用于增强鸡胴体实例分割的合成数据扩充

Yihong Feng, Chaitanya Pallerla, Xiaomin Lin, Pouya Sohrabipour Sr, Philip Crandall, Wan Shou, Yu She, Dongyi Wang

***Abstract*—家禽工业主要由肉鸡生产驱动，已发展成为世界上最大的动物蛋白行业。屠宰场和家禽加工厂的质量控制、食品安全和运营效率关键在于加工线上对鸡尸自动检测。然而，在这些快节奏的工业环境中，开发用于实例分割等任务的强大深度学习模型通常受限于需要耗时的大规模真实世界图像数据集获取和标注。为此，我们提出了第一个生成逼真、自动标记的鸡尸体合成图像的管道，这些图像在不同的姿态下呈现。我们还推出了一个新的基准数据集，包含 300 张注释过的真实世界图像，专门为家禽分割研究而整理。利用这些数据集，本研究探索了合成数据和自动数据注释在增强鸡尸体实例分割方面的有效性，特别是在加工线上缺乏真实标注数据的情况下。对包含不同比例合成图像的小型真实数据集（60 张鸡尸体图像）在著名实例分割模型中进行评估：YOLOv11-seg、Mask R-CNN（使用 R50 和 R101 骨干网络）以及 Mask2Former。结果显示，合成数据显著提升了所有模型对鸡尸体的分割性能。YOLOv11-seg 始终实现了最高的准确率。值得注意的是，容量较大的模型（R101）和基于转换器的架构（Mask2Former）从中获得了更大的收益，尤其是在使用大量合成数据时。此外，还观察到模型特定的合成数据与真实数据的最佳比例。本研究强调了合成数据增强作为缓解数据稀缺、减少手动注释工作量、推进家禽加工行业中基于 AI 的鸡尸体自动检测系统开发的可行且有效的策略的重要性。**

***Index Terms*—Synthetic Data, Chicken, Data Augmentation, Instance Segmentation, Blender, Mask-RCNN, YOLOv11.**

在过去的二十年里，受到成熟市场和新兴市场需求增加的推动，家禽已成为全球消费最广泛的动物蛋白。此外，预计在未来十年内，家禽业将继续是全球最大的肉类出口行业。从 2001 年到 2021 年，家禽生产的迅速扩张和消费者需求的增长导致全球家禽进口达到历史新高。美国的家禽业在国内外处于领先地位，这得益于先进的生产结构、现代化的家禽遗传技术、丰富的国内饲料资源以及强劲的消费者需求。2024 年全球家禽销售额达到 702 亿美元，比 2023 年的 674 亿美元增长了 4%。

尽管取得了经济上的成功，家禽业在加工阶段面临着重大挑战，尤其是在屠宰设施中。鸡肉加工厂的工作条件十分严酷，因为环境温度低至 4 °C ，湿度又高至 [1] 。此外，工人往往需要按照每分钟处理 140 只鸡的生产线速度工作，手动将冷冻鸡尸体挂在吊架上，以便后续自动脱骨和包装工作 [2] 。这些重复性和手工操作不仅导致了高发的工作场所伤害事故，也导致了熟练工人的短缺 [3] ，这迫使行业寻找创造性的解决方案，以减轻工人安全方面的顾虑，维持生产效率，并确保产品质量。

Yihong Feng, Pouya Sohrabipour Sr, Dongyi Wang are with the Department of Biological and Agricultural Engineering, University of Arkansas, Fayetteville, AR 72701 USA (e-mail: yihongf@uark.edu, pouyas@uark.edu, dongyiw@uark.edu).

Chaitanya Pallerla, Philip Crandall are with the Department of Food Science, University of Arkansas, Fayetteville, AR 72701 USA (e-mail: pallerla@uark.edu, crandal@uark.edu).

Xiaomin Lin is with Department of Electrical Engineering, University of South Florida, Tampa, FL 33620 USA (e-mail: xlin2@usf.edu).

Wan Shou is with Department of Mechanical Engineering, University of Arkansas, Fayetteville, AR 72701 USA (e-mail: wshou@uark.edu).

Yu She is with Department of Industrial Engineering, Purdue University, West Lafayette, IN 47907, USA (e-mail: yushe@purdue.edu).

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

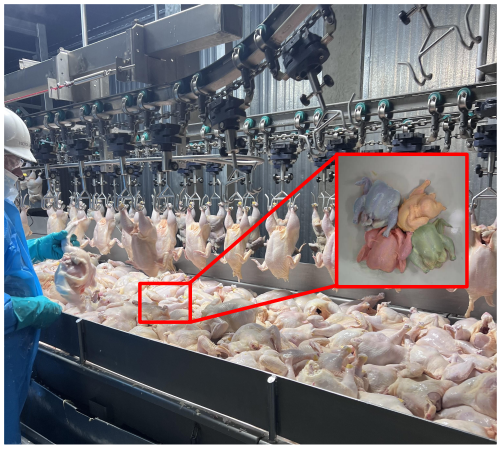

Fig. 1. 家禽屠宰场加工线上的鸡胴体。插图显示了我们数据集中重叠的鸡胴体实例，并覆盖着分割掩码，每种颜色代表一只单独的鸡胴体。

近年来，自动化技术的整合在降低人类暴露在危险环境中的风险、提高生产效率和食品安全方面，为改善家禽加工实践带来了巨大的潜力。在家禽生产阶段，自动化和机器人技术在精准动物管理方面取得了很大进展，诸如监测环境条件、健康管理和蛋的拾取。然而，相比之下，家禽屠宰厂自动化的研究，尤其是在尸体处理方面仍然有限，因为在图像理解和机器人操作方面存在独特挑战。图像理解的挑战包括：(a) 鸡尸体姿势和形状的变化，在传送带上堆积的尸体呈现出多样的姿态，如伸展的翅膀或弯曲的腿；(b) 视觉特征的相似性，因为所有尸体通常堆叠在一起，并且具有相似的颜色和纹理，使得个体尸体的视觉区分变得困难；(c) 遮挡和重叠，尸体在传送带上部分或全部遮挡着彼此；(d) 光照条件的变化，包括阴影和反射，降低了图像质量；如果没有智能的视觉理解，机器人系统将难以以达到行业标准所需的精度和准确性来操作鸡尸体。本研究集中于图像理解部分，旨在提高这些复杂环境中的图像识别准确性和系统的鲁棒性。

I. 相关工作

机器视觉技术的集成通过提供智能输入能力显著推进了自动化系统。图像实例分割和目标检测是基于视觉的工业自动化中的两个主要任务。对于目标检测，YOLO (you only look once) 模型由于其在整个农食品行业中的速度和鲁棒性而在实时目标检测中获得了广泛关注。在葡萄栽培中，采用 YOLOv5s 与跟踪算法实现了每秒 50.4 帧实时葡萄簇计数，准确率为 84.9 % [4] 。对于机器人除草，Multimodule-YOLOv7-L 模型在每秒 37.3 帧的情况下实现了 97.1 % mAP，用于生菜识别和杂草严重性分类 [5] 。YOLO 模型也被有效应用于在多种条件下检测苹果、柑橘

和其他水果 [6] 。在家禽养殖中，YOLOv8x-DB 支持自由笼养母鸡的扬尘行为检测 [7] ，而轻量级 YOLO-Claw 网络实现了对于肢体健康评估的 97.1 % mAP [8] ，这为持续监控和健康评估生长中的肉鸡提供了精准的福利监测。

虽然像 YOLO 这样的模型能够快速识别和定位物体，但许多农业和食品加工应用需要更细致的理解，不仅要求检测，还要求精确的像素级划分。这对详细空间信息的需求使实例分割模型成为关注的焦点。

实例分割模型，如 Mask R-CNN 及其变体，已被广泛应用于农业中，用于识别和计数水果、监测植物健康以及检测害虫等任务。这些任务不仅要求检测和定位物体，还需要能够区分同一类别的个体对象。这种像素级的分类在涉及遮挡、形状变化和重叠物体的复杂环境中至关重要。由于视觉复杂性和商业价值，苹果和草莓是研究中常见的基准作物。例如，一个定制的 Mask R-CNN 结合了可变形卷积和注意力模块，在遮挡和光照条件变化的情况下实现了对苹果的分割 [9] 。在田间之外，任务常常要求处理物品和质量控制，而简单的检测是不够的。例如，Mask R-CNN 已被有效应用于多种光照条件下的食品识别，展示了其在加工环境固有视觉挑战中的适应性 [10] [11] 。更广泛地讲，实例分割支持自动化质量控制，例如识别餐桌用橄榄中的缺陷 [12] ，并通过提供精确的位置和形状信息，甚至在物品重叠时，促进了对柔软、可变形食品如鸡肉片的机器人分拣 [13] 。这些应用展示了实例分割在提供详细的像素级理解以推动食品加工自动化和精确化方面的重要性。

尽管基于深度学习的视觉模型具有有效性，它们的部署和性能很大程度上依赖于大规模标注数据集。在家禽加工中，特别是难以获得像素级标注，这是因为尸体具有复杂的视觉特征，例如不规则形状、重叠的四肢、柔软可变形的组织和不同的姿势。此外，可变的照明、湿滑表面的反射以及羽毛或血液的污染加剧了保持一致的物体识别的难度 [14] 。图 1 展示了这些挑战的一个代表性例子，显示了一张来自家禽加工线的典型顶视图，其中多个尸体部分重叠，姿势不规则，带有阴影效果。即使对于人工标注者来讲，准确勾勒每个身体部位进行分割也是一项劳动密集且容易出错的任务。这些问题突显出没有先进的数据策略时实现鲁棒实例分割的困难。

为了应对数据收集和标注的挑战，有不同的图像增强策略可以帮助提高模型的鲁棒性，同时通过将单个对象暴露于多种方向和尺度来解决重叠目标对象的问题。传统图像增强方法，如翻转、旋转、裁剪和缩放，有效地扩大了训练数据集的规模 [15] 。然而，这些仿射变换技术在引入新的、多样的对象实例方面能力有限，并不能模拟对象外观的显著变化，例如质地变化、遮挡或复杂变形 [16] ，[17] 。这些限制在家禽加工领域尤为明显，现实世界的不可预测性显著，而传统的增强方法无法捕捉复杂的胴体处理条件。

合成数据生成为扩展神经网络训练的数据集提供了一种新颖且更先进的方法。有多种生成合成数据的方法。3D 建模和渲染：这些方法依赖于基于物理的渲染（PBR）引擎从头开始创建合成场景，模拟照明、纹理和相机角度 [18], [19] 。最近的进展已转向神经渲染方法，如神经辐射场（NeRF），这种方法通过多视角建模体积场景以合成真实感的新颖视角 [20] ，以及 3D 高斯分层（3DGS），它使用各向异性高斯原语来高效实时地渲染表面 [21] 。其他应用包括草莓植物病害检测系统，其中合成图像能够促进模型训练，以识别在不同环境条件下的各种病理阶段 [22] 。这些神经方法可以更忠实地复制对象级细节、动态几何和真实的遮挡。通过利用这些方法，研究人员可以生成更复杂的合成数据集，从而更准确地反映现实世界的多样性和光照条件。

基于这些先进的视觉渲染技术，基于物理的模拟模仿现实世界的动态 [23] ，这为合成数据生成提供了另一条途径，重点在于再现对象和环境之间的动态行为和交互，而不仅仅是它们的视觉外观 [24] 。这种方法通过模拟现实世界的物理规律生成数据，确保合成内容不仅在视觉上合乎逻辑，而且其行为，比如物体运动、碰撞、变形、流体动力学和环境变化，都遵循物理逻辑。因此，研究人员可以创建涵盖各种复杂场景、极端情况或交互的庞大数据集，这在实际生产环境中将难以复制 [25] [26] ，从而增强机器学习模型的鲁棒性和泛化能力。

这两种方法的结合尤其强大：基于物理的模拟可以生成具有复杂动态行为和物理交互的三维场景和事件，而像 3DGS 这样的神经渲染技术可以高效地从任意新的视角以高保真度渲染这些动态的、物理上合理的场景。这种协同效应使得生成的合成数据在视觉真实感和行为真实性方面都能达到高标准，为训练更为健壮的视觉模型提供了宝贵的数据资源。

此外，利用生成模型进行的风格迁移和图像合成可以为训练目的产生新的数据变体 [27] 。生成对抗网络（GAN）[28] 是一种生成框架，旨在合成具有与训练集数据相似视觉特征的新数据 [29], [30] 。在农业领域，增强数据集的 GAN 已被用于诸如杂草分割 [31] 、苹果检测 [32] 和牡蛎识别 [33] 等任务。GANs 提高了模型的准确性和鲁棒性，解决了数据稀缺性和环境多变性等挑战。然而，GANs 也存在局限性，包括模式崩溃，即模型生成的输出多样性有限，以及由于对抗过程而导致的训练不稳定。此外，GAN 生成的数据可能会引入偏差，可能影响模型在实际应用中的准确性 [34] ；

另一个著名的生成框架是扩散模型 [35] [36] 。这些模型通过一个两阶段的过程来运行：系统地向数据中添加噪声，然后训练一个神经网络来去除噪声，通过将添加的噪声迭代地转化为连贯的数据。扩散模型因其生成高保真和多样化图像的能力而受到认可 [37] ，在样本质量和训练稳定性上常常超过其他模型。这使得扩散模型在某些特殊应用中很有前景，例如生成用于水产养殖的合成牡蛎 [38], [39] 或扩充罕见植物病害或生产缺陷的数据集。然而，这些模型的主要限制是其缓慢的迭代采样过程，与单次生成方法相比，这可能在计算上更加密集 [40] 。继续解决这些采样速度的挑战并探索其在不同科学领域中的应用价值。

尽管大多数合成数据生成方法显著增加了训练数据集的数量和视觉真实性，但它们仍未解决深度学习中最为持久的困难之一：耗时且费力的人工标注任务。即使有合成图像可用，许多应用仍然需要手动标注蒙版或边界框以确保模型的准确性。这使得高效地扩大数据集规模变得困难。

在本文中，我们提出了一个实用的框架，该框架能够自动生成标注的合成数据，并在各种配置下将其与真实数据集结合，以增强实例分割性能，初步示例来自家禽加工行业。我们评估了三个领先的深度学习模型：Mask R-CNN、Mask2Former 和 YOLOv11，在五个数据集设置下进行，包括一个真实的鸡胴体数据集基线和四个增加多达 1000 张 Blender 生成的合成图像的数据集。Blender 是一个 3D 渲染平台，允许对场景参数进行精确控制并自动生成准确的标注，与 GANs 或扩散方法等其他生成模型相比，提供了更高的结构保真度和标注可靠性，这对处理非刚性、解剖结构多样和堆叠物体的任务尤为重要。我们的研究聚焦于家禽加工的重新挂载阶段，此时完整的肉鸡尸体在离开

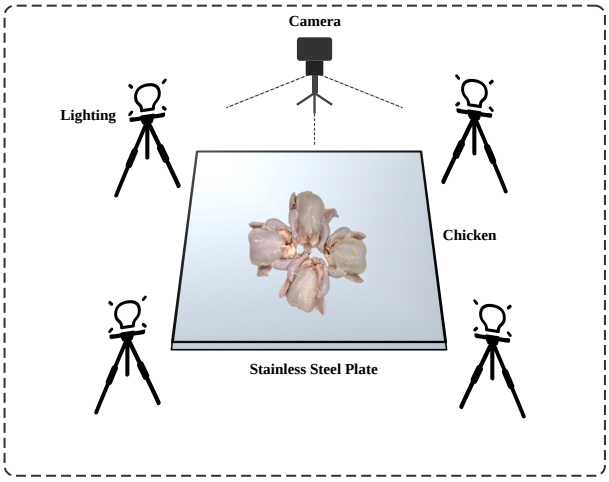

Fig. 2. 真实鸡数据采集系统

冷水冷却机时通常堆积在输送带上，必须重新挂载到一个在上方移动的挂钩线上。至关重要的是，在下一阶段使用机器人进行准确识别堆积家禽尸体的方向、拾取单个尸体并通过踝部挂在移动的挂钩线上时，可靠的实时分割将是可用的。本研究提出了一个从仿真到现实（Sim2Real）的框架，用于生成合成训练数据，目标是分割在密集配置中被遮挡的鸡尸体。此项工作支持更广泛的提升工人人身安全、效率、一致性和家禽生产系统可扩展性的努力，通过数据高效的解决方案来实现。

本研究详细阐述了真实世界和合成鸡胴体分割数据集的收集过程，随后提出了一种数据集增强方法以提高家禽实例分割的性能。在本研究中，深度神经网络在一小部分带注释的真实世界数据和从 Blender（一种计算机图形软件）生成的大规模合成数据的组合上进行了训练。在这个实验中，真实世界的数据是使用一个定制的硬件系统收集的，如图 2 所示，然而，获取带有注释的数据集需要大量时间和资源。为了解决这一挑战，本研究采用了一种“从模拟到现实”的策略，依赖于特定任务和数据高效的解决方案。通过使用 Blender 模拟解剖上合理的胴体识别和位置，并生成高保真度的真实掩码，我们的框架直接解决了这一特定数据集中的关键障碍：注释瓶颈和视觉复杂性。对竞争性深度学习模型的比较分析为该方法提高分割准确性和增强实际可行性的能力提供了有力支持。

A. 真实世界数据集收集

我们使用一个定制的图像采集系统构建了一个真实世界的家禽数据集，该系统设计用于捕捉多样化的场景。为了确保样本的一致性，所有家禽尸体均来自与同一供应商网络关联的本地超市，该网络遵循标准化的包装和冷链程序，这有助于减少尸体外观的变化。为了模拟各种真实世界的条件，家禽尸体被随机排列在不同的位置组合中，包括堆叠和紧密放置的配置。如图 2 所示，图像采集系统配备了可控的照明，以在图像捕获过程中保持一致的光照强度。家禽尸体的背景为不锈钢板，模拟了家禽生产中使用的不锈钢食品接触表面。总共收集了 300 张真实世界的图像，这些图像包含了不同尸体覆盖和捕捉角度的图像。这些图像构成了我们数据集的基线，为我们在不同条件下的分割模型的健壮性能评估提供了依据。每张图像都使用 LabelMe [41] 手动标注了像素级的实例分割蒙版。这 300 张标注的真实世界图像集合随后被划分为不同的训练集、验证集和测试集，以开发和评估我们的分割模型。更多的细节在 C 节中讨论。

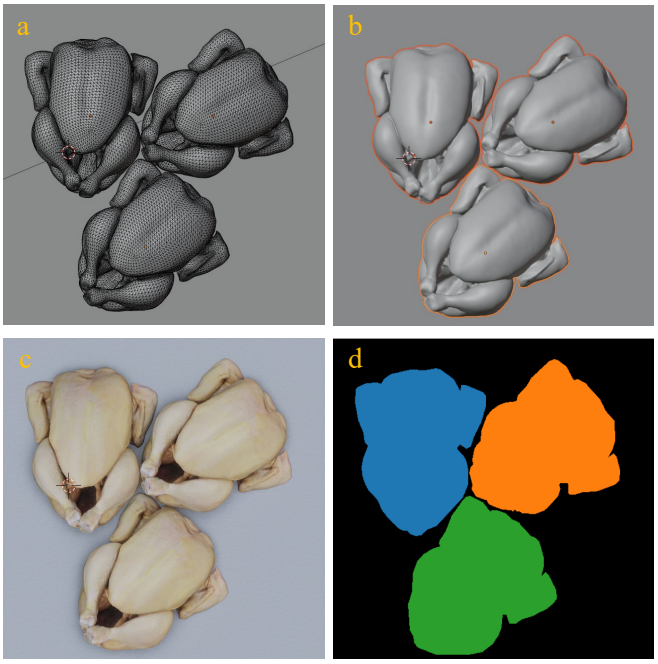

Fig. 3. 我们的合成数据生成工作流程概述。(a) 网格准备：高质量的 3D 鸡模型被导入 Blender 并布置在不重叠的布局中。在线框模式下检查网格分辨率和拓扑结构。(b) 几何细化：在实心视口阴影下调整表面光滑度、方向和模型边界，以确保在渲染之前的几何清晰。(c) RGB 渲染：应用真实感光照、纹理和材料，通过 Cycles 引擎生成逼真的合成图像。(d) 掩码生成：为每个实例分配唯一的 ID 以生成相应的分割掩码。这些自动生成的掩码用作训练实例分割模型的基准。

B. 合成数据集生成

Blender 是一款开源的 3D 创作软件，具有广泛的建模、渲染和动画功能 [42] 。通过使用 Blender 等工具，研究人员可以生成与现实场景非常相似的合成数据集 [18] [33] [19] 。其 Python API 可用于通过改变相机位置、光照条件和对象属性来自动生成合成数据集。在渲染过程中会自动生成真实标签，这使得研究人员可以快速构建具有准确注释的大规模合成数据集。该软件的基于物理的渲染引擎可以模拟逼真的光照、阴影和材质特性，使合成图像更具现实条件的代表性。

Blender 在合成数据生成方面已经在多种农业应用中展示了相当高的效率。Barth 等人利用 Blender 通过基于实验证实的形态测量对 3D 植物模型进行系统随机化，生成了一个包含 10,500 张合成辣椒（Capsicum annuum）图像的全面数据集。这个数据集专门设计用于复制温室环境中的视觉条件，并促进精准农业中的实例分割任务 [43] 。Blender 生成的合成数据能够通过有效捕捉复杂的真实世界变化，显著提升模型在食品加工应用中的性能，这一点已被多项研究强调。例如，Ummadisingu 等人展示了这一点在辅助配餐机器人系统的食品分割中的增强效果 [44] ，而 Jonker 等人则展示了其在自动化机器人系统处理鸡肉片中的优势 [13] 。

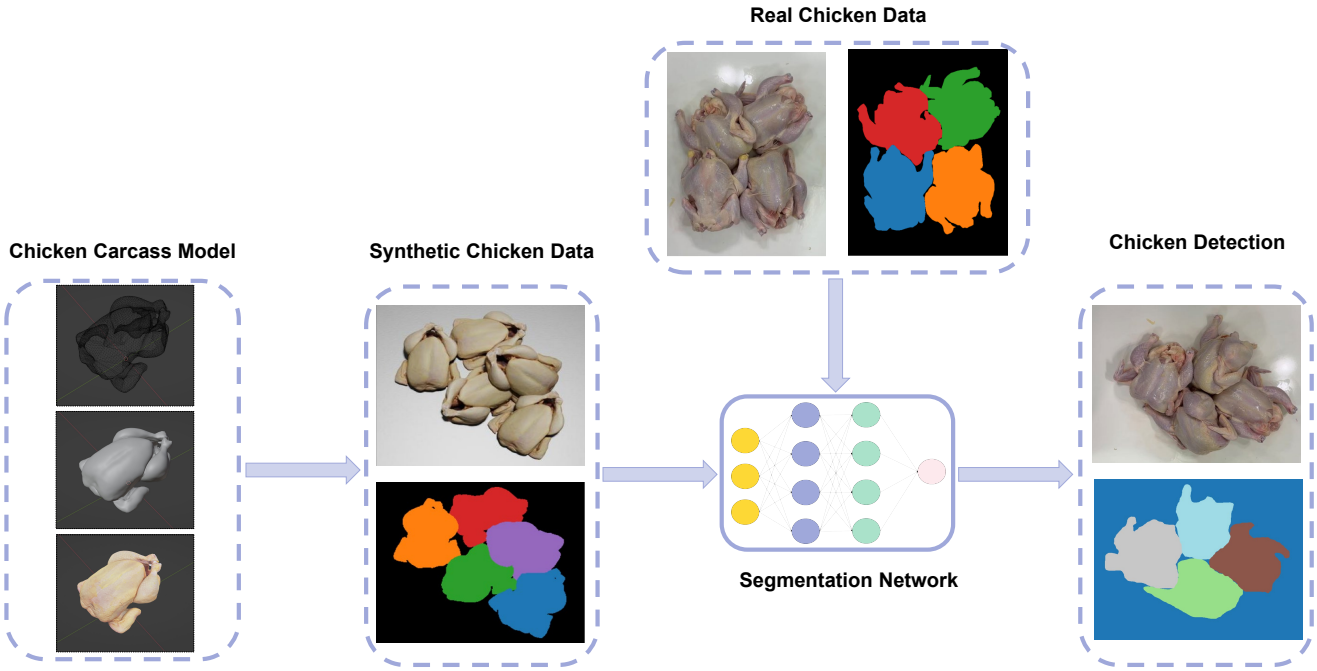

Fig. 4. 合成数据生成和模型训练的概述。该过程始于一个高保真度的 3D 鸡胴体模型，该模型用于生成大规模的合成鸡数据集。然后，将这些合成数据与我们的真实训练数据以不同数量结合，形成一系列不同的混合数据集。每个数据集随后用于训练和评估深度学习分割网络（包括具有 ResNet-50/101 骨干网络的 Mask R-CNN、Mask2Former 和 YOLOv11-seg）。这种全面的方法允许系统性地研究合成数据在不同架构中对性能的影响。然后，训练好的网络可以在新的、未见过的真实世界图像上执行稳定的实例分割（鸡检测），即使在杂乱的场景中也能准确识别和描绘每个单独的胴体。

尽管先前的应用已经展示了 Blender 在农业背景下的能力，但家禽加工提出了独特的挑战，这些挑战在很大程度上尚未得到解决。与前述涉及静态植物模型或食品项目的应用不同，家禽尸体处理具有来自每个尸体的复杂解剖变化。我们的方法利用 Blender 的功能来开发全面的训练数据集，这些数据集捕捉了工业家禽加工中固有的复杂变化，包括在重新悬挂操作中通常遇到的多样尸体方向和解剖重叠的真实模拟 [45] 。这代表着一个超越现有应用的重大进步，因为它解决了通常在食品加工环境中遇到的生物产品动态和形态复杂的特性。

在家禽加工厂中，鸡的尸体通常在检查和处理过程中被摊开或堆叠在平面上。我们的研究着重于解决家禽尸体实例分割中特有的挑战，例如重叠和紧密排列，使得视觉模型难以区分个别尸体。这些复杂的空间排列对人工注释者也构成了巨大的挑战，使得人工标注既耗时又容易出错，这突显了 Blender 方法的优势。

在虚拟仿真中，尸体的排列方式从孤立到紧密聚集或重叠。如图 3 所示，每个尸体在掩码标注中被视为单独的输入，而背景保持均匀以避免不必要的复杂性。选择顶视图角度是为了反映人类操作者在生产环境中查看尸体的实际应用场景。使用这种方法，共生成了 1000 张合成图像，每张图像包含 RGB 图像和对应的 COCO 格式实例掩码。这个合成数据集提供了坚实的训练数据，确保模型能够有效地处理重叠和紧密排列的尸体。

本文实现了 Mask R-CNN [46] 、Mask2Former [47] 和 YOLOv11-seg [48] 作为我们分析的基线实例分割模型。系统地评估了每个模型以确定在工业家禽加工应用中最佳的性能特征。以下是一些实现细节：

对于 Mask R-CNN，模型使用了两种不同的骨干结构（ResNet-50-FPN 和 ResNet-101-FPN）进行部署，以评估网络深度对分割性能的影响。这两种骨干网络都整合了特征金字塔网络（FPN），以分层方式结合多尺度特征，增强在不同分辨率层次上的语义表示。区域提议网络（RPN）通过滑动窗口分析生成候选对象区域，提出并改进锚点以定位潜在的胴体实例。这些优化后的提议随后推动了预测头的三重任务：分类、边界框回归和掩膜生成，从而为禽类胴体产生精确的每个实例的分割掩膜。对于 Mask2Former，模型采用了 Swin Transformer Tiny（Swin-T）骨干，它提供了一种具有移动窗口的层次化表示以实现高效的自注意力计算。此模型用变压器架构替代了传统的卷积骨干，提供了在保持计算效率的同时捕获全局上下文的增强能力。对于 YOLO 模型，采用了最新的 YOLO 架构（YOLOv11）以及其分割能力，它提供了一种单阶段检测方法，与传统双阶段检测器相比，具有更好的速度-精度权衡。

TABLE I
利用真实和合成数据集训练的鸡检测模型。

Setting	Real Train	Real Val	Real Test	Synthetic Train
Real_Baseline	60	60	180	0
Rea+Syn-250	60	60	180	250
Rea+Syn-500	60	60	180	500
Rea+Syn-750	60	60	180	750
Rea+Syn-1000	60	60	180	1000

所有模型的训练都使用一致的超参数以确保公平比较：所有模型训练 200 个 epoch，批量大小为 8，图像分辨率为 640×640 像素，利用阿肯色州高性能计算中心（AHPCC）

提供的 NVIDIA A100 GPU（40GB 内存），以及包括随机翻转（概率 =0.5）、随机裁剪（512×512）和归一化的数据增强。对于 Mask R-CNN 和 Mask2Former，使用逐步学习率调度方案，在 500 次迭代中进行线性预热，并在第 30、60 和 100 个 epoch 时进行学习率衰减。Mask R-CNN 使用带动量（0.9）和权重衰减（0.0001）的 SGD 进行优化，而 Mask2Former 则采用 AdamW 优化器，根据不同组件的需要在变压器架构 [49] [47] 中设置参数化学习率。对于 YOLOv11，默认的优化器配置为初始学习率 0.01，动量 0.937，权重衰减 0.0005。

实验设计采用了一种系统的方法来评估合成数据增强对模型性能的影响。300 幅真实世界图像被随机分配到训练集（20 %）、验证集（20 %）和测试集（60 %），使用固定的随机种子以确保可重复性。我们在训练集中分配较少的数据，而在测试集中分配更多的数据，以证明使用合成数据进行模型训练的有效性。如表 I 所示，为了评估合成数据的贡献，设置了五种真实图像和合成数据的训练组合。基线模型仅在 60 幅真实世界图像上进行训练，另外四种设置在相同数量的真实数据基础上分别补充 250、500、750 和 1000 幅合成图像。验证集（60 张图像）和测试集（180 张图像）保持不变，均仅包含真实的家禽加工图像，确保公平评估合成数据增强对模型在实际工业条件下泛化的影响。

为了评估家禽胴体分割模型的性能，使用了 PyTorch 中提供的 COCO 数据集评估插件，这是实例分割任务的标准。每个模型处理 RGB 图像并生成实例预测，包括分割掩码、置信值和单个家禽胴体的定位边界。性能评估主要集中在使用掩码重叠分析的实例分割质量上。我们利用预测与真实注释之间的空间重合率来进行计算，计算方式如下：

$$\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of union}}. \tag{1}$$

用于评估的主要指标是平均精度均值（mAP），它提供了一个全面衡量分割准确度的标准。分析框架报告了三个主要指标变体：AP $_{50}$（交并比阈值为 0.5 时的平均精度）、AP $_{75}$（交并比阈值为 0.75 时的平均精度）以及 AP（在所有交并比阈值从 0.5 到 0.95 之间的平均精度）。这些指标分别针对实例掩码和边界框进行计算，以独立评估分割和定位性能。

II. 结果

表 II 总结了四个实例分割框架的定量性能，包括使用 ResNet-50 (R50) 和 ResNet-101 (R101) 主干的 MaskR-CNN、Mask2Former 以及在五种不同数据集设置下训练的 YOLOv11-Seg。仅包含真实数据的配置用 Real_Baseline 表示，而 Real+Syn- N（ $N \in \{250, 500, 750, 1000\}$ ）表示在同样的真实数据集中添加了合成图像。所有模型都在测试集中的相同 180 张真实世界图像上进行了评估，使用 COCO 风格的 mAP，该 mAP 在 IoU 阈值由 0.50 到 0.95 的区间内计算的平均精度。AP $_{50}$ 和 AP $_{75}$ 分别用于捕获低和高 IoU 表现。

A. 总体趋势

对于所有模型，添加合成图像始终能提升检测和分割指标。然而，这种改进的程度和最佳合成数据量依赖于模型架构以及具体的评估指标。例如，Mask2Former 模型在基准较低的情况下，合成数据带来了显著的相对提高。更为稳健的基准模型如 YOLOv11-seg 也得益于合成数据，尤其是在大量合成图像的情况下，这些模型的高性能得到了进一步增强。并不是总能随着更多合成数据的增加而线性提升；对于某些模型，在添加一定数量的合成数据之后，出现收益递减甚至性能略微下降的现象。比如，Mask R-CNN R50 在超过 750 张合成图像后。

B. 合成与真实比例的影响

真实训练数据集设定为 60 张图像。实验通过添加 0、250、500、750 或 1000 张合成图像来探索合成与真实数据的比例，分别对应于 0:60（基线）、约 4.17:1（Syn-250）、8.33:1（Syn-500）、12.5:1（Syn-750）和 16.67:1（Syn-1000）的比例。对于 Mask R-CNN（R50），大约 12.5:1 的比例在 bbox_mAP（0.795）和 segm_mAP（0.765）方面表现最佳。对于 Mask R-CNN（R101），在大约 16.67:1 的最大比例下达到了最高的 bbox_mAP（0.805）和 segm_mAP（0.766），这表明其比 R50 模型能更有效地利用更多的合成数据。Mask2Former 在最高 8.33:1 的比例中显示出显著提升，获得了 bbox_mAP 为 0.652 和 segm_mAP 为 0.737。进一步增加到 12.5:1 获得了最高的 bbox_mAP（0.672），而 segm_mAP 在 16.67:1 时达到相似水平或峰值为 0.737。YOLOv11-Seg 在不断增加的合成与真实比例下一直受益。最佳 bbox_mAP（0.8570）出现在 12.5:1 的比例下，而最高的 segm_mAP（0.8325）以及其他一些关键指标（bbox_mAP_50、segm_mAP_50、segm_mAP_75）在 16.67:1 的比例下达到峰值。

这些结果表明，不同的模型能力和结构对合成数据比例的反应不同。当添加更多合成数据时，像 R50 这样的轻量级模型停止显著改进，但更深层次或基于变压器的结构继续从更大的合成数据集中受益。

C. 特定模型性能分析

1) Mask R-CNN（R50 和 R101）以及骨干网络比较：

- Mask R-CNN (R50)：基线 *bbox_mAP* 为 0.740，*segm_mAP* 为 0.746。添加合成数据通常会提高性能，其中 “Real+Syn-750” 设置产生了最佳结果（*bbox_mAP* 0.795，*segm_mAP* 0.765）。但当合成图像增加到 1000 时，性能略有下降。
- Mask R-CNN (R101)：基线 *bbox_mAP* (0.750) 比 R50 略高，但 *segm_mAP* (0.733) 较低。R101 骨干网络通常比 R50 从更多的合成数据中受益，其在使用 1000 张合成图像时达到其峰值 *bbox_mAP* (0.805) 和 *segm_mAP* (0.766)。
- 骨干网络比较（R50 vs. R101）：较深的 R101 骨干网络相比 R50 骨干网络，在最大成绩上略有提升（使用 1000 张合成图像时分别达到 0.805 和 0.766，而 R50 的最大成绩是使用 750 张合成图像时达到 0.795 *bbox_mAP* 和 0.765 *segm_mAP* ）。R101 似乎也比 R50 更好地利用了大量的合成数据（在 1000 张图像时达到峰值），而 R50 在 750 张图像时达到峰值。然而，R101 在分割任务中的基准性能仅略低于 R50。在各自最佳合成数据设置下，R101 的表现略优于 R50。

D. 边界框检测和实例分割质量

这些度量改进的定性影响在图 5 中可视化。顶部一行直接比较了三种架构的最佳配置。虽然所有模型性能都很好，

TABLE II
使用不同数量的合成数据训练的实例分割模型的性能

Model	Setting	bbox_mAP	bbox_mAP_50	bbox_mAP_75	segm_mAP	segm_mAP_50	segm_mAP_75
Mask R-CNN (R50)	Real_Baseline	0.740	0.963	0.866	0.746	0.965	0.863
	Real+Syn-250	0.761	0.962	0.861	0.736	0.954	0.838
	Real+Syn-500	0.784	0.959	0.883	0.755	0.959	0.865
	Real+Syn-750	0.795	0.962	0.896	0.765	0.961	0.877
	Real+Syn-1000	0.793	0.960	0.886	0.759	0.950	0.871
Mask R-CNN (R101)	Real_Baseline	0.750	0.960	0.869	0.733	0.956	0.836
	Real+Syn-250	0.795	0.962	0.883	0.751	0.952	0.856
	Real+Syn-500	0.773	0.961	0.884	0.751	0.951	0.857
	Real+Syn-750	0.776	0.964	0.874	0.753	0.963	0.860
	Real+Syn-1000	0.805	0.962	0.881	0.766	0.953	0.875
Mask2Former	Real_Baseline	0.434	0.712	0.437	0.523	0.795	0.597
	Real+Syn-250	0.530	0.776	0.548	0.669	0.899	0.739
	Real+Syn-500	0.652	0.896	0.694	0.737	0.933	0.835
	Real+Syn-750	0.672	0.896	0.715	0.726	0.925	0.809
	Real+Syn-1000	0.661	0.880	0.694	0.737	0.927	0.816
YOLOv11-seg	Real_Baseline	0.835	0.969	0.890	0.784	0.959	0.846
	Real+Syn-250	0.845	0.965	0.891	0.817	0.958	0.874
	Real+Syn-500	0.842	0.966	0.888	0.817	0.964	0.883
	Real+Syn-750	0.857	0.966	0.906	0.828	0.965	0.891
	Real+Syn-1000	0.852	0.972	0.896	0.833	0.971	0.898

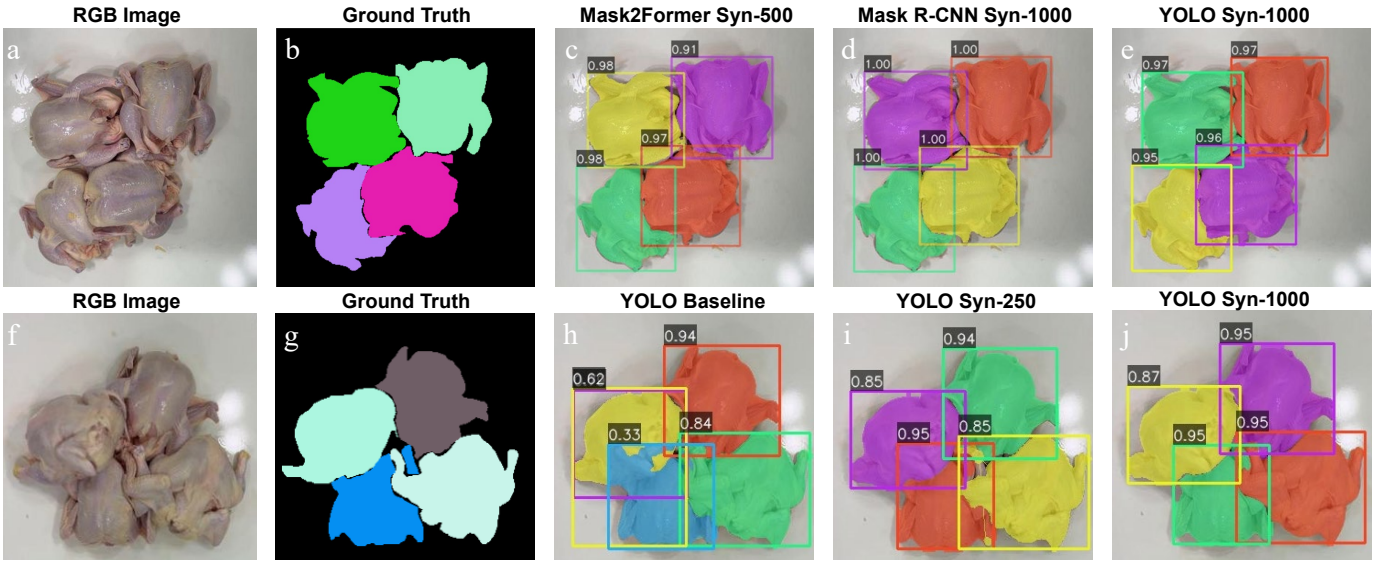

Fig. 5. 比较每个模型（采用 ResNet-101 骨干网络的 Mask-RCNN、Mask2Former 和 YOLOv11-seg）在最佳训练比例下获得的边界框和实例分割结果，以及 YOLOv11-seg 逐渐增大的合成集的结果。对两张真实鸡胴体图像的示例预测。(a, f) 输入 RGB 图像。(b, g) 真实标签的实例蒙版。(c–e) 三种基线架构对原图像 (a) 的预测，每种架构均显示出其最高 COCO mAP;(0.50 : 0.95) 的训练配置：(c) Mask2Former 使用真实 + Syn-500 训练，(d) Mask R-CNN (ResNet-101) 使用真实 + Syn-1000 训练，以及 (e) YOLOv11-seg 使用真实 + Syn-1000 训练。(h–j) 来自 YOLOv11-seg 模型对图像 (f) 的预测，随着合成与真实比例的增加进行训练：(h) 仅真实的基线，(i) 真实 + Syn-250，以及 (j) 真实 + Syn-1000。掩码按实例进行了颜色编码，并覆盖了相关联的边界框和置信度得分。

但细节上的显著差异显而易见。例如，聚焦于左下角的尸体，YOLOv11-seg (e) 的掩膜轮廓对于真实边缘的契合度最紧密，而 Mask R-CNN (R101) (d) 和 Mask2Former (c) 的掩膜在该区域表现出一定程度的欠分割，未能捕捉到物体的完整细节。

图中底部一行的 5 提供了合成数据价值的清晰定性说明。尽管基线模型 (h) 检测到所有实例，但其分割质量和置信分数明显不足，特别是对于两个靠近的较低尸体。通过添加合成数据，该性能得到了显著改善。当增加到 1000 张合成图像 (j) 时，这两个较低尸体的掩码描绘变得更加精确，且它们的置信分数从 (h) 中的 0.33 和 0.84 剧增至在 (j) 中均为 0.95。这直观地表明，合成数据不仅提升了模型的检测置信度，同时也有效地优化了复杂重叠场景中的分割边界。

通常来说，边界框检测（*bbox_mAP* ）与实例分割（*segm_mAP* ）相似，但存在一些差异：

- 对于 Mask R-CNN (R50), 最佳合成数据量 (Syn-750) 在 *bbox_mAP* 和 *segm_mAP* 中是相同的。
- 对于 Mask R-CNN (R101)，1000 张合成图像在 *bbox_mAP* 和 *segm_mAP* 都产生了最佳效果。
- 对于 Mask2Former，*bbox_mAP* 在 750 张图像时达

到了峰值，而 $segm\_mAP$ 在 500 和 1000 张合成图像时达到了峰值，这表明分割的最佳点或平台略有不同。

- 对于 YOLOv11-seg，$bbox\_mAP$ 在 750 张合成图像时达到峰值，而 $segm\_mAP$（及相关的 $segm\_mAP_{75}$）持续提升至 1000 张合成图像，表明分割质量需要更精确的定位，可能会从较大的合成数据集和该模型中获得更多益处。

在所有模型和设置中使用合成数据时，mAP_{50} 评分（包括 bbox 和 segm）通常都非常高，对于 Mask R-CNN 和 YOLOv11-seg，通常超过 0.95。这表明所有模型在检测和分割较容易的实例（IoU 阈值为 0.5）方面都变得非常熟练。这些来自合成数据的改进在更严格的 mAP_{75} 度量和总体 mAP（0.50:0.95）中通常更加明显，突出了合成数据帮助提高模型精确定位能力。例如，YOLOv11-seg 的 $segm\_mAP_{75}$ 从 0.8461（基线）增加了 5.72 % 至 0.8975（使用 1000 张合成图像）。

E.

结果总结

添加合成数据对所有测试的实例分割模型普遍有利，提升了边界框检测和实例分割性能。YOLOv11-seg 表现出最高的整体性能，能够有效利用多达 1000 张合成图像以取得顶尖成绩，特别是在分割指标上。这种定量优势在定性分析中也有所体现，其中 YOLOv11-seg 相比其他架构始终生成更精确的分割掩码。Mask R-CNN 模型也显示出一致的改进，其中 R101 骨干网络比 R50 从更大量的合成数据中获益更多。虽然 Mask2Former 的起点较低，但其表现出显著的相对增益，这突显了合成数据对该架构的价值。值得注意的是，对于某些模型而言，找到了合成数据与真实数据的最佳平衡点，超过此点后性能提升趋于稳定，或在某些个别情况下略有下降。合成数据的理想数量及其带来的性能提升取决于模型，且这些改进通常在更高的 IoU 阈值下更为明显。视觉证据强烈支持这一点，定性地展示了合成数据如何将低置信度、不精确的检测转变为高置信度、精确的分割，特别是在存在对象重叠的复杂情况下。

在本文中，当通过合成图像增强真实训练数据时，各种实例分割模型的性能改善可归因于多个因素，这些因素与先前发表的文献中的发现相符，并在某些情况下对这些发现进行了细化。首先，引入合成数据可能提高了训练样本的多样性。这使得模型能够接触到比我们有限的真实数据集（60 张图像）中更广泛的鸡尸体的方向、遮挡和背景变化。这种增强的多样性通常有助于模型更好地泛化到未见过的数据，并减少对小型真实数据集的特定特征的过拟合。例如，合成数据可以提供大量具有挑战性的场景实例，如在密集簇中或在异常灯光下的鸡尸体，这在有限的现实世界图像中可能很少或没有。这一原则在其他研究中得到了很好的支持。例如，Richter 等人的早期工作展示了使用游戏引擎的合成数据进行语义分割的性能提升 [50]，Tremblay 等人展示了使用领域随机化进行目标检测的益处 [51]。最近，Vanherle 等人开发了一个基于高斯斑点的流水线，专门用于生成高质量、上下文感知和多样化的合成数据用于实例分割，强调其在创建多样化训练实例和缩小域间差距方面的作用 [52]。此外，Eli 等人关于面部解析的研究 [53] 和 Jordan 等人在零样本分类中的研究强调了合成数据集内容的多样性对于模型泛化的关键性甚至超过达到完美照片真实感，这一观念与我们的发现一致，其中多样化的合成鸡尸体数据证明是有益的 [54]。

在我们的实验中，YOLOv11-seg 模型表现优异，一直达到最高的准确度指标，这可以归因于其先进的架构 [55]，该架构有效地将高效的目标检测与分割功能相结合。其预训练步骤可能也有助于其强大的特征提取。这一观察结果与 YOLO 系列的一般趋势一致，即较新的迭代版本在检测和分割任务中不断突破性能界限，这一点在关于 YOLO 架构演变的各种评论和基准研究中都有记录 [56]。

此外，拥有 R101 骨干网络的 Mask R-CNN 模型相对于其 R50 对应版本能够更有效地利用更大比例的合成数据，这表明骨干网络的容量起着重要作用。像 R101 这样更深的网络具有更高的模型容量，使其能够从更大且更具多样性的数据集中学习更复杂的特征和模式，例如那些通过合成图像增强的数据集，而不会很快饱和。这与深度学习中的基本概念相符。更近期的由 Yu 等人在远程感知图像分类领域的领域适应方面的研究也发现，当适应新领域时，ResNet-101 比 ResNet-50 获得了更高的平均准确性提升（5.22 % 对 4.99 %），这表明更深的架构确实能够更好地利用来自不同数据分布的信息。

在使用增强数据集进行训练时，Mask2Former 表现出显著的相对性能提升，这一点值得特别讨论。这个发现可能与其基于 transformer 的架构有关。transformer 模型通常被描述为“数据贪婪型的”，它们的注意力机制尤其可能从合成数据提供的增加的多样性和数量的训练实例中受益。这使得它们可以学习更为稳健和具有普遍性的表征。这一观察也在计算机视觉领域对 transformer 架构的更广泛研究中得到了呼应，例如 Dosovitskiy 等人提出的 Vision Transformer (ViT)，其表明 transformer 在特别是预训练于大规模数据集并随后在多样化数据上进行微调时可以达到显著的性能 [57]。

在测试的几个模型中，确定了一个合适的合成数据与真实数据的比例平衡，超过这个比例后分割性能会达到瓶颈，或者在一些孤立的情况下略微下降。这一观察表明了一个关键的权衡。虽然合成数据提供了宝贵的多样性，但过度依赖它可能适得其反，尤其是在合成图像与真实图像之间存在明显“域间隙”的情况下。在这种情况下，会出现两个不同的问题。首先，模型可能开始对合成数据过拟合，学会识别渲染伪影或其他仅在人工图像中存在而不在真实场景中出现的模式。其次，大量的合成例子可能会削弱训练过程中有限的真实数据的影响，导致模型忽视或“遗忘”那些仅在真实图像中存在的微妙但重要的特征。这种细微的行为，与我们特定的模型如 YOLOv11-seg 和 Mask2Former 在某个测试点上表现出的持续性优势形成对比，而某些研究可能报告合成数据收益更快地减少。近期研究支持了这一观察。例如，Regina 等人在其无人机检测研究中发现，通过少量份额（5-10 % ）的真实数据对预训练于合成数据的模型进行微调，能够显著提升性能，从而突出了真实数据作为“锚”的重要性 [58]。Chang 等人发现，在多目标跟踪中，60-80 % 的真实数据可以被合成数据替代而不损失性能，强调了灵活数据生成器在缩小域间隙方面的作用 [59]。相反，过度依赖合成数据，特别是当它与目标领域不完全对齐或缺乏足够的多样性时，可能对模型性能有害。李等人明确指出，直接使用来自扩散模型的合成数据可能会由于特征分布的不一致而降低性能 [60]，而 Tremblay 等人提到，低质量的合成样本可能会阻碍学习 [51]。Ilia 等人的语言模型研究也警告称，当纯粹依赖合成数据递归训练时会出现“模型崩溃”，建议在与真实数据混合时，合成数据的比例应有上限，以避免性能下降 [61]。这表明，最佳数据组合以及对合成数据的容忍度高度依赖于模型的

具体情况和任务，受到模型架构、容量、合成数据的质量和多样性以及合成与真实领域差距性质等因素的影响。

鉴于在禽类领域数据获取和标注方面存在着广泛记录的挑战，这些特定于模型的洞察力对于利用合成数据尤为重要。我们的研究结果为这些挑战提供了一种潜在的缓解策略，符合合成数据在相关农业领域（如精准农业和食品加工）中的成功应用。通过展示合成数据在鸡实例分割中的有效使用，我们的工作为支持合成数据在推动农业领域人工智能应用方面的作用提供了越来越多的证据，这些领域往往面临数据匮乏的瓶颈问题。

III. 限制与未来工作

尽管结果很有前景，本研究有几个需要确认的局限性。首先，我们的真实数据集（用于训练的 60 张图像）规模相当小。虽然这强调了在数据稀缺情境中合成数据的实用性，它也可能夸大合成增强的影响，因为仅使用真实数据的基线性能可能是次优的。其次，尽管我们的合成数据被设计成多样化，但其现实性和特定特征仍是对现实世界变量的近似。生成过程可能存在固有偏差，或者可能缺乏真实图像中存在的某些细微差异，潜在地创造出一个领域差距，从而限制在真正未见过的真实世界数据上的性能。第三，我们的研究局限于一组特定的实例分割模型。其他架构，包括不同的 YOLO 变体或新兴的分割模型，可能对合成数据增强有不同的反应。最后，我们的评估是在与真实训练数据相似环境中得出的特定测试集上进行的。对训练模型在完全不同植物环境中的泛化能力仍需进行全面测试。

从我们的研究发现和局限性出发，有几条未来研究的方向：1）高级合成数据生成：研究更复杂的合成数据生成技术，例如生成对抗网络（GANs）、扩散模型或基于物理的模拟，这些技术可以产生更逼真和多样化的训练样本，可能减少领域差距。2）现实世界的鲁棒性测试：在多样化的真实世界鸡肉加工生产线环境中广泛测试训练模型，以评估它们的泛化能力和实际部署准备情况。3）探索合成与真实数据的比例：对不同的合成与真实数据比例的影响进行更细致的分析，潜在开发自适应策略以确定特定模型架构和数据集特征的最佳混合。4）成本效益分析：进行正式的成本效益分析，比较生成和策划高质量合成数据所需的资源与获取和注释更多真实数据的成本和努力。5）领域适应技术：探索无监督或半监督的领域适应技术，以进一步缩小合成数据域与真实数据域之间的差距，旨在提高当真实标记数据极其稀缺时的模型性能。

IV.

致谢 这项工作得到了美国农业部（USDA）国家食品与农业研究所（NIFA）通过国家机器人计划（NRI）3.0 与国家科学基金会（NSF）合作的奖项支持，奖号为 2023-67021-39072、2023-67022-39074 和 2023-67022-39075。该研究还得到了阿肯色州高性能计算中心的支持，该中心通过多个国家科学基金会的拨款和阿肯色州经济发展委员会的资助。

References

[1] I. De Medeiros Esper, P. J. From, and A. Mason, “Robotisation and intelligent systems in abattoirs,” *Trends in Food Science & Technology*, vol. 108, pp. 214–222, Feb. 2021.

[2] G. Ren, T. Lin, Y. Ying, G. Chowdhary, and K. Ting, “Agricultural robotics research applicable to poultry production: A review,” *Computers and Electronics in Agriculture*, vol. 169, p. 105216, Feb. 2020.

[3] N. F. Dias, A. S. Tirloni, D. Cunha dos Reis, and A. R. P. Moro, “The effect of different work-rest schedules on ergonomic risk in poultry slaughterhouse workers,” *Work*, vol. 69, no. 1, pp. 215–223, 2021.

[4] L. Shen, J. Su, R. He, L. Song, R. Huang, Y. Fang, Y. Song, and B. Su, “Real-time tracking and counting of grape clusters in the field based on channel pruning with yolov5s,” *Computers and Electronics in Agriculture*, vol. 206, p. 107662, 2023.

[5] R. Hu, W.-H. Su, J.-L. Li, and Y. Peng, “Real-time lettuce-weed localization and weed severity classification based on lightweight yolo convolutional neural networks for intelligent intra-row weed control,” *Computers and Electronics in Agriculture*, vol. 226, p. 109404, 2024.

[6] Z. Liu, R. R. D. Abeyrathna, R. M. Sampurno, V. M. Nakaguchi, and T. Ahamed, “Faster-yolo-ap: A lightweight apple detection algorithm based on improved yolov8 with a new efficient pdwconv in orchard,” *Computers and Electronics in Agriculture*, vol. 223, p. 109118, 2024.

[7] B. Paneru, R. Bist, X. Yang, and L. Chai, “Tracking dustbathing behavior of cage-free laying hens with machine vision technologies,” *Poultry Science*, vol. 103, no. 12, p. 104289, 2024.

[8] D. Wu, Y. Ying, M. Zhou, J. Pan, and D. Cui, “Yolo-claw: A fast and accurate method for chicken claw detection,” *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108919, 2024.

[9] D. Wang and D. He, “Fusion of Mask RCNN and attention mechanism for instance segmentation of apples under complex background,” *Computers and Electronics in Agriculture*, vol. 196, p. 106864, 2022.

[10] Y. Li, X. Xu, and C. Yuan, “Enhanced mask r-cnn for chinese food image detection,” *Mathematical Problems in Engineering*, vol. 2020, no. 1, p. 6253827, 2020.

[11] S. P. Mohanty, G. Singhal, E. A. Scuccimarra, D. Kebaili, H. Héritier, V. Boulanger, and M. Salathé, “The food recognition benchmark: Using deep learning to recognize food in images,” *Frontiers in Nutrition*, vol. 9, p. 875143, 2022.

[12] M. Macías-Macías, H. Sánchez-Santamaria, C. J. Garcia Orellana, H. M. González-Velasco, R. Gallardo-Caballero, and A. García-Manso, “Mask r-cnn for quality control of table olives,” *Multimedia Tools and Applications*, vol. 82, no. 14, pp. 21 657–21 671, 2023.

[13] M. Jonker, “Robotic Bin-Picking Pipeline for Chicken Fillets with Deep Learning-Based Instance Segmentation using Synthetic Data,” 2023.

[14] I. Attri, L. K. Awasthi, T. P. Sharma, and P. Rathee, “A review of deep learning techniques used in agriculture,” *Ecological Informatics*, p. 102217, 2023.

[15] D. Lewy and J. Mańdziuk, “An overview of mixing augmentation methods and augmentation strategies,” *Artificial Intelligence Review*, vol. 56, no. 3, pp. 2111–2169, 2023.

[16] B. Hüttenrauch, “Limitations of data augmentation and outlook,” in *Targeting Using Augmented Data in Database Marketing: Decision Factors for Evaluating External Sources*. Springer, 2016, pp. 279–290.

[17] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, “Learning data augmentation strategies for object detection,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer, 2020, pp. 566–583.

[18] S. I. Nikolenko, *Synthetic Data for Deep Learning*. Springer, 2021, vol. 174.

[19] A. Gaur, C. Liu, X. Lin, N. Karapetyan, and Y. Aloimonos, “Whale detection enhancement through synthetic satellite images,” in *OCEANS 2023-MTS/IEEE US Gulf Coast*. IEEE, 2023, pp. 1–7.

[20] B. Mildenhall, P. Srinivasan, M. Tancik, J. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” in *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, 2021, pp. 1–12.

[21] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3d gaussian splatting for real-time radiance field rendering,” in *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, 2023, pp. 1–14.

[22] K. Aghamohammadesmaeilketabforoosh, "Enhancing Strawberry Disease and Quality Detection: Integrating Vision Transformers with Blender-Enhanced Synthetic Data and SwinUNet Segmentation Techniques," 2024.

[23] A. Kar, A. Prakash, M.-Y. Liu, E. Cameracci, J. Yuan, M. Rusiniak, D. Acuna, A. Torralba, and S. Fidler, "Meta-sim: Learning to generate synthetic datasets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4551–4560.

[24] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.

[25] M. Lin, X. Wang, Y. Wang, S. Wang, F. Dai, P. Ding, C. Wang, Z. Zuo, N. Sang, S. Huang *et al.*, "Exploring the evolution of physics cognition in video generation: A survey," *arXiv preprint arXiv:2503.21765*, 2025.

[26] C. Gan, J. Schwartz, S. Alter, D. Mrowca, M. Schrimpf, J. Traer, J. De Freitas, J. Kubilius, A. Bhandwaldar, N. Haber *et al.*, "Threedworld: A platform for interactive multi-modal physical simulation," *arXiv preprint arXiv:2007.04954*, 2020.

[27] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[30] Y. Karabatis, X. Lin, N. J. Sanket, M. G. Lagoudakis, and Y. Aloimonos, "Detecting olives with synthetic or real data? olive the above," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 4242–4249.

[31] Y. Tian, G. Yang, Z. Wang, E. Li, and Z. Liang, "Detection of apple lesions in orchards based on deep learning methods of CycleGAN and YOLOV3-dense," *Journal of Sensors*, vol. 2019, no. 1, p. 7630926, 2019.

[32] M. Fawakherji, V. Suriani, D. Nardi, and D. D. Bloisi, "Shape and style GAN-based multispectral data augmentation for crop/weed segmentation in precision farming," *Crop Protection*, vol. 184, p. 106848, 2024.

[33] X. Lin, N. J. Sanket, N. Karapetyan, and Y. Aloimonos, "Oysternet: Enhanced oyster detection using simulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5170–5176.

[34] C. Wang and Z. Xiao, "Lychee surface defect detection based on deep convolutional neural networks with gan-based data augmentation," *Agronomy*, vol. 11, no. 8, p. 1500, 2021.

[35] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[37] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[38] X. Lin, V. Mange, A. Suresh, B. Neuberger, A. Palnitkar, B. Campbell, A. Williams, K. Baxevani, J. Mallette, A. Vera *et al.*, "Odyssee: Oyster detection yielded by sensor systems on edge electronics," *arXiv preprint arXiv:2409.07003*, 2024.

[39] B. Campbell, A. Williams, K. Baxevani, A. Campbell, R. Dhoke, R. E. Hudock, X. Lin, V. Mange, B. Neuberger, A. Suresh *et al.*, "Is ai currently capable of identifying wild oysters? a comparison of human annotators against the ai model, odyssee," *Frontiers in Robotics and AI*, vol. 12, p. 1587033, 2025.

[40] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[41] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1, pp. 157–173, 2008.

[42] M. Denninger, M. Sundermeyer, D. Winkelbauer, D. Olefir, T. Hodan, Y. Zidan, M. Elbadrawy, M. Knauer, H. Katam, and A. Lodhi, "Blenderproc: Reducing the reality gap with photorealistic rendering," in *16th Robotics: Science and Systems, RSS 2020, Workshops*, 2020.

[43] R. Barth, J. IJsselmuiden, J. Hemming, and E. J. Van Henten, "Data synthesis methods for semantic segmentation in agriculture: A Capsicum annuum dataset," *Computers and electronics in agriculture*, vol. 144, pp. 284–296, 2018.

[44] A. Ummadisingu, K. Takahashi, and N. Fukaya, "Cluttered Food Grasping with Adaptive Fingers and Synthetic-Data Trained Object Detection," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 8290–8297.

[45] P. Sohrabipour, C. K. R. Pallerla, A. Davar, S. Mahmoudi, P. Crandall, W. Shou, Y. She, and D. Wang, "Cost-effective active laser scanning system for depth-aware deep-learning-based instance segmentation in poultry processing," *AgriEngineering*, vol. 7, no. 3, p. 77, 2025.

[46] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[47] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[48] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov11: Robust, fast and efficient object detection," *arXiv preprint arXiv:2308.11562*, 2023.

[49] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[50] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 102–118.

[51] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 969–977.

[52] B. Vanherle, B. Zoomers, J. Put, F. Van Reeth, and N. Michiels, "Cut-and-splat: Leveraging gaussian splatting for synthetic data generation," *arXiv preprint arXiv:2504.08473*, 2025.

[53] E. Friedman, A. Lehr, A. Gruzdev, V. Loginov, M. Kogan, M. Rubin, and O. Zvitia, "Knowing the distance: Understanding the gap between synthetic and real data for face parsing," *arXiv preprint arXiv:2303.15219*, 2023.

[54] J. Shipard, A. Wiliem, K. N. Thanh, W. Xiang, and C. Fookes, "Boosting zero-shot classification with synthetic data diversity via stable diffusion," *arXiv preprint arXiv:2302.03298*, vol. 3, no. 5, 2023.

[55] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[56] R. Sapkota and M. Karkee, "Comparing yolov11 and yolov8 for instance segmentation of occluded and non-occluded immature green fruits in complex orchard environment," *arXiv preprint arXiv:2410.19869*, 2024.

[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[58] T. R. Dieter, A. Weinmann, S. Jäger, and E. Brucherseifer, "Quantifying the simulation–reality gap for deep learning-based drone detection," *Electronics*, vol. 12, no. 10, p. 2197, 2023.

[59] C.-J. Chang, D. Li, S. Moon, and M. Kapadia, "On the equivalency, substitutability, and flexibility of synthetic data," *arXiv preprint arXiv:2403.16244*, 2024.

[60] X. Li, X. Tan, Z. Chen, Z. Zhang, R. Zhang, R. Guo, G. Jiang, Y. Chen, Y. Qu, L. Ma *et al.*, "One-for-more: Continual diffusion model for anomaly detection," *arXiv preprint arXiv:2502.19848*, 2025.

[61] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "Ai models collapse when trained on recursively generated data," *Nature*, vol. 631, no. 8022, pp. 755–759, 2024.