arXiv:2305.03051v1 [cs.CV] 4 May 2023

可控视觉-触觉合成

Ruihan Gao Wenzhen Yuan Jun-Yan Zhu
卡内基梅隆大学



图 1：我们的方法根据输入草图合成视觉与触觉输出，并在触觉设备（例如 TanvasTouch 屏幕 [15]）上呈现物体。我们的系统使用户能够看到视觉图案，同时感受指尖上的材料纹理。请访问我们的项目页面，观看数据采集和用户交互演示 https://visual-tactile-synthesis.github.io/。

摘要

深度生成模型在内容创作中有着广泛应用，例如平面设计、电子商务和虚拟试穿。然而，当前研究主要关注合成逼真的视觉输出，往往忽略触觉等其他感官模态，这限制了与用户的物理交互。在这项工作中，我们利用深度生成模型打造多感官体验：用户在触觉表面上滑动手指时，可以触摸并看到合成的物体。主要挑战在于视觉与触觉感知之间存在显著的尺度差异，以及缺乏从触觉感知数据到触觉渲染设备的明确映射。为弥合这一差距，我们使用 *GelSight* 传感器采集高分辨率触觉数据，并创建了一个新的视触觉服装数据集。随后，我们开发了一种条件生成模型，可根据单张草图合成视觉与触觉输出。我们从图像质量和触觉渲染准确度两方面评估了该方法。最后，我们提出了一套流程，可在基于静电吸附的触觉设备上渲染高质量的视觉与触觉输出，带来沉浸式体验，并支持具有挑战性的材料和可编辑的草图输入。

1. 引言

过去几年，深度生成模型和神经渲染技术推动的内容创作取得了显著进展 [31, 60]。 [46, 72]近期研究能够在多种用户控制方式下合成逼真的图像，例如用户草图 [29]，文本提示 [56], 和语义图 [52]。然而，大多数研究都专注于合成视觉输出，忽略了触觉等其他感官输出。

在现实生活中，人们通过视觉与触觉来探索物体。购买服装时，我们会观察它们，以感知其形状和外观，并触摸它们，以预想穿着时的感受。一次触摸就能揭示材料的粗糙度、硬度和局部几何形状。多模态感知输入使人们能够更全面地了解目标物体，从而提升在线购物和快速原型制作等用户体验。此外，这也为内容创作开辟了新的可能性，例如可触摸的虚拟现实和电影。

在这项工作中，我们旨在拓展内容创作的能力。我们提出一种新的问题设定——可控视觉-触觉合成，根据用户输入的草图或文本，合成高分辨率图像和触觉反馈输出。我们的目标是让人们在虚拟环境中探索物体时获得更沉浸式的体验。

视觉-触觉合成面临两大挑战。首先，由于感知尺度存在显著差异，现有生成模型难以联合建模视觉与触觉输出：视觉能让我们整体感知周围环境，而触觉只能提供局部细节的有限感知。其次，目前尚无数据驱动的端到端系统能够在触觉显示器上有效呈现采集到的触觉数据，因为现有触觉渲染系统严重依赖人工设计的触觉模式 [5, 3, 34, 62]。

为应对这些挑战，我们提出了一种基于表面纹理和表面形貌的触觉材料建模系统。首先，我们使用高分辨率触觉传感器 Gel-Sight [85, 76]采集目标物体的高分辨率表面几何形状，作为训练数据。为了生成能够根据用户输入渲染材料的视觉与触觉输出，我们提出了一种新的条件对抗学习方法，能够从不同尺度的多模态数据中学习。不同于以往的工作 [29, 77]，我们的模型从视觉图像提供的密集监督以及一组采样局部触觉图像块提供的稀疏监督中学习。在推理过程中，我们根据新的草图设计生成密集的视觉与触觉输出。随后，我们使用 TanvasTouch 触觉屏幕 [15]渲染模型的视觉与触觉输出。TanvasTouch 设备在常规视觉屏幕上显示视觉输出，并使用电粘附技术 [69] 根据摩擦图渲染不同纹理的力反馈。当人们在屏幕上滑动手指时，可以感受到不断变化的摩擦力分布所呈现的纹理 [7]。

我们收集了一个空间对齐的视觉-触觉数据集，名为 TouchClothing，其中包含20件服装，包括裤子和衬衫，材质和形状各异。我们通过自动化指标和用户研究，从图像质量和感知真实感两方面评估了模型。实验结果表明，如图1所示，我们的方法能够成功融合草图提供的整体结构与由布料材质决定的局部细粒度纹理。此外，我们展示了系统支持的草图编辑和基于文本的编辑应用，可生成供人们观看和触摸的新服装设计。我们的代码和数据可在网站上获取：https://visual-tactile-synthesis.github.io/。

2. 相关工作

视觉与触觉。研究表明，利用视觉和触觉输入进行多模态感知与学习，对多种计算机视觉和机器人学应用都很有效，例如估计材料属性 [87, 89, 86, 88], 物体抓取与操作 [39, 11, 10, 80, 73, 90, 43],物体识别 [41, 71], 未来帧预测 [82], 以及面向下游任务的表征学习 [32, 36, 82]。现有大多数研究都致力于改进识别与学习系统，而我们的目标是合成视觉-触觉输出，用于内容创作和虚拟现实应用。近期的一些研究尝试根据视觉输入预测触觉输出 [40, 9, 8, 12]。我们不打算从一种模态预测另一种模态，而是希望根据用户的草图和文本描述，同时合成两种模态的输出。

纹理的触觉渲染。触觉渲染是指生成模拟触摸感受的物理信号并将其传递给人，通常涉及用于建模的软件和用于渲染的物理硬件。尽管对此主题已有大量研究，但高分辨率材料纹理的渲染仍然是一项挑战 [6, 16]。一类研究 [59, 17, 18]使用动觉触觉设备来渲染单点时序信号。用户握住类似笔的触控笔并在平面上滑动时，会感受到振动力信号。渲染过程中缺乏空间分辨率，限制了触觉渲染的真实感。先前的研究还提出在基于静电吸附的设备上渲染纹理 [67, 83, 49, 4], , 但这类设备只能渲染均匀纹理或粗略物体形状。相比之下，我们提出使用 TanvasTouch 设备 [15] ，渲染服装物体的精细局部几何形状和材料纹理。该设备利用静电吸附产生可编程空间分布摩擦力，使用户能够通过手指在触摸屏上滑动来感受纹理。使用这种新设备能够增强用户对纹理和局部几何形状的真实感受。

深度生成模型。 以往的工作 [33, 22, 75, 19, 26,70, 60]实现了各种内容创作应用，例如文本到图像合成 [64, 56, 57, 84], 虚拟试穿 [24, 37, 1], 和风格迁移 [93, 63, 42]。大多数现有工作侧重于生成单模态视觉输出，如图像、视频 [25], 和三维数据 [50]。一些无条件生成对抗网络可合成两个领域的输出，例如图像和语义标签 [2, 92, 38, 74], 或 RGB-D数据 [78, 48]。虽然上述工作从潜向量中采样多模态输出，但它们无法进行控制。相比之下，我们的方法能够根据用户输入控制多模态合成。

图像到图像转换。 多种方法采用条件生成模型 [22]，将一个域的图像转换为另一域的图像 [29, 93, 28, 47, 63, 45, 14]。它们广泛应用于跨模态预测任务，例如草图到照片 [29, 65] 和标签到图像 [77, 52, 94]。相比之下，给定用户输入，我们的模型能够学习合成不同空间尺度下的两种模态输出。我们的方法也有别于以往工作：我们只需稀疏监督，就能学习合成密集触觉输出。

3. 数据采集与硬件

为开发我们的多模态合成方法，我们构建了一个新的空间对齐的视觉触觉数据集 TouchClothing，其中包含20件服装



图 2：**中的物体TouchClothing 数据集**。我们的数据集由 20 件具有不同形状（衬衫、夹克、短裤、裤子等）和各种面料（牛仔布、灯芯绒、亚麻布、摇粒绒等）的服装组成。请放大以查看更多细节。

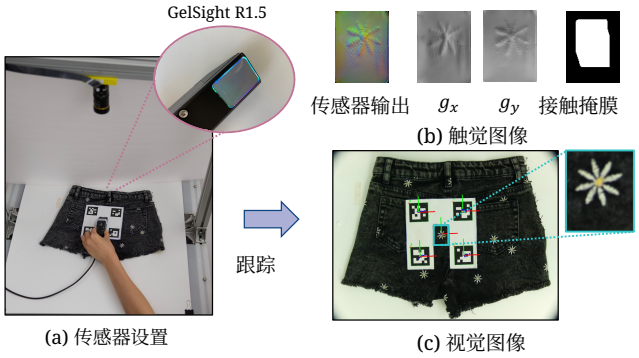


图 3：**视觉-触觉数据采集装置**。(a) 我们的装置包括一台 PiCamera RGB 相机、一个 GelSight R1.5 高分辨率触觉传感器，以及用于追踪传感器相对位姿的 Aruco 标记。(b) 我们展示采集到的触觉数据，包括原始传感器输出、推导出的 x 和 y 方向表面梯度 g_x 和 g_y，以及计算得到的接触掩膜。(c) 我们在视觉图像中定位与触觉数据相对应的边界框。

如图所示2。这些服装涵盖了市场上常见的各种面料，如牛仔布、灯芯绒、亚麻布、摇粒绒和羊毛。该数据集可用于在线购物和时装设计应用。对于每件服装，我们获取一张 1，280 × 960 张视觉图像，以捕捉物体的整体外观，以及 ∼200 个触觉图像块（32 × 32 像素），这些图块从物体表面稀疏采样得到。我们追踪传感器接触区域的三维坐标，并将其投影到二维视觉图像上，以实现空间对齐。最后，我们提取轮廓，作为每张视觉图像的输入草图。请访问网站获取我们的数据集。下面将详细介绍我们的采集流程。

视觉-触觉数据采集装置。 图 3 展示了我们用于采集对齐的视觉-触觉数据的装置，其中每个服装物体都用胶带固定在平面台上。我们使用安装在顶部铝杆上的 Pi RGB 相机拍摄俯视图，并通过手动按压 GelSight 传感器 [85, 76]在物体网格模式中的不同位置，记录数百个触觉图像块。我们的装置能够从每个物体采集多样的图块，包括具有均匀纹理的平坦缝纫图案、口袋边缘等局部几何变化，以及花形装饰等随机分布的特征。

GelSight 触觉传感器。GelSight 传感器 [85, 76] 是一种基于视觉的触觉传感器，采用光度立体视觉，以数十微米的高空间分辨率测量接触几何形状。本文使用的是由 Wang 等人 [76]改进的 GelSight R1.5。其感测区域为 32mm × 24mm（$H \times W$），像素分辨率为 320×240，相当于每像素 100 微米。传感器输出 RGB 图像，可将其转换为表面梯度，并用于重建三维高度图。

视觉-触觉对应关系。为了计算 GelSight 传感器相对于相机的相对位置，我们在 GelSight 上安装四个 Aruco 标记，并运行随机抽样一致性算法 [20]来跟踪其三维姿态。这样，我们便能将接触区域的三维坐标投影到二维视觉图像上，并确定每个触觉块的边界框坐标。示例数据见图4。

触觉数据预处理与接触掩膜。每个触觉输出代表 GelSight 传感器在服装上的一次触摸，此时只有感测区域的一小部分与服装接触。我们发现，使用原始数据训练模型时会出现明显的伪影。因此，我们屏蔽非接触区域，仅使用接触区域来改进模型。具体而言，我们将触觉输出从 320×240 下采样至 104×78 （每像素约 300 微米），使其与图像分辨率相匹配，然后通过对高度图进行阈值处理，为每个触觉图块创建接触掩膜。我们通过启发式方法将阈值设为高度图数值的第 75 百分位数，并进行膨胀处理，以避免漏检。我们依据接触掩膜采样 32×32个图块，作为最终的触觉数据。每种服装采集约 200 个图块，覆盖图像区域的 1/6 。

草图图像。我们遵循 pix2pix 中描述的流程 [29] ，从视觉图像中获取草图。我们首先使用 DexiNed 网络提取粗略轮廓 [54] ，然后手动移除较小的边缘，以获得细轮廓。

用于触觉渲染的 TanvasTouch。TanvasTouch [15] 是一种触觉屏幕，可为手指接触渲染分布式摩擦图。它将屏幕表面与手指之间的空气间隙建模为一个电容器。当手指在屏幕上滑动时，屏幕下方变化的电压会在手指中感应出微小电流，从而使人感受到变化的

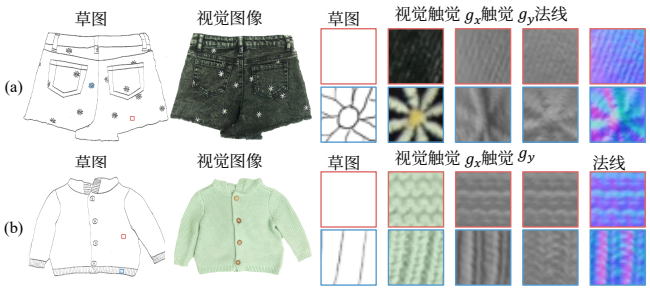


图4：**我们TouchClothing 数据集中的示例数据**。对于每个物体，我们展示输入草图、视觉图像和两个触觉图像块。对于每个触觉图像块，我们展示相应的草图裁剪区域、视觉裁剪区域，以及采集到的触觉数据，包括 x 和 y 方向上的表面梯度（g_x, g_y）和表面法线图。草图中的颜色编码边界框标记了各触觉图像块的位置，并体现了视觉数据与触觉数据之间显著的尺度差异，这使得我们的条件合成任务颇具挑战性。

摩擦力。该设备以灰度摩擦图作为输入，以调节屏幕上的电压分布。屏幕会同时显示视觉图像并渲染触觉信号，从而产生视觉-触觉耦合输出。

4. 方法

由于视觉与触觉的感受野差异很大，视觉-触觉合成颇具挑战。相机能够捕捉物体的颜色和形状等全局特征，而触觉传感器只能捕捉小块区域内的局部信息，例如边缘和材料纹理。现有的条件生成模型并不适用于此任务，因为它们假设所有输入的尺度都相对相同。

为应对这一挑战，我们提出了一种新的多模态条件生成对抗网络，该网络通过全局视觉监督和稀疏局部触觉监督进行学习。如图 5所示，我们的模型以单张草图为输入，合成空间对齐的视觉-触觉输出。我们在第 4.1节对该任务进行定义，并在第 4.2节介绍学习目标。我们在第 4.3 节描述网络设计，并在第 4.4节讨论如何在 TanvasTouch 触觉设备上渲染视觉与触觉输出。

4.1. 视觉-触觉合成

<escape>我们针对每个物体分别训练一个模型，并将视觉-触觉合成任务表述为一种条件式单图像生成建模任务 [51, 68, 66],，这种方法即使在单张图像上训练模型，也已展现出灵活的编辑能力。具体而言，给定一张尺寸为x $H \times W$ 的草图，其中 H 和 W 分别表示图像的高度和宽度，我们旨在学习一个函数，将输入草图x 映射为两个空间对齐的输出：一张 RGB 视觉图像 y_I和一个触觉输出 y_T。</escape>

该草图 x 是一种轮廓图，用于勾勒物体的外形，并捕捉其粗尺度的边缘和图案。例如，在图 4(a) 中，一条短裤的草图展示了短裤的整体形状、凹槽和腰带的位置，以及局部装饰图案。在实际操作中，我们遵循 Isolaetal. [29] 的方法，使用 DexiNed [54]和边缘细化来提取草图。图 4展示了一条短裤和一件毛衣的草图、视觉图像和触觉图像示例。

视觉图像 y_I 是由相机捕获的 RGB图像。触觉输出 $y_T = (g_x, g_y)$ 是表示表面在 x 和 y方向上梯度的双通道图像。它们可以使用公式 **n** 将其转换为法线，见公式 1 ，然后通过泊松积分将其转换为高度图 [85]。由于触觉输出是通过校准网络将 GelSight 原始输出（RGBXY）映射到表面梯度 (g_x, g_y) [76], 获得的，因此其对传感器坐标中的局部噪声和位置偏移更加稳健。它也不易受泊松积分后高度图可能产生的积分误差影响。因此，我们的条件生成对抗网络使用 (g_x, g_y) 作为触觉输出格式。

$$\mathbf{n} = \frac{(g_x, g_y, -1)}{\sqrt{g_x^2 + g_y^2 + 1}}, \quad g_x = \frac{n_x}{n_z}, \quad g_y = \frac{n_y}{n_z}. \tag{1}$$

生成的视觉与触觉输出可用于时装设计和触觉渲染等应用。在这项工作中，我们在 TanvasTouch屏幕上渲染服装，让人们可以同时看到并触摸它。

4.2. 学习目标

在这项学习任务中，我们面临两个主要挑战。首先，我们必须从密集视觉图像和稀疏触觉监督中学习，同时考虑尺度差异。其次，由于我们需要在单个高分辨率示例上学习一个合成网络，因此训练数据有限。为应对这些挑战，我们提出以下学习目标。

视觉合成损失。为了合成一张以用户草图为条件的真实图像 y_I x，我们优化视觉生成器 G_I 和视觉判别器 D_I，使其匹配真实草图-图像对的条件分布。我们优化以下极小极大目标 [29, 47]:

$$\begin{aligned} V(G_I, D_I, x, y_I) = \mathbb{E}_{x,y_I}[\log D_I(x, y_I)] \\ + \mathbb{E}_x[\log(1 - D_I(x, G_I(x)))]. \end{aligned} \tag{2}$$

遗憾的是，由于采用单图像训练设置，上述对抗损失会导致训练不稳定。为适应有限的数据集规模，我们使用了视觉辅助判别器 Dclip [35]，该判别器由冻结的 CLIP 特征提取器 [55] 和一个小型可训练多层感知机头组成。视觉辅助损失可以减少小规模数据集中的过拟合问题，并合成与

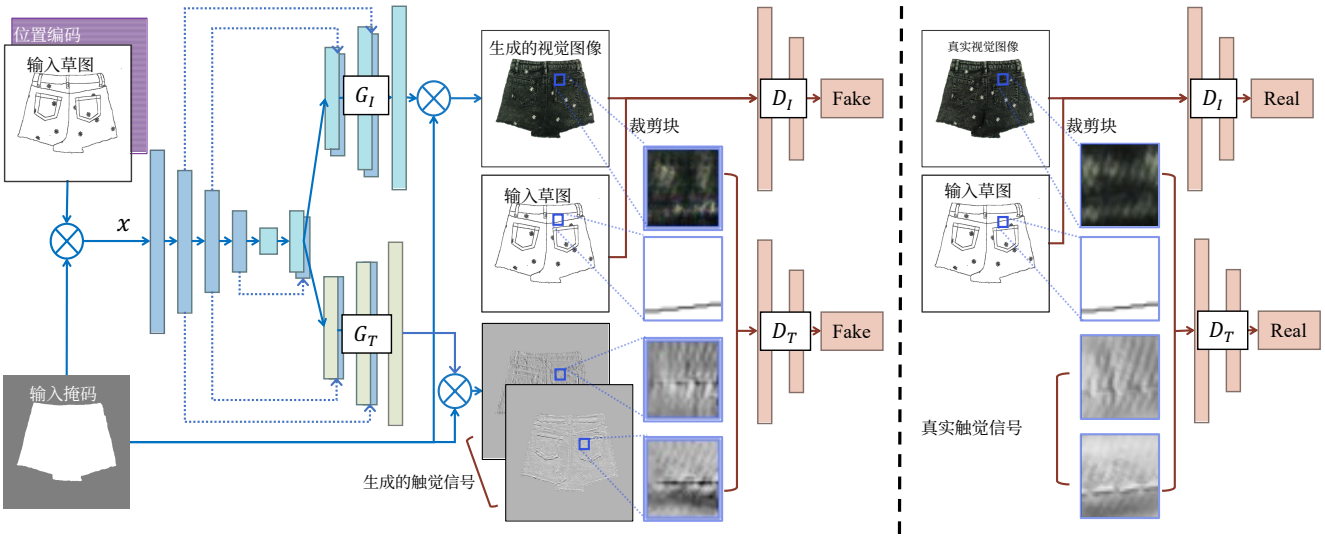


图5：**概述**。生成器：给定用户草图、其前景掩码以及像素坐标的位置编码，我们将它们输入双分支生成器。两个分支共享编码器和解码器的前四层，随后分开，分别合成视觉结果和触觉结果。判别器：我们将完整的视觉图像输入视觉判别器 D_I，并将图块输入触觉判别器 D_T。D_I 以草图为条件，而 D_T 则以草图裁剪块和视觉裁剪块为条件。

人类感知更加匹配的视觉图像。我们的对抗损失包括：

$$\mathcal{L}_{\text{cGAN}} = V(G_I, D_I, x, y_I) + V(G_I, D_{\text{clip}}, x, y_I). \quad (3)$$

<escape>为进一步稳定生成对抗网络的训练，我们引入了一种基于重建的损失。在这里，我们结合使用逐像素 L1 距离和基于 CNN 特征的感知损失（LPIPS）[91], ，因为它们能够生成更清晰的图像 [29]，并提高与真实图像的感知相似度。</escape>

$$\begin{aligned}\mathcal{L}_{\text{rec}}(G_I, x, y_I) = \mathbb{E}_{x,y_I}[\mathcal{L}_{\text{LPIPS}}(y_I, G_I(x))] \\ + \lambda_1 \mathbb{E}_{x,y_I}[\|y_I - G_I(x)\|_1],\end{aligned} \quad (4)$$

其中，λ_1 用于平衡感知损失和 L1 损失。视觉输出的最终目标函数可写为：

$$\mathcal{L}_I = \mathcal{L}_{\text{cGAN}} + \mathcal{L}_{\text{rec}}. \quad (5)$$

触觉合成损失。遗憾的是，我们无法直接使用上述损失函数来合成触觉输出，因为我们不再能够访问完整尺寸的触觉真实数据。此外，视觉辅助损失不适用于触觉数据和小图块，因为视觉辅助判别器D_{clip}是在大规模自然图像集合上预训练的。

相反，我们通过数百个触觉图像块的监督，学习一个全尺寸触觉生成器 G_T。在这里，我们将采样位置 p 处对应的（草图、图像、触觉）图像块记为（x^p, y_I^p, y_T^p）。生成器G_T 一次性合成全尺寸触觉输出，而我们的图块级判别器 D_T学习判断每个图块对是真实还是伪造，目标函数如下：

$$\begin{aligned}V(G_T, D_T, x, y_I, y_T) = \mathbb{E}_{x,y_I,y_T,p}[\log D_T(x^p, y_I^p, y_T^p)] \\ + \mathbb{E}_{x,p}[\log(1 - D_T(x^p, G_I^p(x), G_T^p(x)))],\end{aligned} \quad (6)$$

其中，$G_I^p(x)$ 和 $G_T^p(x)$ 分别表示合成视觉输出和触觉输出的裁剪块。为降低训练内存占用和复杂度，我们不将梯度反向传播至 G_I。

除了标准的非饱和 GAN 目标外，我们还使用基于判别器特征的特征匹配目标 [77]，因为与预训练 CLIP 模型相比，判别器能更好地适应触觉领域。此外，我们还加入了图块级重建损失。触觉合成分支的最终损失可写为：

$$\mathcal{L}_T = \lambda_{\text{GAN}} V(G_T, D_T, x, y_I, y_T) + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}(G_T, x^p, y_T^p). \quad (7)$$

图块采样。我们采样两种类型的图块。对于具有配对的触觉真实数据的图块，我们可以同时使用重建损失和对抗损失。不过，我们只有 200 个训练图块。为了进一步增加训练图块的数量，我们还会随机采样没有配对真实值的图块。我们只尝试最小化第二项log($1 - D_T(x^p,\ G_I^p(x)\ G_T^p(x))$)，即触觉对抗损失（公式 6）中的第二项，因为它只取决于合成图块。

完整目标函数。我们的最终目标函数为

$$G_I^*, G_T^* = \arg \min_{G_I, G_T} \max_{D_I, D_T, D_{\text{clip}}} (\mathcal{L}_I + \mathcal{L}_T). \quad (8)$$

通过网格搜索选择权重，使各项损失的尺度相近，最终取值为 $\lambda_1 = 100$,$\lambda_{\text{GAN}} = 5$, $\lambda_{\text{rec}} = 10$。网格搜索仅针对随机选取的一个物体进行一次，所得参数将用于数据集中的所有物体。

在第 5 节中，我们仔细评估了对抗损失和图像重建损失对我们最终模型性能的影响。

4.3. 训练细节

下面我们将介绍生成器和判别器的网络架构以及其他训练细节。

网络架构。 我们使用 U-Net [61] 作为生成器的骨干网络，并从解码器的中间层分出两个分支， G_I和 G_T。这样，视觉与触觉输出共享用于全局结构的相同编码，同时在每个像素位置保留模态特定的细节。对于判别器，我们对视觉判别器 [29, 77]和触觉判别器 D_I 均采用多尺度 PatchGAN D_T，因为研究表明，多尺度 PatchGAN 能够改善结果的精细细节。

位置编码和物体掩码。由于草图通常包含大面积的均匀纹理区域，我们使用正弦位置编码（SPE） [81]对像素坐标进行编码，并在网络输入处将位置编码与草图拼接起来。我们还提取物体掩码，并用它从输入和输出中去除背景。因此，网络的最终输入是拼接后的草图和位置编码特征的掩码版本。更多训练细节请参阅附录A。

4.4. 触觉渲染

合成视觉与触觉输出后，我们使用以下渲染流程在 TanvasTouch 触觉屏幕上呈现它们，使用户能够看到并感受到物体。具体而言，我们将视觉图像直接显示在屏幕上，并将双通道触觉输出$(g_x,\ g_y)$ 转换为 TanvasTouch 所需的灰度摩擦图。正如 Manuel 等人 [44] 和 Fiesen 等人 [21], 所指出的，人类在表面触觉解释中对轮廓和高频强度变化十分敏感。受此启发，我们首先计算梯度$z = g_x^2 + g_y^2$的 $z \in [0, 1]$平方幅值，然后应用非线性映射函数 $z' = \log_{10}(9 \times z + 1),\ \ z' \in [0, 1]$ 以增强对比度，最后将其调整为 TanvasTouch 屏幕的尺寸，作为最终的摩擦图。我们通过经验发现，这有助于利用电黏附力增强纹理的触感。

5. 实验

下面介绍我们的主要结果。有关数据采集和用户交互视频，请访问我们的网站。

评估指标。 我们根据合成输出与 TouchClothing 数据集真实数据之间的相似性来评估我们的方法。对于视觉与触觉输出，我们报告 LPIPS 指标 [91]，用于衡量感知真实感，因为先前的工作 [91, 30] 已表明，与 PSNR 和 SSIM [79] 相比，LPIPS 指标更符合人类感知。我们还使用单图像 Fr´echet Inception 距离（SIFID） [66] 来衡量纹理相似性，这在先前的工作 [66, 53] 中得到了广泛应用。由于该数据集每个物体仅包含一张视觉图像，因此我们在已见过的草图上评估视觉重建的 LPIPS，并在未见过的草图上评估纹理一致性的 SIFID，以考察泛化能力。除自动化指标外，我们还开展了人类偏好研究。

基线方法。 据我们所知，本文首次研究以草图输入为条件的视觉-触觉合成。因此，我们将图像到图像转换视为一项相似任务，并将我们的方法与多种条件生成对抗网络进行比较，包括 pix2pix [29],pix2pixHD [77] 和 GauGAN [52]。pix2pix 是最常用的图像转换网络之一；pix2pixHD 使用由粗到细生成器和多尺度判别器来处理高分辨率图像合成；GauGAN 则采用空间自适应去归一化层。pix2pixHD 和 GauGAN 均使用感知损失、条件生成对抗网络损失以及基于生成对抗网络的特征匹配损失进行训练。

对于基线方法，我们为触觉输出 g_x和 g_y增加两个通道，使输出通道数量从 3 增至 5。视觉与触觉输出分别输入两个判别器，且两个判别器都以草图输入为条件。由于触觉真值仅提供图块数据，我们将草图和视觉图像中对应的区域裁剪成图块，并使用草图-视觉-触觉图块对训练网络。我们采用与我们的方法相同的数据增强量，并遵循原始工作中的默认参数。推理时，我们输入完整的草图图像，以获得全尺寸的视觉与触觉输出，因为全卷积网络能够泛化到不同尺寸的输入。

定量比较。如表1所示，我们的方法在所有指标上均大幅优于所有基线方法。我们的方法使视觉 LPIPS 降低了 50% 以上，使触觉 LPIPS 降低了约 30%。正如 $5\times$ 和 $2\times$ 所示，我们的结果生成了更加逼真且忠实的纹理，分别体现在视觉与触觉输出的 SIFID 值更低。这表明我们的方法在视觉与触觉合成方面均具有优势。我们注意到，在大多数指标上，pix2pix 的表现优于 pix2pixHD 和 GauGAN。这可能是因为所有基线方法都需要配对数据集，而在我们的实验中，配对数据的分辨率较低（32×32），不适用于 pix2pixHD 和 GauGAN。

定性结果。图6 展示了与基线方法进行定性比较的示例。对于每种方法，第一行展示完整尺度的视觉输出；第二行

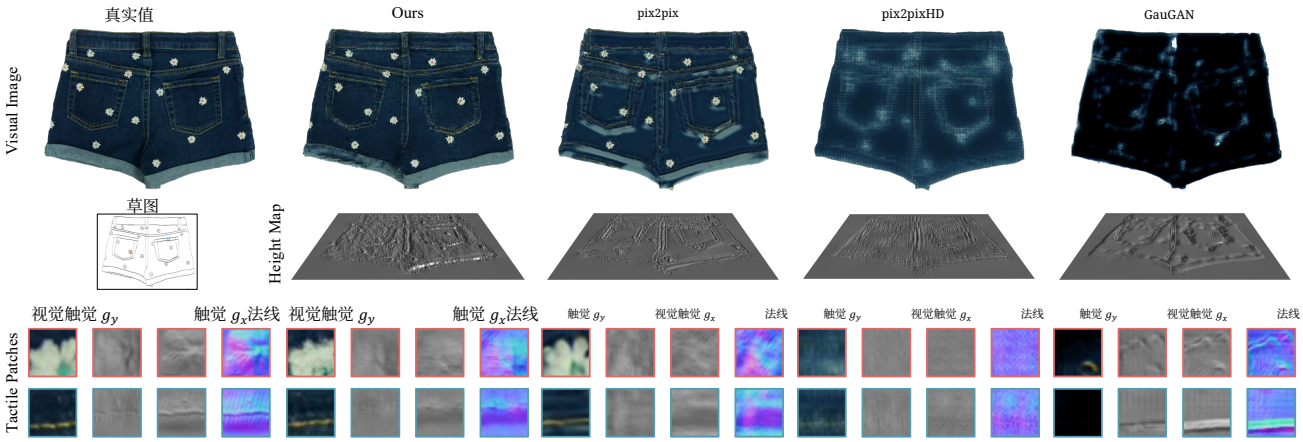


图6：**与基线方法的定性比较**。我们将我们的方法与 pix2pix [29], pix2pixHD [77], 和 GauGAN [52]进行比较。对于每种方法，我们展示视觉输出（上方）以及高度图的渲染结果（中间）。我们还展示了两个以颜色编码边界框标示的放大图块（下方），并配有对应的视觉裁剪区域、触觉表面梯度和法线图。

方法	视觉		触觉	
	LPIPS↓	SIFID↓	LPIPS↓	SIFID↓
Ours	**0.070**	**0.029**	**0.676**	**0.104**
Pix2pix [29]	0.173	0.115	1.028	0.247
Pix2pixHD [77]	0.161	0.289	0.753	0.458
GauGAN [52]	0.189	0.252	1.034	0.286

表1：**基线比较**。我们的方法在感知真实感（由 LPIPS [91] 衡量）和纹理一致性（由 SIFID [66]衡量）两方面均优于所有基线。

方法	视觉		触觉	
	LPIPS↓	SIFID↓	LPIPS↓	SIFID↓
Ours	**0.070**	**0.029**	**0.676**	**0.104**
我们的方法（不含 $\mathcal{L}_{\text{cGAN}}$）	0.113	0.115	0.687	0.107
我们的方法（不含 $\mathcal{L}_{\text{rec}}$）	0.084	0.079	1.035	0.260

表2：**损失分量消融研究**。我们将完整方法与两个变体进行比较：我们的方法（不含 $\mathcal{L}_{\text{cGAN}}$）（不含条件生成对抗网络损失）和 $\mathcal{L}_{\text{rec}}$（不含图像重建损失）。在多个指标方面，我们的完整方法优于这些变体。

展示重建的 3D 高度图；第三行以视觉、灰度 g_x、 g_y 以及导出的表面法线形式展示了一些采样图块。我们的方法能够成功捕捉显著的几何特征，例如凹槽和花纹装饰，以及材料纹理的局部几何细节。相比之下，基线方法只能捕捉部分显著的几何特征，却会遗漏局部纹理细节，并产生颜色伪影。

使用未见过的草图图像进行生成。我们的视觉-触觉合成模型仅使用单张草图图像进行训练，也能泛化到新的草图输入，使用户能够编辑和定制草图，从而快速进行设计和原型制作。

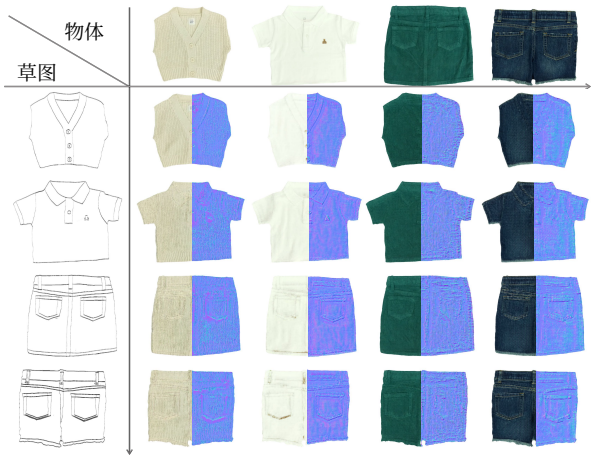


图7：**草图与材质替换**。我们的模型能够为已知和未见过的草图合成视觉与触觉图像。对于每个输出，左半部分显示视觉输出，右半部分显示触觉输出的法线图。

由于我们针对每个物体分别训练一个模型，因此在图 7 中展示了使用未见过物体的草图得到的测试结果。每一行对应一张测试草图，每一列代表一个基于单个物体训练的模型。我们通过在左侧展示视觉图像、在右侧展示法线图来呈现结果。对于每个模型，视觉与触觉输出都能很好地对齐，并保留精细尺度的材质纹理细节。我们的方法能够适应新草图输入的全局几何信息，包括边缘和凹槽。

基于文本的视觉-触觉合成。我们还将方法扩展到同时给定草图和文本提示时合成视觉-触觉输出。我们使用 DALL·E2 [56] 来生成原始草图的变体，然后将编辑后的草图输入我们的条件生成模型。图 8展示了使用文本提示进行基于文本的合成示例。

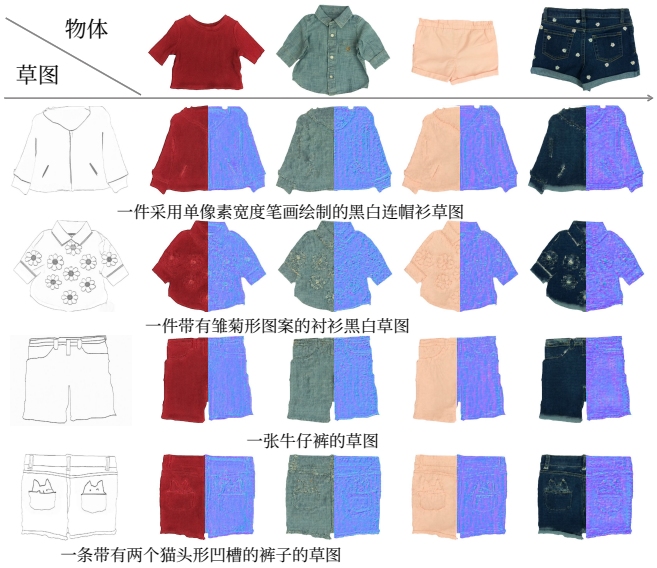


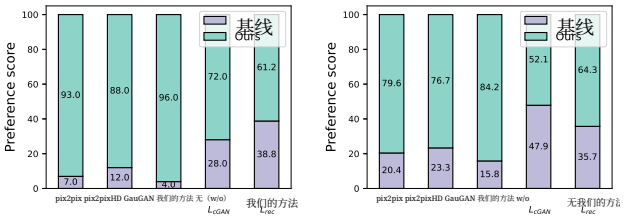
图8： **基于文本的视觉-触觉合成**。 我们使用 DALL·E2 [56]，这一文本生成图像模型，通过基于文本的图像修补来修改原始草图设计。随后，我们使用编辑后的草图合成视觉输出（左）和触觉输出（右）。每张图像下方均列出了文本提示。

即使仅使用单张草图进行训练，我们的模型也能合理地泛化到笔画和形状各异的未见过的草图，同时捕捉原始材料的视觉和触觉特征。

消融研究。 我们针对每个损失分量进行了消融研究，以考察它们对训练目标的影响。表2显示，在视觉和触觉合成中同时移除对抗损失或重建损失，都会导致LPIPS误差和 SIFID指标上升。从定性结果来看，移除对抗损失后图像会过于平滑，而移除重建损失后则会出现棋盘格伪影。更多视觉结果请见附录 B。

视觉图像的人类感知研究。 我们使用亚马逊土耳其机器人（AMTurk）开展人类感知研究。我们进行配对测试，问题为：“你认为哪张图像更真实？”每位用户先进行五轮练习，随后进行 30 轮测试，以评估我们的方法与 pix2pix、pix2pixHD、GauGAN、我们的方法（去除$\mathcal{L}_{\text{cGAN}}$）以及我们的方法（去除 $\mathcal{L}_{\text{rec}}$）的表现。所有样本均经过随机抽取和排列，我们共收集了 1,500 份回答。如图9a所示，相较于所有基线方法，参与者更倾向于选择我们的方法，甚至胜过我们的方法（去除 $\mathcal{L}_{\text{cGAN}}$ ）和我们的方法（去除 $\mathcal{L}_{\text{rec}}$），这表明各项损失项都十分重要。

触觉渲染的人类感知研究。我们还按照以往研究的惯例开展了人类感知研究，以评估生成的触觉输出的感知保真度 [23, 13]。我们并排呈现两种不同的



(a) 视觉用户研究的柱状图。 (b) 触觉用户研究的柱状图。

图9：**人类感知研究**。在每组配对比较中，无论是视觉还是触觉输出，相较于基线方法，参与者更偏好我们的方法（≥ 50%）。

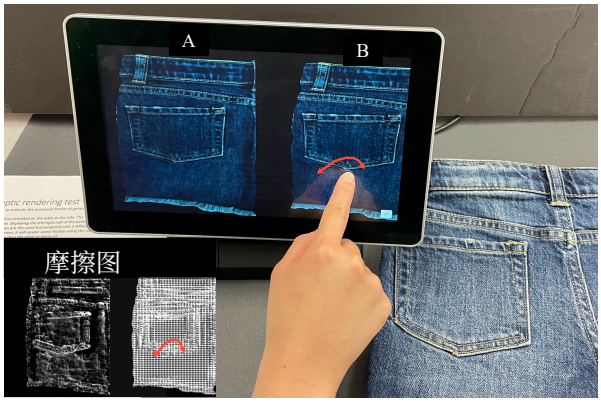


图10：**用户研究的实验设置**。我们进行了一项 A/B测试，比较我们的方法与其中一种基线方法的触觉输出。左下角显示渲染后的触觉信号（摩擦图）。真实服装放在一侧供参考。

触觉输出，并在 TanvasTouch 屏幕上将它们与相同的真实视觉图像并排展示，然后询问参与者：“您觉得哪一侧更符合真实物体的材质？”共有 20 人参与实验，其中男性 13 人、女性 7 人，平均年龄为 24.1 岁（标准差：2.1 ）。图10 展示了实验设置示例，更多详情见附录A。如图9b所示，参与者明显更青睐我们的方法，而非所有其他基线方法（随机猜测的概率为50%）。76.7% 的参与者更喜欢我们的方法，而非 pix2pixHD；与 pix2pix 和 GauGAN 相比，我们的方法优势更大，分别赢得了79.6% 和 84.2% 参与者的青睐。用户较难区分消融模型，但我们的方法仍优于 Ours w/o$\mathcal{L}_{\text{cGAN}}$ 和 Ours w/o $\mathcal{L}_{\text{rec}}$，分别领先52.1% 和64.3% 。用户研究的结果与表1中使用各种指标进行的定量评估结果一致。

6. 讨论与局限性

在这项工作中，我们提出了一种新方法，可根据草图和文本等用户输入，自动合成视觉与触觉图像。我们使用高分辨率触觉传感器 GelSight 捕捉物体的高保真局部几何形状。随后，我们提出了一种新的条件生成对抗网络模型，仅给定一个即可生成视觉与触觉输出。



图11：**失败案例**。当训练物体与测试草图呈现的图案差异很大时，我们的方法表现不佳。例如，给定一件带有大型字母图案的连帽衫，我们的模型能够重建训练物体（第1列），但无法为花朵图案填充颜色（第2列），也无法在领口（第3列）或腰带（第4列）周围生成均匀的纹理。

草图图像。最后，我们介绍了一种在 TanvasTouch 触摸屏上呈现视觉与触觉输出的流程。

我们的视觉-触觉合成方法可用于不同的材料和物体，让用户在探索虚拟物体时获得更身临其境的体验。

局限性。首先，如图11所示，独特的图案（如封闭字母）仍然难以处理。我们的模型无法泛化到其他用户草图。其次，由于触觉属于主动感知，渲染效果受特定硬件限制。在本研究中，表面触觉设备擅长渲染服装这类主要为平面且带有细致纹理的对象。然而，在同一设备上渲染表面法线变化显著的三维物体（如苹果）仍然具有挑战性。

最后，由于触觉数据是在静态触摸过程中采集的，且渲染设备主要针对摩擦力，因此我们能够很好地渲染粗糙度，但渲染柔软度的能力有限。

社会影响。 用于触觉渲染的可控视觉-触觉合成是一项尚未得到充分探索的新研究问题。我们迈出了应对建模挑战的第一步，并将模型部署到最新硬件上。最终，我们希望这项工作能推动生成模型在在线购物、虚拟现实、远程临场和远程操作等应用中实现多模态合成。

致谢。 感谢 Sheng-Yu Wang、Kangle Deng、Muyang Li、Aniruddha Mahapatra 和 Daohan Lu 帮助校对稿件。我们也感谢 Sheng-Yu Wang、Nupur Kumari、GauravParmar、George Cazenavette 和 Arpit Agrawal 提供宝贵的意见并参与讨论。此外，感谢 Yichen Li 、 Xiaofeng Guo 和 Fujun Ruan 协助搭建硬件。Ruihan Gao 获得 A*STAR 国家科学奖学金（博士）。

参考

[1] Badour Albahar、Jingwan Lu、Jimei Yang、Zhixin Shu、Eli Shechtman 和 Jia-Bin Huang。Pose with style：使用条件式 StyleGAN 进行保留细节的姿态引导图像合成。《ACM图形学汇刊》（TOG），40 11，2021 6 1 2 ( ): – 。[2] Samaneh Azadi、Michael Tschannen、Eric Tzeng、Sylvain Gelly、Trevor Darrell 和 Mario Lucic。语义瓶颈场景生成。arXiv 预印本 arXiv:1911.11357，2019。2[3] Elham Beheshti、Katya Borgos-Rodriguez 和 Anne Marie Piper。通过触觉反馈显示器支持亲子协作学习。载于 ACM 国际儿童交互设计会议（IDC），2019。2[4] Amit Bhardwaj、Hojun Cha 和 Seungmoon Choi。使用随机森林对黏弹性可变形物体的法向交互进行数据驱动的触觉建模。《IEEE 机器人与自动化快报》（RA-L），4(2):1379–1386，2019。2[5] Jeremy Birnholtz、Darren Gergle、Noah Liebman 和 Sarah Sinclair。感知到的意识：探究移动式可变摩擦触觉显示器在感知信息方面的应用。载于移动设备与服务人机交互国际会议（MobileHCI），2015。2[6] S´er´ena Bochereau、Stephen Sinclair 和 Vincent Hayward。使用表面显示器再现虚拟触觉纹理时的知觉恒常性。《ACM 应用感知汇刊》（TAP），15(2)，2018。2[7] David Arthur Burns、Roberta L Klatzky、Michael A Peshkin 和 J Edward Colgate。纹理的空间感知取决于长度尺度。载于 IEEE 世界触觉大会（WHC），2021。2[8] Shaoyu Cai、Lu Zhao、Yuki Ban、Takuji Narumi、Yue Liu 和 Kening Zhu。基于 GAN 的图像到摩擦生成，用于织物材料的触觉模拟。《计算机与图形学》，102:460–473，2022。2[9] Shaoyu Cai、Kening Zhu、Yuki Ban 和 Takuji Narumi。使用结合特征匹配与感知损失的残差融合 GAN 进行视觉触觉跨模态数据生成。《IEEE 机器人与自动化快报》（RA-L），6(4):7525–7532，2021。2[10] Roberto Calandra、Andrew Owens、Dinesh Jayaraman、Justin Lin、Wenzhen Yuan、Jitendra Malik、Edward H Adelson 和 Sergey Levine。不止于感觉：学习利用视觉与触觉进行抓取和重新抓取。《IEEE 机器人与自动化快报》（RA-L），3(4):3300–3307，2018。2[11] Roberto Calandra、Andrew Owens、Manu Upadhyaya、Wen- zhen Yuan、Justin Lin、Edward H Adelson 和 Sergey Levine。成功的感觉：触觉感知是否有助于预测抓取结果？机器人学习大会（CoRL），2017。2[12] Guanqun Cao、Jiaqi Jiang、Ningtao Mao、Danushka Bolle- gala、Min Li 和 Shan Luo。Vis2hap：通过跨模态生成实现基于视觉的触觉渲染。IEEE 国际机器人与自动化会议（ICRA），2023。2[13] Afonso Castic.o 和 Paulo Cardoso。基于电吸附的触觉设备中用于纹理比较的可用性测试。《多模态技术与交互》，6(12):108，2022。8[14] Jooyoung Choi、Sungwon Kim、Yonghyun Jeong、Youngjune Gwon 和 Sungroh Yoon。ILVR：用于 de-

去噪扩散概率模型。发表于 IEEE 国际计算机视觉大会（ICCV），2021。2[15] J. Edward Colgate 和 Michael Peshkin。Tanvas——表面触觉技术与产品。https://tanvas.co/，2022。1、2、3[16] Heather Culbertson 和 Katherine J Kuchenbecker。在创建逼真的触觉虚拟表面时，匹配物理摩擦、硬度和纹理的重要性。IEEE触觉学汇刊（ToH），10(1)，2016。2[17] Heather Culbertson、Joseph M Romano、Pablo Castillo、Max Mintz 和 Katherine J Kuchenbecker。利用工具介导接触加速度数据创建逼真触觉虚拟纹理的改进方法。发表于 IEEE触觉研讨会（HAPTICS），2012。2[18] Heather Culbertson、Juliette Unwin、Benjamin E Goodman 和 Katherine J Kuchenbecker。基于不受约束的工具—表面交互生成触觉纹理模型。发表于 IEEE 世界触觉大会（WHC），2013。2[19] Laurent Dinh、Jascha Sohl-Dickstein 和 Samy Bengio。使用 real nvp 进行密度估计。arXiv 预印本 arXiv:1605.08803，2016。2[20] M. Fischler 和 R. Bolles。随机抽样一致性算法：一种用于模型拟合的范式及其在图像分析和自动制图中的应用。《ACM 通讯》，24(6):381–395，1981。3[21] Rebecca Fenton Friesen、Roberta L Klatzky、Michael A Peshkin 和 J Edward Colgate。构建可导航的精细纹理设计空间。IEEE触觉学汇刊（ToH），14(4)，2021。6[22] Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。神经信息处理系统会议（NeurIPS），2014。2[23] Roman V Grigorii、Roberta L Klatzky 和 J Edward Colgate。通过宽带摩擦调制，以数据驱动方式回放自然触觉纹理。IEEE触觉学汇刊（ToH），15(2):429–440，2021。8[24] Xintong Han、Zuxuan Wu、Zhe Wu、Ruichi Yu 和 Larry S Davis。Viton：一种基于图像的虚拟试穿网络。发表于 IEEE 计算机视觉与模式识别大会（CVPR），2018。2[25] Jonathan Ho、William Chan、Chitwan Saharia、Jay Whang、Ruiqi Gao、Alexey Gritsenko、Diederik P Kingma、Ben Poole、Mohammad Norouzi、David J Fleet 等。Imagen video：使用扩散模型生成高清视频。arXiv 预印本 arXiv:2210.02303，2022。2[26] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。发表于神经信息处理系统会议（NeurIPS），2020。2[27] Xun Huang 和 Serge Belongie。使用自适应实例归一化实现实时任意风格迁移。发表于 IEEE 国际计算机视觉大会（ICCV），2017。14[28] Xun Huang、Arun Mallya、Ting-Chun Wang 和 Ming-Yu Liu。利用专家乘积生成对抗网络进行多模态条件图像合成。欧洲计算机视觉会议（ECCV），2021。2[29] Phillip Isola、Jun-Yan Zhu、Tinghui Zhou 和 Alexei A Efros。《使用条件对抗网络进行图像到图像转换》。载于 IEEE计算机视觉与模式识别会议（CVPR），2017 年。1、2、3、4、5、6、7、13[30] Justin Johnson、Alexandre Alahi 和 Li Fei-Fei。《用于实时风格迁移和超分辨率的感知损失》。载于 欧洲计算机视觉会议（ECCV），2016 年。6[31] Tero Karras、Samuli Laine 和 Timo Aila。《用于生成对抗网络的基于风格的生成器架构》。载于 IEEE计算机视觉与模式识别会议（CVPR），2019 年。1[32] Justin Kerr、Huang Huang、Albert Wilcox、Ryan Hoque、Jef-frey Ichnowski、Roberto Calandra 和 Ken Goldberg。《通过视觉与触觉的自监督表征学习，实现对可变形表面的主动滑动感知》。arXiv 预印本 arXiv:2209.13042，2022 年。2[33] Diederik P Kingma 和 Jimmy Ba。《Adam：一种随机优化方法》。载于 国际学习表征会议（ICLR），2015 年。2、13[34] Roberta L Klatzky、Amukta Nayak、Isobel Stephen、Dean Dijour 和 Hong Z Tan。《静电摩擦显示器上模式信息的检测与识别》。IEEE触觉学汇刊（ToH），12(4):665–670，2019 年。2[35] Nupur Kumari、Richard Zhang、Eli Shechtman 和 Jun-Yan Zhu。《集成现成模型以进行 GAN训练》。载于 IEEE计算机视觉与模式识别会议（CVPR），2022 年。4[36] Michelle A Lee、Yuke Zhu、Krishnan Srinivasan、Parth Shah、Silvio Savarese、Li Fei-Fei、Animesh Garg 和 Jeannette Bohg。《理解视觉与触觉：面向高接触任务的多模态表征自监督学习》。载于 IEEE 国际机器人与自动化会议（ICRA），2019 年。2[37] Kathleen M Lewis、Srivatsan Varadharajan 和 Ira Kemelmacher-Shlizerman。《TryOnGAN：通过分层插值实现身体感知试穿》。ACM图形学汇刊（TOG），40(4):1–10，2021 年。2[38] Daiqing Li、Junlin Yang、Karsten Kreis、Antonio Torralba 和 Sanja Fidler。《生成模型下的语义分割：半监督学习与强大的域外泛化》。载于 IEEE计算机视觉与模式识别会议（CVPR），2021 年。2[39] Rui Li、Robert Platt、Wenzhen Yuan、Andreas ten Pas、Nathan Roscup、Mandayam A Srinivasan 和 Edward Adelson。《使用 GelSight触觉感知定位和操作小型部件》。载于 IEEE/RSJ智能机器人与系统国际会议（IROS），2014 年。2[40] Yunzhu Li、Jun-Yan Zhu、Russ Tedrake 和 Antonio Torralba。《通过跨模态预测连接触觉与视觉》。载于 IEEE计算机视觉与模式识别会议（CVPR），2019 年。2、15[41] Justin Lin、Roberto Calandra 和 Sergey Levine。《学习通过触觉识别物体实例：基于多模态匹配的触觉识别》。载于 IEEE 国际机器人与自动化会议（ICRA），2019 年。2[42] Ming-Yu Liu、Thomas Breuel 和 Jan Kautz。《无监督图像到图像转换网络》。神经信息处理系统会议（NeurIPS），2017 年。2

[43] Yiyue Luo、Yunzhu Li、Pratyusha Sharma、Wan Shou、Kui Wu、Michael Foshey、Beichen Li、Tom´as Palacios、Antonio Tor-ralba 和 Wojciech Matusik。《使用共形触觉织物学习人类—环境交互》。Nature Electronics，4(3):193–201，2021 年。2[44] Steven G Manuel、Roberta L Klatzky、Michael A Peshkin 和 James Edward Colgate。《表面触觉解释中的巧合规避原理》。美国国家科学院院刊，112(8)，2015 年。6[45] Chenlin Meng、Yutong He、Yang Song、Jiaming Song、Jiajun Wu、Jun-Yan Zhu 和 Stefano Ermon。《SDEdit：使用随机微分方程进行引导式图像合成与编辑》。载于 国际学习表征会议（ICLR），2022 年。2[46] Ben Mildenhall、Pratul P Srinivasan、Matthew Tancik、Jonathan T Barron、Ravi Ramamoorthi 和 Ren Ng。《NeRF：将场景表示为神经辐射场以合成视图》。载于 欧洲计算机视觉会议（ECCV），2020 年。1[47] Mehdi Mirza 和 Simon Osindero。《条件生成对抗网络》。arXiv 预印本 arXiv:1411.1784，2014 年。2、4[48] Atsuhiro Noguchi 和 Tatsuya Harada。《RGBD-GAN：通过 RGBD图像合成，从自然图像数据集中进行无监督三维表征学习》。国际学习表征会议（ICLR），2020 年。2[49] Reza Haghighi Osgouei、Sunghwan Shin、Jin Ryong Kim 和 Seungmoon Choi。《用于电振动显示器数据驱动纹理渲染的逆向神经网络模型》。载于 IEEE 触觉研讨会（ HAPTICS ），第 270–277 页。IEEE，2018 年。2[50] Jeong Joon Park、Peter Florence、Julian Straub、Richard New-combe 和 Steven Lovegrove。《DeepSDF：学习用于形状表征的连续有符号距离函数》。载于 IEEE计算机视觉与模式识别会议（CVPR），2019 年。2[51] Taesung Park、Alexei A Efros、Richard Zhang 和 Jun-Yan Zhu。《用于非配对图像到图像转换的对比学习》。载于 欧洲计算机视觉会议（ECCV），2020 年。4[52] Taesung Park、Ming-Yu Liu、Ting-Chun Wang 和 Jun-Yan Zhu。《通过空间自适应归一化实现语义图像合成》。载于 IEEE计算机视觉与模式识别会议（CVPR），2019 年。1、2、6、7[53] Taesung Park、Jun-Yan Zhu、Oliver Wang、Jingwan Lu、Eli Shechtman、Alexei Efros 和 Richard Zhang。《用于深度图像操控的交换式自编码器》。神经信息处理系统会议（NeurIPS），33:7198–7211，2020 年。6[54] Xavier Soria Poma、Edgar Riba 和 Angel Sappa。《密集极值Inception网络：迈向稳健的卷积神经网络边缘检测模型》。载于 IEEE/CVF冬季计算机视觉应用会议（WACV），2020 年。3、4[55] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。《通过自然语言监督学习可迁移视觉模型》。载于 国际机器学习会议（ICML），2021 年。4、14

[56] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。基于 CLIP 潜变量的层次化文本条件图像生成。arXiv 预印本 arXiv:2204.06125，2022。1、2、7、8[57] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。零样本文本到图像生成。发表于国际机器学习会议（ICML），2021。2[58] Benjamin A Richardson、Yasemin Vardar、Christian Wallraven 和 Katherine J Kuchenbecker。学习感知纹理：根据不受约束的手指与表面交互预测感知相似性。IEEE触觉学汇刊（ToH），2022。13[59] Joseph M Romano 和 Katherine J Kuchenbecker。利用接触加速度数据创建逼真的虚拟纹理。IEEE触觉学汇刊（ToH），5(2)：109–119，2011。2[60] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。利用潜在扩散模型进行高分辨率图像合成。发表于IEEE计算机视觉与模式识别会议（CVPR），2022。1、2[61] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。U-Net：用于生物医学图像分割的卷积网络。发表于国际医学图像计算与计算机辅助干预会议（MICCCAI），2015。6、13[62] Bushra Sadia、Ayberk Sadic、Mehmet Ayyildiz 和 Cagatay Basdogan。利用触摸屏静电振动显示触觉图形的探索策略。《国际人机交互研究》，160：102760，2022。2[63] Chitwan Saharia、William Chan、Huiwen Chang、Chris Lee、Jonathan Ho、Tim Salimans、David Fleet 和 Mohammad Norouzi。Palette：图像到图像扩散模型。发表于ACM SIGGRAPH，2022。2[64] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S Sara Mahdavi、Rapha Gontijo Lopes 等。深度语言理解驱动的逼真文本到图像扩散模型。神经信息处理系统会议（NeurIPS），2022。2[65] Patsorn Sangkloy、Jingwan Lu、Chen Fang、Fisher Yu 和 James Hays。Scribbler：利用草图和颜色控制深度图像合成。发表于IEEE计算机视觉与模式识别会议（CVPR），2017。2[66] Tamar Rott Shaham、Tali Dekel 和 Tomer Michaeli。SinGAN：从单幅自然图像学习生成模型。发表于IEEE国际计算机视觉会议（ICCV），2019。4、6、7[67] Sunghwan Shin、Reza Haghighi Osgouei、Ki-Duk Kim 和 Seungmoon Choi。利用频率分解神经网络对各向同性触觉纹理进行数据驱动建模。发表于 IEEE 世界触觉大会（WHC），2015。2[68] Assaf Shocher、Shai Bagon、Phillip Isola 和 Michal Irani。InGAN：捕捉并重新定位自然图像的“DNA”。发表于IEEE国际计算机视觉会议（ICCV），2019。4[69] Craig Shultz、Michael Peshkin 和 J Edward Colgate。将触觉、听觉和超声波力应用于人类

指尖：利用宽带静电吸附。IEEE触觉学汇刊（ToH），11(2)：279–290，2018。2[70] Yang Song、Jascha Sohl-Dickstein、Diederik P Kingma、Abhishek Kumar、Stefano Ermon 和 Ben Poole。通过随机微分方程进行基于分数的生成建模。国际学习表征会议（ICLR），2021。2[71] Subramanian Sundaram、Petr Kellnhofer、Yunzhu Li、Jun-Yan Zhu、Antonio Torralba 和 Wojciech Matusik。利用可扩展的触觉手套学习人类抓握的特征。Nature，569(7758)：698–702，2019。2[72] A. Tewari、O. Fried、J. Thies、V. Sitzmann、S. Lombardi、K. Sunkavalli、R. Martin-Brualla、T. Simon、J. Saragih、M. Nießner、R. Pandey、S. Fanello、G. Wetzstein、J.-Y. Zhu、C. Theobalt、M. Agrawala、E. Shechtman、D. B Goldman 和 M. Zollh¨ofer。神经渲染的最新进展。CoRR，2020。1[73] Stephen Tian、Frederik Ebert、Dinesh Jayaraman、Mayur Mudigonda、Chelsea Finn、Roberto Calandra 和 Sergey Levine。凭触觉操作：基于触觉的深度预测模型控制。发表于IEEE 国际机器人与自动化会议（ICRA），2019。2[74] Nontawat Tritrong、Pitchaporn Rewatbowornwong 和 Supasorn Suwajanakorn。重新利用生成对抗网络进行单样本语义部件分割。发表于IEEE计算机视觉与模式识别会议（CVPR），2021。2[75] Aaron Van den Oord、Nal Kalchbrenner、Lasse Espeholt、Oriol Vinyals、Alex Graves 等。利用 PixelCNN 解码器进行条件图像生成。神经信息处理系统会议（NeurIPS），29，2016。2[76] Shaoxiong Wang、Yu She、Branden Romero 和 Edward Adelson。GelSight 楔形传感器：利用紧凑型机器人手指测量高分辨率三维接触几何形状。发表于IEEE 国际机器人与自动化会议（ICRA），2021。2、3、4[77] Ting-Chun Wang、Ming-Yu Liu、Jun-Yan Zhu、Andrew Tao、Jan Kautz 和 Bryan Catanzaro。利用条件生成对抗网络进行高分辨率图像合成与语义操作。发表于IEEE计算机视觉与模式识别会议（CVPR），2018。2、5、6、7[78] Xiaolong Wang 和 Abhinav Gupta。利用风格与结构对抗网络进行生成式图像建模。发表于欧洲计算机视觉会议（ECCV），2016。2[79] Zhou Wang、Alan C Bovik、Hamid R Sheikh 和 Eero P Simoncelli。图像质量评估：从误差可见性到结构相似性。IEEE图像处理汇刊（TIP），13(4)：600–612，2004。6[80] Youngsun Wi、Pete Florence、Andy Zeng 和 Nima Fazeli。Virdo：可变形物体的视觉-触觉隐式表征。IEEE 国际机器人与自动化会议（ICRA），2022。2[81] Rui Xu、Xintao Wang、Kai Chen、Bolei Zhou 和 Chen Change Loy。位置编码作为生成对抗网络中的空间归纳偏置。发表于IEEE计算机视觉与模式识别会议（CVPR），2021。6[82] Fengyu Yang、Chenyang Ma、Jiacheng Zhang、Jing Zhu、Wenzhen Yuan 和 Andrew Owens。触摸即走：学习来自人类采集的视觉与触觉。神经信息处理系统会议（NeurIPS），2022。2[83] Sunghoon Yim、Seokhee Jeon 和 Seungmoon Choi。对可变形物体的黏弹性和摩擦响应进行数据驱动的触觉建模与渲染。IEEE触觉学汇刊（ToH），9(4):548–559，2016。2[84] Jiahui Yu、Yuanzhong Xu、Jing Yu Koh、Thang Luong、Gunjan Baid、Zirui Wang、Vijay Vasudevan、Alexander Ku、Yinfei Yang、Burcu Karagol Ayan 等。扩展自回归模型，实现内容丰富的文本到图像生成。《机器学习研究汇刊》（TMLR），2022。2[85] Wenzhen Yuan、Siyuan Dong 和 Edward H Adelson。GelSight：用于估计几何形状和力的高分辨率机器人触觉传感器。《传感器》，2017。2，3，4[86] Wenzhen Yuan、Rui Li、Mandayam A Srinivasan 和 Edward H Adelson。使用 GelSight 触觉传感器测量剪切与滑移。载于 IEEE 国际机器人与自动化会议（ICRA），2015。2[87] Wenzhen Yuan、Mandayam A Srinivasan 和 Edward H Adelson。使用 GelSight 触觉传感器估计物体硬度。载于 IEEE/RSJ智能机器人与系统国际会议（IROS），2016。2[88] Wenzhen Yuan、Shaoxiong Wang、Siyuan Dong 和 Edward Adelson。连接视觉与触感：关联实体材料的视觉和触觉属性。载于 IEEE计算机视觉与模式识别会议（CVPR），2017。2[89]Wenzhen Yuan、Chenzhuo Zhu、Andrew Owens、Mandayam A Srinivasan 和 Edward H Adelson。使用深度学习和 GelSight 触觉传感器进行与形状无关的硬度估计。载于 IEEE 国际机器人与自动化会议（ICRA），2017。2[90] Qiang Zhang、Yunzhu Li、Yiyue Luo、Wan Shou、Michael Foshey、Junchi Yan、Joshua B Tenenbaum、Wojciech Matusik 和 Antonio Torralba。通过触觉感知对手物交互进行动态建模。载于 IEEE/RSJ智能机器人与系统国际会议（IROS），2021。2[91] Richard Zhang、Phillip Isola、Alexei A Efros、Eli Shechtman 和 Oliver Wang。深度特征作为感知度量的惊人有效性。载于 IEEE计算机视觉与模式识别会议（CVPR），2018。5，6，7[92] Yuxuan Zhang、Huan Ling、Jun Gao、Kangxue Yin、Jean-Francois Lafleche、Adela Barriuso、Antonio Torralba 和 Sanja Fidler。DatasetGAN：以最少人工投入高效构建标注数据工厂。载于 IEEE计算机视觉与模式识别会议（CVPR），2021。2[93] Jun-Yan Zhu、Taesung Park、Phillip Isola 和 Alexei A Efros。使用循环一致对抗网络进行非配对图像到图像转换。载于 IEEE计算机视觉与模式识别会议（CVPR），2017。2[94] Peihao Zhu、Rameen Abdal、Yipeng Qin 和 Peter Wonka。SEAN：使用语义区域自适应归一化进行图像合成。载于 IEEE计算机视觉与模式识别会议（CVPR），2020。2

附录

在本附录中，我们提供了更多实验结果和详细信息。

在第 A 节中，我们介绍了有关触觉感知研究方案、网络架构和训练流程的更多实验细节。在第 B 节中，我们展示了主要方法的更多结果和分析、针对各个损失分量的更多消融研究，以及与视觉-触觉跨模态预测基线的比较。请查看我们的网站视频 https://visual-tactile-synthesis.github.io/，了解数据采集和用户交互演示。

A. 更多实验细节

触觉用户研究方案。数据通过心理物理实验以线下方式收集，请查看我们的网站视频，了解用户交互和用户研究设置的演示。共有 20 人参与实验，其中男性 13 人、女性 7 人，平均年龄为 24.1 岁（标准差：2.1）。实验流程已获所在机构的机构审查委员会（IRB）批准。所有参与者均已提供知情同意，并获得每小时 15 美元的报酬。

具体而言，每轮实验中，参与者会看到桌边的一件真实服装，以及 TanvasTouch屏幕上的两种渲染结果。这两种渲染结果的视觉外观相同（真实视觉图像），但触觉输出不同。其中一种由我们的方法生成，另一种则由一种基线方法生成（pix2pix、pix2pixHD 或 GauGAN）。

参与者需要用惯用手的食指在 TanvasTouch屏幕和真实衣物上滑动，力度和速度不限，并可自由来回切换。在触摸屏上，两种渲染结果并排显示，呈现衣物相同的一半（左侧或右侧）；参与者可以自由探索两侧，并在一分钟内选出更逼真的一种。

每次实验开始前，参与者需要完成一场培训，其中包括对 TanvasTouch 设备的简要介绍，以及快速演示如何渲染 Tanvas 公司设计的官方演示应用 TanvasIntro 应用所提供的均匀纹理®。这些步骤有助于参与者熟悉设备的工作方式，以及他们可以期待的渲染反馈类型。

我们从`TouchClothing` 数据集中抽取一个服装物体进行预热，并将其余未见过的物体留作测试，遵循 Richard- son 等人的方法。 [58]预热和测试均遵循上述相同的实验流程，但预热环节不设时间限制，以确保参与者有足够的时间探索设备。为防止用户疲劳，我们随机选择 5 每次测试会话均针对 19 个未见过的物体进行实验，并报告平均结果。每次实验持续约 60 分钟。

网络架构。 我们采用 pix2pix [29]中的符号来描述网络架构。令 Ck 表示一个包含k 个滤波器的卷积-BatchNorm-LeakyReLU 层，其中 LeakyReLU 的斜率为 0.2。CTk 表示一个 dropout 率为 50% 的转置卷积-BatchNorm-ReLU 层。所有卷积均为 4×4 空间滤波器，步幅为 2。编码器中的卷积将特征图下采样 2 倍，而解码器中的卷积则将特征图上采样 2 倍。

编码器-解码器架构由以下部分组成：**编码器：**

C10-C20-C40-C80-C80-C80-C80-C80
解码器：
CT80-CT160-CT160-CT160-CT160-CT80-CT40-CT20

G_I 和 G_T 共用编码器以及解码器的前四层。解码器的最后一层之后，会应用一个卷积，将特征图映射为最终输出（视觉合成为 3 个通道，触觉合成为 2 个通道），随后接入 Tanh函数。作为上述符号表示的例外，编码器中的第一个 C10 层不应用 BatchNorm。我们采用 U-Net架构 [61]，在编码器的每一层 i 与解码器的第 $n-i$ 层之间设置跳跃连接，其中 n 为总层数。

判别器架构遵循 pix2pix [29]，由 C64-C128-C256-C512 组成。最后一层之后，应用一次卷积将特征图映射为一维输出，随后经过 Sigmoid函数。与上述记法不同的是，编码器中的第一个 C64 层不应用 BatchNorm。

训练详情。 我们为数据集中的每件服装分别训练一个模型。为了扩充数据集，我们对视觉图像和草图图像（960×1280）进行填充至 1800×1800，然后随机裁剪为 1536×1536，同时确保草图保持闭合轮廓。我们为每件服装手动收集 300 个触觉图像块，总共收集 60,000 个图块。我们按 6:2:2 的比例将它们划分为训练集、验证集和测试集。我们使用 Adam求解器 [33]，学习率为 $lr = 0.001$、$\beta_1 = 0$ 和 $\beta_2 = 0.99$ 进行优化，并采用“编码维度=4”作为主文第 4.3 节所述空间位置编码的参数。所有实验均在单块 24GB NVIDIA RTX A5000 GPU 上进行。训练耗时 16 小时，推理平均耗时 0.19 秒，可同时生成尺寸为 1536×1536 的视觉与触觉输出。

B. 补充实验结果

主方法的更多定性结果。 图 15至 19 展示了我们的方法与其他方法对比的定性结果

方法	视觉		触觉	
	LPIPS↓	SIFID↓	LPIPS↓	SIFID↓
Ours	0.070	0.029	**0.676**	0.104
我们的方法（不含 $\mathcal{L}$条件生成对抗网络，视觉）	0.116	0.336	0.686	0.107
我们的方法（不含 $\mathcal{L}$条件生成对抗网络，触觉）	0.070	0.040	0.677	0.104
我们的方法（不含 $\mathcal{L}$重建损失，视觉）	0.078	0.073	**0.676**	**0.103**
我们的方法（不含 $\mathcal{L}$重建损失，触觉）	**0.064**	**0.017**	1.021	0.255

表3：**损失分量消融研究**。我们分别针对视觉合成和触觉合成，单独考察对抗损失与重建损失，并研究它们对输出性能的各自影响。我们观察到视觉与触觉重建之间存在权衡，也就是说，去除触觉重建损失会使视觉合成性能略有提升；然而，我们的完整模型在视觉-触觉合成方面取得了最佳性能。

以及数据集中全部20个物体的基线方法。对于每个物体，最左侧一列展示真实图像和输入草图，随后依次展示我们的方法、pix2pix、pix2pixHD 和 GauGAN 生成的视觉与触觉输出（以三维高度图的形式呈现）。

额外的消融研究。 这里我们在图中展示损失消融研究的定性结果20。如正文第5节“消融研究”和表2所述，我们同时对视觉与触觉合成中的 $\mathcal{L}_{\text{cGAN}}$ 和 $\mathcal{L}_{\text{rec}}$ 进行消融。图 20 显示，从定性结果来看，移除对抗损失会使视觉图像过于平滑，触觉输出的连贯性也会降低；而移除重建损失则会导致触觉输出趋于平凡，并在视觉图像中产生棋盘格伪影。我们展示了两个示例，但在数据集中的所有物体上都观察到了类似效果。

此外，我们还分别研究了对抗损失和重建损失对视觉合成和触觉合成的影响。

遵循正文第4.2节中的符号表示，我们将视觉和触觉合成的对抗损失与重建损失分别简写如下，并通过消融各个组成部分来研究它们对最终输出的影响。

$$\begin{aligned}\mathcal{L}_{\text{cGAN (visual)}} &= V(G_I, D_I, x, y_I) + V(G_I, D_{\text{clip}}, x, y_I)\\ \mathcal{L}_{\text{cGAN (tactile)}} &= V(G_T, D_T, x, y_I, y_T)\\ \mathcal{L}_{\text{rec (visual)}} &= \mathcal{L}_{\text{rec}}(G_I, x, y_I)\\ \mathcal{L}_{\text{rec (tactile)}} &= \mathcal{L}_{\text{rec}}(G_T, x^p, y_T^p).\end{aligned} \tag{9}$$

我们在表3 中展示了定量结果，并在图20中展示了定性结果。

<escape>**针对未见过的草图进行测试的更多定性结果**。与正文中的图7类似，图21 展示了使用我们训练好的模型对未见过的草图进行测试的其他结果。我们的模型能够为已见过和未见过的草图合成视觉与触觉输出。图</escape>

21a 展示视觉图像，图21b 展示以法线图形式呈现的触觉输出。每一列对应一种物体材质，

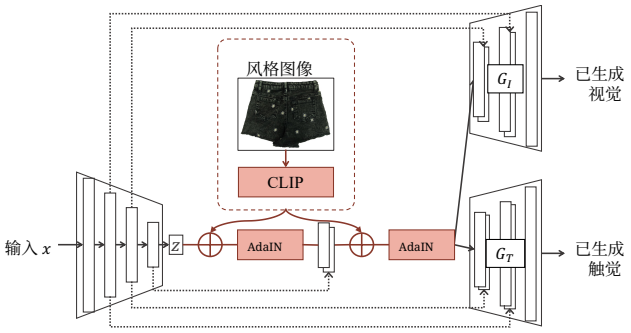


图12：**跨物体模型**。我们从RGB图像中提取CLIP图像编码，并通过AdaIN层将其与草图输入的特征图融合。

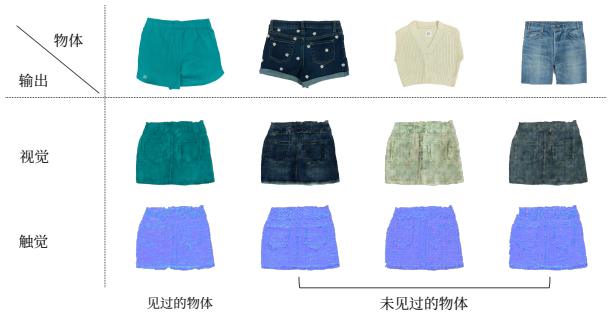


图 13：**跨物体模型**。我们使用 15 个物体训练模型，并使用具有不同风格/材质图像的未见过的草图进行测试。我们展示了一个已见过的图像（第 1 列）、来自我们数据集的两个未见过的图像（第 2 列和第 3 列），以及一个在线图像（第 4 列）的结果。

每一行对应一个草图输入。对角线上的输出表示在训练中相同物体上的测试结果，而非对角线上的项目表明我们的模型能够泛化到未见过的草图。

跨物体模型。 在我们的主论文中，我们需要为每个物体单独训练一个模型。相比之下，我们现在在多种物体和材质上训练一个跨物体模型。为了编码视觉外观和触觉纹理，我们从真实值视觉图像中提取 CLIP 图像嵌入 [55]，并使用自适应实例归一化（AdaIN）对来自草图输入的特征图进行调制 [27]。我们在潜在瓶颈处以及解码器分支之前的每个公共层中将它们融合。

图 12展示了网络架构，图 13展示了一些定性结果。我们的跨物体模型能够为已见过的材料（第 1 列）合成合理的视觉与触觉输出。对于未见过的材料，它能够为某些物体合成协调一致的输出，例如牛仔裤（第 2 列），但无法为其他物体保留颜色。

其他基线方法。 由于目前关于条件视觉-触觉合成的研究较少，我们还将我们的方法与

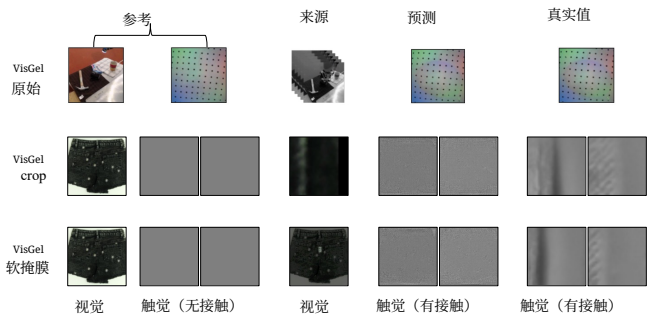


图14：**VisGel 实现方式的不同变体**。 我们实现了两种变体：*VisGel*（裁剪）和*VisGel*（注意力）。四列分别显示参考帧（输入）、源图像（输入）、预测（输出）和真实值。这两种变体都无法像我们的方法一样提供可比较的输出。

VisGel [40], 这一近期提出的视觉-触觉跨模态预测方法进行比较。VisGel旨在根据视觉输入合成合理的触觉信号，反之亦然，即使用条件对抗网络根据触觉信号推断视觉图像。它针对每个方向（即视觉到触觉和触觉到视觉）分别训练一个网络。它们的问题设定与我们不同，因为我们的目标是根据用户草图合成视觉与触觉输出。VisGel还以视频作为输入，而我们使用单张草图。

尽管如此，我们采用 vision2touch 网络，并使用配对视觉-触觉图像块（32×32）对其进行训练。在 VisGel 的原始设置中，作者拍摄机器人在桌面上与物体交互的视频。该方法使用视频的第一帧和第一条触觉读数（无交互、无触觉接触）作为参考帧。为适应设置上的差异，我们尝试了两种变体。两者均使用完整的视觉图像作为视觉参考，并使用无接触时的触觉图像作为触觉参考。

对于第一种变体 *VisGel*（裁剪），我们从完整图像中裁剪出视觉图像块，并将该图块作为源输入。对于第二种变体 *VisGel*（软掩膜），我们将完整的视觉图像作为源输入，将其与软注意力图相乘；该注意力图在图块位置的值最高，为 1，再用高斯模糊对其进行膨胀，并将其他位置的最小值设为 0.1。图 14 展示了两种变体及原始 VisGel 设置的参考帧和源输入帧示例。

遗憾的是，对于这两种变体，网络训练都无法收敛到有意义的结果。可能有几个原因。首先，原始工作可以通过跟踪机器人的运动来学习注意力图，从视觉图像中找到接触点。他们的输入视频序列提供了额外的环境信息，用于建立视觉图像与触觉图像块之间的相关性。而我们的数据在视觉帧中只有一个物体，因此网络难以理解对应关系。其次，原始工作中的网络是在大规模数据集（165 个物体和 250 万帧）上训练的，因此不适用于我们的单图像数据集（1 个物体和 $\sim$250 个数据点）。更重要的是，它无法根据用户草图同时合成视觉与触觉输出。



图15：**我们的方法与基线方法在数据集中全部20个物体上的定性结果**。对于每个物体，我们在最左侧一列展示真实图像和输入草图，随后依次展示我们的方法、pix2pix、pix2pixHD 和 GauGAN 生成的视觉与触觉输出（以三维高度图的形式呈现）。

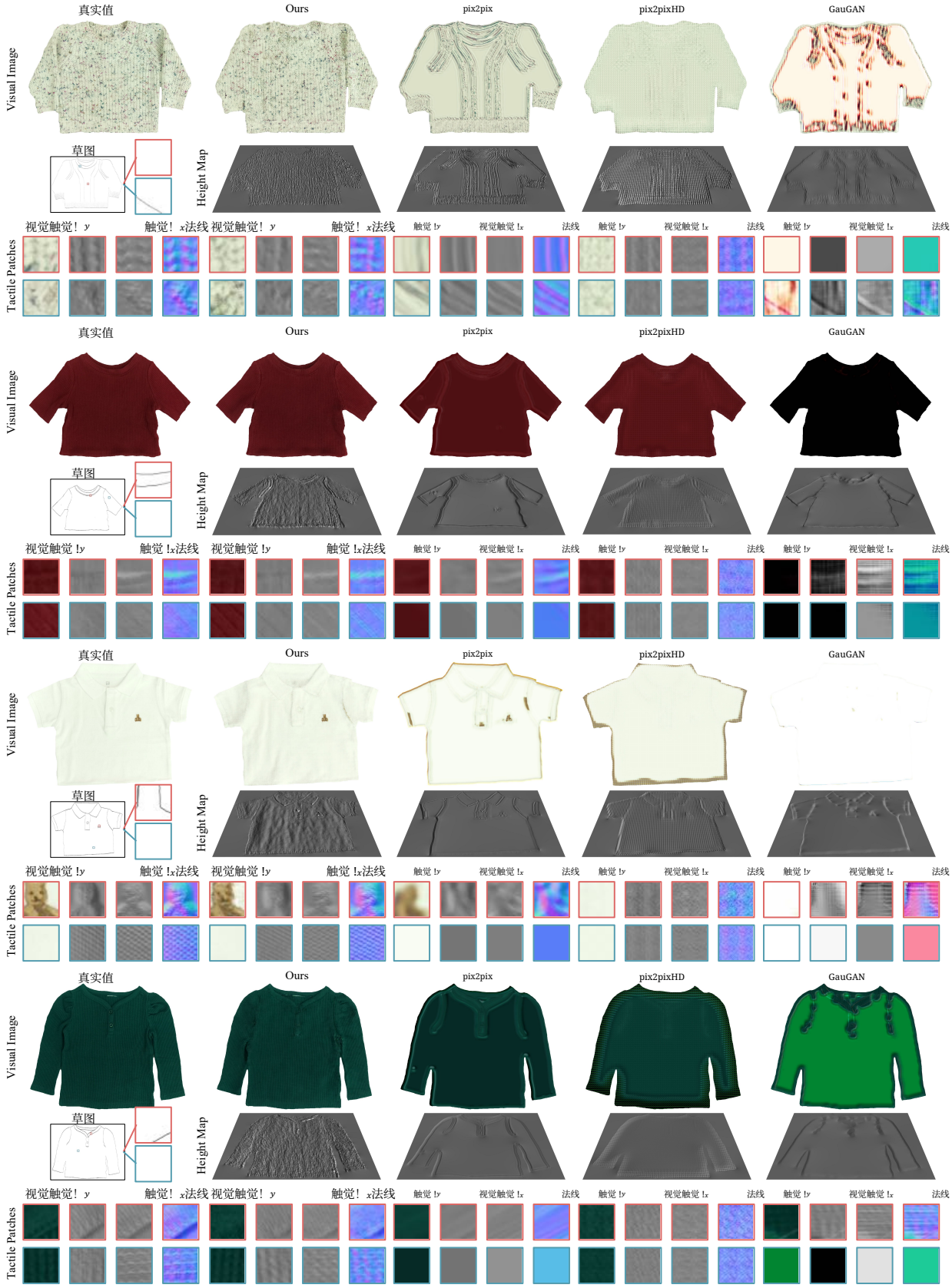


图16： **我们的方法与基线方法在数据集全部20个物体上的定性结果（续).**



图17： 我们的方法与基线方法在数据集全部20个物体上的定性结果（续）).



图18： **我们的方法与基线方法在数据集全部 20 个物体上的定性结果（续）**).

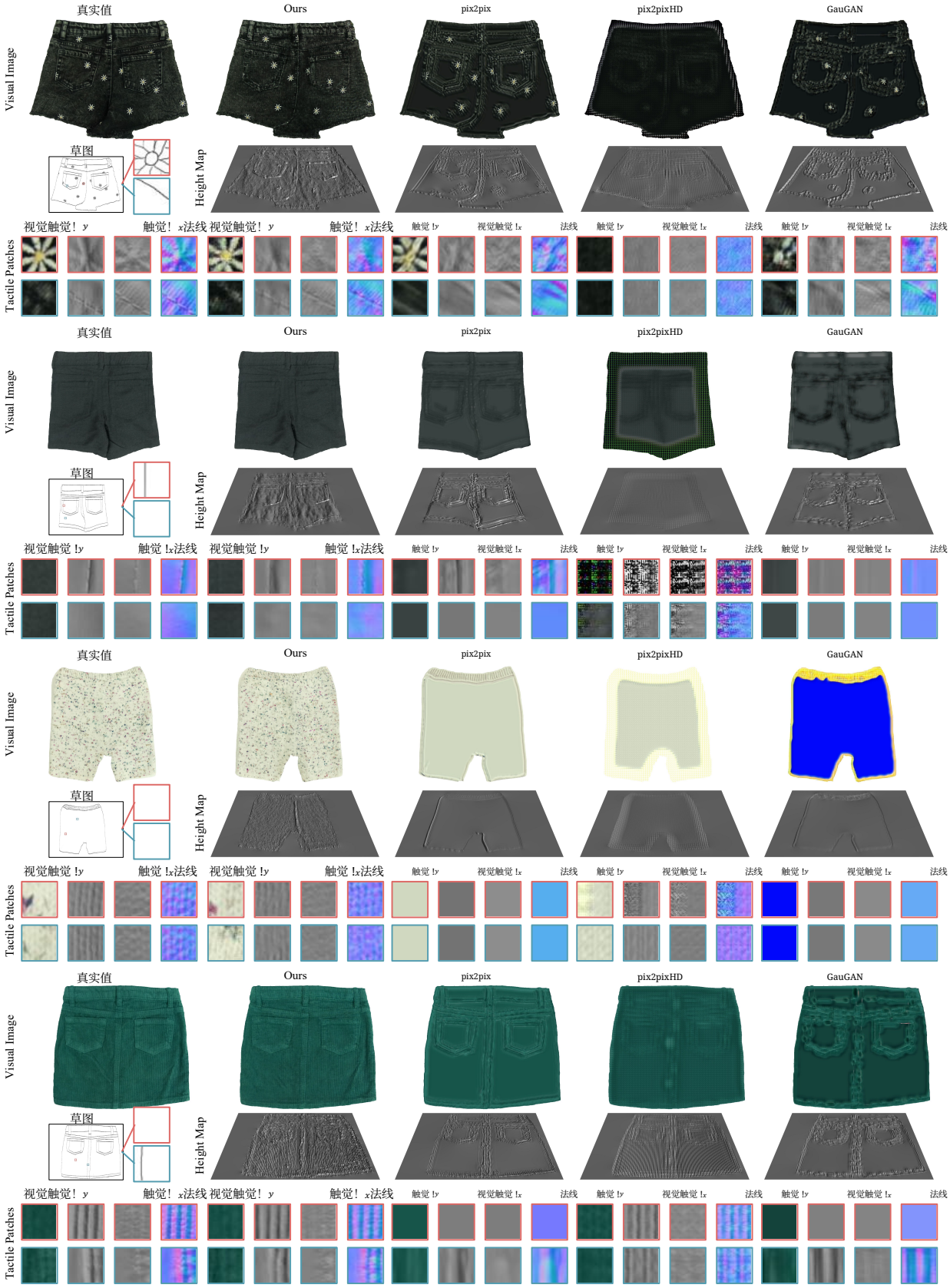
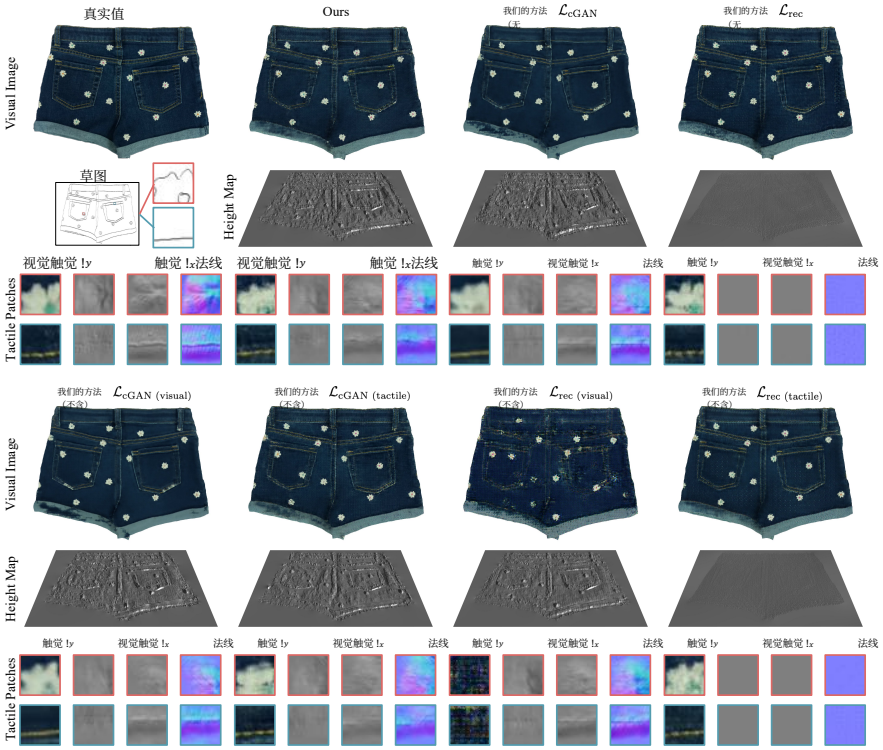
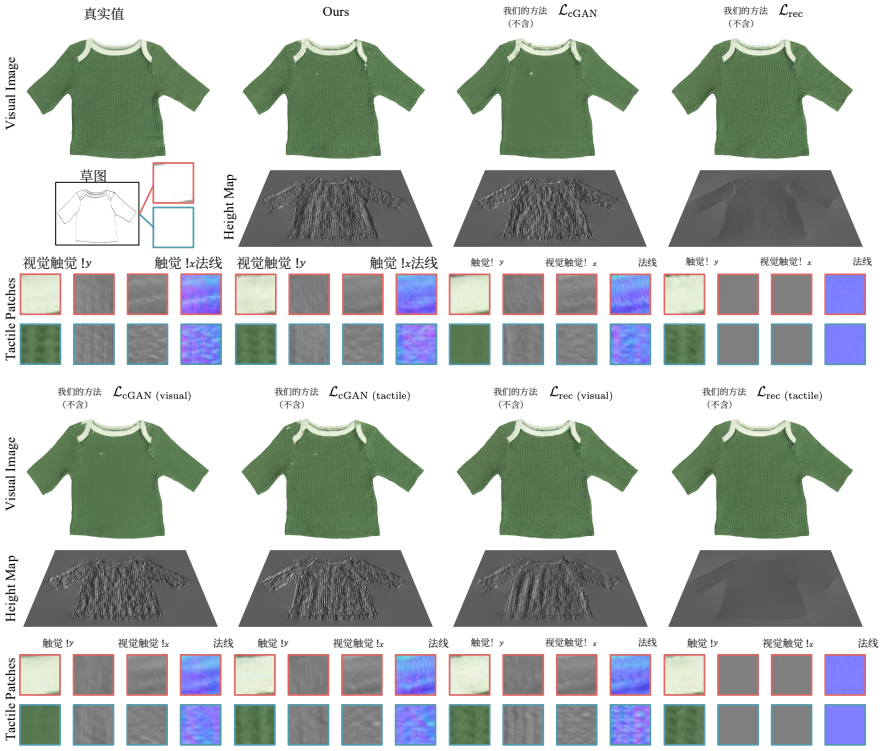


图 19： **数据集中全部20个物体上我们的方法与基线方法的定性结果对比（续).**



(a) 一条带有花形装饰的牛仔裤示例。

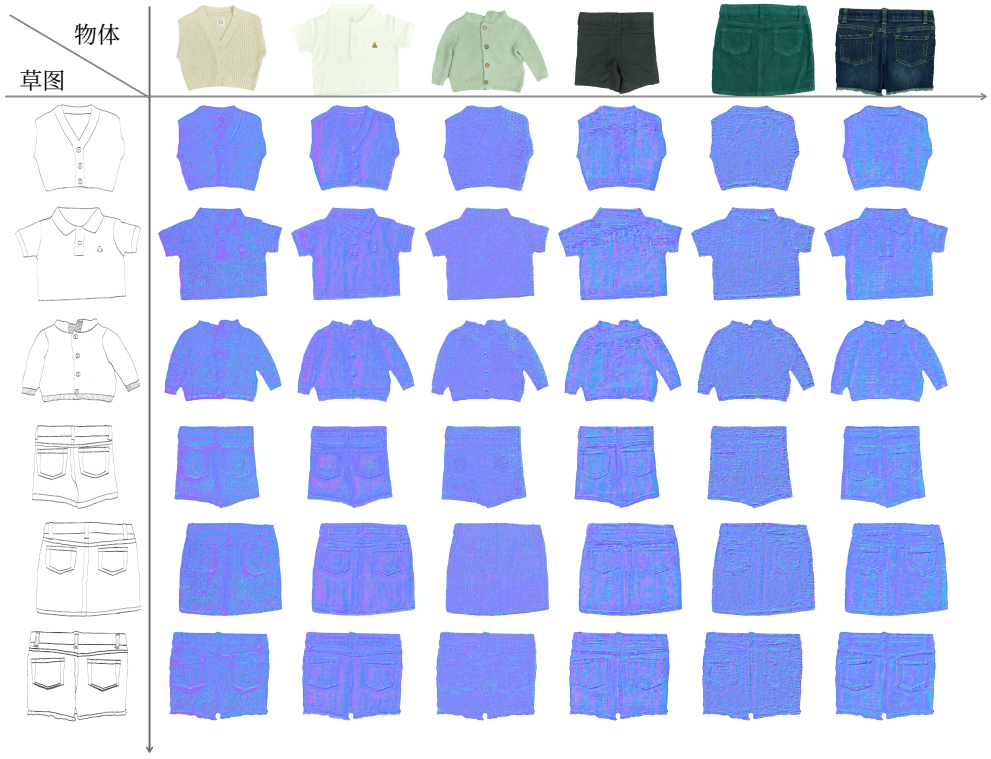


(b) 一件绿色衬衫的示例。

图20：**损失消融研究的定性比较**。我们将我们的方法与六种消融版本进行比较：去除 $\mathcal{L}_{\text{cGAN}}$ 的我们的方法、去除 $\mathcal{L}_{\text{rec}}$ 的我们的方法、去除 $\mathcal{L}$ 条件生成对抗网络（视觉）的我们的方法、去除 $\mathcal{L}$ 条件生成对抗网络（触觉）的我们的方法、去除 $\mathcal{L}$ 重建损失（视觉）的我们的方法，以及去除 $\mathcal{L}$ 重建损失（触觉）的我们的方法。对于每种方法，我们从上到下展示视觉输出、触觉输出（以高度图显示）以及两个放大图块。网络的输入草图显示在最左上方的列中，其中两个颜色编码边界框标示了两个图块的位置。总体而言，移除对抗损失会导致输出过度平滑，而移除重建损失则会引入无意义的触觉输出和视觉伪影。



（a）在不同草图上进行测试的视觉输出。



（b）在不同草图上进行测试的触觉输出（以法线图格式显示）。

图 21：**交换草图和材质的更多结果**。我们的模型能够为已知和未见过的草图合成视觉和触觉结果。对于每组草图-材质组合，我们在 (a) 中展示视觉输出，在 (b) 中以法线图的形式展示触觉输出。