LIST: Learning Implicitly from Spatial Transformers for Single-View 3D Reconstruction

Mohammad Samiul Arshad and William J. Beksi
Department of Computer Science and Engineering
The University of Texas at Arlington, Arlington, TX, USA
mohammadsamiul.arshad@mavs.uta.edu, william.beksi@uta.edu

Abstract

Accurate reconstruction of both the geometric and topological details of a 3D object from a single 2D image embodies a fundamental challenge in computer vision. Existing explicit/implicit solutions to this problem struggle to recover self-occluded geometry and/or faithfully reconstruct topological shape structures. To resolve this dilemma, we introduce LIST, a novel neural architecture that leverages local and global image features to accurately reconstruct the geometric and topological structure of a 3D object from a single image. We utilize global 2D features to predict a coarse shape of the target object and then use it as a base for higher-resolution reconstruction. By leveraging both local 2D features from the image and 3D features from the coarse prediction, we can predict the signed distance between an arbitrary point and the target surface via an implicit predictor with great accuracy. Furthermore, our model does not require camera estimation or pixel alignment. It provides an uninfluenced reconstruction from the input-view direction. Through qualitative and quantitative analysis, we show the superiority of our model in reconstructing 3D objects from both synthetic and real-world images against the state of the art. Our source code is publicly available to the research community [13].

1. Introduction

Constructing a truthful portrayal of the 3D world from a single 2D image is a basic problem for many applications including robot manipulation and navigation, scene understanding, view synthesis, virtual reality, and more. Following the work of Erwin Kruppa [11] in camera motion estimation and the recovery of 3D points, researchers have attempted to solve the 3D reconstruction issue using structure from motion [33, 16, 28], and visual simultaneous localization and mapping [8, 27]. However, the main limitation of such approaches is that they require multiple observations of the desired object or scene from distinct viewpoints with shared features. Such a multi-view formulation allows for integrating information from numerous images to compensate for occluded geometry.

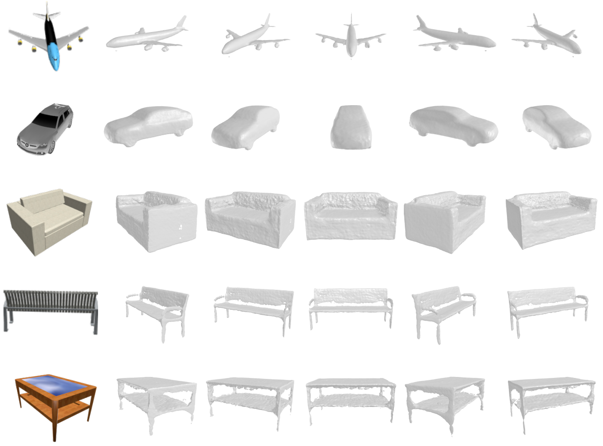

Fig. 1: Five unique views of objects reconstructed by LIST from a single RGB image. Not only does our model accurately recover occluded geometry, but also the reconstructed surfaces are *not influenced* by the input-view direction.

Reconstructing a 3D object from a single image is a more difficult task since a sole image does not contain the whole topology of the target shape due to self-occlusions. Researchers have tried both explicit and implicit techniques to reconstruct a target object with self-occluded parts. Explicit methods attempt to infer the target shape directly from the input image. Nevertheless, a major drawback of such approaches is that the output resolution needs to be defined in advance, which constrains these techniques from achieving high-quality results. Recent advances in implicit learning offer a solution to reconstruct the target shape in an arbitrary resolution by indirectly inferring the desired surface through a distance/occupancy field. Then, the target surface is reconstructed by extracting a zero level set from the distance/occupancy field.

Implicit 3D reconstruction from a single view is an active area of research where one faction of techniques [18, 3] encode global image features into a latent representation and learn an implicit function to reconstruct the target. Yet, these approaches can be easily outperformed by simple retrieval baselines [32]. Therefore, global features alone are not sufficient for a faithful reconstruction. Another faction leverages both local and global features to learn the target implicit field from pixel-aligned query points. However, such methods rely on ground-truth/estimated camera parameters for training/inference [35, 12], or they assume weak perspective projection [25, 9].

To address these shortcomings we propose LIST, a novel deep learning framework that can reliably reconstruct the topological and geometric structure of a 3D object from a single RGB image, Fig. 1. Our method *does not depend on weak perspective projection, nor does it require any camera parameters during training or inference*. Moreover, we leverage both local and global image features to generate highly-accurate topological and geometric details. To recover self-occluded geometry and aid the implicit learning process, we first predict a coarse shape of the target object from the global image features. Then, we utilize the local image features and the predicted coarse shape to learn a signed distance function (SDF).

Due to the scarcity of real-world 2D-3D pairs, we train our model on synthetic data. However, we use both synthetic and-real world images to test the reconstruction ability of LIST. Through qualitative analysis we highlight our model's *superiority in reconstructing high-fidelity geometric and topological structure*. Via a quantitative analysis using traditional evaluation metrics, *we show that the reconstruction quality of LIST surpasses existing works*. Furthermore, *we design a new metric to investigate the reconstruction quality of self-occluded geometry*. Finally, we provide an ablation study to validate the design choices of LIST in achieving high-quality single-view 3D reconstruction.

2. Related Work

In this section we summarize pertinent work on the reconstruction of 3D objects from a single RGB image via implicit learning. Interested readers are encouraged to consult [7] for a comprehensive survey on 3D reconstruction from 2D images. Contrary to explicit representations, implicit ones allow for the recovery of the target shape at an arbitrary resolution. This benefit has attracted interest among researchers to develop novel implicit techniques for different applications. Dai *et al*. [5] used a voxel-based implicit representation for shape completion. DeepSDF, introduced by Park *et al*. [23], is an auto-decoder that learns to estimate signed distance fields. However, DeepSDF requires test-time optimization, which limits its efficiency and capability.

To further improve 3D object reconstruction quality, Littwin and Wolf [14] utilized encoded image features as the network weights of a multilayer perceptron. Wu *et al*. [34] explored sequential part assembly by predicting the SDFs for structural parts separately and then combining them together. For self-supervised learning, Liu *et al*. [15] proposed a ray-based field probing technique to render the implicit surfaces as 2D silhouettes. Niemeyer *et al*. [21] used supervision from RGB, depth, and normal images to reconstruct rich geometry and texture. Chen and Zhang [3] proposed generative models for implicit representations and leveraged global image features for single-view reconstruction. For multiple 3D vision tasks, Mescheder *et al*. [18] developed OccNet, a network that learns to predict the probability of a volumetric grid cell being occupied.

Pixel-aligned approaches [25, 26, 9, 1] have employed local query feature extraction from image pixels to improve 3D human reconstruction. Xu *et al*. [35] incorporated similar ideas for 3D object reconstruction. To enhance the reconstruction quality of surface details, Li and Zhang [12] utilized normal images and a Laplacian loss in addition to aligned features. Zhao *et al*. [37] exploited coarse prediction and unsigned distance fields to reconstruct garments from a single view. Duggal and Pathak [6] proposed category specific reconstruction by learning a topology aware deformation field. Mittal *et al*. introduced AutoSDF [19], a model that encodes local shape regions separately via patch-wise encoding. However, these prior works rely on weak perspective projection and the rendering of metadata to align query points to image pixels. In contrast, LIST does not require any alignment or rendering data, and it recovers more accurate topological structure and geometric details.

3. Implicit Function Learning from Unaligned Pixel Features

Given a single RGB image of an object, our goal is to reconstruct the object in 3D with highly-accurate topological structure and self-occluded geometry. We model the target shape as an SDF and extract the underlying surface from the zero level set of the SDF during inference. To train our model we employ an image and query point pair (x_i, Q_i), where Q_i is a set of 3D coordinates (query points) in close vicinity to the surface of the object with a measured signed distance and x_i is a rendering of the object from a random viewpoint. An overview of the our framework is presented in Fig. 2. The details of each component are provided in the following subsections.

3.1. Query Features From Coarse Predictions

Consider an RGB image $x_i \subset X \in \mathbb{R}^{H \times W \times 3}$ of height H and width W. We propose a convolutional neural

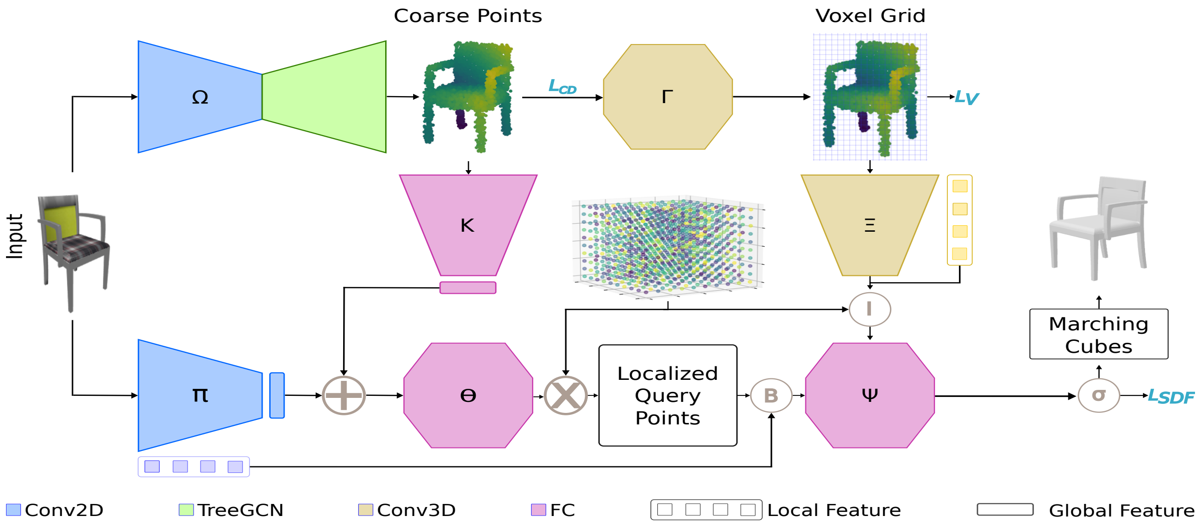

Fig. 2: To reconstruct the target object from a single RGB image, LIST first predicts the coarse topology from the global image features. Simultaneously, local image features are used to extract local geometry at the given query locations. Finally, an SDF predictor (Ψ) estimates the signed distance field (σ) to reconstruct the target shape. Note that images and colors are for visualization purposes only.

encoder-decoder Ω_ω, parameterized by weights ω, to extract latent features from the image and predict a coarse estimation $\dot{y}_i^{x_i}$ of the target object. Concretely,

$$\Omega_\omega(x_i) \coloneqq \dot{y}_i^{x_i} \mid \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{N\times 3}, \tag{1}$$

where $\dot{y}_i^{x_i}$ is a point cloud representation of the target and N is the resolution of the point cloud. Note that the subscript i indicates i-th sample and the superscript x_i designates the source variable. For high-performance point cloud generation, we utilize tree structured graph convolutions (TreeGCN) [29] to decode the image features.

We use the coarse prediction $\dot{y}_i$ as a guideline for the topological structure of the target shape in a canonical space. To extract query features from this coarse prediction, first we discretize the point cloud in an occupancy grid $\dot{u}_i^{\dot{y}_i} \in 1^{M\times M\times M}$ of resolution M. However, the coarse prediction may contain gaps and noisy points that may impair the reconstruction quality. To resolve this, we employ a shallow convolutional network $\Gamma_{\ddot{o}}$ parameterized by weights $\ddot{o}$ to generate a probabilistic occupancy grid from $\dot{u}_i^{\dot{y}_i}$,

$$\dot{v}_i^{\dot{u}_i} \coloneqq \Gamma_{\ddot{o}}(\dot{u}_i^{\dot{y}_i})\colon 1^{M\times M\times M} \to [0,1]^{M\times M\times M}. \tag{2}$$

Specifically, our aim is to find the neighboring points of $\dot{y}_i$ with a high chance of being a surface point of the target shape.

Although it is possible to regress the voxel representation directly from the global image features [4, 30, 9], learning a high-resolution voxel occupancy prediction requires a *significant* amount of computational resources [9]. Moreover, we empirically found that point cloud prediction followed by voxel discretization achieves better accuracy on diverse shapes rather than predicting the voxels directly.

Next, a neural network Ξ_ξ, parameterized by weights ξ, maps the probabilistic occupancy grid (2) to a high-dimensional latent matrix through convolutional operations. Then, our multi-scale trilinear interpolation scheme I extracts relevant query features f_C at each query location q_i from the mapped features. More formally,

$$f_C \coloneqq I(\Xi_\xi(\dot{v}_i^{\dot{u}_i}), Q_i). \tag{3}$$

In addition to q_i, we also consider the neighboring points at a distance d from q_i along the Cartesian axes to capture rich 3D features, i.e.,

$$q_j = q_j + k \cdot \hat{n}_j \cdot d, \tag{4}$$

where $k \in \{1, 0, -1\}$, $j \in \{1, 2, 3\}$, and $\hat{n}_j \in \mathbb{R}^3$ is the j-th Cartesian axis unit vector.

3.2. Localized Query Features

The coarse prediction and query features f_C can aid the recovery of the topological structure of the target shape. Nevertheless, relevant local features are also required to recover fine geometric details. To achieve this, prior arts assume weak perspective projection [25, 9] or align the query points to the image pixel locations through the ground-truth/estimated camera parameters [35, 12]. Predicting the camera parameters is analogous to predicting the object pose from a single image, which is itself a hard problem in computer vision. It involves a high chance of error and a computationally expensive training procedure. Furthermore, the error in the pose/camera estimation may lead to the loss of geometric details in the reconstruction.

To overcome these limitations, we obtain insight from spatial transformers [10] and leverage the spatial relationship between the input image and the coarse prediction. Via the coarse prediction, which portrays an object from a standard viewpoint and the query points that delineate the coarse predictions, it is possible to localize the query points to the local image features. This is done by predicting a spatial transformation with the aid of global features from the input image and the coarse prediction as follows.

First, we define a convolutional neural encoder Π_π, parameterized by weights π, to encode the input image into local $(l_\pi^{x_i})$ and global $(z_\pi^{x_i})$ features. Concretely,

$$\Pi_\pi(x_i) \coloneqq \{l_\pi^{x_i}, z_\pi^{x_i}\}. \tag{5}$$

Concurrently, a neural module K_κ encodes the coarse prediction $\dot{y}_i^{x_i}$ into global point features. Using global features from both the image and the coarse prediction, the spatial transformer Θ estimates a transformation to localize the query points in the image feature space. Then, localized query points $\tilde{Q}_i$ are generated by applying the predicted transformation to Q_i,

$$\Theta_\theta(z_\pi^{x_i}, K_\kappa(\dot{y}_i^{x_i}), Q_i) \coloneqq \tilde{Q}_i \mid \mathbb{R}^{N\times 3} \to \mathbb{R}^{N\times 2}. \tag{6}$$

Finally, a bi-linear interpolation scheme $\mathcal{B}$ extracts the local query features f_L from the local image features $l_\pi^{x_i}$,

$$f_L \coloneqq \mathcal{B}(l_\pi^{x_i}, \tilde{Q}_i). \tag{7}$$

Note that the point encoder K_κ and the localization network Θ are designated to ensure an accurate SDF prediction. Therefore, we do not use any camera parameters during training and we optimize these neural modules directly with the SDF prediction objective. This has the following benefits: (i) *additional modules or training to predict the projection matrix and object pose from a single image are not required;* (ii) *reconstructions are free from any pose estimation error, which boosts reconstruction accuracy*.

3.3. Signed Distance Function Prediction

To estimate the final signed distance Δ_i, we combine the coarse features f_C with the localized query features f_L and utilize a multilayer neural function defined as

$$\Psi_\psi(f_C, f_L) \coloneqq \begin{cases} \mathbb{R}^-, & \text{if } q_i \text{ is inside the target surface} \\ \mathbb{R}^+, & \text{otherwise.} \end{cases} \tag{8}$$

3.4. Loss Functions

We incorporate the chamfer distance (CD) loss and optimize the weights ω to accurately estimate the coarse shape of the target. More specifically,

$$\mathcal{L}_{\text{CD}}(y_i, \dot{y}_i) = \sum_{a\in\dot{y}_i} \min_{b\in y_i} ||a-b||^2 + \sum_{b\in y_i} \min_{b\in\dot{y}_i} ||b-a||^2, \tag{9}$$

where $y_i \in \mathbb{R}^{N\times 3}$ is a set of 3D coordinates collected from the surface of the object and $\dot{y}_i \in \mathbb{R}^{N\times 3}$ is the estimated coarse shape. To supervise the probabilistic occupancy grid prediction, we discretize y_i to generate the ground-truth occupancy $v_i^{y_i} \in 1^{M\times M\times M}$. The neural weight $\ddot{o}$ is then optimized by the binary cross-entropy loss,

$$\mathcal{L}_V(v_i, \dot{v}_i) = -\frac{1}{|v_i|}\Sigma(\gamma v_i \log \dot{v}_i + (1-\gamma)(1-v_i)\log(1-\dot{v}_i)), \tag{10}$$

where γ is a hyperparameter to control the influence of the occupied/non-occupied grid points. To optimize the SDF prediction, we collect a set of query points Q_i within distance δ of the target surface and measure their signed distance σ_i. The estimated signed distance is then guided by optimizing the neural weights ξ, π, θ, and ψ through

$$\mathcal{L}_{\text{SDF}} = \frac{1}{|Q_i|}\Sigma(\sigma_i - \Delta_i)^2. \tag{11}$$

3.5. Training Details

We incorporate a two-stage procedure to train LIST. In the first stage, we only focus on the coarse prediction from the input image x_i and optimize the weights ω through $\mathcal{L}_{\text{CD}}$. Then, we freeze ω after convergence to a minimum validation accuracy and start the second stage for the SDF prediction. During the second stage, we jointly optimize $\ddot{o}$, ξ, π, κ, θ, and ψ through the combined loss $\mathcal{L} = \mathcal{L}_V + \mathcal{L}_{\text{SDF}}$. LIST can also be trained end-to-end by jointly minimizing $\mathcal{L}_{CD}$ with $\mathcal{L}_V$ and $\mathcal{L}_{\text{SDF}}$. However, we found the two-stage training procedure easier to evaluate and quicker to converge during experimental evaluation. To reconstruct an object at test time, we first densely sample a fixed 3D grid of query points and predict the signed distance for each point. Then, we use the marching cubes [17] algorithm to extract the target surface from the grid.

4. Experimental Evaluation

In this section, we describe the details of our experimental setup and results. Additional information, including implementation details, can be found in the supplementary material.

4.1. Datasets

Similar to [12] and [19], we utilized the 13-class subset of the ShapeNet [2] dataset to train LIST. The renderings and processed meshes from [35] were used as the input view and target shape. We trained a single model on all 13 categories. Additionally, we employed the Pix3D [31] dataset to test LIST on real-world scenarios. The train/test split from [36] was used to evaluate on all 9 categories of Pix3D. Following [36], we preprocessed the Pix3D target shapes to be watertight for training.

To prepare the ground-truth data, we first normalized the meshes to a unit cube and then sampled 50 k points from the surface of each object. Next, we displaced the sampled points with a Normal distribution of zero mean and varying standard deviation. Lastly, we calculated the signed distance for every point. To supervise the coarse prediction and probabilistic occupancy grid estimation, we sub-sampled 4 k points from the surface via farthest point sampling. Further details regarding the data preparation strategy can be found in the supplementary material.

4.2. Baseline Models

For single-view reconstruction via synthetic images, we compared against the following prior arts: IMNET [3], and D^{2}IM-Net [12]. IMNET does not require pose estimation. However, the reconstruction only unitizes global features from an image. D^{2}IM-Net extracts local features by aligning the query points to image pixels through rendering metadata and it uses a pose estimation module during inference.

For single-view reconstruction from real-world images, we evaluated against TMN [22], MGN [20], and IM3D [36]. TMN deforms a template mesh to reconstruct the target object. MGN and IM3D perform reconstruction through the following steps: (i) identify objects in a scene, (ii) estimate their poses, and (iii) reconstruct each object separately.

4.3. Metrics

We computed commonly used metrics (e.g., CD, intersection over union (IoU), and F-score), to evaluate the performance of LIST. The definitions of these metrics can be found in the supplementary material. Nonetheless, these traditional metrics *do not* differentiate between visible/occluded surfaces since they evaluate the reconstruction as a whole. To investigate the reconstruction quality of occluded surfaces, we propose to isolate visible/occluded surfaces based on the viewpoint of the camera and evaluate them separately using the traditional metrics. A visual depiction of this new strategy is presented in Fig. 3.

To measure the reconstruction quality of occluded surfaces, we first align the predicted/ground-truth meshes to their projection in the input image using the rendering metadata. Then, we assume the camera location as a single source of light and cast rays onto the mesh surface by ray casting [24]. Next, we identify the visible/occluded faces through the ray-mesh intersection and subdivide the identified faces to separate them. Note that the rendering metadata is only used to evaluate the predictions. Finally, we sample 100 k points from the separated occluded faces to compute the CD$_{\text{os}}$, and voxelize the sampled points to compute the IoU$_{\text{os}}$ and F-Score$_{\text{os}}$.

In our implementation, we set the canvas resolution to 4096×4096 pixels and generated one ray per pixel from the camera location. It is important to note that ray casting and computing ray-mesh intersections are computationally demanding tasks. Therefore, to manage time and resources, we chose five sub-classes (chair, car, plane, sofa, table) to evaluate occluded surface reconstruction.

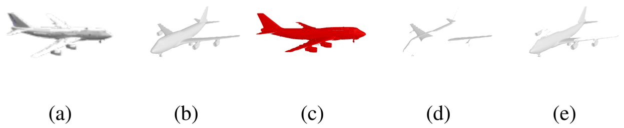

Fig. 3: To evaluate the reconstruction quality of occluded surfaces, we first align the reconstructed shape (b) with the input image (a) and cast rays onto the surface (c). Next, we identify the (red) faces that intersect with the rays via ray-mesh intersection and separate the reconstructed mesh into (d) visible and (e) occluded areas.

4.4. Single-View 3D Reconstruction Evaluation

4.4.1 Single-View 3D Reconstruction from Renderings of Synthetic Objects

In this experiment we performed single-view 3D reconstruction on the test set of the ShapeNet dataset. The qualitative and quantitative results are displayed in Fig. 4 and Table 1, respectively. In comparison to the baselines, the topological structure and occluded geometry recovered by LIST are considerably better. For example, in row 3 all of the baselines struggle to reconstruct the tail of the airplane and they fail to estimate the full length of the wings. In row 5, none of the baselines were able to recover the occluded part of the table. In contrast, LIST not only recovers the structure, but it also maintains the gap in between. Moreover, notice that in row 2 D^{2}IM-Net fails to resolve the directional view ambiguity and imprints an arm shaped silhouette on the seat rather than reconstructing the arm. This indicates a strong influence of the input-view direction in the reconstructed surface. Conversely, LIST can resolve view-directional ambiguity and provide a reconstruction that is uninfluenced by the input-view direction. As shown in Table 1, LIST outperforms all the other baseline models.

We also evaluated LIST against the baselines on occluded surface recovery by partitioning the reconstructions using our proposed metric. The results are recorded in Table 2. LIST outperformed all the baselines hence showcasing the superiority of our approach in reconstructing occluded geometry. Furthermore, LIST provides a stable reconstruction across different views of the same object as shown in Fig. 5. However, the use of ground-truth rendering data instead of the estimated data improved the reconstruction quality. This indicates the source of the problem to be the sub-optimal prediction of the camera pose. Nonetheless,

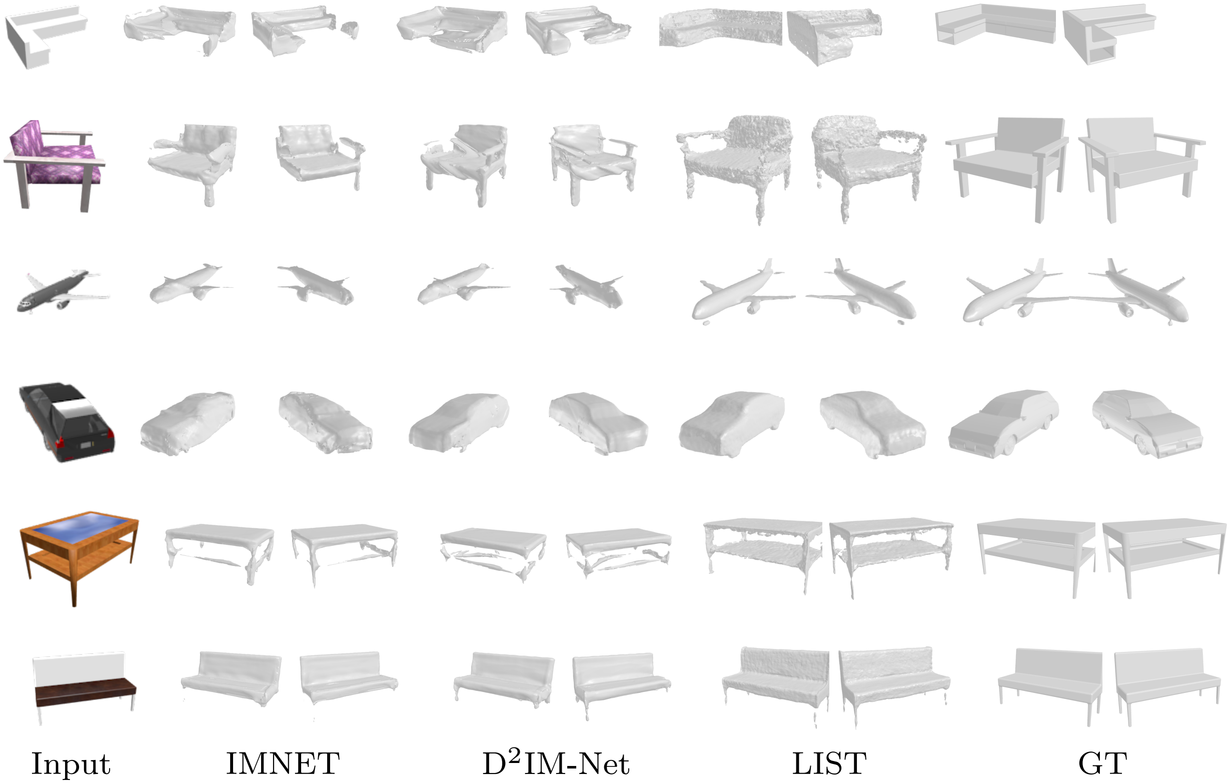

Fig. 4: A qualitative comparison between LIST and the baseline models using the ShapeNet [2] dataset. Our model recovers *significantly better* topological and geometric structure, and the reconstruction is not tainted by the input-view direction. GT denotes the ground-truth objects.

		plane	bench	cabinet	car	chair	display	lamp	speaker	rifle	sofa	table	phone	boat	Mean
CD↓	IMNET	18.95	17.34	15.17	10.86	14.72	16.77	83.64	33.41	10.33	13.35	19.32	9.16	15.24	21.40
	D^2IM-Net	13.25	**12.51**	9.47	7.83	11.31	15.33	**34.08**	17.62	8.55	12.34	14.26	8.11	15.73	13.87
	LIST	**12.13**	13.49	**7.45**	**1.04**	**9.20**	**13.65**	47.31	**16.75**	**7.32**	**9.92**	**11.14**	**7.91**	**15.78**	**13.31**
IoU↑	IMNET	39.43	44.65	49.25	55.75	51.22	53.34	29.26	50.66	46.43	51.12	41.63	52.79	49.61	47.31
	D^2IM-Net	45.44	48.45	48.60	53.58	53.13	52.72	**32.45**	51.75	50.76	53.35	45.17	53.06	52.89	49.33
	LIST	**49.03**	47.57	**56.29**	**65.57**	**52.70**	**57.34**	24.80	**55.34**	**52.42**	**56.79**	**47.90**	**58.98**	**54.35**	**52.23**
F-score↑	IMNET	48.87	31.78	**44.34**	48.78	41.45	48.32	21.23	48.29	52.92	44.12	45.21	51.52	52.31	44.54
	D^2IM-Net	51.37	**36.76**	43.49	51.77	45.56	50.82	**29.57**	51.93	56.25	48.34	47.23	54.84	52.73	47.74
	LIST	**52.46**	36.39	42.51	**53.12**	**46.62**	**51.78**	22.88	**52.67**	**58.24**	**50.52**	**49.62**	**56.89**	**53.58**	**48.25**

Table 1: Quantitative results using the ShapeNet [2] dataset for various models. The metrics reported are the following: chamfer distance (CD), intersection over union (IoU), and F-score. The CD values are scaled by 10^{-3}.

LIST is free from any such complication as our framework does not require any explicit pose estimation.

4.4.2 Single-View 3D Reconstruction from Real Images

In this experiment we evaluated single-view 3D reconstruction on the test set of the Pix3D dataset. The qualitative and quantitative results are provided in Fig. 6 and Table 3, respectively. The baseline results were obtained from the re-

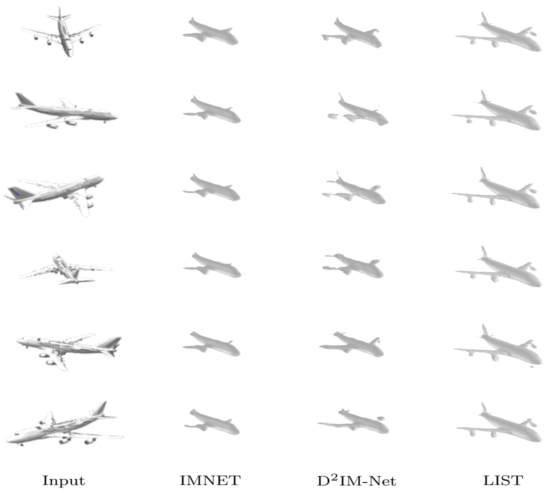

Fig. 5: A qualitative comparison between LIST and the baseline models using distinct views of the same object. Not only can our model both maintain better topological structure and geometric details, but it also provides a reconstruction that is stable across different views of the object.

		plane	car	chair	sofa	table	Mean
$CD_{os}\downarrow$	IMNET	24.11	13.34	15.47	24.34	26.86	20.82
	D^{2}IM-Net	26.23	13.44	13.59	20.45	23.45	19.43
	LIST	**18.93**	**6.57**	**12.66**	**18.44**	**21.76**	**15.67**
$IoU_{os}\uparrow$	IMNET	45.63	46.87	38.32	45.87	39.02	43.14
	D^{2}IM-Net	48.44	50.33	49.43	50.32	42.22	48.14
	LIST	**53.15**	**55.37**	**51.25**	**55.22**	**43.17**	**51.63**
F_{os}-score$\uparrow$	IMNET	40.93	46.94	44.43	46.84	45.64	44.95
	D^{2}IM-Net	47.21	50.73	48.89	49.15	47.72	48.73
	LIST	**50.33**	**52.55**	**49.34**	**51.02**	**48.11**	**50.27**

Table 2: A quantitative evaluation of the occluded surfaces of reconstructed synthetic objects via our evaluation strategy. The metrics reported are the following: chamfer distance (CD_{os}), intersection over union (IoU_{os}), and F_{os}-score. The CD_{os} values are scaled by 10^{-3}.

spective papers. Compared to other methods our approach generates the most precise 3D shapes, which results in the lowest average CD and F-score. Notice that in Fig. 6, rows 3 and 4, only LIST can accurately recover the back and legs of the chair. Additionally, LIST reconstructions provide a smooth surface, precise topology, and fine geometric details.

4.5. Ablation Study

4.5.1 Setup

To investigate the impact of each individual component in our single-view 3D reconstruction model, we performed an ablation study with the following network options.

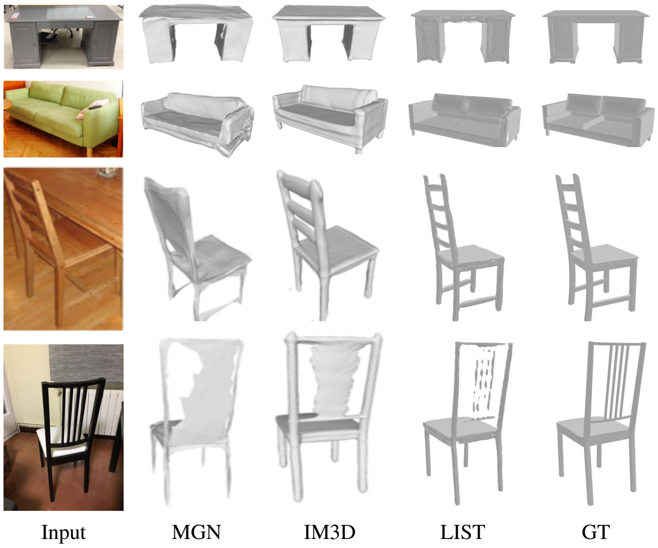

Fig. 6: Single-view reconstruction using real-world images from the Pix3D [31] test set (best viewed zoomed in).

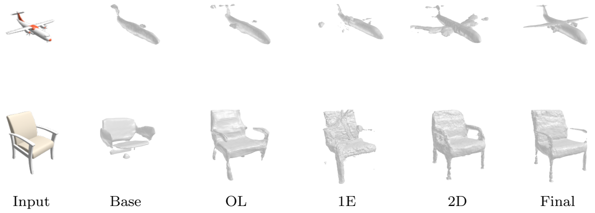

Fig. 7: Qualitative results obtained from the ablation study using different network settings.

- *Base:* A version of LIST that predicts the signed distance utilizing only global image features and coarse predictions.
- *OL:* An improved *Base* version that uses the probabilistic occupancy from the coarse prediction and occupancy loss.
- *1E:* A version of LIST where local and global image features from the same encoder are used for both coarse prediction and localized query feature extraction.
- *2D:* LIST with two separate decoders to estimate the signed distance from local and global query features. The final prediction is obtained by adding both estimations.
- *EC:* We train LIST without the localization module and use a separate pose estimation module similar to [12] to predict the camera parameters. The estimated

		bed	bookcase	chair	desk	sofa	table	tool	wardrobe	misc	Mean
CD↓	TMN	7.78	5.93	6.86	7.08	4.25	17.42	4.13	4.09	23.68	9.03
	MGN	5.99	6.56	**5.32**	5.93	3.36	14.19	3.12	3.83	26.93	8.36
	IM3D	**4.11**	3.96	5.45	7.85	5.61	11.73	2.39	4.31	24.65	6.72
	LIST	5.81	**1.74**	6.11	**3.87**	**2.08**	**1.68**	**1.99**	**0.80**	**5.16**	**4.36**
IoU↑	LIST	45.61	39.54	41.15	59.68	67.34	49.12	27.82	43.87	34.72	46.77
F-score↑	LIST	58.18	67.22	60.01	78.34	70.14	69.19	46.48	75.70	39.14	65.66

Table 3: A quantitative evaluation of the occluded surfaces of reconstructed real-world objects using our evaluation strategy. The metrics reported are the following: chamfer distance (CD_{os}), intersection over union (IoU_{os}), and F_{os}-score. The CD_{os} values are scaled by 10^{-3}.

	Base	OL	1E	2D	EC	Final
CD↓	11.35	9.64	10.72	8.48	7.89	**7.32**
IoU↑	51.34	53.95	51.40	55.23	55.10	**56.83**
F-score↑	43.11	48.06	45.92	51.37	51.33	**52.75**

Table 4: Quantitative results obtained from the ablation study using different network settings.

camera parameters were used to transform the query points during inference.

To maximize limited computational resources, we focused on the most diverse five sub-classes (chair, car, plane, sofa, table) of the ShapeNet dataset for this ablation study. The qualitative and quantitative results of the experiments are recorded in Fig. 7 and Table 4 respectively.

4.5.2 Discussion

In the ablation experiments the *Base* version was able to recover global topology, but it lacked local geometry. As shown in Fig 7, the probabilistic occupancy and optimization loss helped recover some details in the *OL* version. Conversely, the performance decreased slightly after the inclusion of local details in the single-encoder version (*1E*). We hypothesize that the task of query point localization, while estimating the coarse prediction, overloads the encoder and hinders meaningful feature extraction for the signed distance prediction. To overcome this issue, we used a separate encoder for the coarse prediction and query point localization. The dual-decoder version (*2D*), performed similar to the final model. Nonetheless, we found that the geometric details had a thicker reconstruction than the target during qualitative evaluation. This motivated the fusion of features rather than predictions in the final version.

We also ablated the localization module using estimated camera parameters during training and inference. As shown in Table 4, the final version of LIST outscores the version employing estimated camera (*EC*) parameters. This indicates that our localization module with an SDF prediction objective is more suitable for single-view reconstruction compared to a camera pose estimation sub-module. More importantly, this removes the requirement for pixel-wise alignment through camera parameters for local feature extraction. Note that the *EC* reconstruction appears qualitatively similar to the others and was therefore omitted in Fig. 7.

4.6. Limitations and Future Directions

Although LIST achieves state-of-the-art performance on single-view 3D reconstruction, there are some limitations. For example, the model may struggle with very small structures. We speculate that this is due to the coarse predictor failing to provide a good estimation of such structures. Please see the supplementary material for examples of failed reconstruction results. Another shortcoming is the need for a clear image background. LIST can reconstruct targets from real-world images, yet it requires an uncluttered background to do this. In the future, we will work towards resolving these issues.

5. Conclusion

In this paper we introduced LIST, a network that implicitly learns how to reconstruct a 3D object from a single image. Our approach does not assume weak perspective projection, nor does it require pose estimation or rendering data. We achieved state-of-the-art performance on single-view reconstruction from renderings of synthetic objects. Furthermore, we demonstrated domain transferability of our model by recovering 3D surfaces from images of real-world objects. We believe our approach could be beneficial for other problems such as object pose estimation and novel view synthesis.

Acknowledgments

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing software, computational, and storage resources that have contributed to the research results reported within this paper.

References

[1] Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, and Kwan-Yee K Wong. Jiff: Jointly-aligned implicit face function for high quality single view clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2729–2739, 2022. 2

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 4, 6

[3] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2, 5

[4] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision*, pages 628–644. Springer, 2016. 3

[5] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5868–5877, 2017. 2

[6] Shivam Duggal and Deepak Pathak. Topologically-aware deformation fields for single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1536–1546, 2022. 2

[7] Kui Fu, Jiansheng Peng, Qiwen He, and Hanxiao Zhang. Single image 3d object reconstruction based on deep learning: A review. *Multimedia Tools and Applications*, 80(1):463–498, 2021. 2

[8] Jorge Fuentes-Pacheco, José Ruiz-Ascencio, and Juan Manuel Rendón-Mancha. Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review*, 43(1):55–81, 2015. 1

[9] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pages 9276–9287, 2020. 2, 3

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 28, 2015. 3

[11] Erwin Kruppa. *Zur Ermittlung eines Objektes aus zwei Perspektiven mit innerer Orientierung*. Hölder, 1913. 1

[12] Manyi Li and Hao Zhang. D2im-net: Learning detail disentangled implicit fields from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10246–10255, 2021. 2, 3, 4, 5, 7

[13] https://github.com/robotic-vision-lab/Learning-Implicitly-From-Spatial-Transformers-Network. 1

[14] Gidi Littwin and Lior Wolf. Deep meta functionals for shape representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1824–1833, 2019. 2

[15] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 32, 2019. 2

[16] H Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133–135, 1981. 1

[17] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM Siggraph Computer Graphics*, 21(4):163–169, 1987. 4

[18] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2

[19] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022. 2, 4

[20] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 5

[21] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2

[22] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single rgb images via topology modification networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9964–9973, 2019. 5

[23] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[24] Scott D Roth. Ray casting for modeling solids. *Computer Graphics and Image Processing*, 18(2):109–144, 1982. 5

[25] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2, 3

[26] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2

[27] Muhamad Risqi U Saputra, Andrew Markham, and Niki Trigoni. Visual slam and structure from motion in dynamic environments: A survey. *ACM Computing Surveys*, 51(2):1–36, 2018. 1

[28] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 1

[29] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 3d point cloud generative adversarial network based on tree structured graph convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3859–3868, 2019. 3

[30] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 3

[31] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2018. 4, 7

[32] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2019. 2

[33] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. 1

[34] Rundi Wu, Yixin Zhuang, Kai Xu, Hao Zhang, and Baoquan Chen. Pq-net: A generative part seq2seq network for 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 829–838, 2020. 2

[35] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 32, 2019. 2, 3, 4

[36] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8833–8842, 2021. 4, 5

[37] Fang Zhao, Wenhao Wang, Shengcai Liao, and Ling Shao. Learning anchored unsigned distance functions with gradient direction alignment for single-view garment reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12674–12683, 2021. 2