Implicit Modeling of Non-rigid Objects with Cross-Category Signals

Yuchun Liu, Benjamin Planche, Meng Zheng, Zhongpai Gao, Pierre Sibut-Bourde, Fan Yang, Terrence Chen, Ziyan Wu

United Imaging Intelligence
{yuchun.liu01, benjamin.planche, meng.zheng, zhongpai.gao, fan.yang03, terrence.chen, ziyan.wu}@uii-ai.com

Abstract

Deep implicit functions (DIFs) have emerged as a potent and articulate means of representing 3D shapes. However, methods modeling object categories or non-rigid entities have mainly focused on single-object scenarios. In this work, we propose MODIF, a multi-object deep implicit function that jointly learns the deformation fields and instance-specific latent codes for multiple objects at once. Our emphasis is on non-rigid, non-interpenetrating entities such as organs. To effectively capture the interrelation between these entities and ensure precise, collision-free representations, our approach facilitates signaling between category-specific fields to adequately rectify shapes. We also introduce novel inter-object supervision: an attraction-repulsion loss is formulated to refine contact regions between objects. Our approach is demonstrated on various medical benchmarks, involving modeling different groups of intricate anatomical entities. Experimental results illustrate that our model can proficiently learn the shape representation of each organ and their relations to others, to the point that shapes missing from unseen instances can be consistently recovered by our method. Finally, MODIF can also propagate semantic information throughout the population via accurate point correspondences.

Introduction

In recent years, there has been extensive research into deep implicit functions (DIFs) as a means of representing 3D object categories. In comparison to conventional geometric representations, DIFs exhibit potential in effectively reconstructing geometric intricacies, even for non-rigid objects or object categories with large variability. This potential holds significant advantages for various clinical imaging applications, such as 3D organ reconstruction, anatomical segmentation, and surgical navigation. For example, various medical procedures depend on accurately locating precise organ regions, which implies accurately modeling the whole anatomical structure.

However, methods proposed so far to represent 3D entities (Park et al. 2019; Mescheder et al. 2019; Sitzmann et al. 2020; Sun et al. 2022) are constrained to single-object scenarios. The exploration of multi-object 3D neural representations remains relatively uncharted, and prior research (Zhang et al. 2022) tends to overlook the interrelations between the different categories that constitute each instance. *E.g.*, for organs, their shape and location depends on a variety of criteria such as body pose(Guo et al. 2022), metabolic cycles, *etc*. More importantly, their shapes undergo deformation based on the interactions and pressure they exert on one another. In this context, learning the implicit function of each category in isolation is both ineffective and problematic, as it could potentially lead to shape interpenetration.

Dealing with multiple objects presents several challenges. First, the rigidity and shape variability across different object categories adds complexity to the task of simultaneously modeling various types of surface deformation fields. This challenge is often exacerbated by the scarcity of training data, particularly in the context of medical applications. This scarcity makes the task of achieving shape generalization even more complex. Secondly, the model must jointly account for the interactions among objects. This involves not only capturing their relative positions but also modeling the contact regions occurring between these objects.

To address these challenges, we introduce our Multi-Object Deformed Implicit Field (MODIF) model, designed to learn implicit multi-object shape functions. It incorporates a cross-category refinement mechanism that allows us to capture interactions between objects while still maintaining the accuracy of individual reconstructions. The model facilitates the generation of both point correspondences and separate per-category templates, which can be utilized for extrapolating the shape of a specific object missing in an unseen instance. In summary, the primary contributions of our paper can be outlined as follows:

- We introduce a comprehensive method for representing multiple non-rigid objects using an implicit approach. Our model not only produces accurate reconstructions of shapes but also offers precise predictions of individual object positions.
- We design a cross-category refinement mechanism, incorporating the features from each individual sub-function to generate an overall correction field. With this, an attraction-repulsion loss is formulated to supervise contact regions between objects and to effectively reduce erroneous object interpenetration.
- Our model can generate point correspondences for mul-

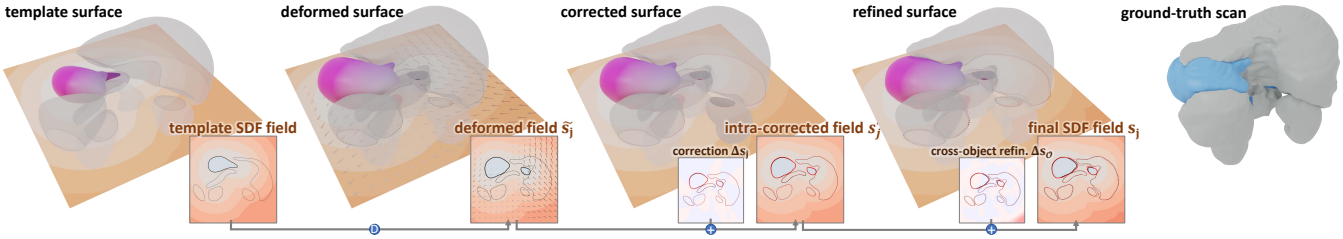

Figure 1: Visualization of intermediary SDF slices (corresponding to the colored `stomach` class) and resulting 3D shapes predicted by our method. Tackling the modeling of multi-object 3D instances, our neural SDF model goes beyond prior work by accounting for object interrelations (both for inference and supervision) to refine its predictions (better viewed zoomed in).

tiple objects simultaneously. We ensure multi-category point correspondence while preserving single-category geometries.

- We provide a solution for reconstructing plausible and consistent shapes when a specific object is missing for new observed instances.
- We evaluate our solution on 3 different datasets over multiple tasks and show that MODIF consistently outperforms state-of-the-art methods, even when the latter are provided with additional supervision.

Related Works

Deep Implicit Functions. DIF methods rely on neural networks to represent continuous 3D shapes and are nowadays considered more flexible and efficient than traditional explicit methods. There has been a proliferation of models (Hao et al. 2020; Duan et al. 2020; Chabra et al. 2020) developed in this direction. DeepSDF (Park et al. 2019) introduces an auto-decoder model to learn a 3D signed distance field (SDF). Occupancy networks (Mescheder et al. 2019) are another popular method, introducing a deep neural network classifier to decide if the point is inside of the object. It inspired various following works (Peng et al. 2020)(Roddick and Cipolla 2020)(Lionar et al. 2021). Despite the impressive shape representation capabilities of DIFs, the reconstruction outcomes often suffer from a lack of intricate details. In response to this issue, (Sitzmann et al. 2020) introduced a periodic activation function to model finer details and improve the overall accuracy over complex scenes.

Template Learning and Point Correspondence. In medical imaging applications, the establishment of point correspondences holds significant importance, as it often becomes necessary to map and compare various anatomical structures. To address this need, several deformation-based DIF methods have been introduced to infer dense correspondences across shapes. Deep Implicit Templates (DIT) (Zheng et al. 2021) utilize an LSTM model to learn conditional deformations and generate templates spanning all shapes. Built upon, Neural Diffeomorphic Flow (NDF) (Sun et al. 2022) employs multiple neural ordinary equation (NODE) blocks to ensure the preservation of topological features during shape deformation; whereas Deformed Implicit Field (DIF) (Deng, Yang, and Tong 2021) offers greater generalization capability. It is worth noting that these methods are tailored for single-object scenarios, while our approach is capable of concurrently generating point correspondences for multiple objects.

Structured Shape Representation. To capture complex 3D shapes, recent endeavors have focused on breaking down instances into simpler, smaller components. Notably, DeepLS (Chabra et al. 2020) uses a grid of independent latent codes to model local structures; and LDIF (Genova et al. 2020) partitions the 3D space into a structured arrangement of learned implicit functions. DMM model (Zhang et al. 2022) presents an implicit dental model, which provides the segmentation labels for individual teeth and gum. Although ImgHUM and DMM are capable of generating distinct subparts, their approaches primarily focus on reconstructing the entire instance rather than comprehensively modeling relationships between these sub-parts. In contrast, our model is specifically designed to address interactions and is thus proficient in circumventing collision problems between objects.

Method

Overview

Our goal is to learn a modular shape representation for a collection $\{\mathcal{O}_i\}_{i=1}^n$ of 3D instances, with each instance $\mathcal{O}_i$ being composed of a set of interrelated and non-overlapping objects $\{\mathcal{O}_{i,j}\}_{j=1}^m$ belonging to m different categories. We define a function $F(\alpha_i) = \mathcal{O}_i$ that maps a latent vector α_i to the corresponding 3D instance, and we jointly learn F and $\{\alpha_i\} \in \mathbb{R}^k$ in a self-supervised manner. Similar to previous works (Park et al. 2019; Alldieck, Xu, and Sminchisescu 2021; Sun et al. 2022; Deng, Yang, and Tong 2021; Zhang et al. 2022), we use signed distance fields (SDFs), which can continuously and implicitly represent surface geometries and be modeled by coordinate-based neural networks. For every point $p \in \mathbb{R}^3$ in the represented domain, an SDF provides a scalar value s corresponding to the distance from p to the closest object surface, with $s < 0$ if p is inside the object and $s > 0$ if p is outside. The explicit surface representation of each shape can be defined as the zero-level set of its SDF, *e.g.*, which can be extracted using the marching-cubes algorithm (noted mc) (Lorensen and Cline 1998). Therefore, our function F can be expressed as:

$$F = \mathrm{mc}\left(\{f(\alpha_i, p) \mid p \sim \Omega\}\right), \tag{1}$$

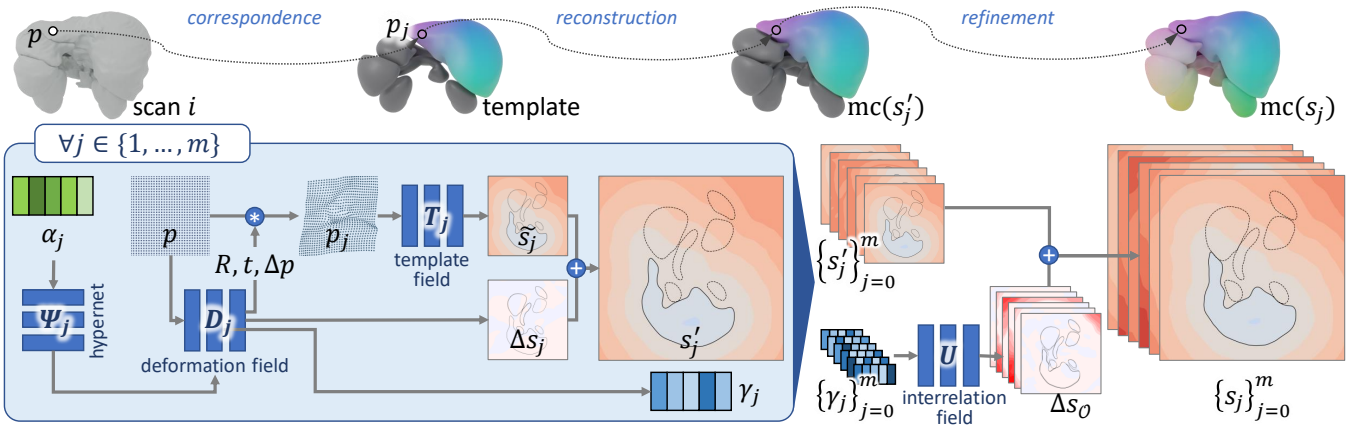

Figure 2: Proposed pipeline, composed of m sub-functions separately modeling the categories composing target 3D instances, and one interrelational refinement field. It achieves accurate instance reconstruction, point correspondence, and object recovery.

with $f : \mathbb{R}^k \times \mathbb{R}^3 \mapsto \mathbb{R}^m$ a novel neural implicit model that, given a point p in the 3D space Ω and a conditioning vector α_i learned from $\mathcal{O}_i$, predicts $s_i = \{s_{i,j}\}_{j=1}^m$, a vector of SDF values w.r.t. each of the m instance shapes.

This formulation differs from most prior arts that output a single distance field per instance. For instances with multiple objects, they either (a) merge their geometries into one, prior to modeling it as a single field (Sun et al. 2022; Zhang et al. 2022); or (b) tackle each object separately, each with their own distinct and uncorrelated latent shape space (Deng, Yang, and Tong 2021). In this work, we aim at learning object-compositional implicit functions that account for interrelations between the m object categories to ensure higher inter- and cross-instance consistency. Inspired by (Alldieck, Xu, and Sminchisescu 2021; Zhang et al. 2022), we decompose the problem into learning m sub-functions f_j:

$$f_j : (\alpha_{i,j}, p) \in \mathbb{R}^{\lfloor k/m \rfloor} \times \mathbb{R}^3 \mapsto (s'_{i,j}, \gamma_{i,j}) \in \mathbb{R} \times \mathbb{R}^l, \quad (2)$$

with $s'_{i,j} \in \mathbb{R}$ the SDF value of p w.r.t. $\mathcal{O}_{i,j}$ estimated by this sub-function and $\gamma_{i,j}$ an l-dimensional feature vector encoding geometrical properties of shape $\mathcal{O}_{i,j}$. Unlike previous work, we encourage signaling between these sub-functions, both in a feed-forward (feature sharing *c.f.* proposed $\gamma_{i,j}$) and back-propagation (cross-category supervision) manners. We further introduce a cross-category refinement neural model U, defined as follows:

$$U : \left(\{\gamma_{i,j}\}_{j=1}^m\right) \in \mathbb{R}^{l \times m} \mapsto \Delta s_{\mathcal{O}_i} \in \mathbb{R}^m, \quad (3)$$

I.e., it generates a m-dimensional residual SDF vector to consolidate and correct the per-module SDF predictions using cross-object signals $\{\gamma_{i,j}\}_{j=1}^m$. The refined SDF predictions thus correspond to $s_i = s'_i + \Delta s_{\mathcal{O}_i}$. In the remainder of this section, we further detail our model, its supervision, and application to multi-object reconstruction and point correspondence (as shown in Figure 2).

Network Structure

Our sub-functions f_j are adapted from DIF-Net (Deng, Yang, and Tong 2021) and composed of 3 neural networks—neural template field T_j, hyper-net Ψ_j, and dual deformation/correction field D_j—to predict $s_{i,j}$ based on instance-specific learned sub-code $\alpha_{i,j}$. The predicted m-dimensional SDF vector is then residually edited by our novel cross-object correction function U.

Template Field T_j. Defined as $T_j : p \in \mathbb{R}^3 \mapsto \widetilde{s_j} \in \mathbb{R}$, it maps 3D points to their signed distances w.r.t. template object j in a reference space. As proposed in (Deng, Yang, and Tong 2021), its neural weights are shared across the whole population, *i.e.*, to learn category-wide shape properties.

Generalized Deformation/Correction Field D_{Ψ_j}. DIF-Net authors propose a function $\widetilde{D}_{\theta_i} : p \in \mathbb{R}^3 \mapsto (\Delta p_i, \Delta s_i) \in \mathbb{R}^{3+1}$ which is conditioned by instance-specific neural weights θ_i and models a point deformation field Δp and a SDF correction field Δs. Instance-specific SDF field s'_i can henceforth be expressed as $s'_i = T(p + \Delta p_i) + \Delta s_i$. Borrowing from DIF meta-learning (Sitzmann et al. 2020), the instance-specific weights θ_i are obtained from a hyper-network $\Psi : \alpha_i \in \mathbb{R}^k \mapsto \theta_i \in \mathbb{R}^{|\Theta|}$. This elegant formulation by DIF-Net authors enables the joint optimization of the models and instance codes α_i, as well as dense correspondence for the modeled shape category. This solution was thus adopted in various subsequent studies (Alldieck, Xu, and Sminchisescu 2021; Zhang et al. 2022).

In this work, each sub-function also relies on its own deformation/correction field D_j and hyper-net Ψ_j, though we propose a more comprehensive formulation of the deformation field, generalized from (Zhang et al. 2022). In order to jointly model several non-rigid objects, one not only needs to model intra-shape deformations, but also positional/scaling changes in objects across various instances. Tackling dental geometry modeling, authors of DMM (Zhang et al. 2022) made relatively simple assumptions about the transformation of their target classes (`gum` and `teeth`) and proposed specific deformation fields for each, to account for their different *rigidness*. Here, we assume that per-object positional information is not provided and that per-category rigidness is not known. Therefore, we define our generalized

deformation/correction function as:

$$D_{\theta_{i,j}} : p \in \mathbb{R}^3 \mapsto (r_{i,j}, t_{i,j}, \Delta p_{i,j}, \Delta s_{i,j}) \in \mathbb{R}^{3+3+3+1}, \tag{4}$$

with the deformation explicitly composed of a rigid transformation $e^{\mathcal{S}_{i,j}}$ defined by the screw-axis $(r_{i,j}; t_{i,j})$ (Park et al. 2021) (*c.f.* formula by Rodrigues 1815) and non-rigid shape transformation $\Delta p_{i,j}$. Hence, the jth SDF value of a point p is expressed as:

$$s'_{i,j} = T(p_{i,j}) + \Delta s_{i,j} \text{ with } p_{i,j} = e^{\mathcal{S}_{i,j}} p + \Delta p_{i,j}. \tag{5}$$

Cross-Category Refinement. Each sub-function receives only local, object-specific information; so their correction field $\Delta s_{i,j}$ cannot account for inter-object relations. This can result in reconstructions with intersecting shapes and disregard for contact regions. To mitigate this issue, we introduce a cross-category correction field $\Delta s_{\mathcal{O}_i}$ modeled by a final network U based on concatenated signals from every sub-function. This correction vector is added to the m SDF values as the final output. In our implementation, each signal $\gamma_{i,j}$ originates from the penultimate activation of D_j, and U is a shallow MLP with sine activation. Figure 1 highlights the significant impact of proposed U on contact regions.

Supervision

Similar to previous DIF solutions, MODIF undergoes two optimization phases. First, the networks are optimized on a training dataset, along with the latent codes corresponding to each training instance (these can be discarded after). Once trained, the model can be used to reconstruct or annotate (via point correspondence) new instances. For each unseen instance, its conditioning code α_i is predicted via a shorter optimization phase, with the functions' weights frozen.

Sub-function Losses. Within each sub-function, we adopt the losses from DIF-Net (Deng, Yang, and Tong 2021) to supervise SDF prediction (L1 accuracy of predicted SDF, correctness of surface normals, enforcing of Eikonal gradient property, cross-instance normal consistency, *etc*.), deformation (smoothness prior), and correction (regularization). We invite readers to access (Deng, Yang, and Tong 2021) for details. We also apply L2 regularization to the latent codes, as suggested in (Park et al. 2019). Additionally, we adapt the centroid loss from (Zhang et al. 2022) to enforce that the centroid $c_{i,j}$ of shape $\mathcal{O}_{i,j}$ computed from post-deformation predictions coincides with the average centroid $\overline{c_j}$ estimated over the training samples. Based on our deformation formulation, the modified centroid loss is:

$$\mathcal{L}_j^{centroid} = \|e^{\mathcal{S}_{i,j}} c_{i,j} + \Delta c_{i,j} - \overline{c_j}\|. \tag{6}$$

Refinement Losses. We supervise the final SDF predictions s_i using the curriculum SDF loss from (Duan et al. 2020). Since the pre-refinement SDF fields are already supervised by the above-mentioned sub-function losses, we use the strictest tolerance and control parameters for the curriculum loss, *i.e.*, $\varepsilon = 0$ and $\lambda = 0.5$.

Similar to the intra-category correction field, we also regularize the cross-category correction:

$$\mathcal{L}^{refreg} = \sum_{p \in \Omega} |\Delta s_{\mathcal{O}_i}(p)|. \tag{7}$$

Finally, we introduce an attraction-repulsion supervision, *i.e.*, contact loss, for off-surface points that lie in contact regions between two or more objects. *I.e.*, we define the set of these *contact points* in training data as: $\mathcal{C} = \{(p, \Gamma) \mid \widehat{s}_{i,j} < \epsilon_c, j \in \Gamma, |\Gamma| \geq 2\}$, with Γ the set of objects that p is close to, $\widehat{s}_{i,j}$ the ground-truth SDF, and ϵ_c a hyper-parameter threshold (the smaller ϵ_c, the thinner the considered contact regions). During optimization, to narrow the boundaries between these surfaces accordingly as well as avoid inter-penetration, we compute the following loss:

$$\mathcal{L}^{\mathcal{C}} = \sum_{(p,\Gamma) \in \mathcal{C}} \sum_{j \in \Gamma} \sigma(|s'_{i,j}|) \tag{8}$$

with $\sigma(s) = 2\,\mathrm{sigmoid}(\lambda^{\mathcal{C}} s) - 1$, and $\lambda^{\mathcal{C}}$ weight controlling the constraint to the output SDF.

The overall optimization objective is defined by linearly combining all the aforementioned losses, applying phase-specific loss weighting (see implementation details).

Experiments

Experimental Protocol

Implementation. For a fair comparison, we use the same MLP architectures as in DIF-Net (Deng, Yang, and Tong 2021) for networks Ψ_j, D_j, T_j; only editing D_j to return feature vector γ_j along with its predictions. We fix the dimensionality of per-object codes $\alpha_{i,j}$ to 128, and the value of $\lambda^{\mathcal{C}}$ to 10^2. Other hyper-parameters are listed in annex. Our model is trained on three NVIDIA RTX A40 GPUs for 300 epochs (~1.5 hours over the WORD dataset).

Datasets. Since our model is focused on modeling sets of non-rigid objects, we opt for three medical shape benchmarks: WORD (Luo et al. 2022), AbdomenCT (Ma et al. 2022) and Multi-Modality Whole-Heart Segmentation (MMWHS) (Zhuang and Shen 2016). For WORD and AbdomenCT, we perform our evaluation on the following $m = 6$ organs: `left-kidney`, `right-kidney` `liver`, `stomach`, `spleen`, and `pancreas`; noting the significant shape variability of some of these classes (*e.g.*, `stomach`). For MMWHS, we consider 4 classes: `right-atrium`, `left-atrium`, `right-ventricle`, and `left-ventricle+left-myocardium` (merged into one shape as the left ventricle is contained inside the myocardium). Each dataset contains the following number of samples: 30 training / 10 testing samples for MMWHS, 100/20 for WORD, and 37/10 for AbdomenCT (after removing 3 cases with livers cropped during scanning).

To generate training data points, we follow the sampling strategy proposed in (Zhang et al. 2022). *I.e.*, for each instance, we sample 200,000 surface points and 250,000 points randomly picked from Ω (normalized to [-1, 1]).

Comparison. We compare to multiple state-of-the-art DIF methods: DeepSDF (Park et al. 2019), DIT (Zheng et al. 2021), DIF-Net (Zhang et al. 2022), NDF (Sun et al. 2022), and DMM (Zhang et al. 2022). Also composed of multiple DIF-Net submodules, DMM is the closest to our proposed model. However, it does not account for class interrelations

	Models	CD ↓ mean / std	CD ↓ med.	EMD ↓ mean / std	EMD ↓ med.	IV ↓ mean
WORD	DeepS.	16.68 / 5.58	17.23	8.24 / 0.83	8.29	11.39
	DIT	25.10 / 8.95	22.46	8.33 / 1.01	8.13	39.08
	DIF	19.01 / 7.11	17.01	8.50 / 1.04	8.29	16.66
	NDF	24.28 / 63.79	**8.46**	8.72 / 3.21	**7.91**	8.96
	Ours	**14.63** / **3.73**	14.36	**8.01** / **0.83**	8.00	**4.24**
MMWHS	DeepS.	7.87 / 3.82	7.17	7.31 / 2.00	7.10	19.79
	DIT	23.95 / 30.57	11.98	8.23 / 2.58	7.94	30.29
	DIF	11.37 / 6.37	10.68	7.63 / 2.50	6.98	6.22
	NDF	**4.37** / 6.83	**2.34**	7.09 / **1.87**	7.38	7.05
	Ours	4.95 / **1.67**	4.83	**6.80** / 1.98	**6.88**	**1.52**
AbdomenCT	DeepS.	41.65 / 12.33	**37.06**	11.13 / 1.10	11.08	20.09
	DIT	66.27 / 13.92	65.74	12.14 / 1.26	11.88	109.35
	DIF	45.67 / 15.78	40.13	10.90 / 0.87	10.90	16.37
	NDF	60.77 / 30.23	56.11	12.37 / 2.54	11.67	56.02
	Ours	**41.60** / **9.31**	41.08	**10.70** / **0.59**	**10.81**	**3.40**

Table 1: Comparison to object-level reconstruction methods on 3 datasets (IV results are in kilo-units). Our method reliably outperforms other solutions, especially in terms of non-interpenetration.

Methods	CD ↓ mean / std	CD ↓ med.	EMD ↓ mean / std	EMD ↓ med.
DeepSDF	108.75 / 29.97	103.68	15.96 / 1.96	15.62
DIT	303.44 / 45.04	297.43	20.15 / 2.64	19.45
DIF	72.96 / 21.47	72.83	14.52 / 2.50	14.57
NDF	128.35 / 78.37	91.92	18.18 / 5.47	17.18
DMM (teeth)	205.71 / 66.57	195.78	17.76 / 3.88	16.96
DMM (gum)	215.40 / 70.08	196.82	18.19 / 3.25	17.17
DMM (ada)	23.62 / 7.03	23.12	10.10 / 1.15	10.19
Ours	**14.63** / **3.73**	**14.36**	**8.01** / **0.83**	**8.00**

Table 2: Comparison to instance-level reconstruction methods on merged instance meshes from WORD dataset.

and its deformation functions are tailored to specific classes (`gum` and `teeth`). For fair comparison, we create a custom version "DMM (ada)" of this pipeline that borrows our proposed deformation formulation.

Metrics. Like prior work (Park et al. 2019), we evaluate the shape reconstruction in terms of Chamfer distance (CD) and earth-mover distance (EMD). We also introduce an intersection volume (IV) metric to measure undesired interpenetration, as the volume (operator $\forcurvearrowright$) of the union of all pairwise 3D shape intersections in an instance:

$$\text{IV}_i = \forcurvearrowright(\mathcal{O}_\cap) \text{ with } \mathcal{O}_\cap = \bigcup_{(a,b)\in\binom{m}{2}} (\mathcal{O}_{i,a} \cap \mathcal{O}_{i,b}). \quad (9)$$

Shape Representativeness

To evaluate representation capability, we challenge the methods to reconstruct instances not seen during training; by fixing the methods' weights, optimizing the latent code corresponding to the new shape(s), and finally measuring the quality of the reconstructed instance (*c.f.* Supervision subsection). Since prior art focuses on outputting a single SDF, we propose two different settings for other methods: *object-level* and *instance-level* reconstructions.

Without:	$\mathcal{L}^{\mathcal{C}}$ & $\Delta s_{\mathcal{O}}$	$\mathcal{L}^{\mathcal{C}}$	$\Delta s_{\mathcal{O}}$	Δp_j	$(r_i; t_i)$	$\emptyset$
CD ↓	15.90	16.36	15.33	15.06	15.44	**14.63**
EMD ↓	8.43	8.67	9.17	8.10	8.27	**8.00**

Table 3: Ablation study on WORD dataset, evaluating our work *without* key contributions (more results in annex).

Object-Level Reconstruction. In this experiment, we train m different instances of each prior method (except DMM which can generate segmented entities), each separately modeling one category. We thus make the assumption that the centroids and scales of each object are available to these models during inference, which is unfair to our method that learns by itself these object transformations.

Instance-Level Reconstruction. In this scenario, we merge the m shapes of each instance into one, and train prior methods to model this global geometry. Compared to ours, these solutions under-perform, losing category information (IV metric thus cannot be computed) and having trouble modeling contact regions.

Results. Tables 1 and Table 2 show that our model achieves the highest reconstruction accuracy on all datasets, despite solving a higher complexity task compared to object- and instance-level prior methods. In comparison, NDF (Sun et al. 2022) can achieve accurate results on objects with stable topological features but occasionally fails for instances with larger deformations. IV results also show that our method effectively reduces cross-object inconsistencies without sacrificing reconstruction accuracy. Qualitative examples are provided in Figure 3 and supplementary material.

Ablation Study

We confirm the significance of our technical contributions (generalized deformation, cross-category shape correction, contact loss) via an ablation study on the WORD dataset (Luo et al. 2022). Figure 1 also provides further insight into the impact of the different algorithmic steps on the predicted shapes. *E.g.*, in the provided sample, we can observe how the `liver`/`stomach` and `liver`/`left-kidney` contact regions are improved by our refinement function.

Applications

Point Correspondences. We showcase the point correspondence capability of MODIF in Figure 4 and in annex. While the mechanisms enabling accurate correspondences are borrowed from the literature (Deng, Yang, and Tong 2021), our method is the only one to simultaneously provide separate per-category template meshes and point correspondences. This means, for instance, that our method can not only accurately extrapolate the shape of any object missing from a new instance (as shown in the next experiment),

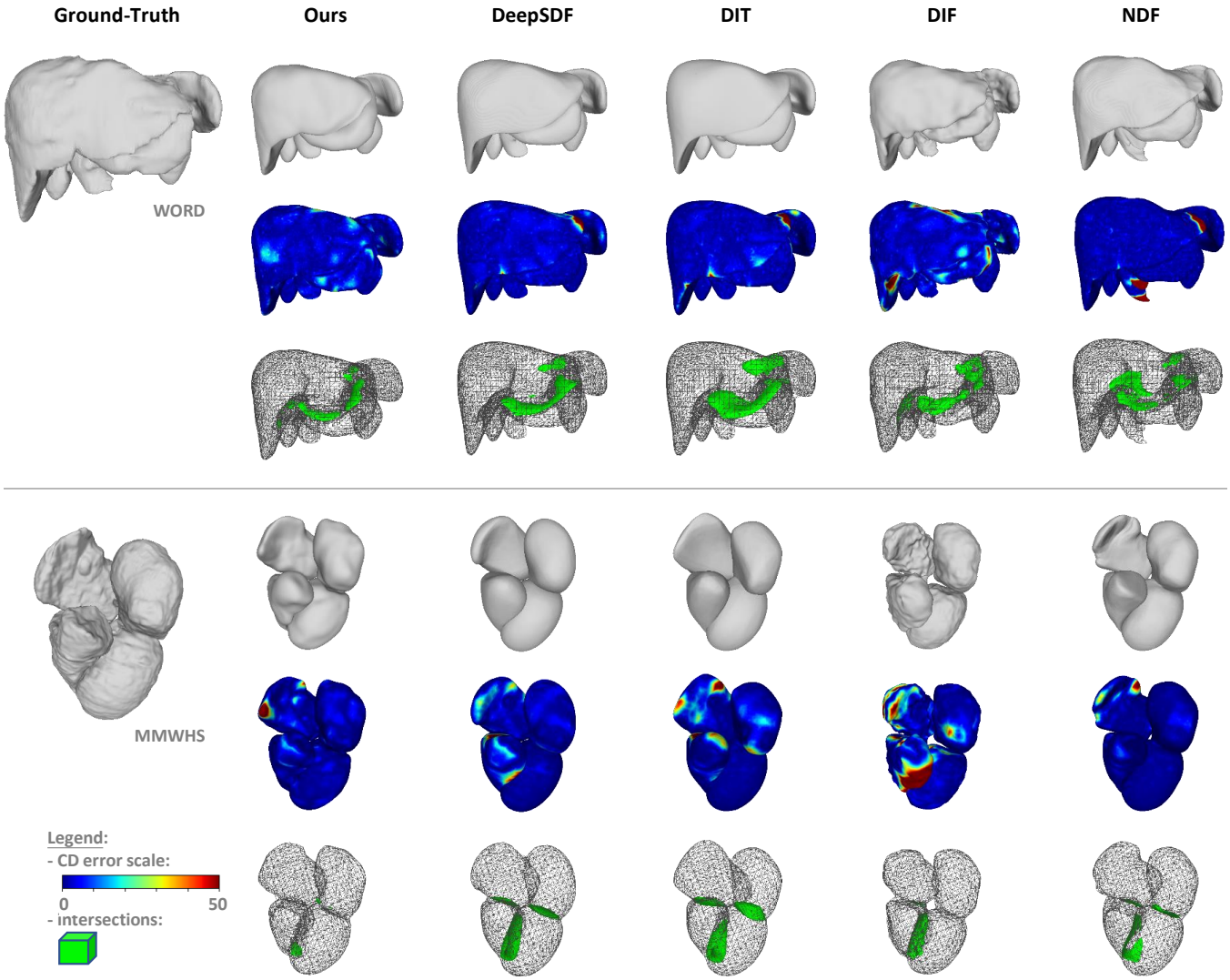

Figure 3: Qualitative comparison to prior art, plotting the reconstruction accuracy (CD-based color coding) and the erroneous intersections (green-colored regions) to highlight the positive impact of proposed inter-objects signals.

but it can also accurately transfer any dense annotation to predicted surfaces. We believe that this may have significant applications, *e.g.*, in clinical data annotation and analysis.

Missing-Object Recovery. Our method uniquely combines per-category shape prediction (via its sub-functions f_j, performing similarly as prior object-level methods) and cross-category relation modeling (*c.f.* refinement U). These properties enable MODIF to tackle an under-studied task: missing-object recovery, *i.e.*, when an instance is missing one of its objects (*e.g.*, organ not properly captured during scanning or not properly segmented by the annotator).

In such cases, MODIF can rely on its knowledge about the missing shape's overall distribution (*c.f.* learned template), as well as information extracted from non-missing objects in the instance, to recover a statistically plausible shape, that is consistent with other present objects (matching scale/positioning, minimal overlap, plausible contact surfaces, *etc.*).

For this task, we consider the MMWHS dataset (more challenging due to the higher ratio of contact regions), removing the `left-myocardium` shape from all test instances. We compare to per-object NDF (best challenger on WORD) and DIF. We adapt NDF inference so that it returns the template shape for the missing organ, with its scale and position adjusted according to the mean rigid transformation between the centroids of predicted non-missing shapes and the centroids of their corresponding templates.

Results in Table 4 demonstrate the superiority of our method, which successfully optimizes the recovered shape to fit with other organs (both in terms of rigid and soft transformation). The table also shows that this recovery does not negatively impact the reconstruction of other present shapes. Note that we also try to apply the instance-level NDF to this recovery task but observed worse results (mean CD = 93.41, mean EMD = 12.63). We provide a qualitative comparison in Figure 5, highlighting the minimized interpenetration.

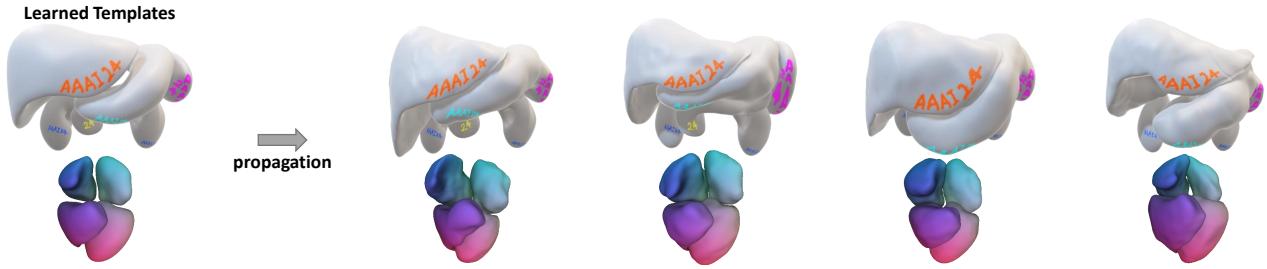

Figure 4: Visualization of propagating annotations by leveraging dense correspondences among reconstructions. The manual surface annotations on the learned template (left) are propagated to various instances (right).

	Models	CD ↓ mean / std	CD ↓ med.	EMD ↓ mean / std	EMD ↓ med.	IV ↓ mean
miss.	NDF	142.19 / 82.29	125.92	11.47 / 3.01	11.12	33.69
	DIF	94.44 / 54.44	79.32	10.84 / 3.20	10.28	16.80
	Ours	**45.52 / 27.06**	**37.74**	**6.46 / 2.09**	**5.73**	**1.20**
pres.	NDF	6.06 / 18.88	**2.19**	**2.81** / 1.87	**2.39**	35.73
	DIF	13.65 / 10.07	12.19	3.96 / 1.42	3.58	12.36
	Ours	**5.77** / **3.01**	5.33	3.11 / **0.61**	3.07	**1.47**

Table 4: Comparison to per-object NDF on the missing-object recovery task (missing `left-myocardium` in MMWHS test instances). We separately report metrics computed over the recovered missing organs ("miss.") and over the reconstructed present organs ("pres.").

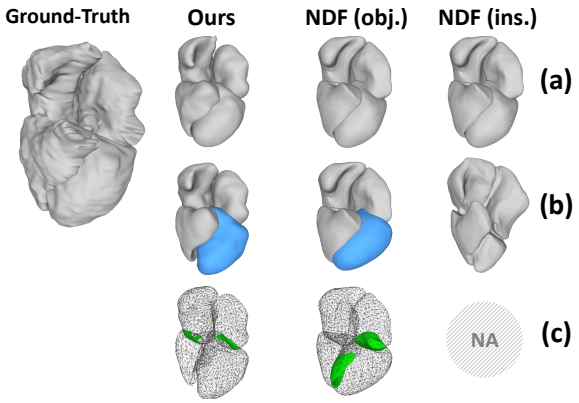

Figure 5: Results for the missing-organ task. Row (a) shows reconstructed results when all organs are provided. Row (b) shows predictions when the left myocardium (blue shape) is missing. Row (c) shows erroneous overlaps in predictions.

Data Augmentation. By learning the general data distribution, our method can generate realistic anatomical digital twins for dataset augmentation, training, *etc*. By tweaking learned latent code values, we are able to generate new synthetic instances, as shown in Figure 6.

Conclusion

We proposed MODIF, a novel implicit neural function to model multi-object 3D instances. Iterating over previous single-category SDF models, our solution relies on a cross-category refinement mechanism and a contact loss to model the correlation between objects. Experiments on various datasets show that our model performs better than prior art in terms of reconstruction accuracy and is highly effective at reducing interpenetration problems. Furthermore, we introduce the task of missing-object recovery from incomplete instance set and demonstrate that our model can predict accurate and consistent shape compared to other methods, thanks to our cross-category signaling.

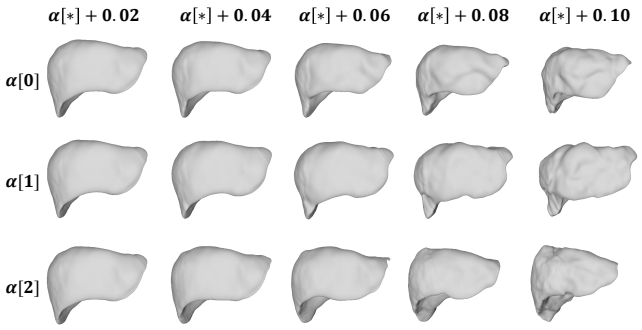

Figure 6: Realistic data augmentation by latent code editing.

Limitations. It should be noted that, similar to previous methods, our model expects class-labeled shapes, which can be a limitation for some scenarios (*e.g*., non-classified instances, open-set instances, *etc*.). Our deformation function may also be considered overly generic. Prior knowledge on specific categories could be injected to better guide the model (*e.g*., borrowing from the extensive literature on non-penetrating non-rigid simulation (Baraff 1993)).

Societal Impact. The above limitation may constrain the adoption of our method to specific use cases, and the statistical modeling of anatomical data may penalize patients with unique conditions (*outliers*). On the other hand, we would argue that the benefits in terms of automated dense data annotation outshine these limitations, with proper safeguards (*e.g*., validation by experts). All in all, we believe that the proposed method can positively impact both the computer-vision and medical communities, by offering a novel model for multi-SDF learning and demonstrating its values on medical benchmarks.

Acknowledgments

This work is supported by the National Key R&D Program of China (2021ZD0111100) and sponsored by Shanghai Rising-Star Program.

References

Alldieck, T.; Xu, H.; and Sminchisescu, C. 2021. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5461–5470.

Baraff, D. 1993. Non-penetrating rigid body simulation. *State of the art reports*.

Chabra, R.; Lenssen, J. E.; Ilg, E.; Schmidt, T.; Straub, J.; Lovegrove, S.; and Newcombe, R. 2020. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, 608–625. Springer.

Deng, Y.; Yang, J.; and Tong, X. 2021. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10286–10296.

Duan, Y.; Zhu, H.; Wang, H.; Yi, L.; Nevatia, R.; and Guibas, L. J. 2020. Curriculum deepsdf. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 51–67. Springer.

Genova, K.; Cole, F.; Sud, A.; Sarna, A.; and Funkhouser, T. 2020. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4857–4866.

Guo, H.; Planche, B.; Zheng, M.; Karanam, S.; Chen, T.; and Wu, Z. 2022. SMPL-A: Modeling Person-Specific Deformable Anatomy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20814–20823.

Hao, Z.; Averbuch-Elor, H.; Snavely, N.; and Belongie, S. 2020. Dualsdf: Semantic shape manipulation using a two-level representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7631–7641.

Lionar, S.; Emtsev, D.; Svilarkovic, D.; and Peng, S. 2021. Dynamic plane convolutional occupancy networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1829–1838.

Lorensen, W. E.; and Cline, H. E. 1998. Marching cubes: A high resolution 3D surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, 347–353.

Luo, X.; Liao, W.; Xiao, J.; Chen, J.; Song, T.; Zhang, X.; Li, K.; Metaxas, D. N.; Wang, G.; and Zhang, S. 2022. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. *Medical Image Analysis*, 82: 102642.

Ma, J.; Zhang, Y.; Gu, S.; Zhu, C.; Ge, C.; Zhang, Y.; An, X.; Wang, C.; Wang, Q.; Liu, X.; Cao, S.; Zhang, Q.; Liu, S.; Wang, Y.; Li, Y.; He, J.; and Yang, X. 2022. AbdomenCT-1K: Is Abdominal Organ Segmentation a Solved Problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6695–6714.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4460–4470.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 165–174.

Park, K.; Sinha, U.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Seitz, S. M.; and Martin-Brualla, R. 2021. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5865–5874.

Peng, S.; Niemeyer, M.; Mescheder, L.; Pollefeys, M.; and Geiger, A. 2020. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 523–540. Springer.

Roddick, T.; and Cipolla, R. 2020. Predicting semantic map representations from images using pyramid occupancy networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11138–11147.

Rodrigues, O. 1815. *De l'attraction des sphéroïdes*. Ph.D. thesis, éditeur non identifié.

Sitzmann, V.; Martel, J.; Bergman, A.; Lindell, D.; and Wetzstein, G. 2020. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33: 7462–7473.

Sun, S.; Han, K.; Kong, D.; Tang, H.; Yan, X.; and Xie, X. 2022. Topology-preserving shape reconstruction and registration via neural diffeomorphic flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20845–20855.

Zhang, C.; Elgharib, M.; Fox, G.; Gu, M.; Theobalt, C.; and Wang, W. 2022. An Implicit Parametric Morphable Dental Model. *ACM Transactions on Graphics (TOG)*, 41(6): 1–13.

Zheng, Z.; Yu, T.; Dai, Q.; and Liu, Y. 2021. Deep implicit templates for 3d shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1429–1439.

Zhuang, X.; and Shen, J. 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Medical image analysis*, 31: 77–87.