國立臺灣大學電機資訊學院資訊工程學研究所

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

基於任務整合與分割策略的 YOLOv7 全景分割系統

Panoptic Segmentation via Tasks Integration and Segmentation-based Strategies on YOLOv7

郭毅遠

I-Yuan Kuo

指導教授: 李明穗 博士、廖弘源 博士

Advisors: Ming-Sui Lee Ph.D.、Hong-Yuan Liao Ph.D.

中華民國 114 年 1 月

January, 2025

doi:10.6342/NTU202404744

國立臺灣大學碩士學位論文
口試委員會審定書

MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

基於任務整合與分割策略的 YOLOv7 全景分割系統

Panoptic Segmentation via Tasks Integration and Segmentation-based Strategies on YOLOv7

本論文係郭毅遠君（學號 R10922A23）在國立臺灣大學資訊工程學系人工智慧碩士班完成之碩士學位論文，於民國 113 年 11 月 21 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Master Program of Artificial Intelligence offered by the Department of Computer Science and Information Engineering on 21 November 2024 have examined a Master's thesis entitled above presented by I-YUAN KUO (student ID: R10922A23) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

李明穗

（指導教授 Advisor）

王建堯

系（所）主管 Director: 陳祝嵩

Acknowledgements

在這篇論文的完成過程中，有許多師長和朋友給予了我無私的指導與支持，讓我能夠順利完成研究。

首先，我由衷地感謝從小栽培我的廖弘源所長。廖老師在我高中時期引領我進入中研院的研究環境，使我年輕時便有機會接觸世界前沿的研究課題，並提點我進行學術研究的方法。我得以在高手雲集的實驗室中自由探索各種不同的研究，一步一步認識人工智慧。廖老師無論是在學業還是人生方面，始終給予我悉心的關懷與支持，對我的成長影響深遠。

同時，我也特別感謝我的指導教授—李明穗教授。李老師在我入學之初，對於校園生活、課程選擇等方面給予我許多寶貴建議。修課期間，老師的「數位影像處理」課程讓我體會到傳統影像處理方法的力量及有趣之處。最令我印象深刻的是，實做 Dehaze 演算法時，我很直覺地認為透過深度學習就能夠解決問題，但實際上許多經典演算法也能產生出色的效果，這改變了我先前固有的觀念。

此外，我要感謝王建堯博士，建堯學長是提出當今最著名物件檢測器—YOLOv7, YOLOv9 的優秀研究者。無論是我研究方向的選擇、技術的探討，還是論文撰寫的建議，他都提供了極大的幫助，讓我能夠一步步克服困難，完成研究。

最後，我也要感謝我的家人，始終給予我無條件的支持與陪伴。

doi:10.6342/NTU202404744

摘要

本研究針對全景分割任務提出了一種新穎且高效的方法，全景分割任務旨在精確區分圖像中的所有前景與背景類別，並辨識同類別中的不同個體。現有方法在邊界預測精度不佳與重複預測問題上仍存在諸多挑戰，且許多先進方法依賴於龐大的網路架構，需大量運算資源，難以滿足實際應用需求。基於 YOLOv7[25] 與 FastInst[7] 的架構，本研究提出三項核心改進策略：(1) Tasks Integration，透過整合多任務的方法，解決傳統 CNN-based 方法邊界預測不精確問題；(2) Segmentation-based Proposal Strategy，有效避免 Query-based 架構中因冗餘 proposals 導致的重複預測；(3) Segmentation-based Intra and Counterfactual Loss，提升特徵的一致性與鑑別性，同時排除潛在誤導性特徵的影響。實驗結果表明，提出的方法顯著提升了模型的預測品質，為全景分割任務提供了一種兼具精度與效率的解決方案。

關鍵字：全景分割、任務整合、反事實注意力、深度學習、圖像分割

doi:10.6342/NTU202404744

Abstract

This study presents a novel and efficient framework for panoptic segmentation, a task aimed at accurately delineating all foreground and background categories in an image while distinguishing individual instances within the same category. Current approaches face persistent challenges, including imprecise boundary predictions and redundant proposals resulting in duplicate predictions. Moreover, many state-of-the-art methods rely on resource-intensive network architectures, making them less practical for real-world applications. Building on the architectures of YOLOv7[25] and FastInst[7], this research introduces three core advancements: (1) Tasks Integration, which unifies multi-task learning to address boundary prediction inaccuracies inherent to traditional CNN-based methods; (2) Segmentation-based Proposal Strategy, which effectively mitigates duplicate predictions by addressing redundancy in Query-based architectures; and (3) Segmentation-based Intra and Counterfactual Loss, which enhances feature consistency and discriminability while suppressing the influence of misleading features. Experimental evaluations demonstrate

doi:10.6342/NTU202404744

that the proposed methodology achieves substantial improvements in prediction quality, offering a robust and efficient solution for panoptic segmentation tasks.

Keywords: Panoptic Segmentation, Tasks Integration, Counterfactual Attention, Deep Learning, Segmentation

doi:10.6342/NTU202404744

Contents

doi:10.6342/NTU202404744

doi:10.6342/NTU202404744

List of Figures

doi:10.6342/NTU202404744

List of Tables

doi:10.6342/NTU202404744

Chapter 1　Introduction

在電腦視覺領域，全景分割 (Panoptic Segmentation) 是一項綜合了語義分割 (Semantic Segmentation) 和實例分割 (Instance Segmentation) 的關鍵任務 [6, 11]，其目標在於精確地區分圖像中的所有前景和背景類別，並辨別同一類別中的不同個體。全景分割技術在自動駕駛、醫療影像分析等領域具有顯著的應用價值，因而吸引了廣泛的研究興趣。隨著深度學習技術的迅速發展，越來越多的研究致力於結合各種網絡架構和方法，以提升全景分割的準確性和效率。

現有的全景分割方法在某些情況下效果顯著，但在應對複雜場景和資源受限的環境時，仍然存在許多不足之處。首先，CNN-based 的方法在圖像邊界處經常出現預測不一致的情況，導致分割結果邊緣不密合或物件重疊等問題，影響了預測品質。其次，許多 Query-based 的方法產生了大量冗餘 proposal，增加了重複預測的風險。這些方法雖然在準確度方面皆有所提升，但通常依賴龐大的架構，需要大量的計算資源，這在實際應用中是一大挑戰。

為了解決上述挑戰，本研究提出了一種新穎的全景分割方法，結合了 YOLOv7[25] 作為主要架構並透過 FastInst[7] 的 Dual-path Transformer Decoder 產生預測結果。YOLOv7 作為當今一種高效的物件檢測器，在目標檢測領域已經證明了其卓越的性能。藉由其輕量的架構及特徵萃取能力，並結合 Dual-path Transformer Decoder 的查詢機制產生高品質的分割結果。

doi:10.6342/NTU202404744

在本研究中，我們基於全景分割任務，設計了一種任務整合策略。透過同時訓練區分個體及區分類別的任務，達到任務間相輔相成的效果，避免任務間的特徵差異造成預測品質的下降。並透過新設計的 Segmentation-based Proposal Strategy 來減少冗餘 proposal，這一策略針對每個預測區域進行更精確的特徵分析，確保每個 proposal 都是獨一無二且具有意義。

此外，我們借鑒反事實注意力機制 (Counterfactual Attention Learning)[21] 的思路，這種方法已經被成功應用於視覺分類任務，其特點在於從真實事件中刪除某些非事實的因素，然後重新進行特徵的建構。並利用其特性推廣到密集預測任務 (Dense Prediction Tasks)，新設計了一種 Segmentation-based Intra and Counterfactual Loss，這種損失函數強化了目標區域內部特徵的相似性，並透過反事實注意力的方式排除誤導性特徵的干擾，進一步提升模型的準確性和穩定性。

綜上所述，提出的方法在有限的運算條件下，不僅提升了全景分割的性能，還能適應各種即時性 (real-time) 的應用場景。透過大量的實驗結果可以總結出下列幾項貢獻: 1.) 此架構透過單一預測分支完成全景分割任務的預測，並可以透過 end-to-end 的方式進行訓練; 2.) 提出的任務整合機制透過相輔相成的訓練方式有效提升分割結果的預測品質; 3.) 透過新設計的 Segmentation-based Proposal Strategy 及 Segmentation-based Intra and Counterfactual Loss，使提出的 proposal 具有意義並加強每個目標區域的特徵表示 (feature representation)，有效減少物件重複預測的問題。

doi:10.6342/NTU202404744

Chapter 2 Related Works

2.1 Panoptic Segmentation

在這個章節中，將會針對我們的目標任務—全景分割 (Panoptic Segmentation) 介紹相關的文獻。全景分割任務是一種結合了語意分割 (Semantic Segmentation) 和實例分割 (Instance Segmentation) 的視覺任務 [6, 11]。語意分割負責將圖片上的每一個像素進行分類，而實例分割則進一步區分同類別物件中的不同個體。全景分割的終極目標是同時區分出畫面中可數的個體作為前景物件 (Things)，例如，人、自行車。以及不可數的區域作為背景 (Stuff)，例如，馬路、天空。

Elharrouss 等人 [6] 對全景分割任務的網路架構進行了分類，提出了四種主要類型，包括 Sharing Backbone、Explicit Connections、Cascade Model 和 One-shot Model。首先，Sharing Backbone 的方法透過共享網路主幹，透過分出兩個獨立的子任務分支，既而進行實例分割和語意分割的預測，最後再整合兩者的預測結果。Kirillov[10] 等人提出的 PanopticFPN 和 Xiong 等人 [31] 提出的 UPSNet 是 Sharing Backbone 的典型代表。這兩種方法均使用 ResNet-FPN[9, 17] 作為主幹，在實例分割分支的部分上使用 Mask R-CNN[8] 的設計，並藉以預測前景物件的邊界框、類別和分割結果 (Segmentation Mask)，並同時在 FPN 結構中添加語意分割分支。PanopticFPN 通過簡單 up-sampling 的方式合併 FPN 所有解析度的特徵，並作為語意分割的輸出。至於 UPSNet 則是使用了 Deformable Convolutional

doi:10.6342/NTU202404744

Network[5] 的結構預測背景，最後再整合兩分支的輸出作為全景分割的預測結果。

然而，分開監督兩個任務的分支再進行整合容易出現語意分割和實例分割結果不匹配的情況，這可能會導致預測結果物件之間出現落差，或是物件彼此重疊的情況。AUNet[15] 和 Auto-Panoptic[30] 則屬於 Explicit Connections 的方法，這類方法通過特定的連接方式讓語意分割和實例分割的兩個分支能夠共享特徵，提高預測結果的一致性。Li 等人 [15] 指出，傳統 Sharing Backbone 的方法忽略前景物件和背景內容之間的潛在關係，並在 AUNet 提出兩種注意力機制，透過模組間的網路結構建立前後景之間的關聯性。Wu 等人 [30] 更進一步在多任務的網路模組上設計 Path-Priority Search Policy 的方法，考慮模組內及模組間的關聯性，並透過最佳的搜索路徑降低模型的運算量。

Cheng 等人 [3] 認為，前述兩種 top-down 的方法容易造成語意分割和實例分割之間的衝突。Panoptic-DeepLab[3] 則透過 bottom-up 的方式，先對圖片進行像素級別分類，然後基於語意分割的分類結果引導後續實例分割的過程。這個方法屬於典型的 Cascade Model，這類方法相較於前述兩種 top-down 類型的方法速度更快，但透過串連的方式先後預測類別與個體，容易導致實例分割的預測結果受限於語意分割的品質。

最近，許多先進的方法 [2, 14, 16, 27, 32, 33] 都屬於 One-shot Model，那就是直接透過相同的網路產生前景及背景的預測結果。這些方法都是基於 query 的形式，隸屬於統一前景與背景預測的方法。例如，DETR[2] 首先引入了 Transformer encoder-decoder[22] 的架構，而這方法與過往的方法不同的是，在檢測目標的過程無須預先定義錨點 (anchor)，且不需要額外的後處理策略。在此方法中，每一個 query 代表一個前景物件或是背景，在訓練時透過匈牙利算法 [12] 將預測結果與標註的正確答案 (ground truth) 進行配對。DETR 透過單個注意力模組直接考慮

doi:10.6342/NTU202404744

前景物件和背景內容之間的關聯性，在預測前後景時則使用相同的方法，上述的作法可以避免語意分割和實例分割結果不匹配的問題。然而，Li 等人 [16] 指出，由於前景物件與背景具有不同屬性，用相同的方法處理前後景不是最理想的做法，且對於區分不同前景物件的差異，引入位置資訊可以顯著提升效果。Panoptic SegFormer[16] 則是使用一個 location decoder，透過前景物件和背景的位置資訊提升分割的品質，並額外設計了一個 mask-wise 後處理策略合併前後景的分割結果。近期，DINO[34] 結合 DAB-DETR[19] 和 DN-DETR[13] 的優勢，在架構中稍微改善了可學習的錨點及去噪訓練 (De-Noising training) 方法，在物件檢測 (Object Detection) 任務上取得良好的成果。然而，Li 等人 [14] 認為 DETR 及其衍伸的架構 [2, 13, 19, 34] 產生分割的方法缺乏與主幹高分辨率的特徵互動，且僅在解碼器的最後一個 Layer 計算分割損失，如此限制分割任務的效能。MaskDINO[14] 通過添加用於分割任務的高分辨率分支對 DINO 進行功能擴充，在全景分割任務上取得優異的表現。過去，許多預測分割結果的方法是直接用檢測器預測物件的邊界框 (Bounding box)[2, 10, 15, 31]，如果分割任務端端仰賴邊界框的預測結果，容易導致分割的品質受限於檢測器的能力。MaX-DeepLab[27] 則是簡化了子任務之間的相依性，直接預測帶有類別的前後景分割結果，並使用提出的 PQ-style loss 同時監督類別與分割的預測結果。Cmt-deeplab[32] 與 kMaX-DeepLab[33] 進一步考慮像素特徵與目標物件的關係，透過像素的聚類方法來執行交叉注意力，如 kMaX-DeepLab 透過 k-means clustering 的方法聚合相似特徵的像素作為 query。

綜上所述，近期基於 query 形式的方法皆在全景分割任務取得顯著的成果。然而，這些方法通常需要極高的運算成本，將其實際應用到不同場合並不是最理想的方法。針對實時應用 (real-time applications) 的需求，本論文提出了基於 YOLOv7[25] 及 FastInst[7] 設計的 One-shot Model。本研究旨在有限的運算條件下，不僅提升全景分割的性能，並且能夠適應各種即時性 (real-time) 的應用場景。

doi:10.6342/NTU202404744

2.2 YOLOv7

YOLO 系列 [1, 24–26] 憑藉其卓越的性能，成為當今最著名的 CNN-based 物件檢測器。無論是物件偵測還是實例分割任務，YOLO 系列的方法都能在有限的運算資源環境中，不使用預訓練模型就能取得出色的表現。YOLOv7[25] 在本研究中扮演關鍵的角色，本節將介紹 YOLOv7 的方法。

Wang[23, 25] 等人設計兩種新的主幹架構—ELAN[23] 和 E-ELAN[25] 並引入 YOLOv7 中。這兩種架構解決了 CNN 網路隨深度增加容易出現的梯度消失或過度膨脹的問題。隨著網路深度的增加，單純堆疊可學習的參數模組並不能顯著提升模型的預測效能，反而會使模型在訓練過程中難以收斂。ELAN 透過改善網路結構中的梯度路徑，使用更少的運算資源產生更豐富的梯度組合。由於該架構設計能夠有效將梯度傳播到每個運算模組，因此解決了由梯度消失導致的訓練退化問題。E-ELAN 在 ELAN 的基礎上進一步引入了 expand、shuffle、merge cardinality 等技術，使網路架構在不破壞原始梯度路徑的情況下，通過更豐富的特徵資訊增強網路的學習能力。

YOLOv7 的特徵整合結構 (Neck) 沿用了 YOLOv4 的 FPN-PAN[1, 24] 輕量結構，並引入了 CSPNet 和 Spatial Pyramid Pooling(SPP) 的設計，更有效地組合梯度，因此能夠進一步透過主幹粹取的多尺度特徵產生細緻的特徵資訊。預測分支 (Head) 引入了 YOLOR[26] 的設計，能夠透過顯性和隱性的資訊，更準確地表達目標的物理特徵。

Wang[25] 等人也在 YOLOv7 中使用了幾種 Trainable bag-of-freebies 的策略。其中，re-parameterization 的策略透過整合訓練過程中被拆分到不同分支的運算模組提升推理 (inference) 階段的運算效能。Coarse-to-fine lead guided assigner 的策

doi:10.6342/NTU202404744

略，則是透過預測主分支及正確答案產生 coarse label 及 fine label 兩種 Soft Label，並利用這些 Soft Label 引導主預測分支及輔助預測分支的學習，使較淺層的輔助分支可以直接學習主分支已學習的資訊，而主分支能夠更專注於尚未學習的剩餘資訊。

在本研究中，我們遵循 YOLOv7 的方法，並將 FastInst 的 Dual-path Transformer Decoder 作為預測分支。透過提出的任務整合方法，能夠有效避免 CNN 架構在不同預測分支進行多任務訓練時所造成的特徵差異。

2.3 FastInst

FastInst[7] 是一種專為實例分割任務設計的輕量化 Query-based 方法。He 等人 [7] 指出，傳統的 CNN-based One-stage 方法能夠實現 end-to-end 的影像分割，而無須產生 Region Proposals。以 SOLO[28] 和 SOLOv2[29] 為例，SOLO 引入實例類別的觀念，區分每個像素所屬的獨立個體，這樣的做法效仿語意分割任務使實例分割任務中每個獨立個體視為像素的分類問題。儘管這種方法具有辨識上的即時性，但在面對複雜的預測場景時，仍然仰賴於人工設計的後處理步驟，例如 SOLOv2 設計的 Mask NMS。相比之下，Query-based 方法，如 DETR[2] 和 Mask2Former[4]，在訓練過程中透過 Bipartite Matching 將預測結果與正確答案關聯，並且無需 Non-maximum Suppression(NMS) 等後處理策略，從而避免模型受限於人工設計的後處理機制。儘管 Mask2Former 透過主幹上的像素編碼器和 masked-attention Transformer decoder 簡化實例分割的架構，但 masked-attention 限制每個 query 的感受野 (receptive field)，可能導致 query 的更新過程達不到最佳效果。此外，這些方法也因其靜態學習 query 的方式，需要使用大量的 Decoder Layer，且像素特徵的編碼需要龐大的架構，這不僅增加了運算資源的需

doi:10.6342/NTU202404744

求，也導致了漫長的更新過程，成為其致命缺點。FastInst 透過其三項關鍵設計—Instance activation-guided queries(IA-guided queries)、Dual-path update strategy 及 Ground truth mask-guided learning，在速度及準確性方面均有優異表現。

在架構設計上，FastInst 繼承了 Mask2Former 的基本思路，使用 ResNet 搭配 PPM-FPN 作為主幹及 Pixel Decoder，並新設計了 Dual-path Transformer Decoder。不同於 Mask2Former 僅透過 CNN 產生 pixel feature 並與靜態可學習的 embeddings 一同作為 Single-path Transformer Decoder 的輸入，FastInst 在 CNN 階段從 PPM-FPN 輸出三種不同尺寸的特徵空間，其中中尺寸及大尺寸的特徵空間分別用作 proposal feature 和 pixel feature。

Instance activation-guided query 的核心概念是從富含潛在目標特徵資訊的 proposal feature 空間中，動態地從 N_a 個具有高信心程度且不相鄰的位置選取特徵向量，這些特徵向量稱為 IA-guided proposals。在訓練過程中，proposal feature 會額外接出一個輔助分支，用以訓練每個被挑選位置的類別資訊，而其餘未被挑選的位置則監督為未標註目標 (unlabeled)。這種方法透過每個位置的機率分布判斷該位置是否存在目標，並將最大機率值視為信心程度。相較於靜態可學習的 embeddings，這種動態選取高語意特徵作為 queries 的方法，更能加速 Transformer Decoder 的收斂速度。然而，當圖像中存在大面積物件或常見類別物件時，可能會在 proposal feature 中產生多個不相鄰的高信心位置，直接從整個 proposal feature 空間動態挑選高語意特徵，可能會導致選出多個指向同一目標的 proposals，造成重複預測的問題。

FastInst 將前述 N_a 個 IA-guided proposals 與 N_b 個輔助用的可學習 embeddings 一同作為 object queries，並和 pixel feature 一同作為 Dual-path Transformer Decoder 的輸入。透過提出的 Dual-path update strategy，同時優化 queries 及像素特徵，減

doi:10.6342/NTU202404744

少對龐大像素特徵編碼器的依賴。He 等人 [7] 更在 Dual-path Transformer Decoder 的類別分支中設計了一種稱為 Ground truth mask-guided learning 的策略。在訓練期間，透過分割結果的正確答案引導每個 query，使其能夠觀察到其預測目標的完整區域，從而克服 masked-attention 造成的限制。這種方法不僅可以更直接地獲得更細緻的像素特徵，也能使每個特徵向量學習到更完整的目標資訊。因此，這不僅提升了模型的效能，還在資源使用上更加高效，進一步降低運算資源的需求。

本研究引用了 FastInst 的思路，將其從實例分割任務擴展到全景分割任務。在提出的方法中，使用 YOLOv7 的架構作為主幹及像素編碼器，並引入 Dual-path Transformer Decoder 作為預測分支。訓練過程中，採用 Ground truth mask-guided learning 策略，並透過提出的 Segmentation-based Proposal Strategy 來避免重複預測的問題。

2.4 Counterfactual Attention Learning

Counterfactual Attention Learning [21] 是一種基於因果推理增強注意力機制學習能力的新穎方法，透過比較事實 (學到的注意力特徵) 與反事實 (非真實的特徵) 對預測結果的影響，使模型能更有效地關注重要且正確的資訊，並避免受到類別不平衡數據的影響。這種方法已成功應用於視覺分類任務，如: 細粒度圖像分類 (Fine-Grained Visual Categorization) 和行人重識別 (Person Re-identification)。

Rao 等人 [21] 指出，傳統注意力機制的訓練多是基於顯性的監督方法，即根據模型的預測結果計算損失函數，並通過反向傳播 (backpropagation) 更新注意力模組的參數。這種訓練方式會忽略預測結果與注意力模組關注特徵之間的因果關係。例如，當數據集中多數樣本是以道路為背景，但分類目標卻是汽車時，模型容易將屬於道路的特徵納入判別汽車的線索，這樣會導致模型學習到誤導性的特

doi:10.6342/NTU202404744

徵，並在推理階段造成偏頗的判斷。

傳統的注意力機制透過學習到的注意力圖 (attention map) 重新加權特徵空間 (feature space)，賦予模型判別空間中重要區域的能力並學習位置之間的關連性，這被稱為事實特徵。而反事實注意力機制則透過隨機產生的虛構注意力圖對特徵空間進行加權，作為反事實特徵，用以強化注意力模組學習的過程。評估注意力好壞的方法在於，若將事實特徵扣除反事實特徵後，分類器仍能從相減過後的特徵空間判斷出物體的類別，則說明模型對該類目標學習到的事實特徵具有鑑別性。

在全景分割任務中，模型需要同時處理前景物體和背景區域的分割，這顯著增加了模型區分不同區域特徵的難度。為了因應這一挑戰，本研究加入一種新設計的反事實注意力機制。反事實注意力通常被應用於分類任務，其特性在於從真實事件中刪除某些非事實的因素，然後重新進行特徵的建構。在本研究中，我們利用這一特性推廣到密集預測任務 (Dense Prediction Tasks) 中。新設計的方法是針對每個目標區域 (Target Segmentation) 進行處理。具體來說，此方法從目標區域的特徵中扣除與該區域最相似的特徵，並要求模型從扣除後的特徵中清晰辨識出其目標區域。這一機制能有效排除誤導性特徵的干擾，幫助模型在訓練過程中更正確地理解前景物體與背景之間的關係，並強化每個目標區域特徵的鑑別性。

doi:10.6342/NTU202404744

Chapter 3　Approach

在本章節中，首先分析全景分割任務在基於 CNN-based 及 Query-based 架構時所面臨的問題。隨後，將詳細介紹本研究的架構以及提出的三種方法，包括 Tasks Integration、Segmentation-based Proposal Strategy，以及 Segmentation-based Intra and Counterfactual Loss。

3.1　Fundamental Problems in Panoptic Segmentation

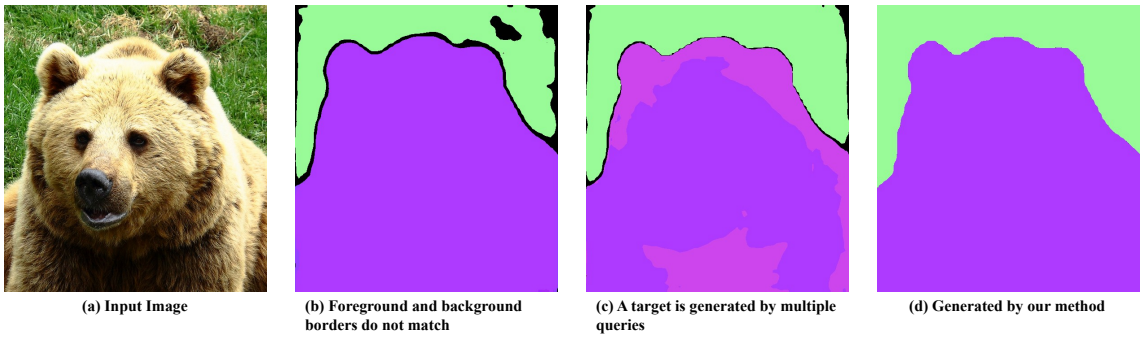


Figure 3.1: 全景分割任務在基於 CNN-based 及 Query-based 架構時所面臨的問題。(a) 原始輸入的影像。(b) 預測的前景與背景邊界不密合。(c) 前景物件被重複預測。(d) 所提出的方法大幅改善預測結果的品質。

CNN-based 的多任務模型架構通常會從主幹 (backbone) 接出不同的分支 (Prediction Head)，並針對不同的任務進行訓練。例如，使用 CNN 模型實現全景分割任務，需要一個實例分割分支來預測前景物件，以及一個語意分割分支來預測背景 [3, 10, 15, 30, 31]。然而，這種分別使用兩個分支來預測前景物件和背景的方法，容易導致物件之間重疊或前景與背景邊界不密合的問題。(參見 Figure

doi:10.6342/NTU202404744

3.1(b))

Query-based 的方法則通過 queries 與像素特徵的 cross-attention 產生前景與背景的分割結果。這種方法的挑戰在於選擇 query 的能力。由於引入位置資訊可以顯著提升前景與背景的預測結果 [16]，這意味著選擇的 query 位置會大幅影響預測品質。然而，當一個物件在畫面中占據較大面積，或該物件屬於數據集中較常見的類別時，容易出現多個 proposal 指向同一物件，導致物件重複預測的情況。(參見 Figure 3.1(c))

針對上述問題，本研究提出一種稱為 Tasks Integration 的策略，簡化模型的輸出並將其整合，以產生不同任務的預測結果，從而改善邊界不密合的問題。此外，也透過提出的 Segmentation-based Proposal Strategy 及 Segmentation-based Intra and Counterfactual Loss 改善重複預測的問題。

3.2 Architecture

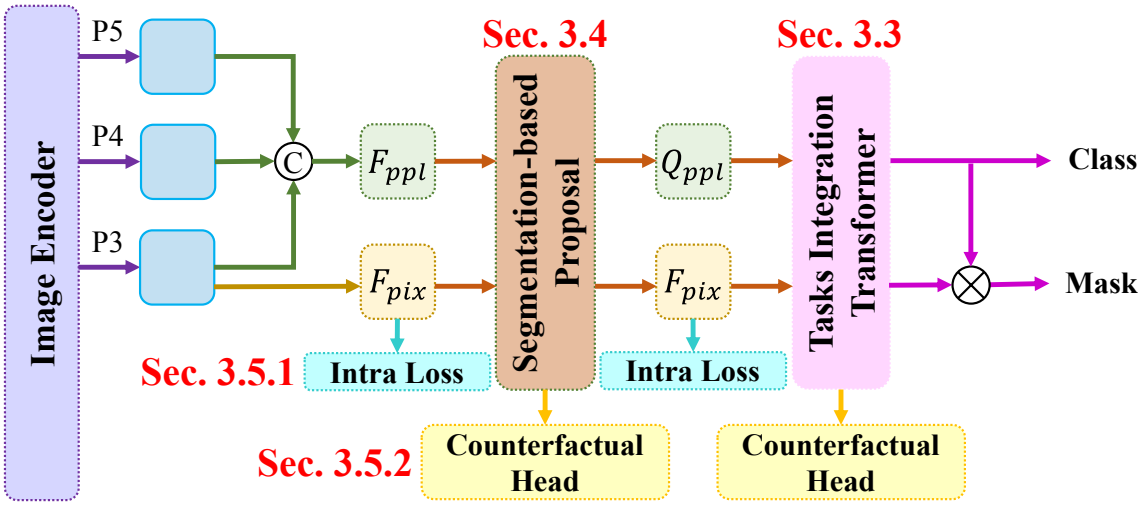


Figure 3.2: 本研究的完整架構如圖所示。P3、P4 及 P5 的三個特徵空間為 YOLOv7 PAN 的輸出；F_{ppl} 代表 Proposal Feature，F_{pix} 代表 Pixel Feature，Q_{ppl} 則代表透過 Segmentation-based Proposal Module 從 F_{ppl} 挑選出的 query。⊗ 代表 cross attention，© 則代表 feature concatenation。

本研究使用完整的 YOLOv7 架構，並將 FastInst 提出的 Dual-path Transformer

doi:10.6342/NTU202404744

Decoder 作為預測分支，在此稱為 Tasks Integration Transformer。其中，Proposal Feature 特徵空間聚合來自 YOLOv7 FPN-PAN 輸出的特徵，Pixel Feature 則是聚合 YOLOv7 用於產生實例分割結果的 prototype feature 與產生語意分割結果的特徵空間。透過 Segmentation-based Proposal Module 從 Proposal Feature 空間中挑選出來的特徵向量會作為 object queries，並一同與 Pixel Feature 作為 Tasks Integration Transformer 的輸入特徵。

提出的任務整合方法將預測分支直接輸出的類別及分割結果視為 Local Predictions，並將這些 Local Predictions 組合成語意分割任務的預測結果，稱其為 Global Prediction。Local Predictions 與 Global Prediction 視為主任務的預測結果，Local Predictions 透過與 FastInst 相同的 bipartite matching 方式配對正確答案並進行監督。主任務的損失函數表示如下:

$$\mathcal{L}_{main} = \lambda_{Class} \times \mathcal{L}_{class} + \lambda_{Mask} \times \mathcal{L}_{mask} + \lambda_{Dice} \times \mathcal{L}_{dice} + \lambda_{Semantic} \times \mathcal{L}_{semantic} \quad (3.1)$$

其中，$\mathcal{L}_{(\cdot)}$ 和 $\lambda_{(\cdot)}$ 分別代表損失函數與其權重。輔助任務包括 Segmentation-based Proposal Strategy 及 Segmentation-based Intra and Counterfactual Loss，Segmentation-based Proposal Module 使用的損失函數表示如下:

$$\mathcal{L}_{proposal} = \lambda_{Class}^{Proposal} \times \mathcal{L}_{Class}^{Proposal} + \lambda_{Mask}^{Proposal} \times \mathcal{L}_{Mask}^{Proposal} + \lambda_{Dice}^{Proposal} \times \mathcal{L}_{Dice}^{Proposal} \quad (3.2)$$

Segmentation-based Intra and Counterfactual Loss 則表示為 $\mathcal{L}_{intra\&counterfactual}$。總損失函數表示如下:

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \mathcal{L}_{proposal} + \mathcal{L}_{intra\&counterfactual} \quad (3.3)$$

doi:10.6342/NTU202404744

3.3 Tasks Integration

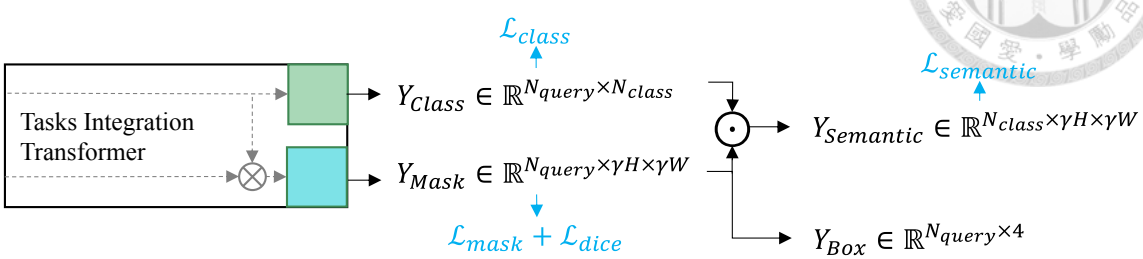


Figure 3.3: 模型的輸出為 N_{query} 張分割結果跟 N_{query} 個對應的類別。語意分割的預測結果透過類別及分割結果內積產生，物件的邊界框為分割結果的最小外接矩形。$\gamma = \frac{1}{8}$，$\otimes$ 代表 cross attention，$\odot$ 則代表內積。

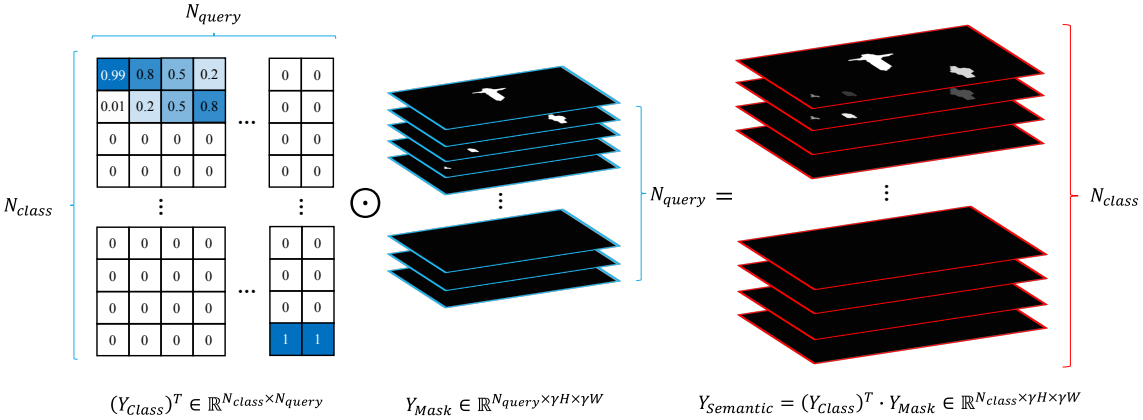


Figure 3.4: 透過類別及分割結果內積產生語意分割預測結果的示意圖。

針對物件重疊或邊界不精確的問題，我們分析了幾種可能的原因。主要原因之一是，當任務之間透過不同分支進行訓練時，即使這些分支以共享特徵的方式促進任務間彼此交互，仍然會存在特徵差異。儘管 Query-based 方法透過注意力機制整合前景和背景的預測，使模型能夠建立前景物體與背景之間的關聯性，但在物體間邊界不明顯或同類背景特徵差異過大的情況下，仍會產生物體分類錯誤、邊界錯誤或前後景邊界不密合等問題。此外，由於任務屬性差異明顯，模型在處理多任務時可能會出現相互牽制的現象。

為了解決這些問題，本研究首先簡化了模型的預測分支，使模型能夠專注於最基本的任務，避免因直接訓練多種屬性差異較大的複雜任務而導致性能受限。複雜任務的預測結果可以基於模型的基本預測結果來組成。(參見 Figure 3.3) 以

doi:10.6342/NTU202404744

全景分割任務為例，基本任務包括判斷畫面中每個像素所屬的個體，以及這些像素集合所屬的類別。給定一張彩色影像 $X \in \mathbb{R}^{3\times H\times W}$，其中 H 與 W 為影像的高度與寬度。將模型的輸出定義為 $Y_{Class} \in \mathbb{R}^{N_{query}\times N_{class}}$ 和 $Y_{Mask} \in \mathbb{R}^{N_{query}\times \gamma H\times \gamma W}$，分別表示目標的類別和分割結果。其中，$\gamma = \frac{1}{8}$，$N_{class}$ 則代表前景和背景的總類別數，輸出共有 N_{query} 對預測結果。在這個階段，模型需要區分前景物體的個體及其類別。由於前景和背景的屬性不同，背景區域只需辨識其類別，無需區分個體。然而，由於背景區域通常佔據畫面的較大部分，容易因特徵差異明顯而被視為不同的個體，即 N_{query} 對預測結果中屬於同類背景的部分可能透過多個 Query 產生。

過去產生背景的方法皆基於語意分割任務，因其特性為僅需區分每個像素的類別，無需區分前景物體的個體，這與背景區域的特徵屬性相符。由於每個預測對已經包含從畫面中區分不同個體的區域及其所屬類別，語意分割任務的預測結果可以基於上述基本任務的預測結果組成，即將每個分割結果透過內積的方式投影到該區域所屬的類別。(參見 Figure 3.4) 語意分割的預測結果可透過下列公式表示:

$$Y_{Semantic} = (\text{softmax}_{\text{dim=1}}(Y_{Class}))^T \cdot \text{softmax}_{\text{dim=0}}(Y_{Mask}) \in \mathbb{R}^{N_{class}\times \gamma H\times \gamma W} \quad (3.4)$$

而對於物件檢測任務的預測結果，可以透過前景分割結果的最小外接矩形來確定物件的邊界框。

上述的 N_{query} 對預測結果可以視為 Local Predictions，每一個 Query 的預測結果代表的可能是一個完整的前景個體，或是背景的部分或完整區域。透過 Local Predictions 直接產生前景物體，但背景部分則為透過 Local Predictions 的結果組合而成的 Global Prediction。這意味著前景物體只允許一個 Query 產生一個個體，但

doi:10.6342/NTU202404744

允許背景區域由多個 Query 產生，這樣的好處是模型不會因為背景不同地方的特徵差異而造成預測結果不完整。由於分割結果的來源相同，這也能避免物件之間重疊或前景與背景邊界不密合的問題。

在訓練過程中，對於 Local Predictions 的結果，類別分支透過 Categorical Cross Entropy Loss 監督，表示為:

$$\mathcal{L}_{class}(\text{softmax}_{\text{dim=1}}(Y_{Class})), Y_{GT_{Class}}) \tag{3.5}$$

其中 $Y_{GT_{Class}} \in \mathbb{R}^{M \times N_{class}}$ 為配對目標類別的正確答案，M 代表畫面中前景物件與背景類別的數量和。分割結果則透過 Binary Cross Entropy Loss 和 Dice Loss 進行監督，分別透過下列公式表示:

$$\mathcal{L}_{mask}(\text{sigmoid}(Y_{Mask})), Y^{\gamma}_{GT_{Mask}}) \tag{3.6}$$

$$\mathcal{L}_{dice}(\text{sigmoid}(Y_{Mask})), Y^{\gamma}_{GT_{Mask}}) \tag{3.7}$$

其中 $Y^{\gamma}_{GT_{Mask}} \in \mathbb{R}^{M \times \gamma H \times \gamma W}$ 為配對目標分割結果寬高縮小為 $\frac{1}{8}$ 的正確答案。Global Prediction 直接透過語意分割的正確答案監督，可表示為:

$$\mathcal{L}_{semantic} = \lambda^{Semantic}_{Mask} \times \mathcal{L}_{mask}(Y_{Semantic}, Y^{\gamma}_{GT_{Semantic}}) + \lambda^{Semantic}_{Dice} \times \mathcal{L}_{dice}(Y_{Semantic}, Y^{\gamma}_{GT_{Semantic}}) \tag{3.8}$$

其中 $Y^{\gamma}_{GT_{Semantic}} \in \mathbb{R}^{N_{class} \times \gamma H \times \gamma W}$ 為語意分割寬高縮小為 γ 的正確答案，$\lambda^{Semantic}_{Mask}$ 和 $\lambda^{Semantic}_{Dice}$ 表示 Binary Cross Entropy Loss 和 Dice Loss 的權重。透過上述的方式組合並訓練多任務，可以提供額外的優勢，即任務間可以產生相輔相成的效果，加速模型收斂的速度與提升預測品質。

doi:10.6342/NTU202404744

3.4 Segmentation-based Proposal Strategy

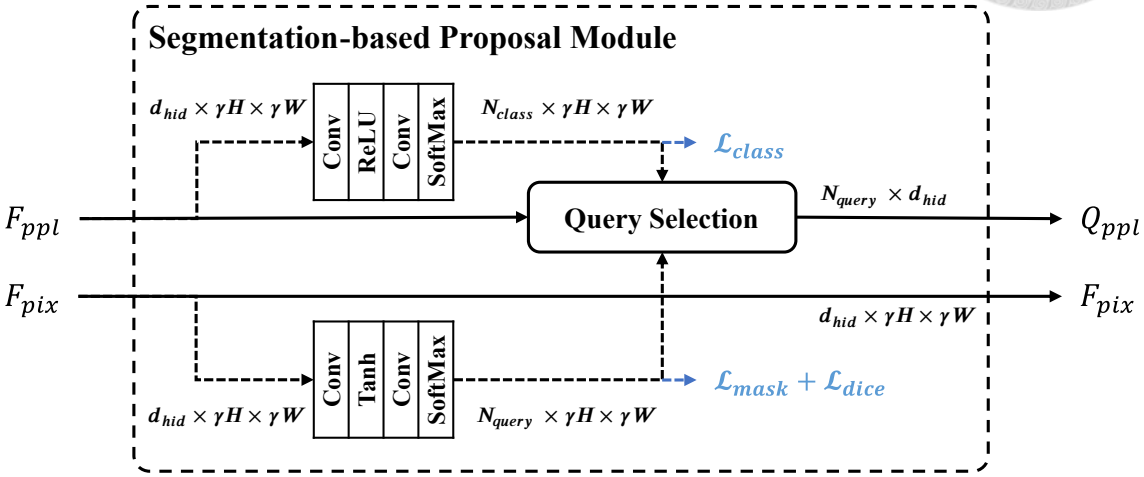


Figure 3.5: Segmentation-based Proposal Module。

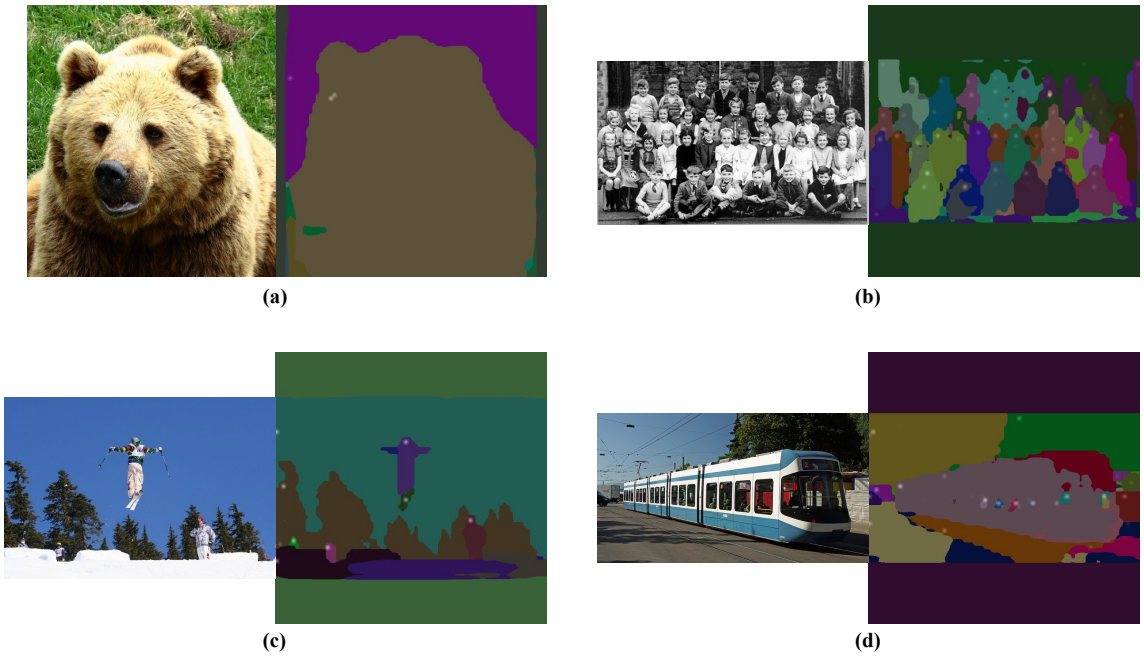


Figure 3.6: Segmentation-based Proposal Strategy 劃分的獨立區域及每個區域中高信心特徵向量的位置。每個區域以不同顏色表示，高信心特徵向量的位置以白點標示，白點的亮度與信心程度成正比，信心越高則白點越亮。

造成 Query-based 架構重複預測的主要解決方法在於避免選取指向同一目標的特徵向量作為 object queries。以 FastInst 為例，雖然其 IA-guided queries 可為 Dual-path Transformer Decoder 提供高語意特徵，但在面對大面積物件或常見的物

doi:10.6342/NTU202404744

件類別時，其基於整個 proposal feature 空間選取特徵的機制未對特徵分布進行細緻區分。若僅以信心程度作為選取條件，導致選取多個指向同一目標的高信心特徵作為 object queries，容易造成重複預測。

針對這一問題，本研究提出了 Segmentation-based Proposal Strategy。該策略透過新增的 proposal segmentation 分支將 proposal feature 劃分為 N_{query} 個獨立區域，每個區域可視為一個潛在目標個體所在的位置集合，表示為 $Y_{Segments}^{Proposal} \in \mathbb{R}^{N_{query} \times \gamma H \times \gamma W}$(參見 Figure 3.5)。在每個區域內，將會動態選取具有最高語意和信心度的特徵向量作為該區域的 proposal (參見 Figure 3.6)。這一過程避免了多個 proposals 指向同一目標的情況。因為每個區域內僅有一個代表性的特徵向量被選取，這樣的區域劃分確保了特徵的相對獨立性，能夠準確反映該區域的目標特徵，從而大幅減少了重複預測的可能性。以下將選出的 proposal 集合表示為 $Q_{ppl} \in \mathbb{R}^{N_{query} \times d_{hid}}$，每個區域所選擇的 proposal 則可透過下列公式表示:

$$Q_{ppl}(n) = F_{ppl}(Location_h(n), Location_w(n)) \in \mathbb{R}^{1 \times d_{hid}} \tag{3.9}$$

$$Location_h(n), Location_w(n) = \underset{(h,w)}{argmax}(Y_{Segments}^{Proposal}(n) \circ \max_{\text{dim=0}}(Y_{Class}^{Proposal})) \in \mathbb{Z}^2 \tag{3.10}$$

其中，$n \in \mathbb{Z} \cap [0, N_{query})$，代表第 n 個獨立區域。$F_{ppl} \in \mathbb{R}^{d_{hid} \times \gamma H \times \gamma W}$ 代表 Proposal Feature 空間，每個位置代表一個候選特徵向量，d_{hid} 則代表特徵向量的維度。$h \in \mathbb{Z} \cap [0, \gamma H)$，$w \in \mathbb{Z} \cap [0, \gamma W)$，$Location_h$ 及 $Location_w$ 分別代表空間高軸及寬軸的座標。$\circ$ 表示 element-wise product。$Y_{Class}^{Proposal} \in \mathbb{R}^{N_{class} \times \gamma H \times \gamma W}$ 代表 Proposal Feature 空間中每個位置的分類預測，可表示為:

$$Y_{Class}^{Proposal} = \text{softmax}_{\text{dim=0}}(\omega_{Class}^{Proposal} \cdot F_{ppl}) \in \mathbb{R}^{N_{class} \times \gamma H \times \gamma W} \tag{3.11}$$

$\omega_{Class}^{Proposal} \in \mathbb{R}^{N_{class} \times d_{hid}}$ 為可學習的參數。$Y_{Class}^{Proposal}$ 透過與 FastInst 相同的訓練方

doi:10.6342/NTU202404744

式，被挑選的 proposal 位置監督其對應目標的類別，其餘未被挑選的位置則監督為未標註目標。與正確答案配對的 Matching Cost 如下公式表示:

$$Cost_{Class}^{Proposal} = \lambda_{Class}^{Proposal} \times \mathcal{L}_{Class}^{Proposal}(Y_{Class}^{Proposal}, Y_{GT_{Class}}) + \lambda_{Location}^{Proposal} \times \mathcal{L}_{Location}^{Proposal} \quad (3.12)$$

並透過 Categorical Cross Entropy Loss 監督，表示為:

$$\mathcal{L}_{Class}^{Proposal}(Y_{Class}^{Proposal}, Y_{GT_{Class}}) \quad (3.13)$$

$\mathcal{L}_{Class}^{Proposal}(\cdot,\cdot)$ 和 $\mathcal{L}_{Location}^{Proposal}$ 分別為 Categorical Cross Entropy Loss 和 Location Cost，當 proposal 位置位於物件範圍內時，Location Cost 為 0，反之則為 1。$\lambda_{Class}^{Proposal}$ 及 $\lambda_{Location}^{Proposal}$ 表示 Categorical Cross Entropy Loss 及 Location Loss 的權重。

Proposal Segments 分支則透過 Supervised 及 Unsupervised 兩種方式進行訓練，$Y_{Segments}^{Proposal}$ 空間中與正確答案 $Y_{GT_{Mask}}^{\gamma}$ 配對到的預測區塊由正確答案進行監督，使其能夠正確表示前景物件的獨立區域，並鼓勵盡可能完整地表達同一類背景。與正確答案配對的 Matching Cost 表示如下:

$$Cost_{Mask}^{Proposal} = \lambda_{Mask}^{Proposal} \times \mathcal{L}_{Mask}^{Proposal}(Y_{Segments}^{Proposal}, Y_{GT_{Mask}}^{\gamma}) + \lambda_{Dice}^{Proposal} \times \mathcal{L}_{Dice}^{Proposal}(Y_{Segments}^{Proposal}, Y_{GT_{Mask}}^{\gamma}) \quad (3.14)$$

$\lambda_{Mask}^{Proposal}$ 和 $\lambda_{Dice}^{Proposal}$ 分別代表 Binary Cross Entropy Loss 和 Dice Loss 的權重。配對到的預測區塊以相同的 Binary Cross Entropy Loss 和 Dice Loss 進行監督，分別表示為:

$$\mathcal{L}_{Mask}^{Proposal}(Y_{Segments}^{Proposal}, Y_{GT_{Mask}}^{\gamma}) \quad (3.15)$$

$$\mathcal{L}_{Dice}^{Proposal}(Y_{Segments}^{Proposal}, Y_{GT_{Mask}}^{\gamma}) \quad (3.16)$$

未被配對到的預測區塊則透過整個網絡的 loss 使其具備判斷應生成區域的能力。

doi:10.6342/NTU202404744

Algorithm 1 Segmentation-based Proposal Strategy 中 Query Selection 的流程。

Input: $F_{\text{ppl}} \in \mathbb{R}^{d_{hid} \times \gamma H \times \gamma W}$, $F_{\text{pix}} \in \mathbb{R}^{d_{hid} \times \gamma H \times \gamma W}$

Output: $Q_{\text{ppl}} \in \mathbb{R}^{N_{\text{query}} \times d_{hid}}$

$proposal\_segmentations\_probs \leftarrow \text{Softmax}_{dim=0}(\text{Conv}(F_{\text{pix}})) \in \mathbb{R}^{N_{\text{query}} \times \gamma H \times \gamma W}$

$proposal\_segmentations \leftarrow \text{Onehot}_{dim=0}(proposal\_segmentations\_probs)$

$proposal\_class\_probs \leftarrow \text{Softmax}_{dim=0}(\text{Conv}(F_{\text{ppl}})) \in \mathbb{R}^{N_{\text{class}} \times \gamma H \times \gamma W}$

$proposal\_class\_onehot \leftarrow \text{Onehot}_{dim=0}(proposal\_class\_probs)$

$confidence\_score \leftarrow \text{Sum}_{dim=0}(proposal\_class\_probs \times proposal\_class\_onehot)$

for $query\_index = 0$ to N_{query} - 1 **do**

  $confidence\_score_{query\_index} \leftarrow$

  $proposal\_segmentations[query\_index, :, :] \times confidence\_score \in \mathbb{R}^{1 \times \gamma H \times \gamma W}$

  $loc\_h, loc\_w \leftarrow Max\_location(confidence\_score_{query\_index})$

  $feature\_vector_{query\_index} \leftarrow Expand\_dim_{dim=0}(F_{\text{ppl}}[:, loc\_h, loc\_w]) \in \mathbb{R}^{1 \times d_{hid}}$

  if Q_{ppl} is None **then**

    $Q_{\text{ppl}} \leftarrow feature\_vector_{query\_index}$

  else

    $Q_{\text{ppl}} \leftarrow \text{Concat}_{dim=0}(Q_{\text{ppl}}, feature\_vector_{query\_index})$

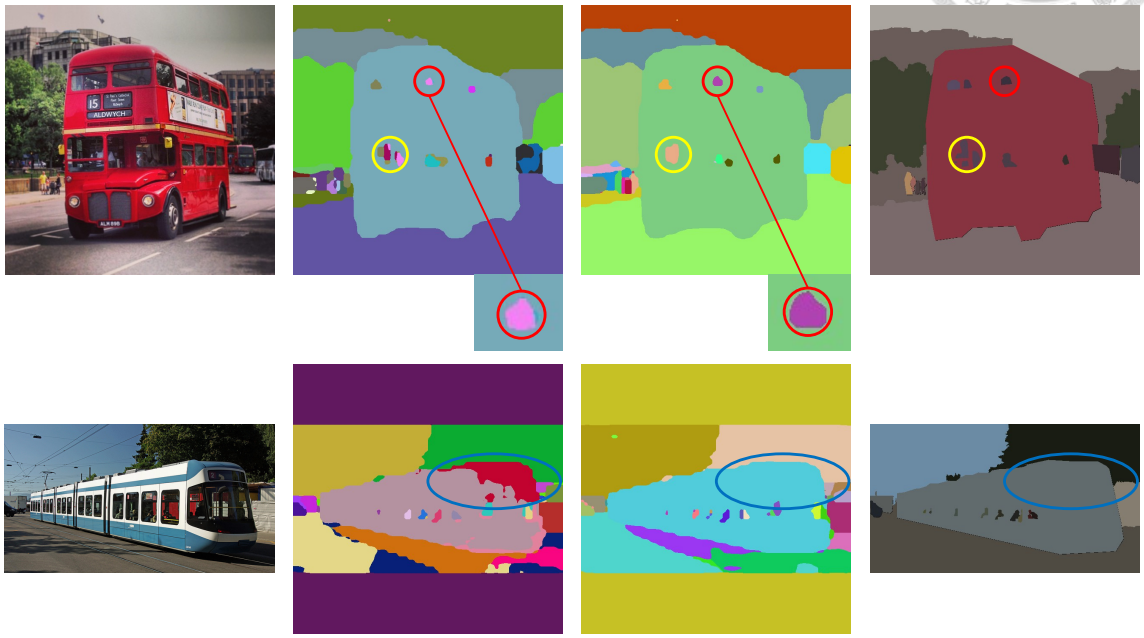
  end if

end for

return Q_{ppl}

doi:10.6342/NTU202404744

3.5 Segmentation-based Intra and Counterfactual Loss



(a) Input Image

(b) Proposal Segmentations predicted by the model trained without Segmentation-based Intra and Counterfactual Loss

(c) Proposal Segmentation predicted by the model trained with Segmentation-based Intra and Counterfactual Loss

(d) Ground Truth

Figure 3.7: 比較有無使用 Segmentation-based Intra and Counterfactual Loss 訓練模型的 Proposal Segmentation 預測結果。

在 Proposal 階段，區域分割的品質對模型最終預測的品質具有重大影響。例如，在上一章節中，本研究提出了 Segmentation-based Proposal Strategy，透過顯性的方式訓練模型，使其具備對獨立區域進行粗分割的能力。然而，僅依靠顯性訓練，當同一目標區域 (Target Segmentation) 內部特徵差異較大或不同目標特徵之間差異較小時，分割結果的品質可能仍不理想。如 Figure 3.7(b) 所示，單一目標被錯誤地判斷為多個個體，增加了重複預測的風險。如果在最終結果的預測階段，模型的預測分支無法產生具鑑別性的分割特徵，則也會直接影響預測品質。

以往的研究中，為了解決不同個體間無法產生鑑別性特徵的問題，許多文獻採用了著名的對比學習 (Contrastive Learning) 的方式。例如，透過 Intra Loss 使同類樣本的特徵在特徵空間中更加相似，而對於不同類別的樣本，則利用 Inter Loss

doi:10.6342/NTU202404744

拉大它們在特徵空間中的距離。然而，因為 Inter Loss 的特性會直接將每一個獨立個體的特徵彼此拉遠，在多類別多物件的全景分割任務引入 Inter Loss 反而容易破壞同類別不同個體物件的特徵。在本研究中，我們設計的 Segmentation-based Intra and Counterfactual Loss 進一步透過隱性方式促使同一目標的內部特徵更加一致，並透過反事實的機制增強不同目標特徵之間的差異，使每個區域隱含具有鑑別性的特徵。正如 Figure 3.7(c) 所示，單一目標在 proposal 階段已經能夠成功被判斷為一個獨立個體。這個損失函數可表示為:

$$\mathcal{L}_{intra\&counterfactual}(F_{pix}, F_{counterfactual}, Y^{\gamma}_{GT_{Mask}}) = \lambda_{intra} \times \mathcal{L}_{intra}(F_{pix}, Y^{\gamma}_{GT_{Mask}}) + \lambda_{counterfactual} \times \mathcal{L}_{counterfactual}(F_{pix}, F_{counterfactual}, Y^{\gamma}_{GT_{Mask}}) \tag{3.17}$$

以下將 Segmentation-based Intra and Counterfactual Loss 分為 Segmentation-based Intra Loss 和 Segmentation-based Counterfactual Loss 兩個部份，公式3.17中的 λ_{intra} 和 $\lambda_{counterfactual}$ 則為兩者的權重。在接下來的章節將詳細介紹這兩部分損失函數的具體設計。

3.5.1 Segmentation-based Intra Loss

Segmentation-based Intra Loss 的設計目的是確保同一目標區域內的像素特徵更加一致。該損失函數的基本思路是，對於每個預測區域，我們期望該區域內部所有的特徵向量在高維特徵空間中具有較小的距離，從而增強同一目標內部特徵的一致性，並有效減少分割過程中同一物體被錯誤分割為多個個體的風險。

具體而言，此方法對每個預測區域內的特徵向量計算均值向量，視為該區域的宏觀特徵，並將每個特徵向量與該均值向量之間的歐氏距離作為損失。損失函

doi:10.6342/NTU202404744

數的定義如下:

$$\mathcal{L}_{intra}(F_{pix}, Y^{\gamma}_{GT_{Mask}}) = \frac{1}{\gamma H \times \gamma W} \sum_{h=0}^{\gamma H-1} \sum_{w=0}^{\gamma W-1} \|F_{pix}(h,w) - F_{mean}(h,w)\|_2 \quad (3.18)$$

其中，$F_{pix} \in \mathbb{R}^{d_{hid} \times \gamma H \times \gamma W}$ 代表像素特徵空間。$F_{mean} \in \mathbb{R}^{d_{hid} \times \gamma H \times \gamma W}$ 透過 $Y^{\gamma}_{GT_{Mask}}$ 將均值向量集合 $S_{mean} \in \mathbb{R}^{M \times d_{hid}}$ 投影回像素特徵空間，該空間中的每一個位置代表該位置所屬目標的均值向量，透過矩陣計算可表示為:

$$F_{mean} = (S_{mean})^T \cdot Y^{\gamma}_{GT_{Mask}} \in \mathbb{R}^{d_{hid} \times \gamma H \times \gamma W} \quad (3.19)$$

每個區域的均值向量 $S_{mean}(m) \in \mathbb{R}^{d_{hid}}$，透過以下公式計算:

$$S_{mean}(m) = \frac{1}{S_{area}(m)} \sum_{h=0}^{\gamma H-1} \sum_{w=0}^{\gamma W-1} Y^{\gamma}_{GT_{Mask}(m)}(h,w) \times F_{pix}(h,w) \in \mathbb{R}^{d_{hid}} \quad (3.20)$$

$$S_{area}(m) = \sum_{h=0}^{\gamma H-1} \sum_{w=0}^{\gamma W-1} Y^{\gamma}_{GT_{Mask}(m)}(h,w) \in \mathbb{Z} \cap [1, \gamma H \times \gamma W] \quad (3.21)$$

其中，$S_{area}(m)$ 代表第 m 個目標區域的像素數量，透過最小化 $\mathcal{L}_{intra}(\cdot,\cdot)$ 可以促使同一目標區域內的特徵向量更加接近，以增強其內部特徵的一致性。

3.5.2 Segmentation-based Counterfactual Loss

Segmentation-based Counterfactual Loss 的設計目的是增加不同目標之間特徵的差異性，使每個區域隱含具有鑑別性的特徵。具體而言，我們期望每個區域的事實特徵減去該區域的反事實特徵後，分類器依然能夠從相減後的特徵空間中準確判定出該目標所屬的完整區域。

換句話說，若分類器能夠透過反事實干預的特徵空間產生完整的目標區域，則說明原本的事實特徵中隱含具有鑑別性的特徵。因此，透過這種方式可以增強

doi:10.6342/NTU202404744

不同目標之間的區別性，提升模型對不同目標的識別能力。損失函數可表示如下:

$$\mathcal{L}_{counterfactual}(F_{pix}, F_{counterfactual}, Y^{\gamma}_{GT_{Mask}}) = \mathcal{L}_{mask}(\omega_{\Theta} \cdot (F_{pix} - F_{counterfactual}), Y^{\gamma}_{GT_{Mask}}) + \mathcal{L}_{dice}(\omega_{\Theta} \cdot (F_{pix} - F_{counterfactual}), Y^{\gamma}_{GT_{Mask}}) \tag{3.22}$$

其中，$\omega_{\Theta} \in \mathbb{R}^{N_{query} \times d_{hid}}$ 為可學習的參數，將特徵空間投影至 N_{query} 個分割結果。$F_{counterfactual} \in \mathbb{R}^{d_{hid} \times \gamma H \times \gamma W}$ 透過 $Y^{\gamma}_{GT_{Mask}}$ 將反事實特徵向量集合 $S_{counterfactual} \in \mathbb{R}^{M \times d_{hid}}$ 投影回像素特徵空間，該空間中的每一個位置代表該位置所屬目標的反事實特徵向量，透過矩陣計算可表示為:

$$F_{counterfactual} = (S_{counterfactual})^{T} \cdot Y^{\gamma}_{GT_{Mask}} \in \mathbb{R}^{d_{hid} \times \gamma H \times \gamma W} \tag{3.23}$$

反事實特徵，即非事實特徵，在 Counterfactual Attention Learning[21] 中被定義為隨機特徵空間。在本研究中，將每個目標區域的反事實特徵向量 $S_{counterfactual}(m) \in \mathbb{R}^{d_{hid}}$ 定義為與該區域宏觀特徵最相似的其他目標區域之宏觀特徵，相似性以最小歐氏距離衡量，公式如下:

$$S_{counterfactual}(m) = S_{mean}(CFIndex(m)) \in \mathbb{R}^{d_{hid}} \tag{3.24}$$

$$CFIndex(m) = \underset{j}{argmin}(\|S_{mean}(m) - S_{mean}(j)\|_2) \in \mathbb{Z} \cap [0, M) \tag{3.25}$$

其中，$m \in \mathbb{Z} \cap [0, M)$，代表第 m 個目標區域。$j \in \mathbb{Z} \cap [0, M)$ 且 $j \neq m$，代表與第 m 個目標區域不同的其他目標區域。透過最小化 $\mathcal{L}_{counterfactual}(\cdot, \cdot, \cdot)$ 使得每個區域在高維特徵空間中更具鑑別性，這樣的設計能夠有效的提高模型處理複雜場景時的準確性和可靠性。

doi:10.6342/NTU202404744

Chapter 4　Experiments

為了驗證所提出方法的有效性，本研究專注於 COCO 資料集 [18] 進行了一系列實驗，並與其他相關方法進行了比較。此外，我們也進行了三種提出方法（Tasks Integration、Segmentation-based Proposal Strategy 以及 Segmentation-based Intra and Counterfactual Loss）的消融研究。最後，透過預測的分割結果與原始圖片進行比對，更直觀的展示預測的品質。

4.1　Implementation Details and Evaluation Protocols

COCO 資料集涵蓋了豐富的場景和物件類別，是用來評估全景分割、語意分割、實例分割、物件檢測和圖片標註 (Caption) 等多項任務演算法的標準資料集。該資料集包含 80 類前景物件及 92 類背景區域，構成了完整且具挑戰性的訓練數據。其中，在全景分割任務的標註資料中，92 類背景經過合併後共計形成 53 類背景區域。在訓練過程中，我們充分利用了 COCO Training set 進行模型的訓練，涵蓋了完整的 172 個類別，確保模型能夠學習到每個類別的特徵。於推論階段，將模型輸出的 92 類背景類別進一步合併，以符合全景分割任務的 53 類背景區域格式。

為了客觀評估本研究提出之全景分割方法的性能，以下實驗皆使用 COCO Validation set 進行評估，並依照標準的全景分割評估指標進行分析，包括 Panoptic

doi:10.6342/NTU202404744

Quality (PQ)、Segmentation Quality (SQ) 和 Recognition Quality (RQ)。其中，PQ 是 SQ 和 RQ 的乘積，SQ 量化了分割結果與其正確答案的擬合程度，而 RQ 則衡量目標物件是否被正確識別。為了更具體地表達前景與背景的預測效果，以下實驗分別使用 PQ_{th} 和 PQ_{st} 來代表前景物件與背景區域的評估指標。PQ 可透過以下公式表示:

$$PQ = SQ \times RQ \tag{4.1}$$

$$SQ = \frac{\sum_{(p,g)\in TP} IoU(p,g)}{|TP|} \tag{4.2}$$

$$RQ = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{4.3}$$

其中，p 代表預測結果，g 代表正確答案。

本研究在 anchor-free 版本的 YOLOv7 [25] 的基礎上進行實驗，訓練策略主要遵循 YOLOv7 的設計，並在 8 個 Nvidia V100 GPU 的環境下進行 60 個 Epoch 的訓練。以下實驗將 batch size 設為 56，透過 AdamW[20] 優化器在不使用 ImageNet pretrained model 的情況下訓練模型，訓練過程使用 YOLOv7 中的 Flat Cosine Learning Rate Scheduler 動態調整學習率，初始學習率為 3×10^{-4}，終端學習率則為 3×10^{-5}。此外，在訓練過程中，預設的輸入影像的大小均調整為 640×640，並採用 $\frac{1}{8}$ 倍輸入影像邊長的特徵空間 (即 80×80) 產生 proposal 以及分割結果。

4.2 Results and Comparisons

為了更具體且全面地評估所提出的方法在全景分割任務中的性能，我們進行了多項實驗，並將結果與現有的主流方法進行了詳細比較，結果如 Table 4.1所示。該表總結了本研究提出的方法與其他經典方法在 COCO Validation set 上的性能差異。本研究在最嚴格的設定下，以最輕量的架構、最小的輸入影像大小、最少

doi:10.6342/NTU202404744

Table 4.1: 本研究提出的方法與其他相關文獻的比較。其中，**Dec.** 代表使用的 Decoder Layer 數量，**Pre.** 代表是否使用了 ImageNet pretrained model 進行訓練，† 代表使用 $\frac{1}{4}$ 倍輸入影像邊長的 proposal 特徵空間。

	Backbone	Input Size	Dec.	Pre.	$PQ\uparrow$	$PQ_{th}\uparrow$	$PQ_{st}\uparrow$	Parms.
CNN-based								
PanopticFPN [10]	ResNet-50	640 or 800	-	O	39.0	45.9	28.7	-
AUNet [15]	ResNet-50	600	-	O	39.6	49.1	25.2	-
UPSNet [31]	ResNet-50	800×1333	-	O	42.5	48.5	33.4	-
Auto-Panoptic [30]	ResNet-50	-	-	O	44.8	51.4	35.0	-
Panoptic-DeepLab [3]	Xception-71	1025	-	O	41.2	44.9	35.7	-
Query-based								
DETR [2]	ResNet-101	(480∼800, <1333)	6	O	45.1	50.5	37.0	61.8M
Panoptic SegFormer [16]	ResNet-50	(480∼800, ≤1333)	6	O	49.6	54.4	42.4	51.0M
Panoptic SegFormer [16]	Swin-L	(480∼800, ≤1333)	6	O	55.8	61.7	46.9	221.4M
MaskDINO [14]	ResNet-50	1024	-	-	53.0	59.1	43.9	52.0M
MaskDINO [14]	Swin-L	1024	-	-	58.3	65.1	48.0	223.0M
MaX-DeepLab [27]	MaX-S	1025	1	X	48.4	53.0	41.5	61.9M
MaX-DeepLab [27]	MaX-L	1025	3	X	51.1	57.0	42.2	451.0M
Cmt-deeplab [32]	Axial-R50	1281	6	O	53.0	57.7	45.9	94.9M
Cmt-deeplab [32]	Axial-R104-RFN	1281	6	O	55.3	61.0	46.6	270.0M
kMaX-DeepLab [33]	ResNet-50	1281	6	O	53.0	58.3	44.9	57.0M
kMaX-DeepLab [33]	ConvNeXt-L	1281	6	O	57.9	64.0	48.6	232.0M
Ours	ELAN	640	1	X	45.1	49.3	38.7	50.8M
Ours	ELAN	960	1	X	46.2	51.1	38.7	50.8M
Ours†	ELAN	640	1	X	46.3	51.2	39.0	52.1M

Table 4.2: 比較本研究提出的方法之消融實驗。粗體字代表模型比較中的優勝者。

	All			Things			Stuff					
Algorithm	$PQ\uparrow$	$SQ\uparrow$	$RQ\uparrow$	$PQ_{th}\uparrow$	$SQ_{th}\uparrow$	$RQ_{th}\uparrow$	$PQ_{st}\uparrow$	$SQ_{st}\uparrow$	$RQ_{st}\uparrow$	**Parms.**	**FLOPs**	**FPS**
A: YOLOv7	37.8	78.1	45.4	45.8	**83.5**	53.5	25.9	69.9	33.1	46.0M	175.9G	120.20
B: A + Fastinst	12.0	49.0	17.3	19.6	67.2	28.2	0.6	21.5	1.0	47.7M	145.5G	100.61
C: B + Tasks Integration	40.8	75.9	51.1	46.3	77.3	58.1	32.4	73.7	40.7	47.7M	145.5G	113.08
D: C + Segmentation-based Proposal	44.4	79.2	54.8	47.9	79.9	59.2	**39.1**	78.1	**48.1**	50.8M	153.3G	20.26
E: D + Intra Loss	44.9	79.0	55.3	49.1	79.6	60.6	38.6	78.2	47.4	50.8M	153.3G	20.27
F: E + Counterfactual Loss	**45.1**	**79.4**	**55.6**	**49.3**	79.6	**60.8**	38.7	**79.1**	47.6	50.8M	153.3G	20.26

的 Decoder layer 數，且不依賴 ImageNet pretrained model 的情況下，依然取得了良好的表現。與 CNN-based 的模型如 PanopticFPN [10]、AUNet [15]、UPSNet [31]、Auto-Panoptic [30] 及 Panoptic-DeepLab [3] 等方法相比，本研究在不依賴大型預訓練模型的情況下，仍能在相同或更小的輸入影像大小下超越這些方法的表現。相較於 Query-based 方法，本研究以更輕量的架構和最少的 Decoder 層數，實現了與這些方法相當的性能，同時在推論階段仍能以 20.26 FPS 的速度運行，展現了良好的效率與實際應用潛力。

接著，Table 4.2比較了所提出方法對最終性能的影響。本次實驗共包含六種不同的 Algorithm，其中 Algorithm A 作為 baseline，Algorithm B 至 Algorithm F 則逐步加入了 FastInst [7] 的 Dual-path Transformer Decoder、Tasks Integration

doi:10.6342/NTU202404744

Strategy、Segmentation-based Proposal Strategy，以及 Segmentation-based Intra and Counterfactual Loss。

首先，Algorithm A 使用了 YOLOv7 [25] 的原始架構，並額外增加了一個語意分割任務的分支。此架構在輸出時會整合主預測分支與語意分割分支，分別產生前景和背景的分割預測結果。Algorithm B 引入了 FastInst 的 Dual-path Transformer Decoder，將架構轉換為 Query-based 的單預測分支，以解決 CNN-based 多任務模型在前後景邊界上不密合的問題 (參見 Figure 4.1(a))。然而，FastInst 從 proposal 空間中選擇 top-k 高信心位置作為 query，卻缺乏區域區分的機制，導致多個高信心的 proposal 指向同一目標，而相對較低信心程度的背景 proposal 無法被選擇，最終出現前景物體重複預測而背景未能正確偵測的情況 (參見 Figure 4.1(b))。Algorithm C 使用了提出的任務整合策略，透過多任務的正確解答全域監督所有 query，使任務間相輔相成，避免了因屬性差異過大而造成的任務牽制。根據 Figure 4.1(c) 和 Table 4.2(C)，該方法有效改善了如 Figure 4.1(a) 中前後景邊界不密合的問題，以及如 Figure 4.1(b) 中冗餘 query 造成前景重複預測但背景無法被偵測的現象。Algorithm D 進一步引入 Segmentation-based Proposal Strategy，透過區分畫面中每個獨立區塊，在每個 proposal segmentation 中只選出一個 proposal，避免高信心 proposal 過度佔用 query 的問題。從 Figure 4.1(d) 和 Table 4.2(D) 可以觀察到，使用這個策略產生 proposal 可以使畫面中的分割結果呈現的更完整，並使許多先前未偵測的區域得以正確預測。最後，Algorithm E 通過 Segmentation-based Intra Loss 強化了各分割區域內的特徵一致性，減少了分割品質不佳的現象，如 Figure 4.1(c)、(d) 中分割結果出現破洞的問題 (參見 Figure 4.1(e))。Algorithm F 則利用 Segmentation-based Counterfactual Loss，促使不同分割區域產生具鑑別性的特徵，顯著提升了分割結果的品質 (參見 Figure 4.1(f))。

此外，針對 Tasks Integration 和 Segmentation-based Intra and Counterfactual

doi:10.6342/NTU202404744

Figure 4.1: 透過視覺化比較本研究提出的方法。

Loss，我們額外設計了兩組消融實驗。Table 4.3比較了在 Tasks Integration 策略中是否使用 Bounding Box Loss 對性能的影響。由於邊界框能引導模型更有效地聚焦於目標物體的範圍，因此在監督邊界框的情況下，分割結果與正確答案的擬合程度通常會高於未使用時的結果，進而提高 SQ。然而，Tasks Integration Transformer 透過原始輸出分割結果的最小外接矩形產生其目標物體的邊界框，這種方法在訓練過程中難以透過邊界框的回歸對其他任務有正面影響，且由於任務屬性的差異，可能會造成任務間相互制約，進而降低物件檢測率，導致 RQ 下降。根據實驗結果，不使用監督邊界框的策略更能使模型專注於主要任務，並因此提升整體性能。

在 Table 4.4中，我們進行了另一組針對提出的 Segmentation-based Intra and Counterfactual Loss 的消融實驗。分別比較了未使用此損失函數、使用傳統反事實

doi:10.6342/NTU202404744

注意力機制 [21]，以及使用該損失函數的效果。傳統的反事實注意力機制雖然可以通過雙分支注意力模組在一定程度上提升模型的性能，但同時增加了計算資源的需求和模型複雜度。相較之下，所提出的方法針對每個不同的分割區域，為其指定對應的反事實特徵，僅依賴損失函數實現分割區域之間的反事實注意能力。這使得在不增加額外參數量的情況下，顯著提升了分割結果的準確性 (參見 Table 4.4(I))。與傳統方法不同，所提出的方法確保了每個區域的反事實特徵具有對應的物理意義，並且進一步加強各分割區域內特徵的一致性，而不同於傳統反事實方法依賴無具體物理意義的全域隨機特徵，因此能夠有效避免隨機特徵對分割表現的潛在干擾，使得模型表現得到更為有效的提升。

最後，Table 4.5比較了使用不同大小的輸入影像及不同尺寸的 proposal 特徵空間對性能的影響，使用較大的輸入影像或較大的 proposal 特徵空間皆對預測結果有正面影響。

Table 4.3: 比較有無監督 Bounding Box Loss 之消融實驗。粗體字代表模型比較中的優勝者。

	All			Things			Stuff		
Algorithm	$PQ\uparrow$	$SQ\uparrow$	$RQ\uparrow$	$PQ_{th}\uparrow$	$SQ_{th}\uparrow$	$RQ_{th}\uparrow$	$PQ_{st}\uparrow$	$SQ_{st}\uparrow$	$RQ_{st}\uparrow$
G: C + Box Loss	38.7	**76.1**	48.7	43.4	**77.5**	54.8	31.6	**74.0**	39.6
C: B + Tasks Integration	**40.8**	75.9	**51.1**	**46.3**	77.3	**58.1**	**32.4**	73.7	**40.7**

Table 4.4: 比較有無使用"Segmentation-based Intra and Counterfactual Loss" 及使用傳統反事實注意力機制之消融實驗。粗體字代表模型比較中的優勝者。

	All			Things			Stuff					
Algorithm	$PQ\uparrow$	$SQ\uparrow$	$RQ\uparrow$	$PQ_{th}\uparrow$	$SQ_{th}\uparrow$	$RQ_{th}\uparrow$	$PQ_{st}\uparrow$	$SQ_{st}\uparrow$	$RQ_{st}\uparrow$	**Parms.**	**FLOPs**	**FPS**
C: B + Tasks Integration	40.8	75.9	51.1	46.3	77.3	58.1	32.4	73.7	40.7	47.7M	145.5G	113.08
H: C + Original CF Branch	41.5	77.4	51.8	46.8	78.8	58.4	33.6	75.4	41.7	53.9M	144.1G	106.00
I: C + Intra & Counterfactual Loss	43.0	78.4	53.3	46.5	79.0	57.8	37.6	77.5	46.5	47.7 M	143.9 G	109.66
F: I + Segmentation-based Proposal	**45.1**	**79.4**	**55.6**	**49.3**	**79.6**	**60.8**	**38.7**	**79.1**	**47.6**	50.8M	153.3G	20.26

Table 4.5: 比較不同影像的輸入尺寸與不同的 proposal 特徵空間之消融實驗。

Algorithm: F		All			Things			Stuff					
Input Size	**Scale**	$PQ\uparrow$	$SQ\uparrow$	$RQ\uparrow$	$PQ_{th}\uparrow$	$SQ_{th}\uparrow$	$RQ_{th}\uparrow$	$PQ_{st}\uparrow$	$SQ_{st}\uparrow$	$RQ_{st}\uparrow$	**Parms.**	**FLOPs**	**FPS**
640	1/8	45.1	79.4	55.6	49.3	79.6	60.8	38.7	79.1	47.6	50.8M	153.3G	20.26
960	1/8	46.2	80.3	56.4	51.1	80.8	62.4	38.7	79.6	47.5	50.8M	345.0G	11.29
640	1/4	46.3	80.7	56.3	51.2	81.2	62.0	39.0	79.9	47.9	52.1M	254.3G	9.34

doi:10.6342/NTU202404744

(a) Images (b) Merged Results (c) Predictions (d) Ground Truth

Figure 4.2: 透過本研究提出的方法產生的分割結果與原始圖片進行比對。

doi:10.6342/NTU202404744

Chapter 5　Conclusions

全景分割任務同時對圖像中的可數物體及背景進行完整的理解與標註，在自動駕駛及醫療影像分析等領域具有顯著的應用價值，準確且高效的分割結果對於提升系統的可靠性至關重要。然而，現有的全景分割方法面臨諸多挑戰，如前後景邊界不一致、重複預測以及高計算資源需求等問題。

針對上述挑戰，本研究提出了一套新穎的全景分割方法。基於 YOLOv7 [25] 的架構，引入了 Tasks Integration Transformer，通過單一預測分支完成全景分割，不僅有效解決了 CNN-based 多分支架構導致的前景與背景邊界不一致問題，還透過全域監督的方式，使多任務預測能相輔相成。並進一步提出了 Segmentation-based Proposal Strategy 和 Segmentation-based Intra and Counterfactual Loss，這些方法使不同目標區域能生成具鑑別性的特徵，並確保生成的 proposal 具有明確的物理意義，有效避免了 Query-based 架構下的重複預測問題，顯著提升了分割結果的品質。在 COCO Validation set 上的實驗結果顯示，所提出的方法在保持輕量化架構的前提下，較 Baseline 模型在 Panoptic Quality (PQ) 指標上提升了 7.3%，並且在推理階段能以 20.26 FPS 的速度運行。

儘管所提出的方法在計算資源與性能之間取得了良好平衡，但仍存在一些限制需要進一步克服。例如，當處理包含大量小物體的極端複雜場景時，模型的分割精度可能下降。此外，為驗證方法的通用性，仍需在其他相關數據集上進行更

doi:10.6342/NTU202404744

廣泛的測試。未來的研究將重點聚焦於提升在複雜場景下小物體的分割表現，並計畫擴展至更多資料集和不同框架，以進一步檢驗通用性及泛化能力。

doi:10.6342/NTU202404744

References

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020.

[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In European conference on computer vision, pages 213–229. Springer, 2020.

[3] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12475–12485, 2020.

[4] 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, pages 1290–1299, 2022.

[5] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 764–773, 2017.

[6] O. Elharrouss, S. Al-Maadeed, N. Subramanian, N. Ottakath, N. Almaadeed, and

doi:10.6342/NTU202404744

Y. Himeur. Panoptic segmentation: A review. arXiv preprint arXiv:2111.10250, 2021.

[7] J. He, P. Li, Y. Geng, and X. Xie. Fastinst: A simple query-based model for real-time instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23663–23672, 2023.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[10] A. Kirillov, R. Girshick, K. He, and P. Dollár. Panoptic feature pyramid networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6399–6408, 2019.

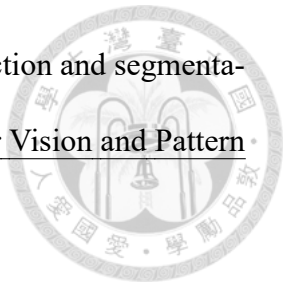
[11] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9404–9413, 2019.

[12] H. W. Kuhn. The hungarian method for the assignment problem. Naval research logistics quarterly, 2(1-2):83–97, 1955.

[13] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 13619–13627, 2022.

[14] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum. Mask dino:

doi:10.6342/NTU202404744



Towards a unified transformer-based framework for object detection and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3041–3050, 2023.

[15] Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, and X. Wang. Attention-guided unified network for panoptic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7026–7035, 2019.

[16] Z. Li, W. Wang, E. Xie, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, and T. Lu. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1280–1289, 2022.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014.

[19] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329, 2022.

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

doi:10.6342/NTU202404744

[21] Y. Rao, G. Chen, J. Lu, and J. Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1025–1034, 2021.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[23] C. Wang, H. Liao, and I. Yeh. Designing network design strategies through gradient path analysis. arxiv 2022. arXiv preprint arXiv:2211.04800.

[24] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. Scaled-yolov4: Scaling cross stage partial network. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pages 13029–13038, 2021.

[25] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7464–7475, 2023.

[26] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao. You only learn one representation: Unified network for multiple tasks. arXiv preprint arXiv:2105.04206, 2021.

[27] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5463–5474, 2021.

[28] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li. Solo: Segmenting objects by locations. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16, pages 649–665. Springer, 2020.

doi:10.6342/NTU202404744

[29] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen. Solov2: Dynamic and fast instance segmentation. Advances in Neural information processing systems, 33:17721–17732, 2020.

[30] Y. Wu, G. Zhang, H. Xu, X. Liang, and L. Lin. Auto-panoptic: Cooperative multi-component architecture search for panoptic segmentation. Advances in neural information processing systems, 33:20508–20519, 2020.

[31] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun. Upsnet: A unified panoptic segmentation network. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8818–8826, 2019.

[32] Q. Yu, H. Wang, D. Kim, S. Qiao, M. Collins, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen. Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2560–2570, 2022.

[33] Q. Yu, H. Wang, S. Qiao, M. Collins, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen. kmax-deeplab: k-means mask transformer. arXiv preprint arXiv:2207.04044, 2022.

[34] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605, 2022.

doi:10.6342/NTU202404744