55.0
Ours	70.0	70.3	40.0	65.8	72.1	50.1	77.0	26.1	43.1	57.2

Table 2: Semantic Correspondence Evaluation on SPair-71k [28]. Our method achieves an overall better keypoint transfer accuracy compared to prior 2D congealing methods and a 2D-matching baseline using the same semantic feature extractor as ours.

D Semantic Correspondence Matching

We provide additional quantitative evaluation of our method on the task of semantic correspondence matching. Given a pair of source and target image $(x_{\text{source}}, x_{\text{target}})$, and given a keypoint in the source image u_{source} , the goal of this task is to find its most semantically similar keypoint u_{target} in the target image.

The matching process using our method is specified as follows. We first map the 2D keypoints being queried to the 3D coordinates in the canonical space, and then project these 3D coordinates to the 2D image space of the target image. Formally, given an image pair $(x_{\text{source}}, x_{\text{target}})$ and a 2D keypoint u_{source} , the corresponding keypoint u_{target} is computed with

$$u_{\text{target}} = \Phi_{x_{\text{target}}}^{\text{rev}} \circ \Phi_{x_{\text{source}}}^{\text{fwd}}(u_{\text{source}}), \quad (17)$$

with notations defined in Sec. 3.3.

For all experiments in this section, for Eq. (10), we set $\lambda_{\ell_2} = 10$ and for simplicity set $\lambda_{\text{smooth}} = 0$.

Dataset. We use SPair-71k [28], a standard benchmark for semantic correspondence matching for evaluation. We evaluate our method on 9 rigid, non-cylindrical-symmetric categories from this dataset. The images for each category may contain a large diversity in object shape, texture, and environmental illumination.

Following prior works [29, 31], we report the Percentage of Correct Keypoints (PCK@ α) with $\alpha = 0.1$, a standard metric that evaluates the percentage of keypoints correctly transferred from the source image to the target image with a threshold α . A predicted keypoint is correct if it lies within the radius of $\alpha \cdot \max(H_{\text{bbox}}, W_{\text{bbox}})$ of the ground truth keypoint in the object bounding box in the target image with size $H_{\text{bbox}} \times W_{\text{bbox}}$.

Baselines. We compare with a 2D-correspondence matching baseline. Formally, for this baseline, for each querying keypoint u_{query} , we compute the keypoint prediction with

$$u_{\text{target}} = \arg \min_u d_{\zeta}^{u,u}(x_{\text{target}}, x_{\text{source}}), \quad (18)$$

where the distance metric d_{ζ} is defined in Eq. (6) and is induced from a pre-trained features extractor f_{ζ} . We use the same DINO feature extractor for our method and this baseline.

We further compare with previous congealing methods, GANgealing [31], which uses pre-trained GAN for supervision, and Neural Congealing [29] and ASIC [11], which are both self-supervised.

Results. Results are shown in Tab. 2. The performance gain over the DINOv2 baseline, which uses the same semantic feature extractor backbone as ours, suggests the effectiveness of 3D geometric consistency utilized by our framework.

Qualitative results are shown in Figure 11. Our method is the only one that performs correspondence matching via reasoning in 3D among all baselines. Such 3D reasoning offers an advantage especially when the relative rotation between the objects from the source and target image is large. Our method transforms the 3D coordinate from source to target in the canonical frame, where the 3D shape guarantees the 3D consistency. In comparison, as shown on the right of Figure 11, the baseline performs 2D matching and incorrectly matches the front of a plane with its rear, and incorrectly matches the front wheel of a bicycle with its back wheel.

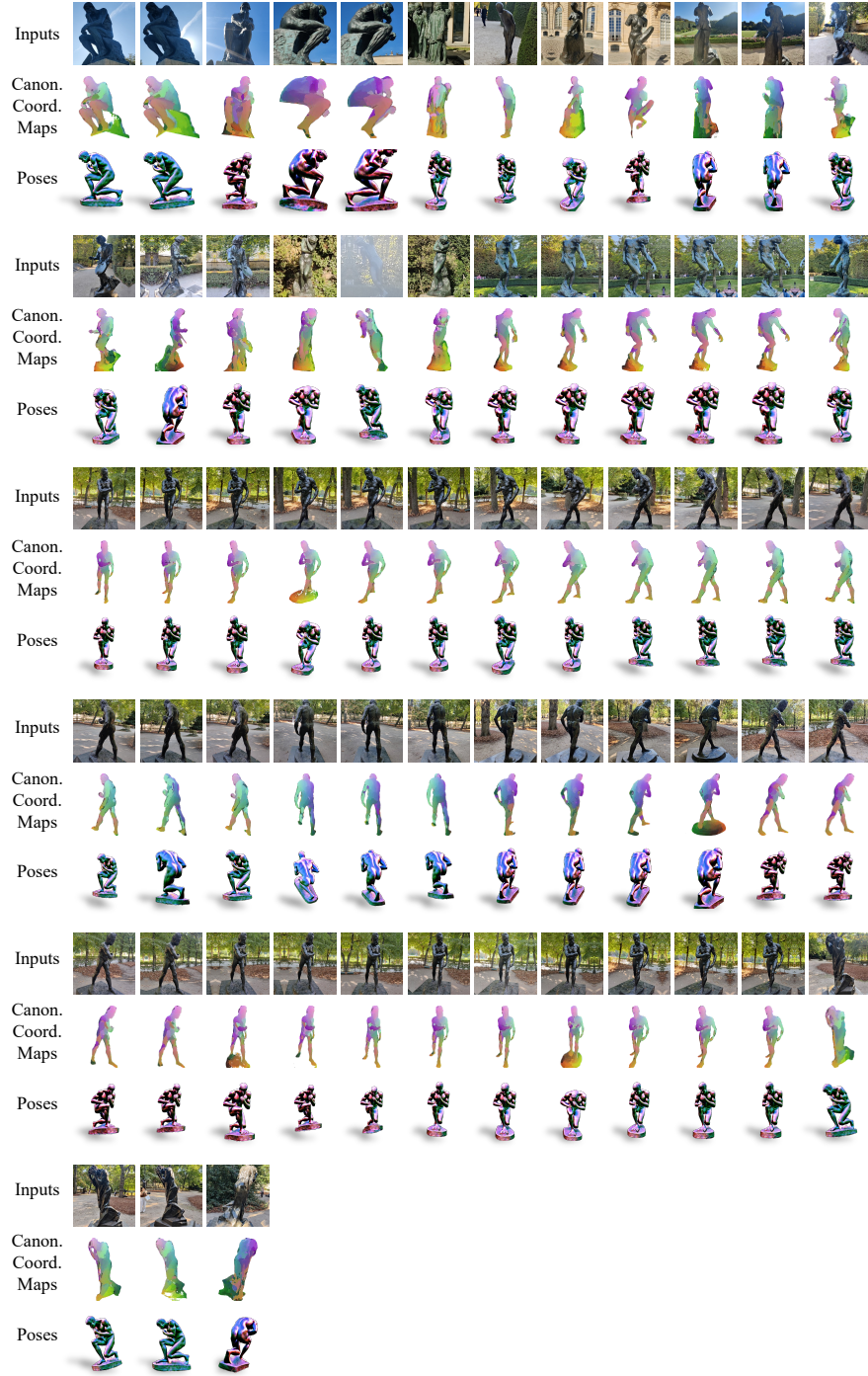


Fig. 10: Results on the Sculpture Dataset.

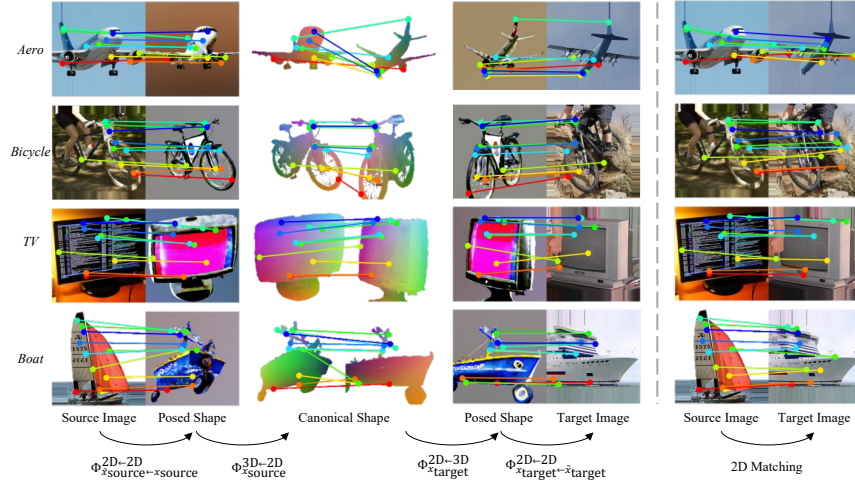


Fig. 11: Semantic Correspondence Matching. The figure shows results on 4 example categories from SPair-71k [28]. To match a given keypoint from the source image, our method first warps the keypoint to the rendered image space (2D-to-2D), then identifies the warped coordinate’s location in the canonical frame in 3D (2D-to-3D), then projects the *same* 3D location to the rendering corresponding to the target image (3D-to-2D), and finally warps the obtained coordinate to the target image space (2D-to-2D). The learned 3D canonical shape serves as an intermediate representation that aligns the source and target images, and it better handles scenarios when the viewpoint changes significantly compared to matching features in 2D.

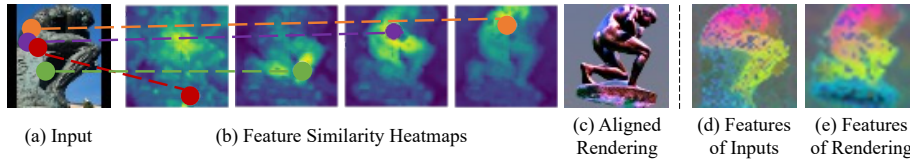


Fig. 12: Feature Visualizations. Despite that DINO features tend to be noisy, our approach assigns a plausible pose to the input, as shown in the aligned rendering.