基于扩散模型的手物交互图像生成与优化

刘 畅[1] 姚剑敏[1,2] 陈恩果[1] 严 群[1]
LIU Chang YAO Jianmin CHEN Enguo YAN Qun

摘 要 手物体交互图像的生成是计算机视觉和人机交互领域的一个重要挑战。准确生成这类图像对于理解物体可供性、改进人机交互系统以及增强虚拟现实体验至关重要。然而，现有方法在处理复杂物体、罕见姿势和遮挡关系时仍面临诸多困难，常常导致生成的手部形状、位置和姿态存在不自然或失真的情况。鉴于此，文章提出了一种基于扩散模型的两阶段方法，用于从单一 RGB 物体图像生成手物体交互图像。该方法包括两个关键组件：MaskNet 和 FusionNet。其中，MaskNet 负责预测手物体交互的空间布局；FusionNet 则基于预测的布局生成详细的交互图像。文章利用扩散模型的强大生成能力，在潜在空间中进行高效的图像生成。实验结果表明，该方法在生成真实、多样化的手物交互图像方面表现出色，在 FID、LPIPS 和 UPR 等指标上均表现良好。此外，该方法还展现出良好的泛化能力，能够处理未见过的物体类别和复杂场景。

关键词 手物交互；图像合成；扩散模型；物体可供性；布局预测

doi：10.3969/j.issn.1672-9528.2025.01.007

0 引言

近年来，图像生成领域取得了巨大的进展，特别是在文本到图像的生成方面。然而，当需要基于特定物体图像来生成逼真的人手物体交互（hand object interaction，HOI）场景时，现有方法仍面临诸多挑战。这项任务需要模型能够理解物体的物理约束、功能语义，以及与人手交互时的遮挡关系等复杂因素。

如何准确生成包含手部与物体交互场景一直是一大挑战。Shan 等人 [1] 提出了通过推断参与交互的手部来实现从人类用手操作的互联网视频中可靠地提取手部状态信息的步骤。Zhang 等人 [2] 使用 DNN 对两个输入深度流中的手部和物体进行分割，并通过预训练的 LSTM[3] 网络根据先前的姿势来预测当前的手部姿势。Hu 等人 [4] 提出了一种新颖的 HOGAN 框架，该框架利用表达模型感知的手和物体表示和利用其固有的拓扑结构来构建统一的表面空间。人机交互和计算机视觉领域中，理解物体的可供性（affordance）对于许多应用至关重要，如机器人操作、增强现实和人机交互系统。而 Ye 等人 [5] 提出的 Affordance Diffusion 方法在从单一物体图像合成手部 - 物体交互图像方面取得了显著进展。该方法利用两阶段扩散模型架构，首先生成交互布局，然后合成详细的交互图像。尽管 Affordance Diffusion 在生成逼真的手部物体交互图像方面表现出色，但仍存在一些局限性。首先，该方法在处理复杂物体或罕见交互姿势时可能会遇到困难；其次，生成的图像有时缺乏细节或存在不自然的手部姿势。

为解决手物交互图像生成的计算效率和生成图像的精细度，本文提出一种改进的两阶段网络架构。主要贡献包括：MaskNet 对与关节无关的手对象交互布局进行采样，FusionNet 根据预测的布局修复图片，合成手抓握对象的图像。两者都建立在大规模预训练扩散模型之上。实验证明，这一改进在多个数据集上显著提升了性能，特别是在处理复杂物体和罕见姿势方面。不仅提高了手部物体交互图像的质量和多样性，更增强了模型在实际应用中的鲁棒性和效率。

1 模型构建

1.1 关键模型

去噪扩散概率模型（denoising diffusion probabilistic models，DDPM）是一种强大的生成模型 [6]，其核心思想是通过逐步添加和移除噪声来学习数据分布$x_0 \sim q(x)$，通常用来生成逼真的图形、音频等数据，如图 1 所示。

1. 福州大学物理与信息工程学院 福建福州 350108
2. 晋江博感公司 福建晋江 362200
[基金项目] 国家自然科学基金（62175032）；福建省杰出青年基金项目（2024J010046）；福建省技术创新重点攻关及产业化项目（2024G020）；闽都创新实验室产学研融合发展专项（2024CXY106）

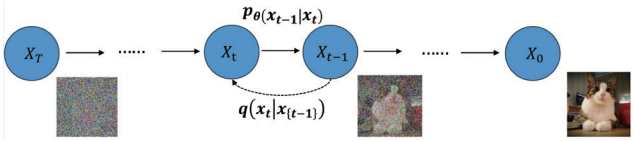

图 1 扩散模型示意图

前向过程逐步向数据添加噪声产生一系列带噪声图片 $x_1, x_2, \cdots, x_T$，加噪过程公式为：

$$q(x_t|x_{t-1}) = N\left(x_t; \sqrt{1-\beta}x_{t-1},\ \beta_t I\right) \tag{1}$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{2}$$

随着 t 的不断增大，最终原始数据 x_0 会逐步失去它的特征。最终当 $T\to\infty$ 时，x_T 趋近于一个各向独立的高斯分布。从视觉上来看，即原本一张完好的照片经过多个加噪步骤后，几乎成为一张完全是噪声的图片。

反向过程如果将上述过程转换方向，即从 $q(x_{t-1}|x_t)$ 中采样，就可以从一个随机的高斯分布 $N(0\sim I)$ 中重建出一个真实的原始样本，即从一个完全杂乱无章的噪声图片中得到一张真实图片。

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{3}$$

$$p_\theta(x_{t-1}|x_t) = N\left(x_{t-1}; \mu_\theta(x_t,t); \sum(x_t,t)_\theta\right) \tag{4}$$

DDPM 的训练目标是最小化预测噪声和实际添加噪声之间的差异，通常使用 L2 损失。

$$L_{\text{DDPM}}[x;c] = E_{(x,c),\ \varepsilon\sim N(0,I),t} \| x - D_\theta(x_t,t,c)\|_2^2 \tag{5}$$

这种方法在生成高质量样本和稳定训练方面表现出色，但采样速度较慢。DDPM 已在图像生成、编辑等多个领域取得显著成果，并衍生出了多种改进版本，如加入条件信息的变体。在本文中基于 DDPM 的原理开发了 MaskNet 和 FusionNet，利用其高质量生成能力和灵活的条件控制来实现手 - 物体交互的精确合成，如图 2 所示。

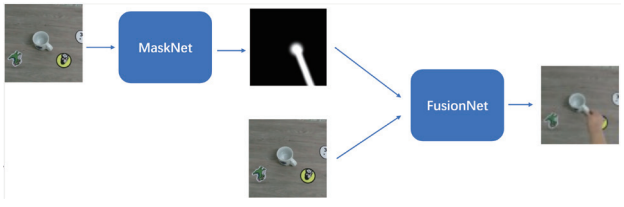

图 2 整体网络结构

1.2 MaskNet

MaskNet 是本文提出的两阶段方法中的第一个网络，其主要目的是生成手 - 物体交互的空间布局。它接收单一的 RGB 物体图像作为输入，输出一个 5 维的布局参数向量 $l := (a, x, y, b_1, b_2)$，其中 a 代表手掌大小，(x, y) 是位置，$\arctan(b_1, b_2)$ 表示接近方向，如图 3 所示。

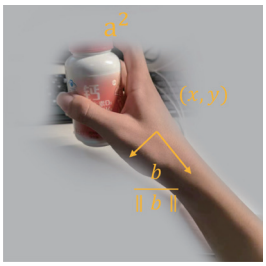

图 3 手部代理示意图

MaskNet 的网络架构基于条件扩散模型，使用 UNet[7] 作为骨干网络，并包含交叉注意力层。它采用无关节的手部代理（hand proxy）来表示基本手部结构，包括手掌和前臂。在扩散过程中，网络在每个步骤接收噪声布局参数 l_t 和物体图像，输出去噪后的布局向量 l_{t-1}。为了进行空间推理，MaskNet 将布局参数映射到图像空间 $M(l_t)$，并将映射后的布局掩码与物体图像连接作为网络输入，如图 4 所示。

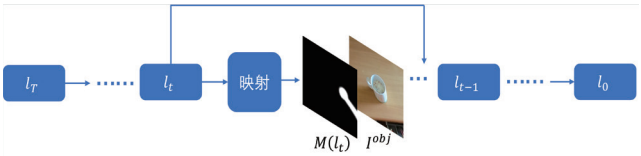

图 4 MaskNet 结构

损失函数包括参数空间的 DDPM 损失，其中 I^{obj} 是只有物体的 RGB 图像：

$$L_{\text{para}} := L_{\text{DDPM}}[l; I^{\text{obj}}] \tag{6}$$

图像空间的损失：

$$L_{\text{mask}} = E_{(l_0,I^{\text{obj}}),\ \varepsilon\sim N(0,I),t} \| M(l_0) - M(\widehat{l_0})\|_2^2 \tag{7}$$

总损失表示为：

$$L_{\text{mask}} + \lambda L_{\text{para}} \tag{8}$$

训练策略同时在参数空间和图像空间应用损失，在早期扩散步骤中，参数空间损失提供更强的训练信号。MaskNet 还支持引导生成，允许在测试时添加额外条件，如指定位置。网络的输入是 7 通道图像，包括 3 通道物体图像，1 通道布局掩码和 3 通道混合图像。噪声布局参数通过空间注意力机制与 UNet 的特征图交互。MaskNet 的独特之处在于将复杂的手物交互问题简化为布局预测任务，并利用扩散模型的强大生成能力来处理这一任务。通过在参数空间和图像空间同时应用损失，MaskNet 能够生成既符合物理约束又多样化的交互布局，为后续的 FusionNet 提供重要的结构信息，从而实现高质量的手物交互图像合成。

1.3 FusionNet

FusionNet 是本文提出的两阶段方法中的第二个网络，其主要目的是通过 MaskNet 生成的布局进行掩膜操作然后修复

图像，进而合成详细的手物交互图像。该网络的关键在于巧妙地设计了一种条件化策略：在 DDPM 的反向扩散过程中，对已知区域使用原始图像信息进行采样，而对未知区域则使用模型生成的内容。为了进一步提高生成质量引入了一种重采样机制 [8]，通过在每个时间步多次重复前向和反向扩散过程，更好地调和生成内容与已知区域的一致性。

将真实图像表示为 x，未知像素表示为 $m \odot x$，已知像素表示为 $(1-m) \odot x$。x_{t-1}^{known} 从非掩膜已知图像区域采样，而 x_{t-1}^{unknown} 则直接从模型中采样。

$$x_{t-1}^{\text{known}} \sim N\left(\sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t) I\right) \tag{9}$$

$$x_{t-1}^{\text{unknown}} \sim N\left(\mu_\theta(x_t, t), \sum (x_t, t)_\theta\right) \tag{10}$$

$$x_{t-1} = m \odot x_{t-1}^{\text{known}} + (1-m) \odot x_{t-1}^{\text{unknown}} \tag{11}$$

FusionNet 的优势体现在可以适应任意形状的 Mask，大大提高了方法的适用性和灵活性；其次，生成的内容在语义上更加合理，细节更加丰富。它接收物体图像 I^{obj} 和 MaskNet 生成的布局参数 l 作为输入，输出完整的手 - 物体交互图像 I^{hoi}，如图 5 所示。

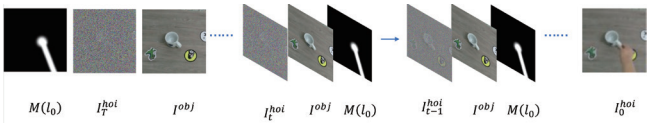

图 5 FusionNet 处理流程

FusionNet 的架构基于潜在扩散模型 [9]，在压缩的潜在空间中进行扩散过程，而不是在原始像素空间中操作。训练策略是从预训练的图像修复模型微调而来，学习在保持物体外观不变的同时生成符合布局的手部细节。

2 实验结果和分析

2.1 数据集

本文使用 HO3Pairs 数据集 [10]，该数据集包含 364 000 对物体 - 只有图像和手 - 物体交互图像。这些图像对是从 HOI4D[11] 数据集中提取的，旨在提供配对的手 - 物体交互图像和仅物体图像，用于模型训练。

2.2 实验条件

本次实验在 autoDL 平台上的实例进行，CPU 型号 Xeon(R) Platinum 8474C，GPU 型号 RTX 4090 24 GB，镜像采用了 PyTorch=1.8.0，去噪概率扩散预训练模型 [12]。

2.3 训练过程

MaskNet 的训练过程主要基于从 HOI4D 数据集构建的 HO3Pairs 数据集，包含 364 000 对物体 - 只有图像和手 - 物体交互（HOI）图像对。训练时，网络以物体 - 只有图像为条件输入，学习预测对应的 HOI 布局参数。整个训练过程从预训练的扩散模型初始化，通过不断优化网络结构和训练策略，最终使 MaskNet 能够从单个物体图像生成合理且多样的手 - 物体交互布局。

FusionNet 的训练过程利用了 HO3Pairs 数据集，该数据集包含配对的手 - 物交互图像和仅物体图像。训练过程中，FusionNet 被实现为一个以图像为条件的扩散模型，在潜在空间中操作。它从预训练的大规模图像修复模型开始，然后进行微调。

2.4 实验结果及分析

图 6 结果表明，本文提出的方法在手 - 物体交互图像合成任务中取得了显著的成果。

图 6 实验结果示意图

如表 1 所示，FID、LPIPS 和 UPR 是评估图像质量的不同指标：FID 比较生成图像与真实图像的整体分布，数值越小越好；LPIPS 衡量两张图像间的感知相似度，数值越小越好；而 UPR 评估单张图像的真实感。它们分别关注多样性、相似度和真实感，适用于不同的评估场景，如图像生成或处理任务的效果评估，数值越大越好。选择合适的指标取决于具体的应用需求和评估目标。

表 1 实验结果

Method	FID	LPIPS	UPR
LDM	64.7	0.510	44.8
Pix2Pix	49.1	0.361	64.2
MaskNet&FusionNet	**44.6**	**0.200**	**70.3**

本文所提出的方法在图像生成和处理任务中展现出显著优势。与现有的 LDM[13] 和 Pix2Pix[14] 方法相比，本文方法在所有评估指标上都取得了最佳表现。具体而言，这一方法在 FID 指标上达到了 44.6，相比 LDM（64.7）和 Pix2Pix（49.1）有明显改善，表明生成图像的整体质量和多样性更接近真实分布。LPIPS 评分为 0.200，大幅低于 LDM（0.510）和 Pix2Pix（0.361），证明本文方法能生成与目标更为相似的图像。在衡量单张图像真实感的 UPR 指标上，本文方法得分为 70.3，远超 LDM（44.8）和 Pix2Pix（64.2），凸显了生成图像的高度真实感和自然度。

这些结果强烈表明，MaskNet 和 FusionNet 的两阶段方法能够生成更加真实和物理上合理的手 - 物体交互图像。

3 总结

本文研究目的在于用单一 RGB 物体图像合成手 - 物体交互图像。该方法的核心是一个两阶段的生成过程，包括 MaskNet 和 FusionNet 两个网络。MaskNet 负责预测手 - 物体交互的空间布局，而 FusionNet 则基于这个布局生成详细的交互图像。实验结果表明，该方法在生成真实、多样化的 HOI 图像方面表现出色，不仅在接触召回率等定量指标上优于基线方法，而且在用户研究中获得了最高的偏好率。更重要的是，该方法展现出了良好的泛化能力，能够处理未见过的物体类别和复杂场景。

尽管本研究取得了显著成果，但仍有几个方向值得进一步探索。未来的研究可以考虑扩展到多视角 HOI 生成，探索生成时序一致的 HOI 图像序列，开发更精细的用户控制机制，直接集成 3D 重建到生成过程中，结合其他模态的信息如触觉或力反馈数据，探索在机器人学习、增强现实、人机交互设计等领域的实际应用，以及进一步优化方法效率以适应实时应用需求。这些拓展将有助于提升该方法的能力和应用范围，为创建更智能、更直观的交互系统铺平道路。总体而言，这项研究为理解物体可供性和人类 - 物体交互开辟了新的途径，为计算机视觉和人机交互领域提供了宝贵的工具和见解，有望在未来产生深远的影响。

参考文献：

[1] SHAN D D, GENG J Q, SHU M, et al. Understanding human hands in contact at internet scale[C/OL]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR).Piscataway:IEEE,2020[2024-05-19].https://ieeexplore.ieee.org/document/9157473.

[2] ZHANG H, BO Z H, YONG J H, et al. InteractionFusion: real-time reconstruction of hand poses and deformable objects in hand-object interactions[J]. ACM transactions on graphics (TOG), 2019, 38(4): 1-11.

[3] YU Y, SI X S, HU C H, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.

[4] HU H Z, WANG W L, ZHOU W G, et al. Hand-object interaction image generation[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. NewYork:CAI,2024: 23805-23817.

[5] YE Y F, LI X T, GUPTA A, et al. Affordance diffusion: Synthesizing hand-object interactions[DB/OL].(2023-05-20)[2024-04-13].https://doi.org/10.48550/arXiv.2303.12538.

[6] HE J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. NewYork: CAI, 2020: 6840-6851.

[7] HUANG H M, LIN L F, TONG R F, et al. UNet 3+: a full-scale connected UNET for medical image segmentation[C/OL]// ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2020[2024-04-17].https://ieeexplore.ieee.org/document/9053405.

[8] LUGMAYR A, DANELLJAN M, ROMERO A, et al. RePaint: inpainting using denoising diffusion probabilistic models[DB/OL].(2022-08-31)[2024-02-19].https://doi.org/10.48550/arXiv.2201.09865.

[9] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[DB/OL].(2022-04-23)[2024-04-11].https://doi.org/10.48550/arXiv.2112.10752.

[10] LIU Y Z, LIU Y, JIANG C, et al. HoI4D: a 4D egocentric dataset for category-level human-object interaction[C/OL]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE,2022 [2024-05-02]. https://ieeexplore.ieee.org/document/9878670.

[11] MENG C L, HE Y T, SONG Y, et al. SDEdit: guided image synthesis and editing with stochastic differential equations[DB/OL].(2022-01-05)[2024-05-09]. https://doi.org/10.48550/arXiv.2108.01073.

[12] HE J J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. NewYork: CAI, 2020: 6840-6851.

[13] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[DB/OL].(2022-04-13)[2024-05-09].https://doi.org/10.48550/arXiv.2112.10752.

[14] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C/OL]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017[2024-04-03].https://ieeexplore.ieee.org/document/8100115.

【作者简介】

刘畅（2000—），男，广东韶关人，硕士研究生，研究方向：深度学习、图像生成等。

姚剑敏（1978—），男，福建莆田人，博士，副研究员，研究方向：人工智能、图像处理、计算机视觉等。

（收稿日期：2024-10-15）