A-DIF: Articulated Deformed Implicit Field: Modeling Articulated Shapes with Learned Dense Correspondence

摘要

本文提出了一种具有潜在空间迁移能力的新的变形隐式场表示方法，用于建模不同拓扑结构的机器手三维形状，并在形状之间产生密集的对应关系。通过 DIF-Net，三维形状由跨类别共享的模板隐式字段表示，以及三维变形字段和专用于每个形状实例的校正字段。利用它们的变形场，可以很容易地建立形状对应关系。通过解耦出两个潜在空间——形状和姿态，网络联合学习两个潜在空间和这些场，而不使用任何对应或零件标签。实验表明，本文的方法可以实现姿态迁移以及形状迁移，并取得不错的重建结果。

关键词：三维重建；隐式表示；姿态迁移；对应关系

1 引言

机器手的形状和拓扑结构在机器人感知领域中起着至关重要的作用，这使得高保真的机器手建模至关重要。机器手是具有不同类别的三维对象，如果可以在同一个类共享一些共同的形状特征和语义对应，可以用来构建一个可变形的形状模型，有利于不同的下游任务如形状理解 [28]，机器人重建 [31] 和物体抓取 [26]。学习具有密集对应的三维形状模型是计算机视觉和图形学中一项长期的任务。然而，现有的工作大多集中在具有一致几何拓扑的对象类，如人脸和身体 [3]。这些对象类别中的形状可以预先对齐，以构建三维模型。最近的基于深度学习的方法直接学习三维对象 [30] 的潜在空间。虽然这些方法可以建模复杂的对象，但它们不处理 3D 形状之间的密集对应。

除了形状间的密集对应，机器手具有高度铰接，复杂的姿态给密集对应的学习带来困难。首先，手的几何形状的变形很难建模。机器手不同于人体，基于剥皮的方法很难为任意查询 [2] 找到准确的剥皮权重，而部分感知的方法通常存在跨部分不一致的问题 [15]。其次，铰接物体通常具有复杂的姿态，这使得姿态在建模中也十分重要。因此，如何在形状对应关系学习中考虑姿态因素也是一个很重要的课题。

受 Deng 等人提出的基于隐式表示的对应关系学习网络 DIF-Net [7] 启发，本文提出了一种具有潜在空间迁移能力的新的变形隐式场表示方法，用于建模不同拓扑结构的机器手三维形状，并在形状之间产生密集的对应关系。通过 DIF-Net，三维形状由跨类别共享的模板隐式字段表示，以及三维变形字段和专用于每个形状实例的校正字段。利用它们的变形场，可以很

容易地建立形状对应关系。除此之外，受到 Mu 等人提出的铰接物体隐式表示场 A-SDF [20] 启发，通过解耦出两个潜在空间——形状和姿态，网络联合学习两个潜在空间和这些场，而不使用任何对应或零件标签，便可以实现姿态和形状这两个潜在空间表示一个复杂的机器手，从而可以实现姿态迁移和形状迁移。

2 相关工作

2.1 基于隐式场的形状表示

现有的大量工作 [4, 6, 8, 11] 都集中于研究高效和准确的三维对象表示。最近的研究表明 [5, 16, 19]，将 3D 对象表示为连续的可微隐式函数，可以以一种记忆高效的方式建模各种拓扑。其基本思想是利用神经网络来参数化一个形状作为三维的决策边界。这些工作大多局限于静态对象和场景的建模 [10]。与之前的工作不同，我们的方法通过学习一个解纠缠的隐式表示，在类别层次上建模连接对象，并在不同类别不同形状不同姿态的机器手上测试我们的模型。

2.2 可驱动的人体部件

针对这个问题，有些工作利用参数网格模型 [18]，通过直接推断形状和关节参数，来估计面部 [25]、手 [9]、人体 [24] 和动物 [17] 的形状和关节。然而，这种参数模型需要专家的大量努力来构建，因此很难推广到大规模的对象类别。为了解决这一挑战，另一些工作 [21, 22, 29] 使用神经网络从数据中学习形状。例如，Niemeyer 等人 [21] 学习了一个隐式向量场，在一个时空空间中分配每个点的运动向量和变形形状。但是，该设计不允许单独控制每个部件。最近，Hoang 等人的 [29] 使用斑块定义了形状，并且可以通过操纵每个斑块定义的外部参数来改变整体形状。然而，学习到的斑块并不对应于零件，且该方法在大变形时失效。相比之下，本文的方法在一般的类人体部件上是类别级的，假设没有零件标签。因此，这些以前的方法不能与本文方法进行直接的比较。此外，本文用一个姿态潜在空间表示建模关节机器手姿态，这允许本文方法可以实现姿态迁移。

2.3 解耦表示

解耦表示侧重于在低维空间中建模复杂的变化，其中单个因素控制不同类型的变化。先前的工作 [1, 12–14, 32] 已经表明，解耦表示对于学习有意义的潜在空间是至关重要的。例如，Zhou 等人 [32] 提出了一种自动编码器架构来解开人的姿态和形状。本文通过实现姿态和形状这两个潜在空间表示一个复杂的机器手，从而可以实现姿态迁移和形状迁移。

3 本文方法

3.1 本文方法概述

我们提出了一种具有潜在空间迁移能力的新的变形隐式场表示方法，用于重建和预测不同姿态下的机器手三维形状。我们的模型将采样的三维点位置、形状代码和姿态代码作为输入，并输出 SDF 值（有符号的距离），测量一个点到最近的表面点的距离。本文关键的见解

是，同一实例的所有形状代码都应该是相同的，独立于它的姿态。我们认为，即使在同一实例的不同姿态的形状看起来很不同，一个好的表示应该在低维空间中捕捉这种可变性，因为零件的几何形状保持不变。图1显示了对我们的方法的概述。

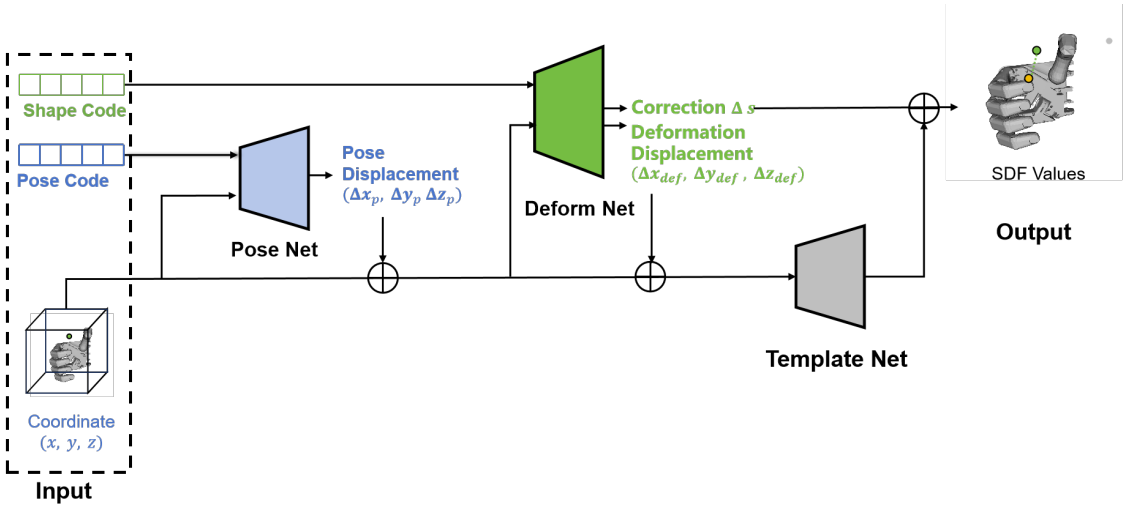

图 1. 方法示意图

3.2 解耦表示模块

假设一个类别的机器手有 M 个姿态，共有 N 个类别。每个实例 $X_{n,m}$ 可以表示为一个形状编码 α_n 和一个姿态编码 β_m。形状代码 α_n 在相同类别不同姿态的实例上共享。在训练期间，我们会为每个实例维护和更新一个形状代码。

设 $x \in \mathbb{R}^3$ 是一个实例的采样点。将 α 和 β 表示为相应的形状和姿态编码。如图1所示，最终用自动解码器结构定义了一个姿态相关的有符号距离函数 f，由模板 SDF 生成网络 T、变形网络 D 和姿态网络 P 组成:

$$f(\alpha, \beta, p) = T(D(P(\beta, x), \alpha)) = s \tag{1}$$

其中，$s \in \mathbb{R}$ 是一个标量 SDF 值（到三维曲面的有符号距离）。SDF 值的符号表示该点是在水密表面的内部（负）还是外部（正）。三维形状由零水平集 $f(\cdot) = 0$ 隐式表示。

3.2.1 模板 SDF 生成网络

为了捕获不同类别机器手的公共结构，本文学习了一个模板 SDF 生成网络 T:

$$T : x \in \mathbb{R}^3 \rightarrow \tilde{s} \in \mathbb{R} \tag{2}$$

它将一个三维点 x 映射到一个标量值 $\tilde{s}$。后者用于通过变形和修正来构造特定对象的 SDF，这将在后面描述。T 的网络权值在不同类中共享，因此它被强制来学习类间的公共模式。我们的模板字段通过其在体积中的不同等表面记录一个类别中的所有结构变化，而基于网格的模板只在其表面有意义。这种差异使我们的方法能够更好地学习具有结构差异的对象的对应关系。

3.2.2 变形网络

为了获得某个类别的 SDF，我们学习了一个变形网络 D 来预测一个变形场，以及在模板场 T 上的一个修正场：

$$D_\omega : x \in \mathbb{R}^3 \rightarrow (v, \Delta s) \in \mathbb{R}^4 \tag{3}$$

其中 $v \in \mathbb{R}^3$ 是变形流，$\Delta s \in \mathbb{R}$ 是标量修正。D 的权值来自一个超参数网络 Ψ，通过输入形状编码 α，输出 D 的权值。因此，D 的权值是特定于形状的，即来自同一个类别的机器手关于 D 的权值是一致的。

3.2.3 姿态网络

给定一个实例的采样点 $x \in \mathbb{R}^3$，我们通过每个实例独有的姿态编码，并利用姿态网络 P 将不同姿态变换到同样的姿态潜在空间中：

$$P_\phi : x \in \mathbb{R}^3 \rightarrow v' \in \mathbb{R}^3 \tag{4}$$

P 的权值来自一个超参数网络 Φ，通过输入姿态编码 β，输出 P 的权值。因此，P 的权值是每个实例独有地。姿态网络 P 的设计见解是因为在进入变形网络 D 中，不同姿态的机器手应该是姿态无关的，这样才能使得变形网络专注于不同的形状：姿态网络专注于不同的姿态，变形网络专注于不同的形状。

总之，我们的模型可以总结性地写成如下公式：

$$f(\alpha, \beta, x) = T(x + P^{v'}_{\Phi(\beta)}(x) + D^{v}_{\Psi(\alpha)}(x + P^{v'}_{\Phi(\beta)}(x))) + D^{\Delta s}_{\Psi(\alpha)}(x + P^{v'}_{\Phi(\beta)}(x)) \tag{5}$$

3.3 损失函数定义

给定一组实例的集合，我们首先应用一个类似于 [27] 的 SDF 回归损失来学习这些形状的 SDF。$f_i(\alpha, \beta, x)$ 是预测的 SDF 值，i 表示实例索引，我们有

$$\begin{aligned} L_{sdf} = \sum_i \Big(& \sum_{x \in \Omega} |f_i(\alpha, \beta, x) - \bar{s}| + \sum_{x \in \mathcal{S}_i} (1 - \langle \nabla f_i(\alpha, \beta, x), \bar{n} \rangle) \\ & + \sum_{x \in \Omega} |\|\nabla f_i(\alpha, \beta, x)\|_2 - 1| + \sum_{x \in \Omega \setminus \mathcal{S}_i} \rho(f_i(\alpha, \beta, x)) \Big) \end{aligned} \tag{6}$$

如同 [23] 中，我们也应用正则化损失来约束学习到地形状和姿态潜在码：

$$L_{shapeReg} = \sum_j^N \|\alpha_j\|_2^2. \tag{7}$$

$$L_{poseReg} = \sum_i^M \|\beta_i\|_2^2. \tag{8}$$

为了促进平滑变形，避免较大的形状变形，我们在变形场上增加了一个简单的平滑损失：

$$L_{smooth} = \sum_i \sum_{x \in \Omega} \sum_{d \in \{X,Y,Z\}} \|\nabla D^{v}_{\Psi(\alpha)}|_d (x + P^{v'}_{\Phi(\beta)}(x))\|_2, \tag{9}$$

这惩罚了变形场沿 X、Y 和 Z 方向的空间梯度。

为了通过隐式场变形而不是修正来促进形状修正，我们将修正场最小化：

$$L_c = \sum_i \sum_{x \in \Omega} |D^{\Delta s}_{\Psi(\alpha)}(x + P^{v'}_{\Phi(\beta)}(x))|. \tag{10}$$

综上所述，整个训练过程可以表述为以下优化问题：

$$\underset{\{\alpha_j\},\{\beta_i\},P,D,T}{\arg\min} \quad L_{sdf} + +w_1 L_{smooth} + w_2 L_c + w_3 L_{shapeReg} + w_4 L_{poseReg} \tag{11}$$

4 复现细节

4.1 与已有开源代码对比

受 Deng 等人提出的基于隐式表示的对应关系学习网络 DIF-Net [7] 启发，本文提出了一种具有潜在空间迁移能力的新的变形隐式场表示方法，用于建模不同拓扑结构的机器手三维形状，并在形状之间产生密集的对应关系。因此，本文参考 DIF-Net 的网络框架，并基于此改进提出新功能，得到自身的网络结构。更详细的说，与 DIF 源代码相比，本文源代码不同的地方在于：

- 实现三维数据采样空间点构建训练集的代码，此代码在 DIF-Net 中没有开源；
- 新增姿态编码以及 PoseNet 网络，并串联整个网络训练流程；
- 优化推理阶段，推理阶段冻结形状编码，只优化姿态编码；
- 构建关于姿态编码的正则化损失，去除法向量一致性损失。

4.2 创新点

本文与基线网络 DIF-Net 相比，创新点主要如下：

- 同一实例的所有形状代码都应该是相同的，独立于它的姿态。即使在同一实例的不同姿态的形状看起来很不同，一个好的表示应该在低维空间中捕捉这种可变性，因为零件的几何形状保持不变。因此，设计形状编码和姿态编码解耦地表示机器手三维模型；
- 优化推理阶段，推理阶段冻结形状编码，只优化姿态编码；
- 实现 DIF-Net 所没有的功能：形状迁移与姿态迁移。

5 实验结果分析

5.1 原始 DIF-Net 复现情况

本文首先复现了 DIF-Net，并通过机器手数据集可视化重建结果，如图2所示。可以看出，DIF-Net 在重建复杂形状的机器手数据集上可以取得较好的重建结果，体现了基线网络的有效性。

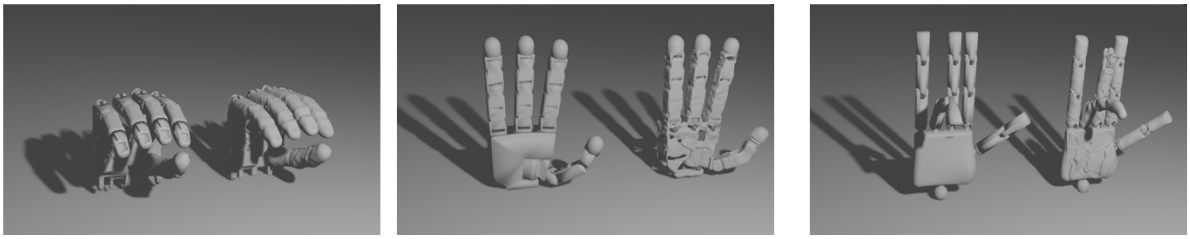

图 2. 原始 DIF-Net 在机器手数据集上的重建结果。每张图中左边是真实值，右边是重建结果。

5.2 A-DIF 复现情况

首先是本文的重建结果，如图3所示。可以看出，本文重建结果整体重建结果会出现粘连情况，可能的原因是在形状编码和姿态编码上没有给更好的先验知识，只是让网络去学习解耦，所以在手指上会出现粘连情况，这也可以看出姿态编码还没能很好地与形状编码解耦出来。细节上的重建效果比 DIF-Net 要好，这是因为本文考虑了同一个类别之间共同优化一个形状编码，使得网络更能学习到类内的公共信息。

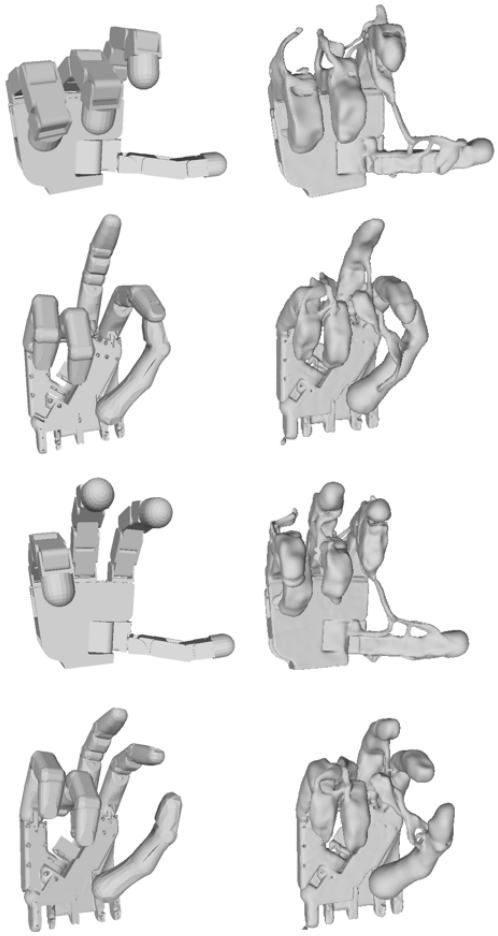

图 3. 本文重建结果示意图。每张图中左边是真实值，右边是重建结果。

接下来是迁移实验。给定一个实例，网络先通过推理阶段得到表示该实例的形状编码和姿态编码，该实例的姿态编码通过与另一个类别的形状编码组合，得到新的实例，结果如图4所

示。实验结果表明，迁移结果整体轮廓可以看出形状上以及姿态上的对应，但也反映出形状编码和姿态编码还没有很好地解耦。

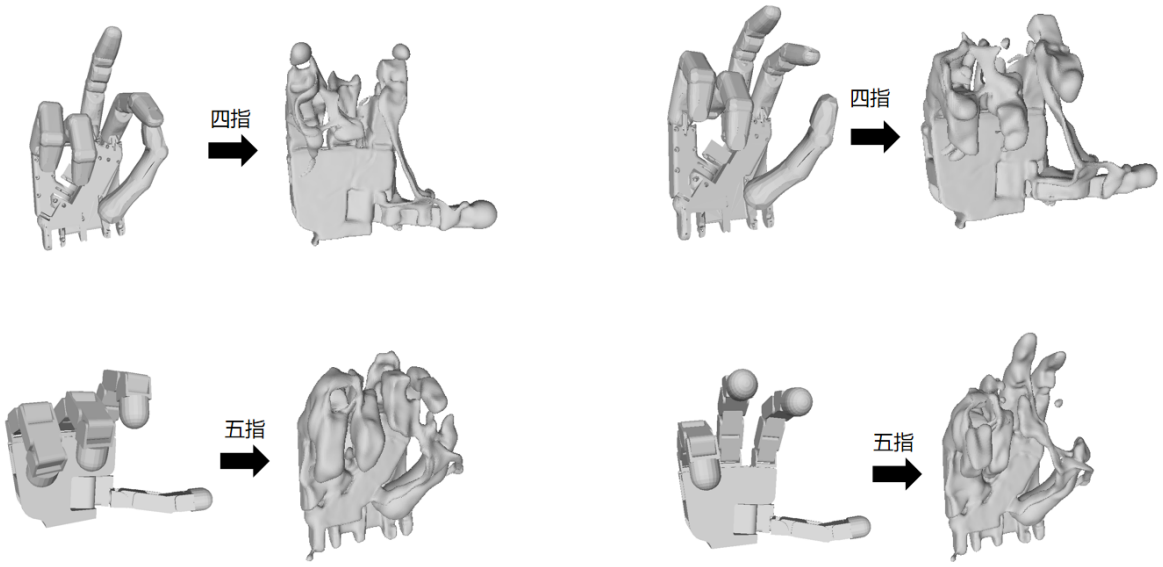

图 4. 本文迁移效果示意图。给定一个由形状编码和姿态编码表示的实例（左边），通过更换另一个类别的形状编码，得到新的实例（右边）。

6 总结与展望

本文提出 A-DIF 网络，一种解耦形状和姿态的新的变形隐式场表示方法，用于建模不同拓扑结构的机器手三维形状，并在形状之间产生密集的对应关系。该方法时基于 DIF-Net 提出的改进方案，通过解耦出姿态编码与形状编码表示一个三维形状，使得 DIF-Net 具有姿态迁移能力。

通过实验表明 A-DIF 在重建结果以及迁移上仍然有不足的地方。比如在形状编码和姿态编码上没有给更好的先验知识，只是约束了同一类别的形状有相同的形状编码，让网络自己通过数据去学习解耦。如何更好地解耦是未来亟需解决的一个问题。

参考文献

[1] Ivan Anokhin, Pavel Solovev, Denis Korzhenkov, Alexey Kharlamov, Taras Khakhulin, Aleksei Silvestrov, Sergey Nikolenko, Victor Lempitsky, and Gleb Sterkin. High-resolution daytime translation without domain labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7488–7497, 2020.

[2] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. *Advances in Neural Information Processing Systems*, 33:12909–12922, 2020.

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023.

[4] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. Bsp-net: Generating compact meshes via binary space partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 45–54, 2020.

[5] 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.

[6] Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnet: Learnable convex decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31–44, 2020.

[7] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10286–10296, 2021.

[8] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[9] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019.

[10] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4857–4866, 2020.

[11] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9785–9795, 2019.

[12] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

[13] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[15] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020.

[16] Amit Kohli, Vincent Sitzmann, and Gordon Wetzstein. Inferring semantic information with 3d neural scene representations. *arXiv preprint arXiv:2003.12673*, 2020.

[17] Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 452–461, 2020.

[18] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[19] 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.

[20] Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, and Xiaolong Wang. A-sdf: Learning disentangled signed distance functions for articulated shape representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13001–13011, 2021.

[21] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5379–5389, 2019.

[22] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12695–12705, 2021.

[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.

[24] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021.

[25] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, pages 704–720, 2018.

[26] Qijin She, Ruizhen Hu, Juzhan Xu, Min Liu, Kai Xu, and Hui Huang. Learning high-dof reaching-and-grasping via dynamic representation of gripper-object interaction. *arXiv preprint arXiv:2204.13998*, 2022.

[27] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

[28] Todor Stoyanov, Robert Krug, Rajkumar Muthusamy, and Ville Kyrki. Grasp envelopes: Extracting constraints on gripper postures from online reconstructed 3d models. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 885–892. IEEE, 2016.

[29] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Carsten Stoll, and Christian Theobalt. Patchnets: Patch-based generalizable deep implicit 3d shape representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 293–309. Springer, 2020.

[30] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016.

[31] Daniel Yang, Tarik Tosun, Benjamin Eisner, Volkan Isler, and Daniel Lee. Robotic grasping through combined image-based grasp proposal and 3d reconstruction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6350–6356. IEEE, 2021.

[32] Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Unsupervised shape and pose disentanglement for 3d meshes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 341–357. Springer, 2020.