一文带你了解基于视觉的机器人抓取自学习(Robot Learning)

“一眼就能学会动作”，或许对人而言，这样的要求有点过高，然而，在机器人的身上，这个想法正在逐步实现中。马斯克（Elon Musk）创立的人工智能公司 Open AI 研究通过 One-Shot Imitation Learning 算法(一眼模仿学习)，让机器人能够复制人类行为。现阶段理想化的目标是人类教机器人一个任务，经过人类演示一次后，机器人可以自学完成指定任务。机器人学习的过程，与人类的学习具有相通之处，但是需要机器人能够理解任务的动作方式和动作意图，并且将其转化为机器人自身的控制运动上。

“机器人学习”是机器人研究的重要方向，其中包含了计算机视觉，自然语言处理，机器人控制等众多技术。机器人抓取(Robotic manipulation/grasping)是机器人智能化发展道路上亟待解决的问题之一。相较于传统的开环控制系统，本文将从基于视觉，基于视觉和语音，基于视觉和触觉三个方向出发，介绍机器人抓取的相关研究进展，并罗列相关的文章供大家查找阅读。

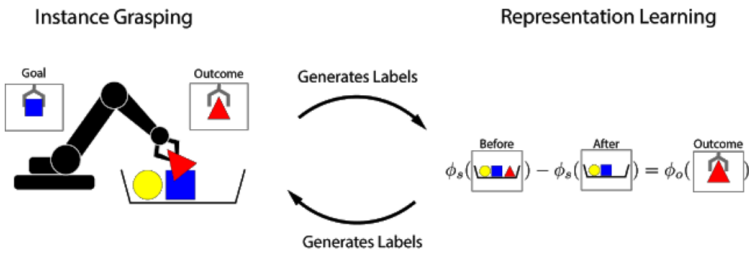
1. 基于视觉信息的机器人抓取学习

Google AI Blog: Grasp2Vec: Learning Object Representations from Self-Supervised Grasping

【论文原文摘要】结构良好的视觉表示可以使机器人学习更快，并且可以提高通用性。在本文中，研究人员研究了在没有人工标记的情况下，如何通过使用自主的机器人与环境的交互获得有效的以物体为中心的表示方法，即可完成机器人操作任务。这种机器人学习的方法可以让机器人收集获取更多的经验，不断完善机器人的认知，从而无需人工干预即可有效地进行缩放。本文中的学习方法是基于对象的永久性：当机器人从场景中删除对象时，该场景的表示会根据被删除对象的特征而随之变化。研究人员根据观察结果会在特征向量之间建立关系，并使用它来学习场景和物体的表示。这些场景和物体可用于识别对象实例，将它们在场景中进行定位，并在机器人从目标箱中检索命令对象时，执行以目标为导向的任务。整体的抓取过程是通过记录场景图像，抓取和移除物体以及记录结果，该抓取过程也可以用于为文中的方法自动收集训练数据。文中实验表明，这种用于任务抓取的自我监督方法明显优于直接增强图像学习方法和先前的表征学习方法。

从小时候开始，即使从未有人明确地教过如何做，人们依旧能够识别并收拾取自己喜欢的物品。根据认知发展研究，这种与世界中的物体相互交互的能力，在人类感知和操纵物体的能力形成的过程中起着重要的作用。通过与周围世界的互动，人们可以通过自我监督来学习：知道自己采取了什么行动，并且从结果中学到了什么知识。在机器人技术中，人们积极研究了这种自我监督型学习，因为它使机器人系统无需大量的训练数据或人工监督即可进行学习。

受对象永久性概念的启发，研究人员提出了 Grasp2Vec，一种用于获取物体表示的简单而高效的算法。Grasp2Vec 算法中尝试抓取任何东西都会获取以下几条信息——如果机器人抓住一个物体并将其抬起，则物体必须在抓取前进入场景。此外，若机器人知道它抓住的物体当前处于夹爪中，就会将其从场景中移除。通过使用这种形式的自监督，机器人可以利用抓取前后的场景视觉变化来学习识别物体。



基于前与 X Robotics 合作的基础上(该项目的任务是让一系列机器人同时学习使用单目相机输入来抓取家用物品)，研究人员使用机械臂“无意间”抓取物体，这种经验使机器人能够学习丰富的图像对象。这些表示可用于获取“有意抓握”的能力，并且机械臂可以拾取用户指定的对象。

在强化学习的框架中，通过“奖励函数”可以衡量任务的成功与否。通过最大化奖励函数，机器人可以从头开始自学各种抓握的技能。如果任务的成功与否可以通过简单的方法来衡量，设计奖励函数就很容易。一个简单的例子是当一个按钮被按下时，该按钮直接向机器人提供奖励。

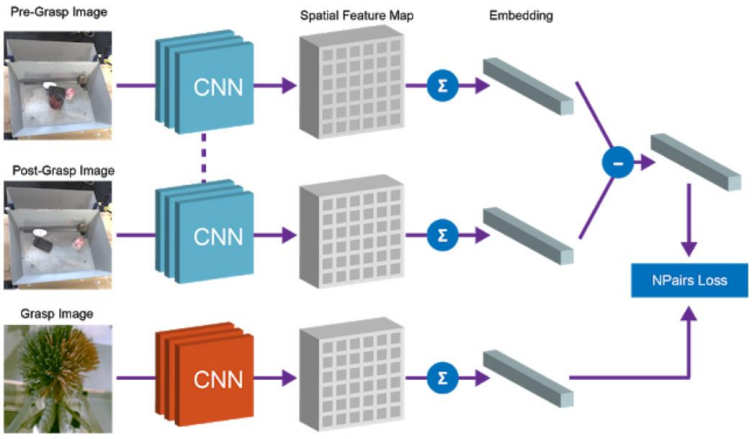
然而，当成功标准取决于对当前任务的“感性理解”时，设计奖励函数的难度就会加大。考虑实例抓取的任务，其中机器人看到的是期望的物体图片。当机器人试图抓住该物体后，将会检查抓取的对象。此任务的奖励函数可以看作物体识别问题：抓住的物体是否与期望相匹配？

为了解决这种识别问题，需要一种感知系统：该系统能从非结构化图像数据中提取有意义的物体概念，并能以无监督的方式学习物体的视觉感知。该研究在数据收集的过程中，利用机器人可以操纵物体移动的优势，提供数据所需的变化因素。通过对物体进行抓取，可以获得 1）抓取前的场景图像；2）抓取后的场景图像；3）抓握物体本身的孤立视图。

研究人员提出了一个从图像中提取“物体集合”的嵌入函数，该函数满足以下减法关系:

objects_before_grasp - objects_after_grasp = grasped_object

文中使用了全卷积架构和简单的度量学习算法来实现这种等式关系，特征图中嵌入抓取前的场景图像和抓取后的场景图像，并将其平均池化后保存到向量中，而“抓取前”和“抓取后”向量的差表示一组物体。该向量和对应的被抓取物体的向量表示之间的等价约束是通过 N-Pairs 目标函数实现的。通过 N-Pairs 目标函数实现该向量和对应的被抓取物体的向量之间的等价约束关系。

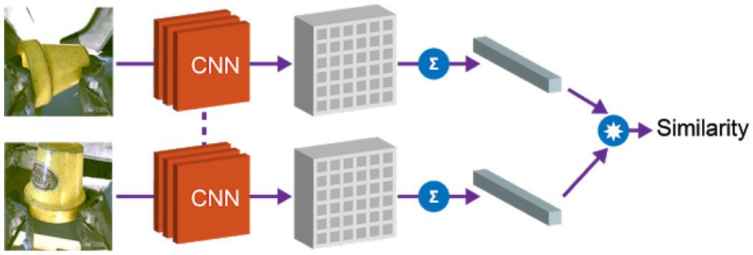


Add caption

训练过后，模型中会出现两个有用的属性。

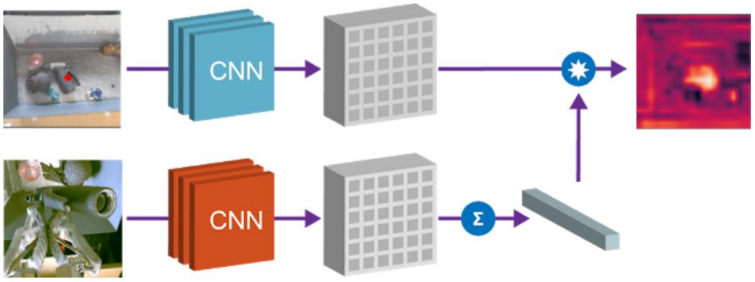
1) 物体相似度

第一个属性是余弦距离，利用向量间的余弦距离对物体进行比较，并确定是否相同。这个属性可以用于实现强化学习的奖励函数，并允许机器人在没有人工提供的标签的情况下学习实例抓取。



2) 目标物体本地化

第二个属性是，可以组合场景空间映射和物体嵌入来本地化图像空间中的“查询对象”。将空间场景的特征图和查询对象的向量相乘，以找到两者之间“匹配”的所有像素。例如下图中的场景，模型可以检测出场景中的多个相应的色块，通过点乘得到的“热图”，可用于规划机器人接近目标物体的方法。



该项目展示了机器人抓取技能如何生成用于学习以物体为中心的表示的数据，并使用表示学习来实现更复杂的技能，例如实例抓取，与此同时保留自主抓取系统中的自监督学习属性。

2. 基于视觉和语音信息的机器人抓取

Improving Grounded Natural Language Understanding through Human-Robot Dialog

【摘要】机器人自然语言理解会需要大量特定性领域和平台的工程量。例如，移动机器人在特定环境中接收操纵者的命令拾取放置物品，人类可以指定语言为某类命令，并将概念词与物体对象的属性进行关联，例如红色这样的概念词。减轻类似工作量的方法是使环境中的机器人能够动态适应，不断学习新的语言构造和感知概念等。在这项工作中，研究人员提出了一种端到端的方法，用于将自然语言命令翻译为离散的机器人动作，并使用对话框共同明确和改善语义和基础概念。研究在 Amazon Mechanical Turk 的虚拟设置上对该目标对象进行训练和评估，并将该智能体转移到现实世界中的物理机器人平台上，进行展示。

随着机器人在家庭、工厂和医院等环境中变得无处不在，人类对有效的人机交互的需求也在不断增长。上述各类场景中会包含特定的词汇和行为启示，例如，打开厨房的灯；把托盘往北移 6 英尺；如果病人的情况有变化，就通知我。因此，预编程机器人的语言理解会需要昂贵的特定性领域和平台的工程。在本文中，研究人员提出并评估了一种机器人智能体，它可以通过与人类对话的方式扩展一个初始状态下资源较少、依靠手工编程的语言理解管道，从而与人类伙伴更好地达成共识。

研究人员结合了通过对话的信号进行更好的语义解析(以前不使用物体的感官表征)和主动学习方法来获取这些概念(以前仅限于对象识别任务)。因此,文中的系统能够执行自然语言命令，例如将一个能发出叮叮当当响声的容器从会议室的休息室移到 Bob 的办公室，其中包含组成语言（例如，语义分析器理解的会议室休息室以及将由其识别的对象的物理性质，如能发出叮叮当当响声的容器）。系统仅用少量的用于语义解析的自然语言数据进行初始化，没有将概念词与物理对象绑定的初始标签，而是需要通过人机对话学习解析和接地。

本文的贡献主要是:1)提出了一种对话策略，仅利用少量初始领域内的训练数据来提高语言理解;2)利用对话问题在现场实时获取感知认识，而不是仅从预先标记的数据或过去的交互过程中获取;3)在一个完整的物理机器人平台上部署对话智能体。

研究人员在 Mechanical Turk 上评估智能体的学习能力和可用性，要求用户通过对话指挥智能体去完成三个任务:导航(由厨房去休息室)，传递(将红色的罐子拿给 Bob),和搬运(将一个空瓶子从厨房休息室转移到爱丽丝的办公室)。研究发现，根据之前对话中提取的信息对智能体进行训练后，它的评价指标会更好。然后，研究人员将经过训练的智能体转移到物理机器人上，并在人机对话中演示它的持续学习过程。

该会话智能体主要通过视觉信息和自然语言结合完成请求。整体主要包括以下几个部分。1)语义解析器：智能体通过获取的单词序列推断任务的语义表示，使用组合类别语法(CCG)形式来进行解析。2)语言接地，根据不同的外部环境，相同的语义也可能会以不同的方式接地。例如，厨房旁边的办公室指的是一个物理位置，但这个位置取决于建筑。3)对话框，人机之间的对话常常从人类用户开始，指示智能体完成某项任务，智能体会对未观察到的真实任务进行建模，并使用来自用户的语言信号推断该任务。该命令由语义解析和基础组件处理，以获得成对的符号和置信状态值。置信状态值通过语义解析（例如，“在北边的办公室的豆荚”中的介词歧义；豆荚还是办公室向北）和语言理解（例如，嘈杂的概念模型）步骤对不确定性进行建模。4)从对话中学习：该智能体通过在完成的对话中引入训练数据来改进其语义解析器，智能体能够将用户的初始命令与确认的动作进行匹配，从会话中学习语义。同时，采用主动学习的方式，从向用户提出的问题中快速扩展感知概念模型，然后在各个用户之间汇总扩展，并且可以将学习到的概念应用于远程测试对象，有助于获取新概念。

会话智能体的组成如下图所示，左侧是将用户的命令进行语义解析，中间为利用已有的地图和概念模型等信息对指令进行接地，右侧是利用对话改进完善智能体的认知模型。

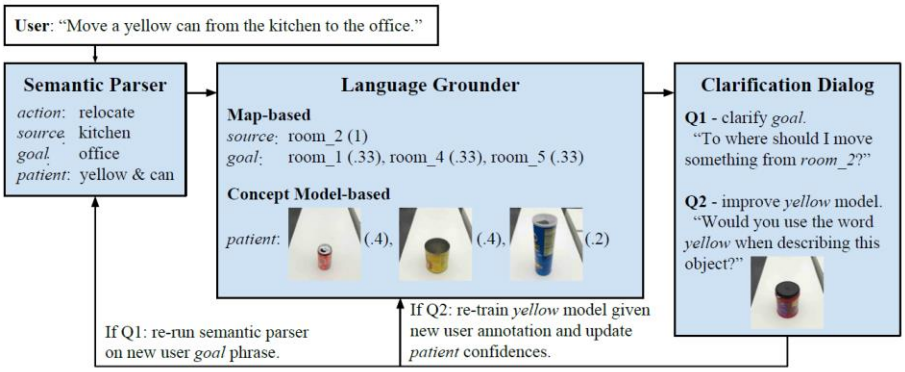
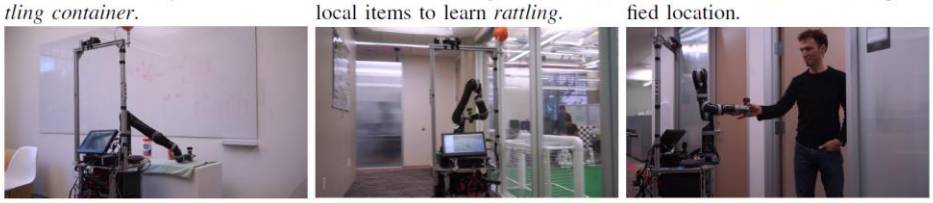


Fig. 2: User commands are parsed into semantic slots (**left**), which are grounded (**center**) using either a known map (for rooms and people) or learned concept models (for objects) to a distribution over possible satisfying constants (e.g., all rooms that can be described as an "office"). A clarification dialog (**right**) is used to recover from ambiguous or misunderstood slots (e.g., *Q1*), and to improve concept models on the fly (e.g., *Q2*).

实验中指定的任务包含：根据用户指示完成到达指定地点，将物品递送给某人，将物品从指定地点移动到目的地。下图为受过训练的智能体采用动态学习的方式实现指定的目标。

R: What should I do?
H: Move a rattling container from the lounge by the conference room to Bob's office.
R: Does rattling refer to a property of an object?
H: Yes.
R: Does rattling mean the same thing as empty?
H: No.
R: I am still learning the meanings of some words.
R: Show me an object you could use the word rattling when describing, or say "none of them."
H: Watch.
R: I am getting the objects on the table into focus... Okay, I see them.
H: This one. *The human picks up a small, full pill jar.*
R: I see.
R: Show me an object you could not use the word rattling when describing, or say "all of them."
H: Watch. This one. *The human picks up a red water bottle.*
R: I see.
R: Show me an object you could use the word rattling when describing, or say "none of them."
H: Watch. This one. *The human picks up a container half-full of beans.*
R: I see.
R: Thanks. Now back to business.
R: You want me to move an item from 3.514 to somewhere?
H: Yes.
R: I should move something from 3.514 to 3.510? [*room numbers*]
H: Yes.

(a) The human says *move a rattling container*. (b) The robot asks questions about local items to learn *rattling*. (c) The robot moves to the specified location.



(d) The robot infers and grasps a *rattling container*. (e) The robot navigates to the specified destination room. (f) The robot hands over the item at the specified destination.

Fig. 5: The *Trained (Parsing+Perception)* agent continues learning on the fly to achieve the specified goal.

Agent	Clarification Questions ↓		
	Navigation (p)	Delivery (p)	Relocation (p)
In	3.02 ± 6.48	6.81 ± 8.69	22.3 ± 9.15
Tr*	$4.05 \pm 8.81(.46)$	$8.16 \pm 13.8(.53)$	$23.5 \pm 6.07(.67)$
Tr	$1.35 \pm 4.44(.11)$	$7.50 \pm 9.93(.72)$	$19.6 \pm 7.89(.47)$

TABLE II: The average number of clarification questions agents asked among dialogs that reached the correct task. Also given are the p-values of a Welch's t-test between the **Tr***ained** (*Perception*) and **Tr***ained* (*Parsing+Perception*) model ratings against the **In***itial* model ratings.

上表比较初始智能体，受过训练(仅感知训练)智能体，受过训练(解析训练和感知训练)的智能体三者的实验情况，衡量的标准是在满足正确的任务规范之前，需要进行的询问的问题的个数，实验显示受过训练(仅感知训练)智能体表现较差，可能是由于对话中的许多形容词和名词的属性没有及时更新。

Agent	Usability Survey (Likert 1-7) ↑		
	Navigation (p)	Delivery (p)	Relocation (p)
In	3.09 ± 2.04	3.20 ± 2.12	3.37 ± 2.17
Tr*	$3.51 \pm 2.05(.09)$	$3.60 \pm 2.09(.12)$	$3.60 \pm 2.08(.37)$
Tr	$\mathbf{3.76} \pm 2.07(.01)$	$\mathbf{3.87} \pm 2.10(.01)$	$\mathbf{3.93} \pm 2.16(.04)$

TABLE III: The average Likert rating given on usability survey prompts for each task across the agents. **Bold** indicates an average **Tr***ained** (*Perception*) and **Tr***ained* (*Parsing+Perception*) model ratings significantly higher than the **In***itial* model ($p < 0.05$) under a Welch's t-test.

上表比较初始智能体，受过训练(仅感知训练)智能体，受过训练(解析训练和感知训练)的智能体三者的实验情况，衡量的标准是用户对智能体表现的定性的评价，主要包括：我将使用这样的机器人来帮助导航到一栋新楼；我将会用这样的机器人为自己或其他人拿取东西；我将会用这样的机器人来将物品从一个地方移到另一个地方。实验显示受过训练(解析和感知)的智能体的表现最好。

该研究提出了一种机器人智能体，其可以利用与人类的对话来扩展自定义的小型化的语言理解资源，利用这些资源既可以将自然语言命令翻译为抽象的语义形式，又可以将物理对象的抽象属性接地。在这项工作中，机器人可以执行的动作可以分解为离散语义角色的元组，但是通常，他们需要推理更多的连续动作空间，并获取新的、与人类对话中看不见的行为和知识。该研究中的智能体可以从人机对话中学习知识，甚至可以处理复杂的形容词和名词之间的依赖和上下文关系。

3. 基于视觉和触觉信息的机器人抓取

Connecting Touch and Vision via Cross-Modal Prediction

【摘要】人类使用视觉、听觉和触觉等多种模式的感觉输入来感知世界。在这项工作中研究了视觉和触觉之间的交叉模式连接。跨模态建模任务的主要挑战在于两者之间在比例上存在显着差异：虽然我们的眼睛一次性就可以感知到整个视觉场景，但人类在任何给定时刻只能触碰感觉到物体的一个小部分。为了连接视觉和触觉，文中合成来自视觉输入的合理的触觉信号，以及想象我们如何与以触觉数据作为输入的对象进行交互。为了实现该目标，研究人员首先为机器人配备了视觉和触觉传感器，并收集了相应视觉和触觉图像序列的大规模数据集。为了缩小规模差距，研究中提出了一个新的条件对抗模型，该模型结合了触摸的规模和位置信息。人类的感知研究表明，本文中的模型可以从触觉数据中产生逼真的视觉图像，

反之亦然。最后，展示了有关不同系统设计的定性和定量实验结果，以及可视化了模型的学习表示。

文中提出了一种跨模态预测方法，用于从触摸预测视觉，反之亦可。研究人员首先将触觉中的程度、规模、范围和位置信息结合在模型中。然后，使用数据平衡的方法多样化其结果。最后，通过考虑时间信息的方法进一步提高准确性。

研究中的模型基于 pix2pix 方法，是一个用于图像到图像任务的条件 GAN 框架。在任务中，生成器接受视觉图像或触觉图像作为输入，并生成一个对应的触觉或视觉图像。而判别器观察输入的图像和输出的图像。在训练中，对判别器进行训练，以分辨合成图片和真实图片之间的差异，而生成器则是用于产生可以欺骗判别器的图片。在实验中，研究人员使用视觉-触觉图像对训练模型。在从触觉还原视觉的任务中，输入触觉图像，而输出是对应的视觉图像。而在视觉预测触觉的任务中，则输入和输出对调。

模型使用编码器-解码器架构用于生成任务。在编码器上分别使用两个 ResNet-18 模型用于输入图像（视觉或触觉图像）和参考的视觉-触觉图像，将两个编码器的向量合并为一个 1024 维向量，将其输入解码器。解码器包括五层标准的卷积神经网络，并在编码器和解码器间加入了跨层连接，研究中使用的判别器为 ConvNets。

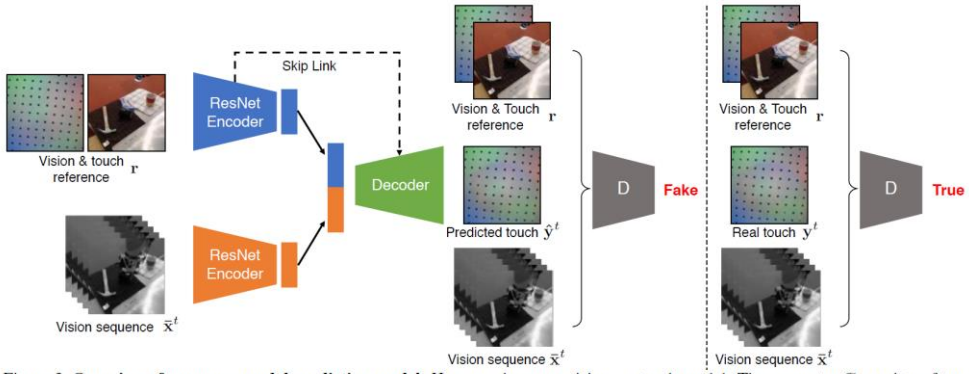


Figure 3. **Overview of our cross-modal prediction model.** Here we show our vision → touch model. The generator G consists of two ResNet encoders and one decoder. It takes both reference vision and touch images $\mathbf{r}$ as well as a sequence of frames $\bar{\mathbf{x}}^t$ as input, and predict the tactile signal $\hat{\mathbf{y}}^t$ as output. Both reference images and temporal information help improve the results. Our discriminator learns to distinguish between the generated tactile signal $\hat{\mathbf{y}}^t$ and real tactile data $\mathbf{y}^t$. For touch → vision, we switch the input and output modality and train the model under the same framework.

研究发现，实验结果不是很好，图片中有严重的视觉伪影，并且生成的结果与输入信号不一致。为解决上述问题，研究人员对基本算法进行修改和完善。首先将触觉和视觉参考图像提供给生成器和判别器，以便该模型只需要学习为交叉模式变化建模，而不是整个信号。其次，为防止模式崩塌，研究人员采取数据重均衡策略帮助生成器生成不同的模式，性能更加健壮。最后，从输入视频的多个相邻帧而不是仅从当前帧中提取信息，从而产生时间相干的输出。

研究人员在一个 KUKA 机械手臂上放置 GelSight 传感器，机械臂背面的三脚架上安装了一个网络摄像头，以捕捉机械臂触摸物体的场景视频，实验中让机械臂去戳弄不同的物体。GelSight 表面有一层薄膜，在接触物体的过程中会发生形变，进而采集到高质量的触觉数据。研究团队总共记录了 195 件物品的 12000 次触碰，这些物品属于不同类别。每个触摸动作包含一个 250 帧的视频序列，产生了 300 万视觉和触觉成对的图像的数据集—VisGel。根据此数据集，当模型辨认到接触位置的形状和材料，与参考图像进行比较，以识别触摸的位置和范围。

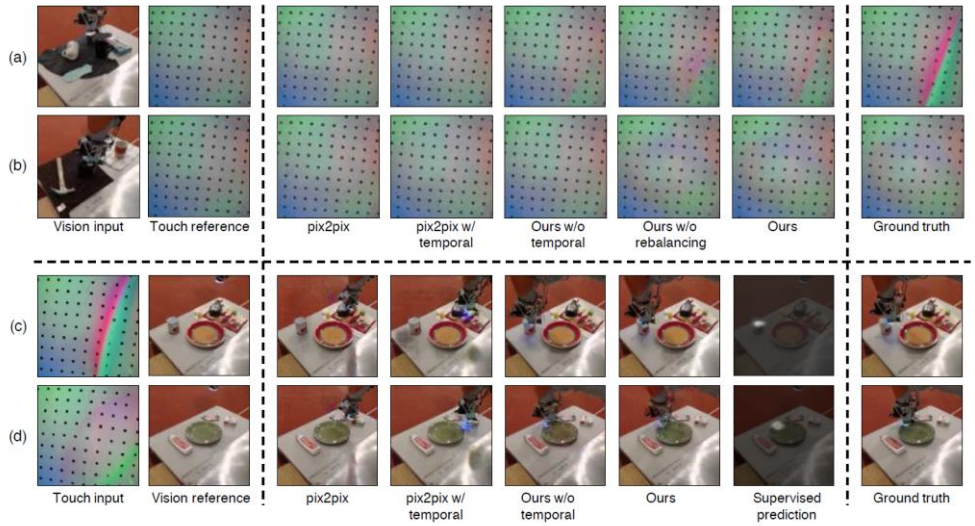


Figure 5. **Example cross-modal prediction results.** (a) and (b) show two examples of vision → touch prediction by our model and baselines. (c) and (d) show the touch → vision direction. In both cases, our results appear both realistic and visually similar to the ground truth target images. In (c) and (d), our model, trained without ground truth position annotation, can accurately predict touch locations, comparable to a fully supervised prediction method.

上图是本文模型和其他基线模型实验结果的可视化对比，该模型可以更好地根据视觉信息预测物体表面的触觉信息，也能够更好地根据触觉信息还原图像表面。

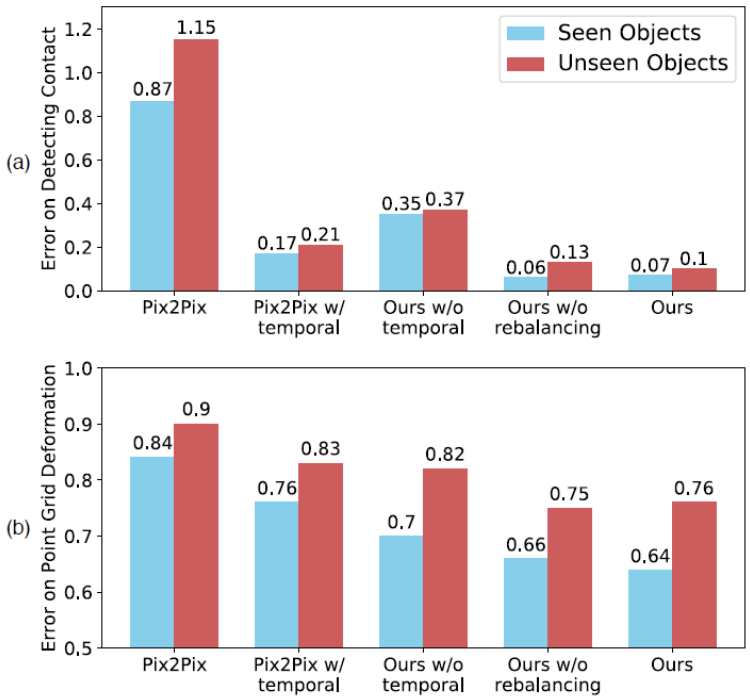


Figure 6. **Vision2Touch - quantitative results. Top:** Errors on detecting the moment of contact. Our method generally performs the best. The use of temporal cues can significantly improve the performance or our model. **Bottom:** Errors on the average markers' deformation. Our method still works best.

上图是从视觉到触觉的量化评测结果。a 图的评价指标是测试机器人是否已经认知到触摸了物体表面的错误数。b 图的评价指标是根据图像还原触觉点位置的失真错误情况。本文中的模型表现优于其它模型。

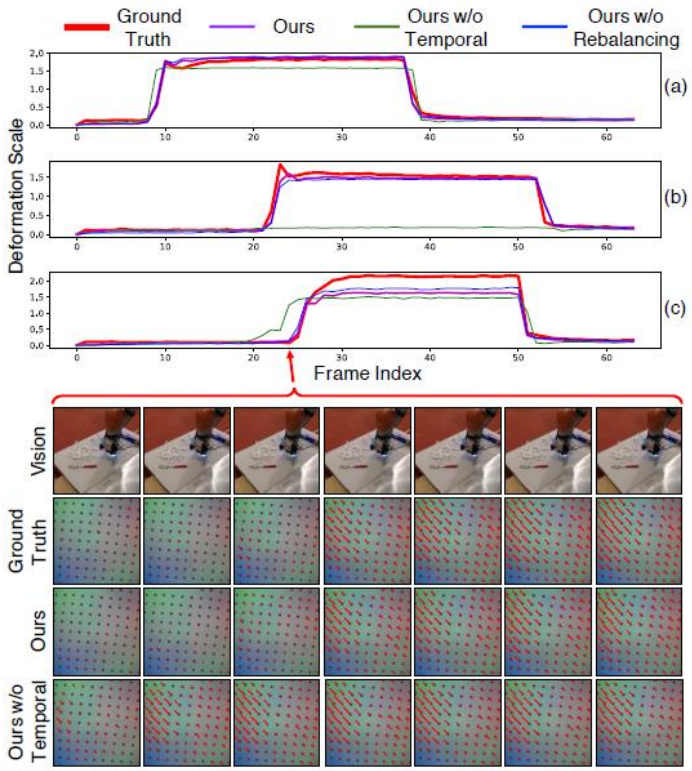


Figure 7. **Vision2Touch - detecting the moment of contact.** We show the markers' deformation across time, determined by the average shift of all black markers. Higher deformation implies object contact with a larger force. **Top:** Three typical cases, where (a) all methods can infer the moment of contact, (b) the method without temporal cues failed to capture the moment of contact, and (c) the method without temporal cues produces misaligned results. **Bottom:** We show several vision and touch frames from case (c). Our model with temporal cues can predict GelSight's deformation more accurately. The motion of the markers is magnified in red for better visualization.

上图是从视觉还原触觉的情况，其中显示了标记随时间的变形，该变形由所有黑色标记的平均位移确定，较高的变形意味着物体以较大的力接触。下图是根据图像还原的触觉点阵信息，为便于增强可视化的效果，图片中的标记的运动以红色放大。

该项工作提出了在视觉和触觉与条件对抗网络之间建立联系。当与外界互动时，人类非常依赖视觉和触觉的感官方式。该模型可以为已知物体和未知物体进行跨模态的预测。研究人员认为在将来，视触交叉的模式可以帮助视觉和机器人技术应用，例如在弱光环境下的物体识别和抓取以及物理场景理解。

发布链接：https://mp.weixin.qq.com/s/-kgnT61IAIuCe1mHu-driQ

相关文章：

1) Vision-based Navigation with Language-based Assistance via Imitation Learning with Indirect Intervention;
2) Vision-based grasp learning of an anthropomorphic hand-arm system in a synergy-based control framework;
3) Visual Curiosity: Learning to Ask Questions to Learn Visual Recognition;

4) Vision-based grasp learning of an anthropomorphic hand-arm system in a synergy-based control framework.

References:

1) Google AI Blog: Grasp2Vec: Learning Object Representations from Self-Supervised Grasping;
2) Improving Grounded Natural Language Understanding through Human-Robot Dialog;
3) Connecting Touch and Vision via Cross-Modal Prediction.