D26-103

高效能點雲神經網路加速器晶片

An Energy-Efficient Point
Cloud Neural Network Accelerator IC

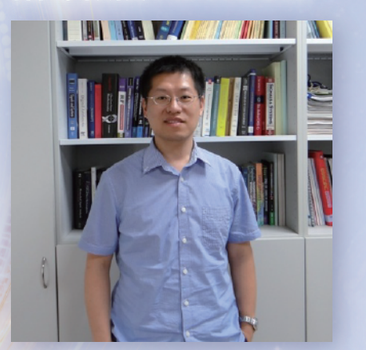


隊伍名稱｜點點點雲 DotDotDot Cloud

隊　　長｜莊承霖 / 臺灣大學電子工程學研究所

隊　　員｜徐以帆 / 臺灣大學電子工程學研究所

指導教授 ADVISOR



楊家驤｜臺灣大學電機工程學系暨電子工程學研究所

美國加州大學洛杉磯分校電機博士，現為臺灣大學電機工程學系暨電子工程學研究所教授。目前擔任IEEE頂尖期刊JSSC與TBioCAS副主編，曾擔任頂尖國際會議(ISSCC、VLSI Symposium)技術議程委員，曾獲ISSCC傑出技術論文獎、ISSCC遠東區最佳論文獎、國科會傑出研究獎、中國電機工程學會傑出電機教授獎、胡正明半導體創新獎、傑出人才發展基金會年輕學者創新獎、臺灣積體電路設計學會傑出年輕學者獎等獎項。

研究領域

AI晶片設計、生醫積體電路訊號處理、通訊基頻積體電路設計。

作品摘要

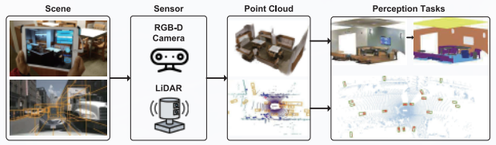
為了將先進科技落實於真實世界，智慧裝置必須具備理解周遭空間、物件與環境變化的能力。次世代邊緣應用，例如智慧眼鏡、機器人與自主移動平台，皆高度仰賴即時且可靠的三維空間解析技術，以支援場景理解、物件辨識、路徑規劃與人機互動等功能。然而，不同應用場域所產生的三維資料在密度、分布與運算型態上存在明顯差異，使得傳統處理方式往往難以在精度、效率與適用性之間取得良好平衡。另一方面，邊緣裝置通常受到晶片面積、功耗與記憶體頻寬限制，進一步提高即時運算的實作挑戰。

針對上述問題，本作品從演算法與硬體架構兩個層面進行協同設計，提出一套面向多樣化場域的高效能空間處理運算設計。此設計透過統一化的資料表示與運算流程，降低不同環境資料型態之間的處理差異，使系統能更有效率地支援複雜情境下的任務。相較於傳統針對特定場景或特定資料型態最佳化的設計，本作品更著重於提升彈性與環境涵蓋性，適合部署於資源受限的邊緣智慧裝置中。

在演算法層面，本作品利用空間資料在局部區域與特徵表現上的相似性，減少重複且低效益的運算，並透過資料重組與近似化策略降低主要運算模組的負擔。同時，設計中亦考量感測資料在空間上的連續性與局部關聯，縮小不必要的資料處理範圍，進而降低整體運算複雜度與資料搬移成本。上述方法在維持任務表現的同時，有效提升運算效率，使高負載的空間模型更適合在邊緣端執行。

在硬體架構層面，本作品根據資料分布與運算特性設計專用加速流程，提升處理單元效率並減少無效運算。晶片內部透過彈性的資料排程、可重構的記憶體配置與高效率的資料復用機制，減少外部存取需求，並改善多個運算階段的延遲。整體設計兼顧運算效率與資料搬移效率，能在有限功耗與面積條件下提供高效能的三維運算能力。

本作品實現了一顆面向次世代邊緣空間理解的點雲神經網路加速器晶片，能支援多樣化場域下的應用，並展現優於過去相關設計的任務表現與能量效率。



圖一 次世代邊緣應用中的環境感知與智慧運算示意圖。

Abstract

To bring advanced technologies into the physical world, intelligent devices must be able to perceive surrounding spaces, objects, and environmental changes. Frontier edge applications, such as smart glasses, robotics, and autonomous mobile platforms, increasingly rely on real-time and reliable spatial awareness to support scene understanding, object recognition, path planning, and human-machine interaction. However, sensing data collected from different application environments can vary significantly in density, distribution, and computational behavior, making it difficult for conventional processing approaches to achieve a balanced trade-off among accuracy, efficiency, and adaptability. Meanwhile, edge devices are often constrained by chip area, power consumption, and memory bandwidth, further increasing the difficulty of implementing real-time workloads.

To address these challenges, this work adopts a joint algorithm-and-hardware design approach and proposes a high-efficiency spatial computing design for diverse operating environments. By introducing a more consistent data representation and processing flow, the proposed design reduces the gap between different types of environmental data and enables more efficient support for tasks in complex scenarios. Compared with conventional designs that are optimized for specific scenes or specific data formats, this work places greater emphasis on flexibility and scene coverage, making it suitable for deployment in resource-constrained edge intelligence platforms.

At the algorithm level, this work leverages common patterns in local spatial regions to reduce repetitive and low-value computations. It also applies data organization and approximation strategies to lower the computational burden of major processing modules. In addition, the design considers the spatial relationships within sensing data to limit unnecessary operations, thereby reducing overall computational complexity and data movement. These methods improve computational efficiency while maintaining task performance, enabling high-workload 3D models to be more practical for edge-side execution.

At the hardware level, this work develops a dedicated acceleration flow based on data distribution and computational characteristics, improving processing-unit efficiency, and reducing ineffective operations. Within the chip, flexible data scheduling, reconfigurable on-chip memory allocation, and efficient data reuse mechanisms are adopted to reduce external access and improve latency across multiple processing stages. The overall design balances computational efficiency and data movement efficiency, providing high-performance spatial processing capability under limited power and area constraints.

This work demonstrates a dedicated edge AI chip for advanced scene-understanding applications. It supports tasks across diverse environments and achieves superior task performance and energy efficiency compared with prior related designs.