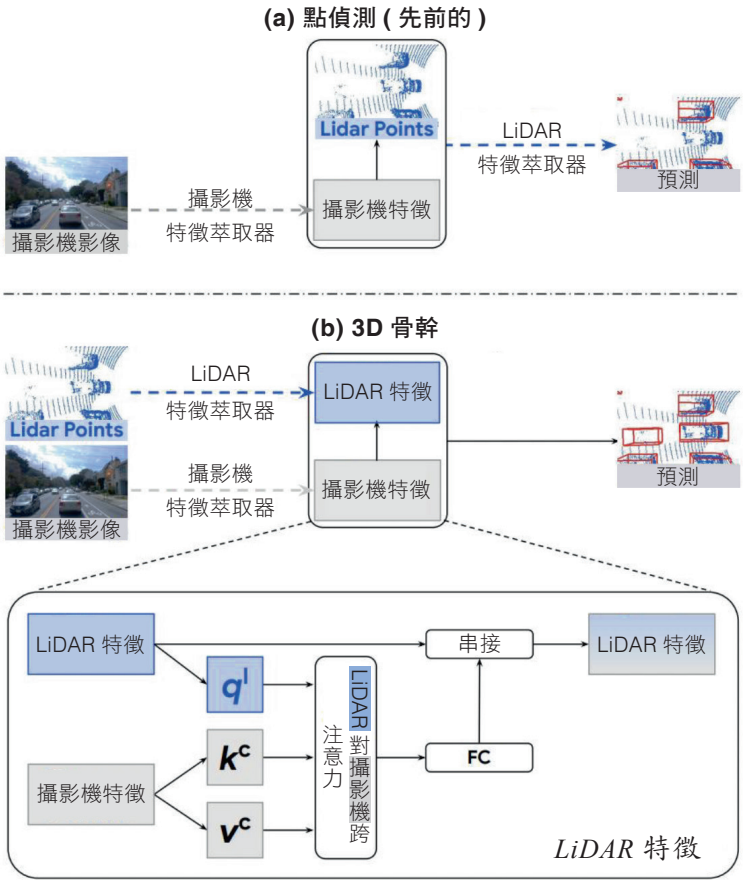

圖 3-15 DeepFusion 概念（b）。
經許可使用，來源：https://arxiv.org/pdf/2203.08195[25]

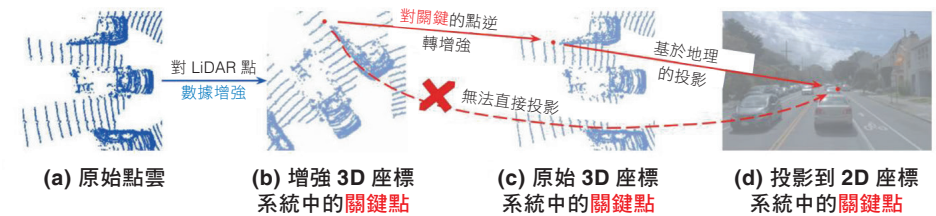

圖 3-16 InverseAug 方法概念。原始 LiDAR 點雲（a）。應用增強後的 LiDAR 點（b）。原始 3D 座標系中的關鍵點（c）。2D 座標系中的投影點（d）。
經許可使用，來源：https://arxiv.org/pdf/2203.08195[25]

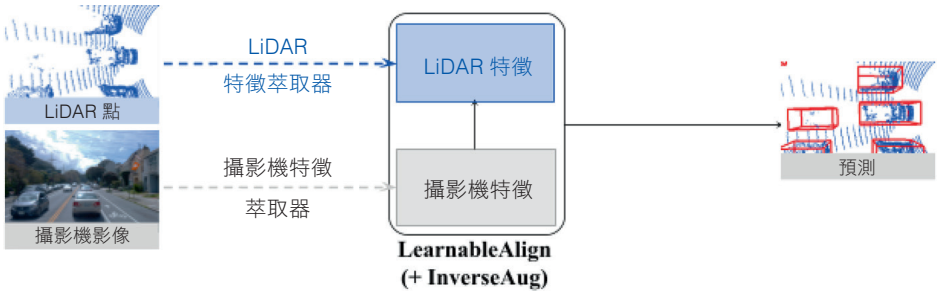

圖 3-19 DeepFusion 工作流程概念。
經許可使用，來源：https://arxiv.org/pdf/2203.08195[25]

BEVFusion

在多感測器融合的基礎上，BEVFusion[27][28] 採用不同方法處理 3D 感知任務。其將來自兩種感測器的輸入轉換為特徵，並轉為環境俯視視角的鳥瞰視角（bird's-eye view，BEV），此統一視圖簡化了數據整合。BEV 編碼器處理結合特徵，供如物體偵測與追蹤等特定任務執行 3D 感知，如圖 3-20 所示。

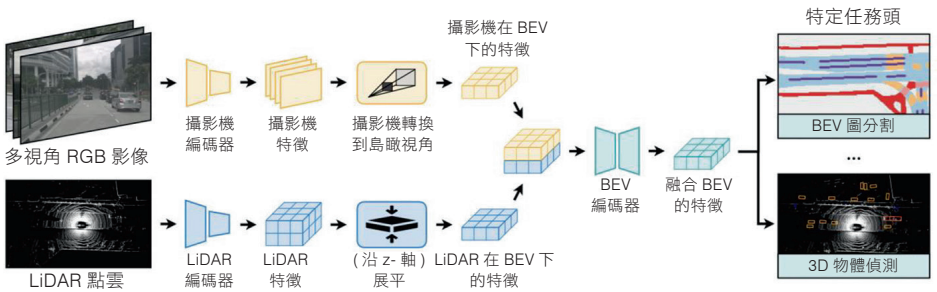

圖 3-20 BEVFusion 從不同輸入萃取特徵並轉為共享 BEV 空間的流程。
經許可使用，來源：https://arxiv.org/pdf/2205.13542[47]

若欲深入了解此方法，建議閱讀原論文 [27]。BEV 的主要優勢在於提供一致的俯視視角，簡化 LiDAR 與攝影機數據的融合。數據放到共享空間，改善空間對齊外，也能進一步提升 3D 感知任務的準確度。

整體來說，感測器融合能整合攝影機、LiDAR 等多感測器數據，發揮互補優勢，提升如偵測、分割與追蹤等常見的機器人感知任務性能。

任務中成功，但訓練的集中就難以涵蓋多樣性。因此，透過學習更好的表徵或推理任務語義以實現通用化的方法，就更具優勢。

當前機器人研究有多個通用化面向：

1. **物體通用化**：模型能否處理未見到的物體？
2. **環境通用化**：模型能否在未見過的環境中行動？
3. **運動通用化**：模型能否創造新的運動？
4. **視角通用化**：模型能否從第三人稱轉至自我視角（或反之）？攝影機的姿態是否重要？
5. **體現通用化**：一種機器人體現的數據，能否改善另一種機器人體現的技能？

此外，機器人控制模型可在符號理解、推理、長程規劃、人類識別、實體安全性等方面進行評估。

端到端機器人控制涵蓋從控制輸出到機器人輸入的低階控制學習問題，梯度從控制輸出流向輸入 [56]。傳統上，這意味著行動可從攝影機影像與其他觀察直接學習而來。如何開發端到端控制機器人的方法？如何將這些模型連結至 AI 的規模化與大型模型趨勢？機器人如何從通用化面向進行衡量與提升？

使用自迴歸 Transformer 的端到端機器人控制

截至 2022 年，規模化大型多任務強化學習模型被視為解決機器人問題的關鍵，但隨後因模仿預訓練在多任務基準上表現優於強化學習方法，因此趨勢轉向大型模仿學習模型。如圖 4-13，Robotics Transformer 1(RT-1)[57] 在這一時期誕生，它是一個早期大規模模仿學習多任務模型，在涵蓋多樣任務的真實世界機器人數據集訓練，算是早期的基礎模型。RT-1 運作如下：

- 以使用者文字指令與機器人擷取的影像序列（歷史）作為 RT-1 模型輸入。文字由凍結（frozen）文字標記器編碼，影像與文字透過 FiLM- 高效網路熔融（fuse）[58][59]。視覺與語言標記的早期熔融對於從影像萃取正確的情境極為重要。

- 這些視覺語言標記由使用注意力聚焦於影像與文字的最相關部分的 TokenLearner[60] 進行壓縮，推論速度提升了三倍。
- 接著由使用自注意力理解影像中物體與文字指令關係的 Transformer 處理標記 [43]，輸出每個自由度的單一控制參數。對於笛卡爾末端執行器控制，輸出為夾爪 3D 位置、3D 旋轉、基底運動等。連續值的行動則透過離散均勻分佈的映射進行標記化，接著再預測對應浮點行動的桶（bucket）編號。（譯按：即連續行動的空間離散化。）
- 模型以交叉熵損失進行訓練，訓練數據（行動輸出與影像 / 文字輸入）由人類遙控機器人加以收集。

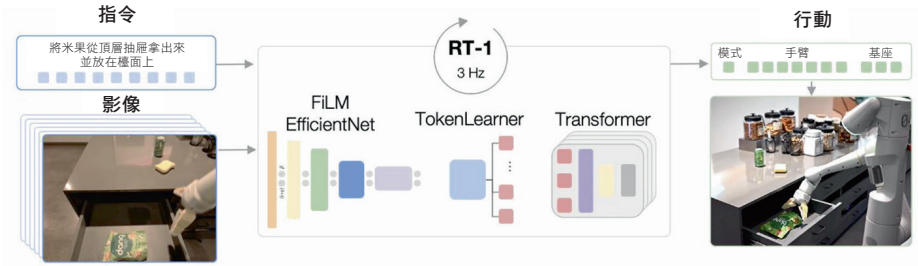

圖 4-13 RT-1 處理影像與自然語言指令以控制機器人行動。
經許可使用，來源：https://arxiv.org/pdf/2212.06817[57]

RT-1 展現了閉迴路控制、持續執行行動直到生成終止行動或達到預定時步數。此工作顯示在給定足夠的領域數據下，基於 Transformer 的模型可適配創建出機器人控制的大型多任務 Transformer。研究還顯示，適當混合類似語言模型領域合成數據的模擬數據可提升性能。RT-1 亦顯示機器人之間的轉移跡象，這可在之後成為更廣泛研究領域。

RT-1 的缺點是其主要用於測試粗笨的操作（取放），難以知曉是否適用於高度靈巧任務。靈巧任務需要類似人類手的精細運動控制與協調，以處理不同材質、大小與形狀的物體。另一缺點是，雖然在領域內的通用化表現出色，但在全新的物體與推理領域外，其通用化能力相當有限。

另一類似 RT-1 的廣泛使用模型是 ACT（行動分段 Transformer）[61]，這是編碼器 - 解碼器模型，其引入建模未來 k 步行動的「行動分段」（action chunking）概念。在任意時間點觀察輸入但預測 k 個行動，因而有效的將

任務行程縮減 k 倍。例如在示範中暫停，由於下一個行動不只取決於狀態還取決於時步，因此這個方法於存在時間相關干擾因素時，此法會優於單步策略。只要干擾因素在分段長度內，行動分段策略是可以恢復的。實際的分段長度對應未來 1 秒行程，若控制頻率為 30Hz，單一觀察就不再如 RT-1 預測的一個行動，而是預測 32 個行動（即 1 秒行程）。訓練分段時亦採用時間性集合（ensemble），如圖 4-14 右側所示，訓練於重疊分段而非不相交分段，因此可實現極密集建模。

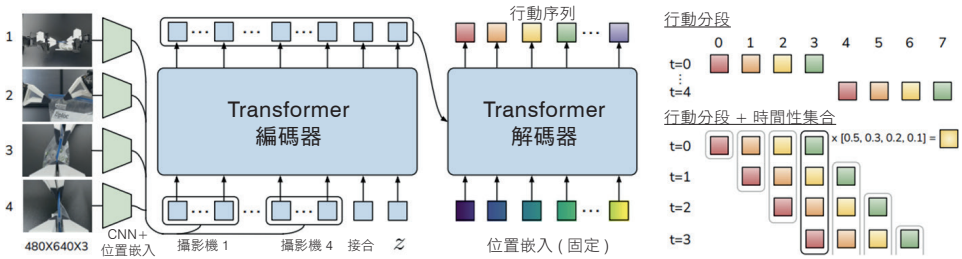

圖 4-14 左側：行動分段編碼器 - 解碼器 Transformer 架構；右側：時間性集合的分段。經許可使用，來源：https://arxiv.org/pdf/2304.13705[61]

RT-2[62] 是在 RT-1 基礎上，使用預訓練視覺語言模型的語義資訊，生成低階機器人行動。

RT-2 使用視覺問答（VQA）數據（影像搭配描述影像情境的問答）與機器人行動數據（影像與文字指令搭配完成任務的機器人行動）。模型首先使用網際網路大規模數據預訓練的視覺語言模型，接著與機器人行動及網際網路大規模數據共同訓練，進而生成視覺語言 - 行動（VLA）模型。其訓練方式如圖 4-15 所示。

從如「機器人該如何拿起蘋果？」等使用者的查詢，推論就開始了。查詢與任務執行場景影像由 ViT 與 LLM 進行處理，ViT 從影像萃取視覺特徵，LLM 則理解查詢的語言部分。接著結合特徵生成行動序列，呈現出機器人的特定平移與旋轉。

論文結果顯示，由這種網際網路與機器人數據集的共同訓練，機器人只能理解網際網路上見過的概念並應用到真實世界中。

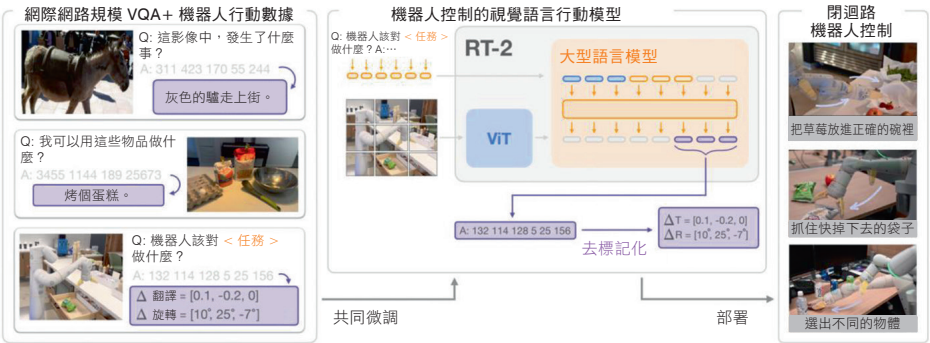

圖 4-15 RT-2 將機器人行動表示為文字標記，與大規模視覺語言數據集共同訓練。經許可使用，來源：https://arxiv.org/pdf/2307.15818[62]

例如機器人能識別名人（「將罐子移給 Taylor Swift」）、簡單數學（「將可樂罐移至 1+2 之和」）與符號（「將可樂罐移往 Google」），甚至還可理解相對概念，如「拿起不同顏色的物體」、「將草莓放入正確的碗中」，這都需要對場景中的選項進行推理。研究指出，網際網路大規模的預訓練帶來了比以往更好的表現。部分結果如圖 4-16 所示。

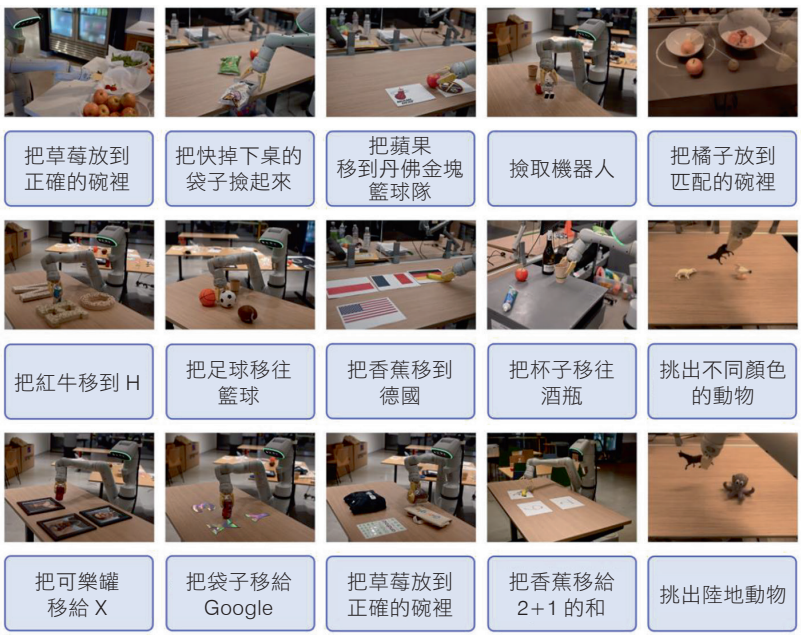

圖 4-16 RT-2 評估結果，機器人理解網際網路上概念如 Google、字母、顏色、Taylor Swift 等。經許可使用，來源：https://arxiv.org/pdf/2307.15818[62]

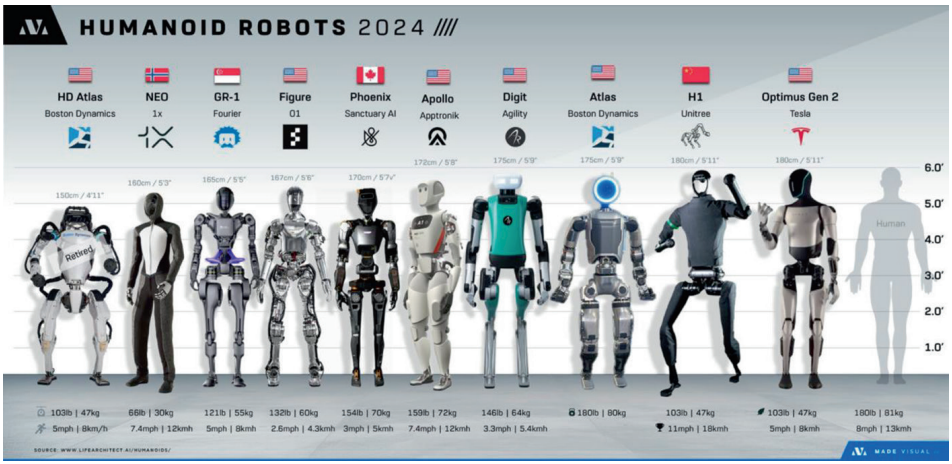

圖 10-1 人形機器人硬體市場。
經許可使用，來源：https://lifearchitect.ai/humanoids/by https://lifearchitect.ai/[32]

這些公司分享了成功開發人形機器人的關鍵要素，包括 [14]：

1. **資金**：開發先進機器人成本高昂，需要支持研究、開發與生產的資金。隨著產業的影響力與成長，投資者對人形機器人公司的資金投入也在增加。例如加州新創 Figure AI 於 B 輪融資 6,750 萬美元，投資者包括 LG、三星與微軟 [15]。溫哥華的 Sanctuary AI 亦獲得用於推進通用人形機器人工作的大量資金，近期投資來自 Accenture Ventures 與 Magna[15]。人形機器人技術的投資增長，凸顯其潛力與市場化所需資源增加。

2. **基礎模型**：大規模 AI 模型使人形機器人能進行大量數據的理解與學習，展現認知能力，自主執行任務並適應新情境。這些通用模型可以微調並用於特定任務。

3. **數據**：數據量越多，機器人越能有效學習如何執行任務、辨識物體與理解人類行為。人形機器人的數據包括影像、影片、文字與真實環境的感測器數據。

4. **機器人**：人形機器人的實體建構包含如感測器與模擬人類行動的致動器等硬體與生物力學。波士頓動力公司與特斯拉在此領域大力投資，開發能舉重或執行複雜動作的靈活機器人。

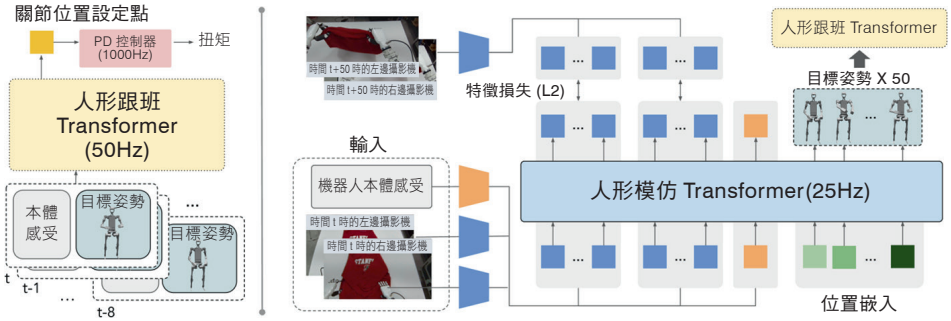

圖 10-2 人形機器人跟班 Transformer 架構。經許可使用，來源：https://arxiv.org/html/2406.10454v。HumanPlus: Humanoid Shadowing and Imitation from Humans by Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn at Stanford University[23]

由第四與第五章談到的方法，從大規模影片數據學習仍是解鎖通用人形機器人智慧的終極目標（Holy Grail）。因策略結構相似，人類操作數據與人形機器人控制的體現差距最小，所以其間的轉移會較為容易。

行走方法

傳統行走的解決方案是採用不依賴機器學習的程式化方法。但本書聚焦機器學習方法，因此本節將討論新穎的實驗性行走方法。近期 Radosavovic 等人 [24] 將人形機器人行走視為預測下一個標記的問題，以部分包含、部分不包含行動輸出的多種數據類型，訓練行走神經網路控制器。其訓練的四類數據集包括：

1. **神經網路策略**：由模擬訓練的 RL 策略所模擬生成的觀察 - 行動對。
2. **基於模型控制器的數據**：由如 Agility Robotics 等人形機器人公司控制器生成的非行動軌跡。
3. **Mocap（動作擷取）數據**：如 KIT[25] 的人類標記數據，此數據透過人形機器人的逆向運動學模型重新定向。
4. **人類姿態的 YouTube 數據**：應用姿態估算人類做事的 YouTube 影片，再透過逆向運動學模型重新定向至人形機器人。