解决方案简介

英特尔锐炫™ 显卡
Intel® Extension for PyTorch
Intel® Extension for TensorFlow
OpenVINO™ 工具套件

英特尔锐炫™ 显卡
赋能医学影像 AI 推理解决方案

概述

随着算法的进一步成熟、算力的提高以及数据的持续积累，人工智能（AI）在医学影像领域得到了广泛的应用。AI 医学影像应用可以有效提高医师诊疗效率与诊断精度，使医学影像的分析技术下沉，缩短患者就诊等待时间，为患者提供更佳的服务。但同时，医学影像 AI 应用也面临计算资源不容易获取、成本高，计算资源容易被单一厂商绑定，算力要求高、硬件无法满足医疗定制化要求等挑战。

为帮助用户应对上述挑战，英特尔推出了基于英特尔锐炫™ 显卡的边缘端 AI 推理解决方案。该方案可充分发挥英特尔锐炫™ 显卡在 AI 推理方面的性能潜力，并可通过 Intel® Extension for PyTorch（IPEX）、Intel® Extension for TensorFlow（ITEX）、OpenVINO™ 工具套件和英特尔® oneAPI 工具包等软件加速实现算法迁移，提升 AI 推理速度。该方案具备低 TCO、产品线丰富、兼容性高、医疗定制化硬件丰富等优势，能够助力合作伙伴快速构建卓越的医学影像 AI 推理方案，有效满足 AI 辅助诊疗等关键性能指标的要求。

背景：边缘 AI 赋能智慧医疗

AI 在医疗健康领域的应用非常广泛，从医学影像、辅助诊断、疾病预测、健康管理，到药物研发等诸多环节，都可发挥关键作用。伴随着深度学习技术的进步，AI 能够帮助应用从海量数据中自动归纳出相关特征，而无需像传统模式根据领域特定知识手工去发现和设计特征。这使得用户能够更加快速地训练出高质量的 AI 医疗模型，同时更加灵活地应对不同场景的辅助诊断需求。

随着 AI 在医疗行业的持续渗透，边缘计算得到了快速发展。借助融合了网络、计算、存储和应用，且在数据源附近部署的边缘终端，边缘人工智能可将 AI 工作流的推理部分从云或数据中心转移到就近部署的边缘计算终端，从而降低时延，节约网络带宽，同时满足隐私和安全等方面的要求。得益于此，边缘 AI 应用近年来在各个医疗领域得到了广泛部署。边缘 AI 应用能够为专业医生提供参考建议，提高医生的病例处理量，同时提高影像分析的准确度，缩短诊断结果报告时间。这对于增强基层医疗机构的诊疗能力，推动疾病的防治有着重要意义。

要推动边缘 AI 系统在医疗行业的落地，需要提供强大的边缘算力，以满足 AI 模型推理对于算力的苛刻需求。同时，医疗行业的边缘 AI 系统还需要化解稳定性、经济性等方面的诸多挑战：

- **计算资源不容易获取、建设成本高：**为了获得更高的灵活性、成本效益，并避免被单一厂商绑定，用户通常希望能够拥有更为广泛的推理算力选项。
- **算力要求高：**医学 AI 应用依赖于深度学习模型的推理，随着模型复杂度的提升以及数据量的增长，系统推理能力面临巨大的考验。此外，由于用户希望能够承载更高的用户并发量，并满足目标应用的检测时间要求，算力挑战进一步增长。
- **需要更适用医疗场景的定制硬件：**伴随着医学影像 AI 应用临床化的发展趋势，AI 推理的部分日渐需要转移到边缘端进行，从而对边缘终端设备也提出了更高的要求。医疗边缘 AI 终端需要面临低噪音、高稳定性、防水、抗菌等苛刻的医疗场景需求，并经过医疗专业认证。

基于英特尔锐炫™ 显卡的医学影像 AI 推理解决方案参考设计

基于在 GPU 等领域的长期技术积累与技术创新，以及活跃的 GPU 生态，英特尔推出了英特尔锐炫™ 系列显卡。该系列显卡具备强大的算力、丰富的软件功能以及可支持算法兼容、移植的生态体系，能够在边缘端加速 AI 应用在千行百业的落地。在医疗行业，英特尔提供了基于英特尔锐炫™ 显卡的医学影像 AI 推理解决方案参考设计，可助力独立软件开发商（ISV）、原始设备制造商（OEM）等伙伴开发出高性能、高性价比、高灵活性的边缘医学影像 AI 推理方案。

该解决方案的参考架构如图 1 所示，其底层硬件为英特尔® CPU 与英特尔® GPU，并能通过 Intel® Extension for PyTorch、Intel® Extension for TensorFlow、OpenVINO™ 工具套件与英特尔® oneAPI 工具包实现 AI 算法的迁移与算法性能的优化，支持医疗行业 AI 算法的高效运行。

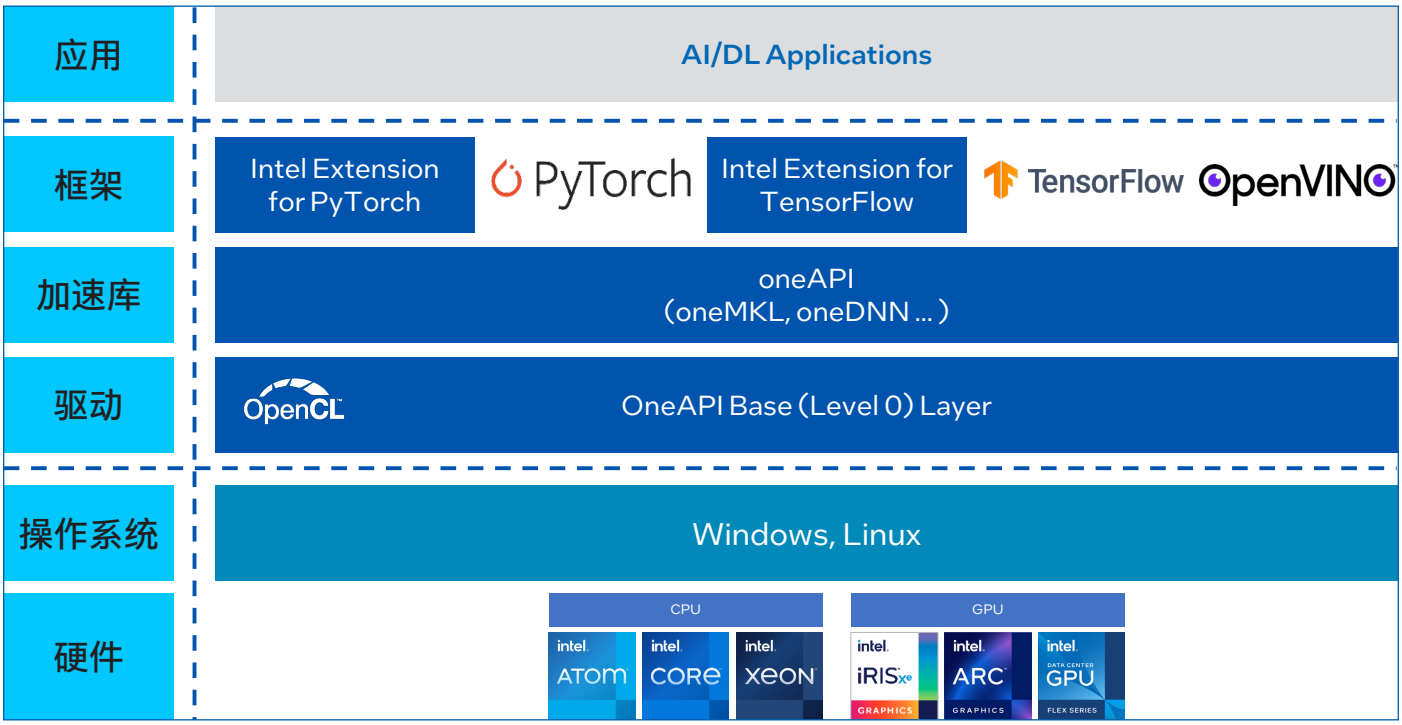

图 1. 基于英特尔锐炫™ 显卡的医学影像 AI 推理解决方案参考架构

高性能的硬件基础

基于英特尔锐炫™ 显卡的医学影像 AI 推理解决方案推荐采用英特尔锐炫™ 显卡和英特尔® 酷睿™ 处理器。其中，英特尔锐炫™ 显卡拥有最高 32 个 X^e 核心、8 个渲染切片，核心最大频率 2400MHz，显存最高可达到 16 GB 256bit GDDR6，英特尔锐炫™ 显卡的 X^e 内核集成了扩展向量引擎 Extended Vector Engine（XVE）和矩阵引擎 Extended Matrix Engine（XMX），能够加速 AI 工作流，在边缘端为 AI 推理提供强大、实时的算力支持。

图 2. 英特尔锐炫™ 显卡

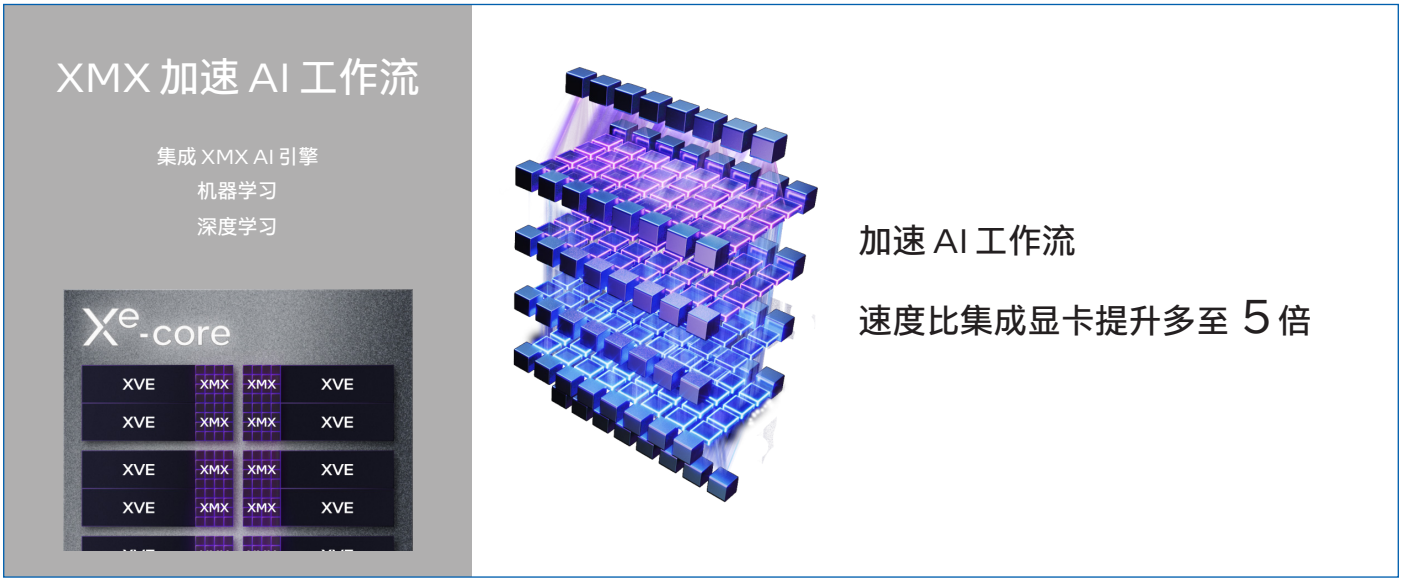

图 3. 英特尔锐炫™ 显卡具备卓越性能[1]

第 13 代英特尔® 酷睿® 处理器凭借全新的性能混合架构重新定义了边缘与终端设备的多核架构。其性能核（即“P 核”）可大幅地提高单线程性能和响应速度，而能效核（即“E 核”）则能够为多任务处理提供可扩展的多线程性能和高效的后台任务卸载。该处理器包含由英特尔® X^e 架构驱动的英特尔锐炬® X^e 显卡，具备高达 96 个执行单元（EU），能够与英特尔锐炫™ 显卡并行执行计算任务。

易于兼容、移植的 AI 算法生态

基于英特尔锐炫™ 显卡的医学影像 AI 推理解决方案依托 Intel® Extension for PyTorch（IPEX）、Intel® Extension for TensorFlow（ITEX）、OpenVINO™ 工具套件和英特尔® oneAPI 工具包等软件工具，能够便捷地实现 AI 算法的移植，同时还能够实现面向英特尔® 硬件的性能优化。

● 使用 Intel® Extension for PyTorch 迁移基于 PyTorch 编写的模型

对于基于 PyTorch 编写的模型，为了实现算法的迁移，并在英特尔® 硬件上获得额外的性能提升，合作伙伴可采用专为 GPU 优化的 Intel® Extension for PyTorch 获取支持。Intel® Extension for PyTorch 是英特尔发起的一个开源扩展项目，它基于 PyTorch 的扩展机制实现，旨在通过提供额外的软件优化充分发挥硬件特性，帮助用户在原生 PyTorch 的基础上显著提升英特尔® 硬件（如 CPU 和 GPU）上的深度学习推理计算和训练性能。通过扩展，PyTorch 用户将能够更充分地发挥英特尔硬件的最新功能，并在第一时间体验软件优化带来的卓越性能和部署便捷性。

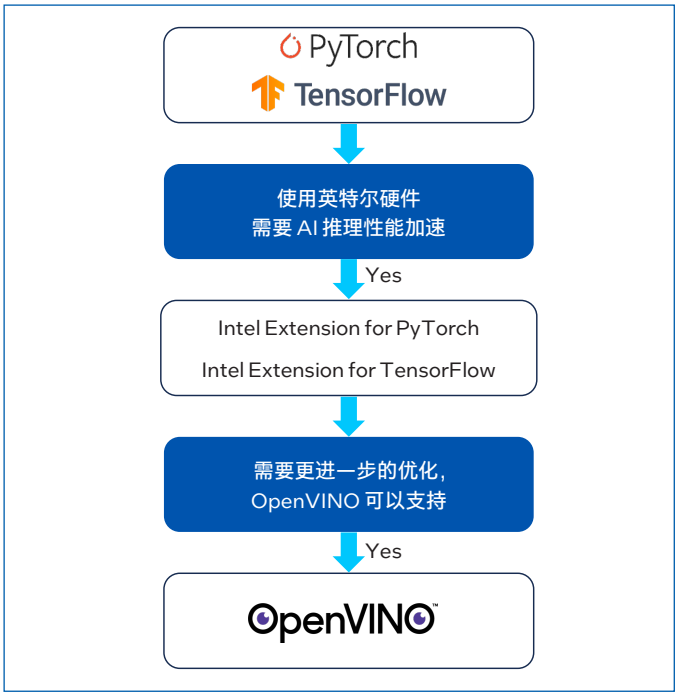

图 4. 英特尔软件生态支持便捷地进行模型迁移与加速

[1] 英特尔技术可能需要启用硬件、软件或激活服务。实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 www.Intel.com/PerformanceIndex

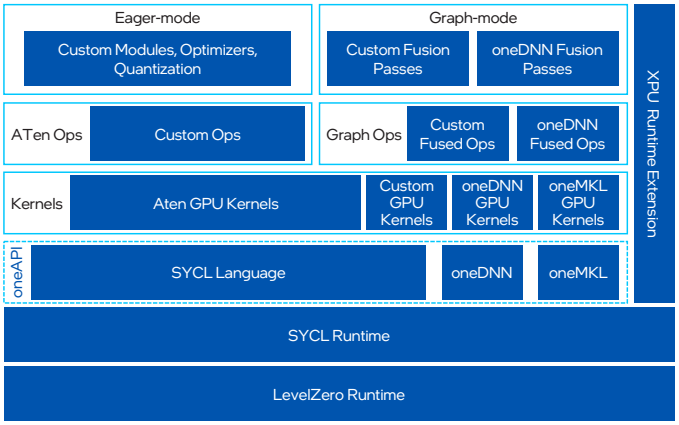

图 5. Intel® Extension for PyTorch 框架

Intel® Extension for PyTorch 支持英特尔® GPU，可以在英特尔® GPU 上启用最新的英特尔软件和硬件 AI 优化，包括将多个优化上传到框架的库存版本中，以获得开箱即用的性能提升。IPEX 扩展的额外性能来自于对 Eager 模式和 Graph 模式的优化。在 Eager 模式下，PyTorch 前端扩展为自定义 Python 模块（如融合模块）、更佳的优化器和 INT8 量化 API。通过扩展的图形融合通道将 Eager 模式模型转换为 Graph 模式，可以获得额外的性能提升。对于设备后端，通过 PyTorch 调度机制可实现并注册优化的操作程序和内核。

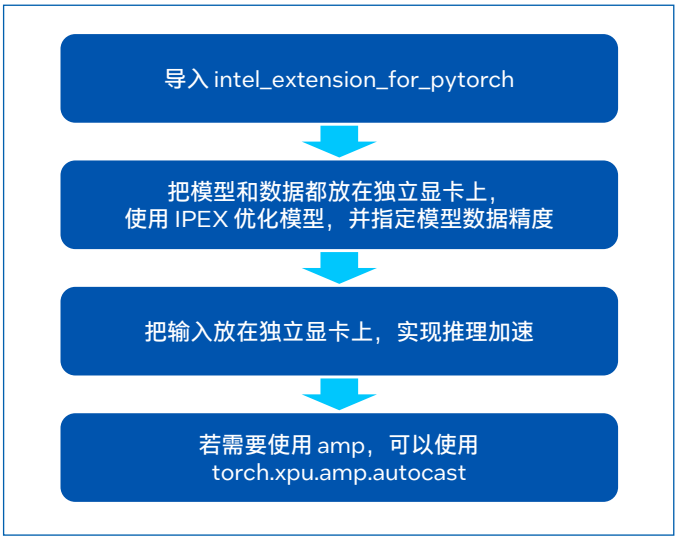

图 6. 采用 Intel® Extension for PyTorch 迁移应用流程

采用 Intel® Extension for PyTorch 的代码修改示例如图 7 所示，通过 Intel® Extension for PyTorch，用户可以快速将代码迁移到多元化的独立显卡中运行，无需繁琐的代码开发工作。对于使用 C++ 语言编写的模型，用户同样也可以使用 Intel® Extension for PyTorch，调用相关的函数库来提供支持。

```
import torch
import torchvision.models as models
############# code changes ###############
import intel_extension_for_pytorch as ipex
############# code changes ###############

model = models.resnet50(pretrained=True)
model.eval()
data = torch.rand(1, 3, 224, 224)

################### code changes ################
model = model.to("xpu")
data = data.to("xpu")
model = ipex.optimize(model, dtype=torch.bfloat16)
################### code changes ################

with torch.no_grad():
    d = torch.rand(1, 3, 224, 224)
    ############################ code changes ####################
    d = d.to("xpu")
    with torch.xpu.amp.autocast(enabled=True, dtype=torch.bfloat16):
    ############################ code changes ####################
        model = torch.jit.trace(model, d)
        model = torch.jit.freeze(model)
        model(data)
```

图 7. 代码修改示例[2]

● 使用 Intel® Extension for TensorFlow 迁移基于 TensorFlow 编写的模型

对于基于 TensorFlow 编写的模型，合作伙伴可采用 Intel® Extension for TensorFlow[3] 进行迁移。Intel® Extension for TensorFlow 是一个高性能深度学习扩展，实现了 TensorFlow PluggableDevice 接口。通过与 TensorFlow 框架的无缝集成，它允许 TensorFlow 开发人员轻松访问英特尔 GPU 和 CPU 等设备。借助英特尔扩展，开发人员可以在零代码更改的情况下在英特尔 AI 硬件上进行 TensorFlow 模型的训练和推理。

Intel® Extension for TensorFlow 构建在 oneAPI 软件组件之上。大多数性能关键图和运算符都通过使用英特尔 oneAPI 深度神经网络（oneDNN）进行了高度优化，这是一个用于深度学习应用程序的开源跨平台性能库。其他运算符则使用 SYCL（一种用于编程加速器和多处理器的 API 核心语言）实现。

[2] 更多代码修改实例，请见 https://github.com/intel/intel-extension-for-pytorch/blob/v2.0.110%2Bxpu/examples/gpu/inference/python/resnet50_torchscript_mode_inference_bf16.py

[3] 更多信息请见 https://www.intel.com/content/www/us/en/developer/articles/technical/introduction-to-intel-extension-for-tensorflow.html

● 使用 OpenVINO™ 工具套件迁移基于 ONNX 或 PaddlePaddle 等框架的模型

对于基于 ONNX 或 PaddlePaddle 等框架编写的模型，可以采用 OpenVINO™ 工具套件[4] 进行迁移。OpenVINO™ 工具套件是用于快速开发应用程序和解决方案，以解决各种任务（包括人类视觉模拟、自动语音识别、自然语言处理和推荐系统等）的综合工具套件。OpenVINO™ 工具套件可实现对基于 ONNX 或 PaddlePaddle 等框架的模型的直接支持，模型优化器 MO 工具可以直接完成对上述模型的离线迁移。

OpenVINO™ 工具套件还为模型带来高的 AI 推理性能提升。该工具套件基于最新一代的人工神经网络，包括卷积神经网络（CNN）、递归网络和基于注意力的网络，可跨英特尔® 硬件扩展计算机视觉和非视觉工作负载，从而大幅提高性能。它通过从边缘到云部署的高性能、人工智能和深度学习推理来为应用程序加速。

针对基于其他框架编写的模型，英特尔还提供了英特尔® oneAPI 工具包[5]，可以助力快速完成迁移。英特尔® oneAPI 工具包是基于新一代标准的英特尔® 软件开发工具，用于跨各种架构构建和部署以数据为中心的高性能应用程序。它能够通过充分利用出色的硬件特性加速计算进程，并全面兼容现有的编程模型和代码库，可确保开发者已经编写的应用能够在 oneAPI 上无缝运行。

英特尔® oneAPI 工具包可以自动将基于其他框架编写的代码迁移为可以在英特尔锐炫™ 显卡上运行的代码。通过这一迁移，用户能够降低 AI 推理任务跨平台开发与迁移的复杂性，提升 AI 模型在异构平台中运行的性能，并充分利用现有的医疗 AI 模型，从而加速医疗 AI 应用的开发。

除此之外，这一解决方案参考设计还提供了丰富的软件功能支持，这可以帮助医疗机构高效处理医疗影像等数据，满足进阶的智慧医疗场景需求。例如，英特尔提供了开源的 OpenGL 驱动，这使得其能够应对医学影像辅助诊断等场景中，三维重建等三维可视化处理需求，让医生可以从多角度清晰了解到各结构之间的空间位置关系。

面向医疗边缘场景的定制化硬件支持

鉴于医学影像 AI 临床化，专科智能化的趋势，AI 推理的部分常与边缘计算形态相结合，这与常规的服务器 + 软件松耦合的使用形式不同，对硬件的设计要求更需要符合医护人员的使用需求，如 AI 终端需要面临低噪音、稳定性、防尘、防水、抗菌等苛刻的医疗场景需求，并经过安规和 EMC 认证。

基于英特尔® 架构的典型医疗专用终端采用卓越的散热方案，具备大散热孔、风扇以及水冷散热，保证 CPU 与 GPU 性能尽情释放，并且保持低噪音，提供安静的医疗办公环境。这些终端支持个性化定制，能够满足客户的产品差异化需求，并可搭载 AI 辅助诊断平台，支持多种附加卡扩展功能，可有效支撑 AI 辅助诊断软件本地化端侧部署的 AI 算力需求。这些终端还具备丰富的 I/O 口，易于连接各种设备，方便部署，能够全面推进智慧 AI 医疗目标的实现。

图 8. 英特尔® 架构医疗专用终端满足各种要求

[4] 更多信息请见 https://www.intel.cn/content/www/cn/zh/developer/tools/openvino-toolkit/overview.html

[5] 更多信息请见 https://www.intel.com/content/www/us/en/docs/dpcpp-compatibility-tool/get-started-guide/2023-0/overview.html

收益

基于英特尔锐炫™ 显卡的医学影像 AI 推理解决方案参考设计利用英特尔硬件 + 软件的生态整合能力，使得其能够在边缘端支持 AI 应用的高效、稳定运行。

对于合作伙伴而言，该方案能够带来以下重要价值：

- **降低 TCO：** 该方案能够充分利用英特尔® CPU、英特尔® GPU 等硬件，以及英特尔® 软件及技术的优势，提升性价比，推动医学影像 AI 应用的普及。
- **获得完整的产品线：** 英特尔可提供涵盖 CPU、GPU、软件工具在内的完整产品线，实现更便捷的方案构建。
- **高兼容性：** 英特尔提供的软件套件能够灵活地跨多个英特尔® 硬件进行扩展，还可通过独立显卡与集成显卡并行运行，进一步提升性能。通过该方案，ISV 和 OEM 能够拥有更多的推理算力选项，避免依赖单一设备所带来的硬件锁定与供应等风险。
- **丰富的定制化硬件：** 英特尔大量的硬件合作伙伴提供了丰富的定制化硬件支持，可满足实时性、可靠性、低噪音、防水抗菌、医疗认证等各种要求。

应用实践

目前，多个合作伙伴已经参考基于英特尔锐炫™ 显卡的医学影像 AI 推理解决方案参考设计，推出了相应的产品与方案。以汇医慧影推出的 AI 骨密度辅助检测系统为例，该系统是基于 CT 胸腹部平扫图像数据的 CT 人工智能骨密度检测软件，可自动进行骨密度测量与分析，并同步给出椎体分析数据、椎间分析数据、腹部组织成分分析。

为了验证该方案的性能表现，汇医慧影进行了测试，受测的骨密度筛查项目共包含 4 个 AI 模型：基于 3D-Resnet-Unet 的脊椎粗分割模型、基于 3D-Unet 的脊椎分割模型、基于 3D-DenseNet 的骨密度回归模型、基于 2D-Unet 的组织分割模型，测试数据共 120 例 CT 平扫数据，扫描部位包括胸部、腹部、胸腹连扫，典型测试样本示意图如图 9 所示：

测试结果显示，纯影像 AI 推理计算时间平均为 9.55 秒[6]。这一测试结果表明，英特尔锐炫™ A770 性能优异，可助力检测系统高效完成骨密度筛查场景中的数据测算与分析任务。

汇医慧影首席执行官柴象飞指出：“借助 AI 筛查，我们能够主动提醒就诊者可能存在的骨密度问题，助力其提早防范。在打造骨密度筛查 AI 方案的过程中，英特尔锐炫™ 显卡提供了强大的算力支撑，让我们能够轻松满足性能目标，同时在灵活性、稳定性等方面表现优异，有助于推动骨密度筛查的大规模应用。”

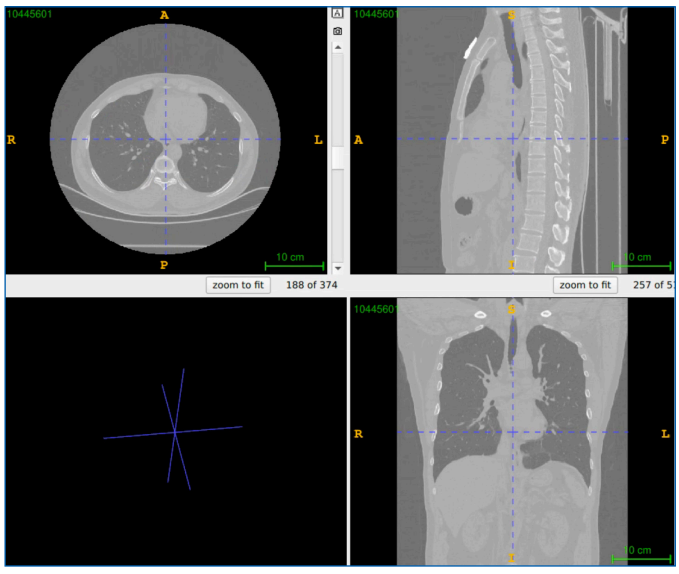

图 9. 骨密度筛查典型测试样本示意图

[6] 数据援引自汇医慧影截至 2023 年 8 月的内部测试结果。测试配置：英特尔® 酷睿™ i9-12900K 处理器（12 核 24 线程），64 GB 总内存，英特尔锐炫™ A770 显卡 (16G)。英特尔技术可能需要启用硬件、软件或激活服务。英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

展望

《“健康中国 2030”规划纲要》提出，实现全民健康的根本目的，要覆盖全生命周期，针对生命不同阶段的主要健康问题及主要影响因素，强化干预，实现全程健康服务和健康保障。基于英特尔锐炫™ 显卡的医学影像 AI 推理解决方案可助力医院提升医学影像阅读能力，降低对于医师、设备等资源的要求，从而提升疾病诊断的效率，帮助就诊者及时采取相应的应对措施。

英特尔将强化 GPU 等关键技术创新，并加强生态合作，一方面利用更多的优化策略，如独立显卡与集成显卡并行等，持续优化方案的性能表现，另一方面探索将英特尔锐炫™ 显卡用于更多的 AI 医疗领域，在更广泛的场景释放该显卡在 AI 加速方面的潜力，帮助医院用户推动智慧化转型。