Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

ISTI Technical Reports

Studio e analisi delle architetture di reti convolutive

Davide Moroni, ISTI-CNR, Pisa, Italy
Oscar Papini, ISTI-CNR, Pisa, Italy
Maria Antonietta Pascali, ISTI-CNR, Pisa, Italy
Gabriele Pieri, ISTI-CNR, Pisa, Italy
Marco Reggiannini, ISTI-CNR, Pisa, Italy

ISTI-TR-2022/022

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

Studio e analisi delle architetture di reti convolutive

Moroni D., Papini O., Pascali M.A., Pieri G., Reggiannini M.

ISTI-TR-2022/022

Questo rapporto tecnico di progetto è il risultato del contributo fornito dal Laboratorio Segnali e Immagini dell'ISTI-CNR per il documento di progetto RTOD-SYS-SDD-010-INT per il progetto RTOD (Real-Time Object Detection mediante Machine Learning basato su tecnologia Low-Power GPU). In particolare, il rapporto studia e discute delle varie possibilità di architetture di reti convolutive che sono state valutate e che potranno essere utilizzate nel contesto del progetto per effettuare delle categorizzazioni di immagini mediante algoritmi di machine learning.

Keywords: Architetture, Reti convolutive.

Citation

Moroni D., Papini O., Pascali M.A., Pieri G., Reggiannini M., Studio e analisi delle architetture di reti convolutive. ISTI Technical Reports 2022/022. DOI: 10.32079/ISTI-TR-2022/022.

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1
56124 Pisa Italy
http://www.isti.cnr.it

Studio e analisi delle architetture di reti convolutive

Davide Moroni, Oscar Papini, Maria Antonietta Pascali, Gabriele Pieri, Marco Reggiannini

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

CNR - Pisa

Sommario

Abstract

Questo rapporto tecnico di progetto è il risultato del contributo fornito dal Laboratorio Segnali e Immagini dell'ISTI-CNR per il documento di progetto RTOD-SYS-SDD-010-INT per il progetto RTOD (Real-Time Object Detection mediante Machine Learning basato su tecnologia Low-Power GPU). In particolare, il rapporto studia e discute delle varie possibilità di architetture di reti convolutive che sono state valutate e che potranno essere utilizzate nel contesto del progetto per effettuare delle categorizzazioni di immagini mediante algoritmi di machine learning.

1 Introduzione

Questo documento (estratto dall' ANNEX al documento RTOD-SYS-SDD-010-INT) fornisce un supporto per lo sviluppo del Detection System.

In dettaglio, lo stato dell'arte descritto nella survey tecnologica RTOD-TN-RSD-005-INT viene qui aggiornato, con particolare riferimento agli algoritmi di object detection, già validati su alcuni dataset di benchmark, e che prendono in considerazione come annotazione dell'oggetto il bounding box (BB) anche nella versione *orientata* (OBB).

In uno scenario di aerial imaging la scelta del bounding box orientato, rispetto ai bounding box "orizzontali" (HBB) (e cioè con i lati paralleli agli assi coordinati dell'immagine), permette di ottenere una predizione più accurata (vedi **Figura -1**) sotto più punti di vista:

- **il bounding box orientato (OBB) fitta meglio l'oggetto come rapporto tra pixel relativi all'oggetto e pixel totali del BB;**
- **l'annotazione per OBB risulta più performante nel caso in cui si abbia un affollamento di oggetti nella stessa scena, perché i BB associati ad oggetti vicini si sovrappongono.**

Figura 1-1: Esempi di bounding box: Oriented BB (sinistra) e Horizontal BB (destra)

Nelle seguenti sezioni verranno descritti in modo sintetico gli algoritmi di object detection allo stato dell'arte più performanti, unitamente ai riferimenti bibliografici delle review più recenti. Sulla scorta di questi dati, INTECS potrà valutare se addestrare da zero dei modelli di object detection, o se importare dei modelli già addestrati e resi disponibili alla comunità scientifica, ed effettuare i propri test per verificare la rispondenza di tali modelli ai requisiti di sistema (principalmente risorse di storage e di calcolo).

2 Architetture di reti

2.1 Tipologia e struttura delle reti analizzate

In questa sezione si richiamano un insieme di architetture di reti neurali capaci di svolgere il task di object detection. Mentre nella precedente survey tecnologica la trattazione era stata più generale, indicando anche architetture di rete adatte alla risoluzione di altri task, inclusi i task di classificazione delle immagini, si riportano ora architetture progettate e addestrate per lo specifico task di object detection, privilegiando quelle che lavorano su dati tipo immagini Google Earth. Si individuano quindi le architetture che risultano più interessanti e promettenti, rispetto ad una implementazione (eventualmente con adattamenti) negli scenari di utilizzo previsti in RTOD.

Nello specifico nel documento RTOD-TN-RSD-005-INT, erano stati già riportati alcuni metodi per la detection di oggetti in immagini, tra cui Yolo v3 e CenterNet, citando brevemente altri metodi proposti in precedenza dalla comunità scientifica.

Negli approcci classici, il task di object detection è spesso ridotto al task di image classification. Questo è il caso, ad esempio, dei metodi basati su *sliding windows*, che controllano la presenza di oggetti mediante image classification utilizzando come input sottofinestre dell'immagine originale di diversa dimensione e aspect ratio, facendole scorrere sull'immagine. Nelle implementazioni naïve, l'onere computazionale dell'object detection è pari quindi al costo delle classificazioni di ogni singola sottofinestra moltiplicato per il numero delle sottofinestre da testare. Tale numero è in generale elevato, aspetto che rende il task difficoltoso in presenza di sistemi di object classification anche di complessità media.

Pertanto, anche i primi metodi, quali l'Haar detector di Viola-Jones Viola, 2001) impiegano euristiche a basso costo computazionale per rigettare sottofinestre in cui gli oggetti non sono evidentemente presenti, abbattendo il numero delle regioni da andare a testare mediante image classification. Si noti che in immagini comuni e, in particolar modo, nei casi di studio di RTOD, il numero delle regioni da testare è in generale ordini di grandezza superiore rispetto al numero di oggetti attesi, quasi sempre inferiore all'ordine delle centinaia.

Le sottofinestre "negative" sono quindi preponderanti rispetto alle finestre "positive", creando uno sbilanciamento tra le classi. Abbattere il numero delle finestre negative a priori è utile sia per ridurre i tempi di calcolo in fase di inferenza sia per focalizzare l'addestramento degli algoritmi di machine learning sugli esempi negativi difficili, i cosiddetti "hard negative", e non su quelli facili, specializzando quindi i modelli. Nell'esempio in Figura -1 una finestra di dimensioni fisse viene fatta scorrere sull'immagine muovendosi in senso orizzontale e verticale di un numero intero di pixel specificato da uno stride orizzontale e verticale. Ogni sottofinestra individuata viene quindi testata per la presenza di un oggetto al suo interno utilizzando un sistema di object classification, eventualmente preceduto da euristiche per rigettare finestre facilmente riconducibili allo sfondo.

Figura 2-1: Esempio dell'approccio sliding windows.

Sulla base di queste considerazioni, si possono individuare due tipologie principali di reti per l'object detection: quelle a doppio stadio e quelle a singolo stadio.

Le prime suddividono il task di detection in due step (stadi): dapprima vengono generate un insieme di regioni candidate corrispondenti a sottofinestre dell'immagine in cui è verosimile ci sia un oggetto di interesse e, quindi, tali regioni vengono testate utilizzando architetture mutuate dal task di object classification producendo in output un vettore di dimensione pari al numero delle classi C da analizzare. L'entrata *i*-esima di tale vettore corrisponde alla probabilità che nella regione ci sia un oggetto di classe *i* (con *i* compreso tra 1 e C). Tra i metodi più noti si cita R-CNN (Girshick 2014), Fast R-CNN (Girshick R. , 2015) Faster R-CNN (Ren, 2015) e R-FCN (Dai, 2016).

Le reti a stadio singolo invece producono contestualmente delle regioni candidate e la loro classificazione, spesso rappresentando in maniera efficiente un numero a priori di oggetti rilevabili con certe scale e aspect ratio prefissati, detti ancore. L'effettivo posizionamento e dimensione delle regioni rappresentate da un'ancora è ottenuto raffinando i dati iniziali dell'ancora mediante la stima di opportuni fattori di offset a mezzo di regressione. A tale tipologia di architettura appartengono Yolo v3 e Centernet già esaminati nella survey tecnologica RTOD-TN-RSD-005-INT. Un ulteriore metodo ben noto è costituito da Single Shot MultiBox Detector (SSD) (Liu 2016), che si basa su una rete convoluzionale feed-forward in grado di produrre un insieme di cardinalità prefissata di bounding box, ciascuno caratterizzato da un vettore di valori di confidenza sulla presenza di oggetti al suo interno. Una procedura di non maximum suppression consente quindi di individuare un sottoinsieme di tali bounding box, corrispondenti agli oggetti presenti nell'immagine. I primi layer della rete SSD si basano su architetture standard utilizzate per la classificazione delle immagini (troncate però prima di qualsiasi livello di classificazione), strati a cui generalmente ci si riferisce con il nome di backbone. Vengono quindi aggiunti dei livelli ausiliari per produrre, anche mediante regressione, dei raffinamenti su posizione e dimensione dei bounding box.

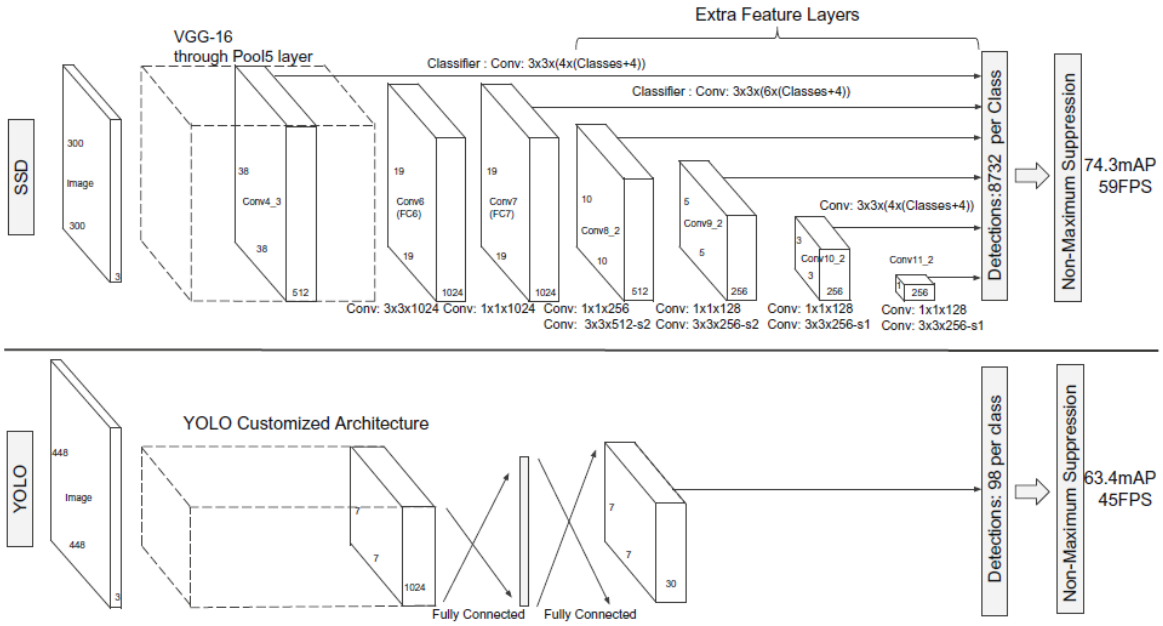

Figura 2-2: Confronto tra due architetture a stadio singolo: YOLO e SSD. I tempi e le performance riportate sono relative a VOC2007 test (Liu 2016)**.**

I metodi a stadio singolo sono stati apprezzati per la loro convenienza computazionale, caratteristica che li rende una scelta ideale per applicazioni in tempo reale, quali quelle previste in RTOD. Tuttavia, le performance registrate sono spesso state inferiori rispetto ai metodi a doppio stadio. A questo fine, molti lavori recenti hanno cercato di portare alcune caratteristiche di successo dei detector a due stadi all'interno dei detector a stadio singolo, senza alterarne la convenienza computazionale. Nel seguito si fa riferimento prevalentemente a tali lavori, concentrandosi infine sui metodi che sono in grado di gestire bounding box orientati (OBB), una caratteristica rilevante in ambito di remote sensing.

2.2 RefineDet

RefineDet (Zhang 2018) è una delle prime architetture a cercare di introdurre alcuni dei benefici dei detector a due stadi in un'architettura a stadio singolo, cercando di coniugare l'alta efficienza in termini computazionali dei detector ad uno stadio con la buona precisione dei detector più complessi. In maggior dettaglio, gli autori osservano che i detector a due stadi, quali ad esempio Faster R-CNN, R-FCN e FPN, posseggono tre vantaggi rispetto ai metodi ad uno stadio: (a) utilizzano una struttura a due stadi con euristica di campionamento per gestire lo squilibrio tra le classi; (b) utilizzano un sistema a cascata in due fasi per ottenere una regressione dei parametri dei bounding box degli oggetti e (c) utilizzano sempre due stadi per classificare gli oggetti.

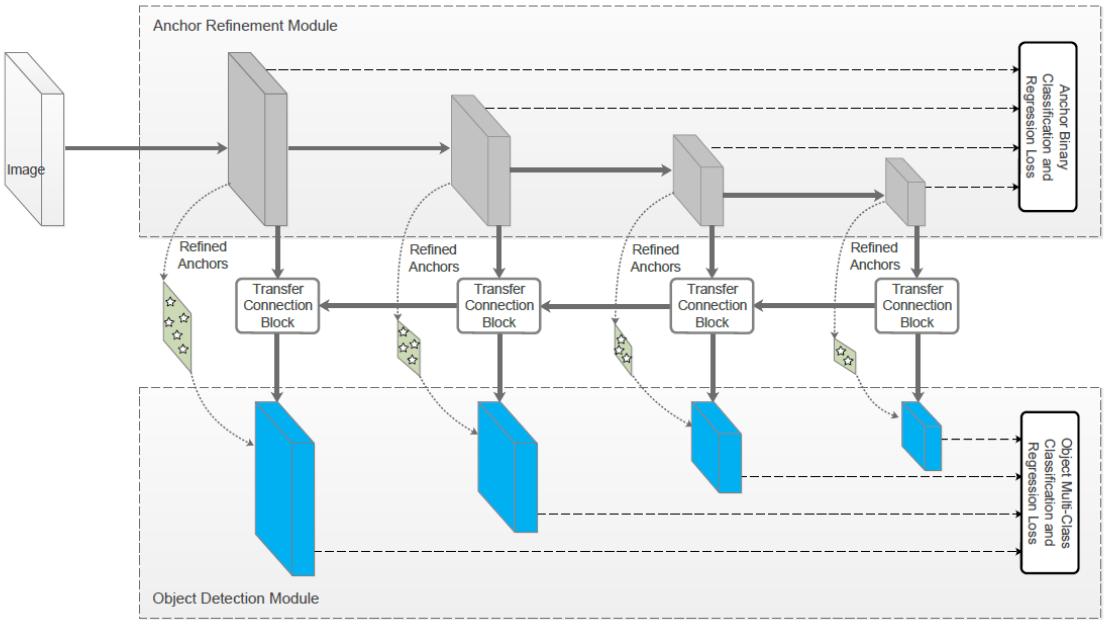

Figura 2-3: Architettura di RefineDet (Zhang 2018)**.**

Similmente a SSD, RefineDet si basa su una rete convoluzionale feedforward che produce un numero fisso di bounding box indicanti la presenza di diverse classi di oggetti. Impiega quindi non maximum suppression per produrre il risultato finale.

A livello architetturale, come mostrato in Figura -3, RefineDet è formato da due moduli interconnessi, cioè l'Anchor Refinement Model (ARM) e l'Object Detection Model (ODM) che mimano la struttura dei detector a due stadi. L'ARM mira a rimuovere le ancore negative in modo da ridurre lo spazio di ricerca per il classificatore e, allo stesso tempo, fornire un'inizializzazione più precisa della posizione e della dimensione delle ancore da passare all'ODM. A sua volta, l'ODM mira a mettere in atto una regressione per identificare le posizioni accurate degli oggetti e i relativi labels della classe a cui appartengono. I due moduli sono paralleli e l'ARM condivide i suoi features con l'ODM tramite un blocco di transfer appositamente creato.

Si noti che l'ARM può essere costruito basandosi su un backbone classico, mutuato da reti per il task di classificazione, rimuovendo gli ultimi strati decisionali e rimpiazzandoli con strutture ausiliarie per ottenere l'output desiderato (nello specifico l'articolo originale impiega VGG-16 e ResNet-101).

Al momento della pubblicazione, RefineDet ha superato i migliori risultati pubblicati in precedenza sia per i detector ad uno stadio sia a due stati con una AP del 41,8% su MS COCO (utilizzando come backbone ResNet-101). In termini di tempo, la rete opera a 40,2 e 24,1 FPS per immagini di dimensione rispettivamente 320x320 e 512x512 su una GPU NVIDIA TITAN X.

2.3 CenterNet

Questa sezione mira ad approfondire CenterNet, ovvero reti basate su rappresentazioni degli oggetti che fanno uso di un punto centrale dello stesso. Prima di analizzare le due varianti esistenti, si introduce la rete CornerNet su cui si basano.

2.3.1 CornerNet

Come RefineDet, anche CornerNet (Law, 2018) è una delle architetture di reti convoluzionali per l'object detection che mira a trasferire alcuni dei benefici dei detector a due stadi (quali R-CNN, Fast R-CNN e Faster R-CNN) ai detector a stadio singolo (quali YOLO, SSD e loro varianti). Come descritto precedentemente, tali detector collocano un insieme denso di ancore sull'immagine per generare un numero sufficiente di bounding box. Tale procedura ha due punti a svantaggio. Da un lato il numero di regioni candidate da vagliare è molto ampio, dall'altro il numero delle regioni che corrispondono a "true positive" è di ordini di grandezza inferiore rispetto alle regioni "negative", fatto che introduce un problema di sbilanciamento delle classi. In CornerNet, anziché cercare le regioni tramite ancore di diversa dimensione e aspect ratio si cercano gli angoli dei possibili bounding box, nella fattispecie, l'angolo in alto a sinistra e l'angolo in basso a destra del bounding box di ciascun oggetto da rilevare. La rete è strutturata in modo da produrre due heatmap corrispondenti alle posizioni più probabili per ciascuno di tali angoli. Una volta ottenute le posizioni candidate è necessario determinare quali delle coppie di angoli superiore sinistro e inferiore destro corrispondono al bounding box di un medesimo oggetto, risolvendo quindi un problema di accoppiamento (matching). Per fare questo, si utilizza un metodo di embedding associativo. La rete stima per ciascun angolo un embedding in uno spazio unidimensionale: se la distanza euclidea tra gli embedding di un angolo superiore sinistro e di un angolo inferiore destro è minore di una certa soglia, i due angoli sono considerati come delimitanti un bounding box di interesse.

In aggiunta, CornerNet introduce un nuovo tipo di layer denominato corner pooling, il cui compito è quello di ridurre l'indeterminazione della localizzazione di un corner rispetto alle features visuali dell'oggetto. Matematicamente, il corner pooling è un massimo eseguito lungo le semirette orientate uscenti da un punto.

CornerNet, implementato utilizzando una variante della rete hourglass (Newell, 2016) come backbone, raggiunge un Average Precision (AP) del 42,1% sul dataset MS COCO.

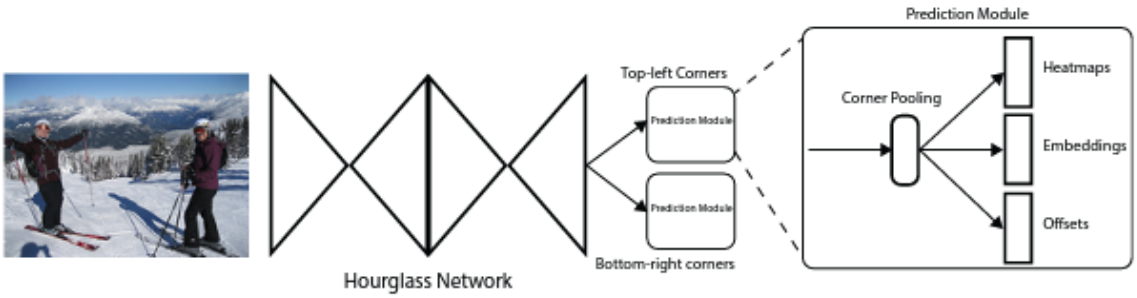

Figura 2-4: Architettura di CornerNet (Law, 2018)**.**

2.3.2 CenterNet (Zhou et al. 2019)

CenterNet, nella variante di Zhou (Zhou X., 2020) è un algoritmo di object detection costruito sulla base di architetture di reti preesistenti come backbone (Hourglass, ResNet, DLA). L'architettura proposta trae motivazione dalla necessità di incrementare le prestazioni degli attuali object detector basati su reti di apprendimento profonde. Allo stato dell'arte i suddetti detector sono incentrati sull'impiego di ancore, template rettangolari con un certo numero di caratteristiche predefinite, utilizzati per delimitare l'area di un oggetto candidato. I parametri che definiscono le ancore (e.g. offset dal centro, dimensioni) fanno parte del modello e pertanto vengono stabiliti durante la fase di training. Tipicamente, il risultato della predizione restituisce una gran quantità di possibili ancore per oggetto, la maggior parte dei quali viene

successivamente scartata. Questo richiede un'ulteriore fase di selezione della predizione migliore (non maximum suppression), comportando un aggravamento della complessità computazionale. CenterNet fornisce una soluzione a queste problematiche mediante una rappresentazione innovativa dell'oggetto tramite un unico punto chiave, permettendo di eliminare dalla sequenza di processamento l'onerosa fase di esclusione dei candidati non idonei. Informazioni geometriche d'interesse (dimensioni dell'oggetto, orientamento) vengono stimate per regressione dall'algoritmo. Le prestazioni del sistema proposto sono competitive rispetto a quelle degli attuali detector più performanti (one-stage e two-stage detectors), in termini di mean Average Precision (mAP), garantendo in più una velocità di esecuzione in tempo reale.

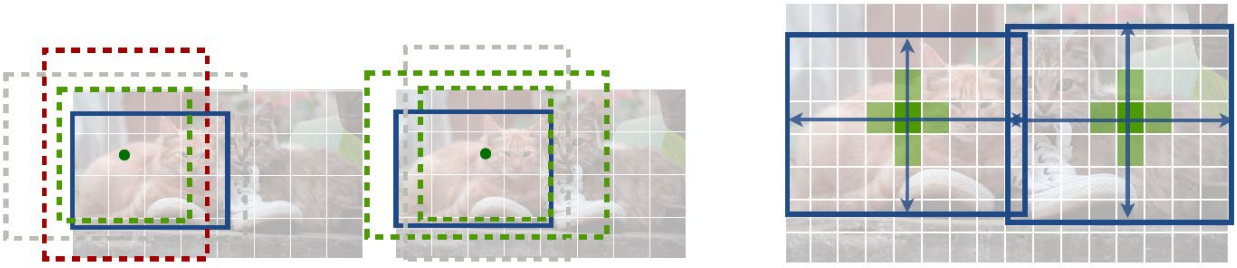

Figura 2-5: A sinistra approccio basato sulle ancore standard. In fase di training le ancore sono considerate positive se hanno un overlap con un oggetto per oltre il 70% (verde), negative se lo intersecano meno del 30% (rosso) e sono ignorate altrimenti (grigio). In CenterNet invece la detection è basata sul pixel centrale e l'estensione della regione è ottenuta per regressione rispetto a tale pixel (Zhou 2019)**.**

2.3.3 CenterNet (Duan et al. 2019)

Duan et al (Duan, 2019) hanno proposto una rete omonima a quella riportata nella sezione precedente, basata anch'essa sull'utilizzo di punti "centrali" all'interno di un oggetto di interesse. Il loro lavoro prende le mosse ed estende l'architettura CornerNet. Nello specifico, in questa variante di Centernet, un oggetto viene rappresentato come una tripla di punti: due corrispondono agli angoli superiore sinistro e inferiore destro, come in CornerNet, mentre il terzo è un pixel centrale dell'oggetto. L'architettura prevede in questo caso la creazione di tre heatmap, una per ciascun punto. Oltre ad altre procedure di pooling, sono introdotte delle euristiche per richiedere che il punto centrale giaccia al centro dell'oggetto in una sottofinestra simile al bounding box con un fattore di similitudine pari ad esempio ad 1:3 o 1:5 (a seconda degli esperimenti). CenterNet511-52 (ossia con risoluzione delle immagini in input pari 511x511 e Hourglass-52 come backbone) fornisce una AP del 41.6% su una singola scala e un AP del 43.5% in modalità multi-scala, migliorando quindi i risultati ottenuti da CornerNet.

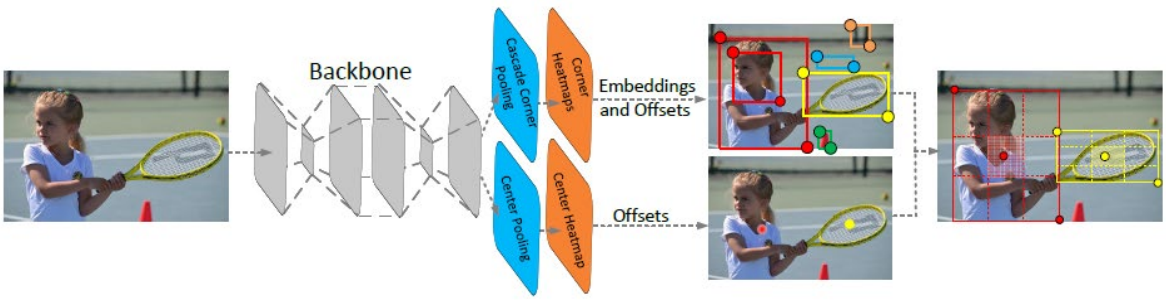

Figura 2-6: Architettura di centernet nella variante di Duan et al. Si noti che in questo esempio viene richiesto che il punto centrale giaccia al centro di una sottofinestra simile con rapporto di similitudine 1:3. (Duan, 2019)

2.4 RetinaNet e R3Det

RetinaNet è un modello single-stage per l'object detection che sta diventando molto popolare nelle applicazioni di remote sensing, in quanto mostra delle buone performance nella detection di oggetti piccoli o di scene affollate.

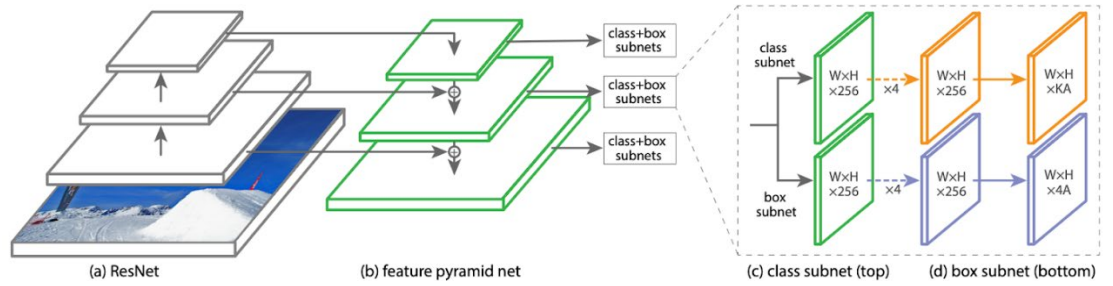

Figura 2-7: Architettura della rete RetinaNet che fa utilizzo di Feature Pyramid Network.

RetinaNet coniuga le Feature Pyramid Networks (FPN) (Lin 2017), progettate per gestire la detection di oggetti di dimensioni molto diverse, con la Focal Loss, introdotta per gestire il problema della class imbalance durante l'addestramento, proprio per i modelli di object detection a single stage (Lin T. Y., 2017). In sostanza, la Focal Loss applica un termine adattivo alla cross entropy loss in modo da forzare un apprendimento più persistente degli esempi negativi (hard negative).

L'architettura RetinaNet, in Figura -7, è costituita da una rete di backbone e da due sottoreti task-specifiche: il backbone è deputato al calcolo di una feature map convolutiva su tutta l'immagine di input (ResNet+FPN); questo step consente di estrarre le deep feature e delle multi-scale pyramid feature (e anchor boxes) dall'immagine di partenza. Queste feature sono poi usate come input per le due sottoreti specifiche (fully convolutional network): una per la classificazione delle anchor box e l'altra per la bounding box regression. Le performance di accuracy ottenute sono comparabili alla rete two-stage Faster R-CNN con FPN, ma con tempi di calcolo minori.

RetinaNet effettua la detection usando bounding box orizzontali; esiste tuttavia una versione per la detection tramite oriented bounding box, la più famosa R3Det (Yang 2021), e applicata, nel remote sensing, per il riconoscimento delle navi da immagini SAR in (Yang R. P., 2021).

2.5 PIoU Loss (Chen et al. 2020) e suo utilizzo nelle architetture per OBB detection

Generalmente i modelli di learning per l'OBB detection usano delle loss che minimizzano l'angolo di rotazione, senza considerare intersection over union, IoU. Per questo motivo, presentano delle limitazioni importanti rispetto al riconoscimento e alla localizzazione di oggetti con un elevato rapporto lunghezza/larghezza, e anche rispetto alle scene che presentano un affollamento di tali oggetti o background complessi.

La loss proposta in (Chen 2020) si pone l'obiettivo di migliorare le prestazioni di tutti gli algoritmi che si basano sulla ricerca di bounding box orientati. La PIoU loss usa sia l'angolo di rotazione del bounding box, sia la IoU (pixel-wise) per effettuare una più accurata regressione sugli oriented bounding box (OBB). Alcuni

degli attuali algoritmi, sia anchor-based (RefineDet con backbone ResNet e VGG) che anchor-free (CenterNet con backbone DLA e ResNet), vengono usati per valutare l'effetto dell'introduzione di questa loss su dataset di benchmark (HRSC2016, DOTA, PASCAL VOC) e anche su un dataset costruito ad hoc (Retail50k). I risultati mostrano un importante aumento delle performance, soprattutto per quanto riguarda i dataset HRSC2016 e Retail50K (dove si guadagna circa 6% di accuracy rispetto a RefineDet e CenterNet).

Una selezione dei valori di performance è riportata in Sezione 3.

2.6 Altre architetture di interesse per l'object detection nel remote sensing

Le architetture descritte nelle sottosezioni superiori sono solo un sottoinsieme di quelle attualmente disponibili e che hanno trovato impiego nel telerilevamento. Innanzitutto, esistono numerose varianti di metodi originariamente ideati per la detection di bounding box standard che hanno come scopo l'estensione al trattamento di bounding box orientati. Questo è ad esempio il caso di Faster R-CNN che è stato oggetto di modifica nel paper (Xia, 2018), il medesimo lavoro in cui è introdotto il dataset DOTA richiamato nel seguito. La modifica si basa sull'estensione dei gradi di libertà della parametrizzazione dei bounding box utilizzata normalmente in Fast e Faster R-CNN. Altri approcci si basano sull'uso di feature piramidali per realizzare reti ad hoc, come nel caso delle reti di features piramidali dense (DFPN), adattate anche al caso oriented (R-DFPN), proposte in (Yang X. S., 2018). In tale paper si mostra che queste reti possono rilevare efficacemente le navi sia in condizioni di mare aperto sia quando si trovano ormeggiate in porto. Tra le caratteristiche di DFPN si ha la buona capacità di gestire aspect ratio anche estremi, come quelli a volte incontrati con oggetti molto allungati come le navi. Rispetto ad altri detector multiscala trattati, DFPN costruisce mappe semantiche di alto livello per tutte le scale mediante connessioni dense, attraverso le quali viene migliorata la propagazione e il riutilizzo delle features.

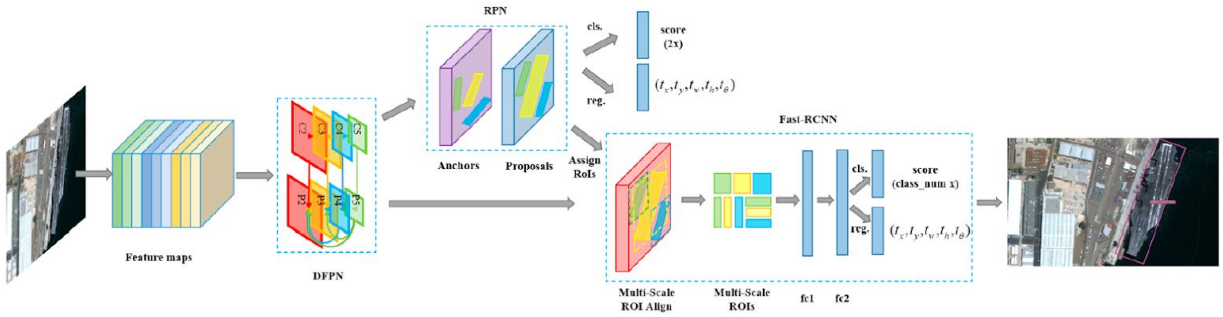

Figura 2-8: Architettura DFPN. I due stadi dell'architettura sono chiaramente discernibili. Entrambi sono preceduti dall'utilizzo di features piramidali (Yang X. S., 2018)**.**

Anche reti mutuate da altri contesti applicativi hanno trovato applicazione nel remote sensing. Paradigmatico è il caso di Rotational Region CNN for Orientation Robust Scene Text Detection (R2CNN) (Jiang, 2017)ideata per rilevare testo orientato in maniera arbitraria all'interno delle immagini e che può essere inteso come una variante di Faster R-CNN estesa alla gestione di oriented bounding box. Similmente le Rotation Region Proposal Networks (RRPN) (Ma, 2018) sono state introdotte per la detection di testo ma la loro applicabilità è di tipo generale. Si tratta in particolare di una architettura di rete per la generazione di regioni di interesse orientate che può essere inserita in diversi approcci a due stadi per la detection di oggetti orientati.

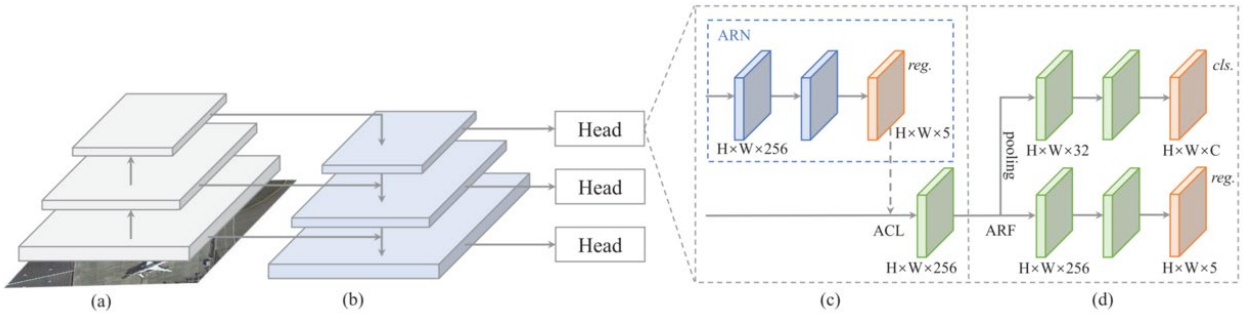

Figura 2-9: Architettura di S^{2}A Net (single-shot alignment network) (Han, 2021)**.**

Nel 2022 è stata pubblicata una nuova architettura di tipo single-shot per la oriented object detection: S^{2}A-Net (Han, 2021). Rispetto alle precedenti, si pone l'obiettivo di migliorare le performance raggiungendo ottimi risultati sia nella classificazione che nell'accuratezza della localizzazione. L'architettura è costruita a partire da un backbone, una rete di tipo feature pyramid (FPN) e da due moduli, Feature Alignment Module (FAM) e un Oriented Detection Module (ODM). FAM e ODM formano una "detection head" che si applica ad ogni scala della piramide di feature. Nel modulo che genera le feature allineate, una rete di refinement delle anchor è deputata alla predizione di anchors ruotate (high quality rotated anchors). Successivamente le feature di input e le anchor predette vengono passate al modulo ODM. Si noti che nella figura si visualizza solo il ramo dell'architettura deputato alla classificazione (*cls.*), e non quello della localizzazione (*reg.*). Nel modulo ODM che è deputato alla codifica delle informazioni di orientazione (attraverso l'estrazione delle feature orientation-sensitive tramite ARF (Zhou Y. Y., 2017), e un successivo pooling che ne seleziona di orientation-invariant) e produce come output delle predizioni (classe dell'oggetto e localizzazione nell'immagine) con un alto punteggio di "confidence". Infine si selezionano le migliori *k* predizioni (ad esempio 2000), e si applica non-maximum suppression per ottenere le detection finali. Nella **Figura -9**: (a) Backbone (ResNet è quello che dà i risultati migliori). (b) Feature pyramid network. (c) FAM. (d) ODM. Questa rete mostra delle performance di mAP su DOTA e su HRSC2016 comparabili se non migliori delle altre architetture descritte, mantenendo dei tempi di inferenza piuttosto contenuti.

Nella sezione successiva dedicata alla valutazione delle performance, ritroveremo una comparazione delle reti qui descritte testate su dataset di interesse.

3 Dataset di benchmarking e performance

3.1 Dataset di benchmarking

Alcuni dataset di interesse sono già stati presentati in RTOD-TN-RSD-005-INT. Riprendiamo la descrizione di alcuni di essi e integriamo con altri dataset di interesse.

VEDAI (Razakarivony, 2016)

VEDAI (Vehicle Detection in Aerial Imagery) è un dataset si focalizza sulla detection e la classificazione dei diversi veicoli terrestri da immagini aeree (RGB o infrarosso) di dimensione 512 x 512 o 1024 x 1024. VEDAI contiene circa 1200 immagini per un totale di 3700 target annotati.

Figura 3-1: Esempio di immagini dal dataset VEDAI.

COWC (Mundhenk, 2016)

Anche COWC (Cars Overhead with Context) si focalizza sui veicoli, ma contiene immagini grayscale e RGB con oltre 32 mila istanze annotate (tramite identificazione del pixel centrale del veicolo). Il dataset contiene anche un gran numero di esempi "negativi".

Figura 3-2: Esempio di immagini dal dataset COWC.

NWPU VHR-10 (Cheng, 2014)

NWPU VHR-10 è un dataset focalizzato sulle immagini di remote sensing ad altissima risoluzione (VHR remote sensing images). NWPU VHR-10 è costituito da circa 800 immagini: 175 a colori, acquisite da Google Earth con risoluzione tra 0.5 m e 2 m; e 85 immagini nell'infrarosso acquisite da dati Vaihingen con risoluzione spaziale di 0.08 m. Il dataset si divide in due: a) immagini "positive", che contengono almeno un oggetto target (650 immagini); b) immagini "negative", che non contengono alcun oggetto target annotato (150 immagini). Tra tutte le immagini si contano: 757 "airplanes", 302 "ships", 655 "storage tanks", 390 "baseball diamonds", 524 "tennis courts", 159 "basketball courts", 163 "ground track fields", 224 "harbors", 124 "bridges", and 477 "vehicles", tutti annotati manualmente con bounding box orizzontali.

Figura 3-3: Esempi di immagini dal dataset NWPU VHR-10.

DOTA (Xia, 2018)

Il DOTA (Dataset of Object deTection in Aerial images) è un dataset di benchmark per object detection che è annotato con bounding box orientati (OBB), e la versione v1.5 è già stata presentata in RTOD-TN-RSD-005-INT, alla sezione 3.3.5. Nel 2021 è stata rilasciata una nuova versione (DOTA-v2.0), ma tutte e tre le versioni disponibili sono attualmente usate dalla comunità scientifica.

DOTA-v1.0 contiene 15 categorie, 2.806 immagini e 188.282 istanze. Training set, validation set e testing set costituiscono rispettivamente la metà, un sesto e un terzo del dataset complessivo.

DOTA-v1.5 contiene le stesse immagini del DOTA-v1.0, ma sono presenti più annotazioni: sono annotate anche le istanze più piccole (meno di 10 pixel) degli oggetti classificati; inoltre contiene una nuova categoria di oggetti: “container crane”, che è presente nel dataset con 403318 istanze.

DOTA-v2.0 (Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges, Ding et al, 2021) include più immagini da Google Earth, GF-2 Satellite e immagini aeree. Sono annotate 18 categorie (rispetto a DOTA-v1.5 sono aggiunte “airport” e ”helipad”), per un totale di 11.268 immagini e quasi due milioni (1.793658) di istanze. Il dataset è diviso in training set, validation set, test-dev set, e test-challenge set. Per limitare il problema dell’overfitting, i testing set sono più popolati di training e validation. La numerosità dei vari set è la seguente: training - 1830 immagini (268.627 istanze); validation - 593 immagini (81.048 istanze); test-dev - 2.792 immagini (353.346 istanze), non ha le annotazioni di ground truth; test-challenge - 6.053 immagini (1.090.637 istanze). Immagini e ground truth sono rese disponibili solo durante le challenge.

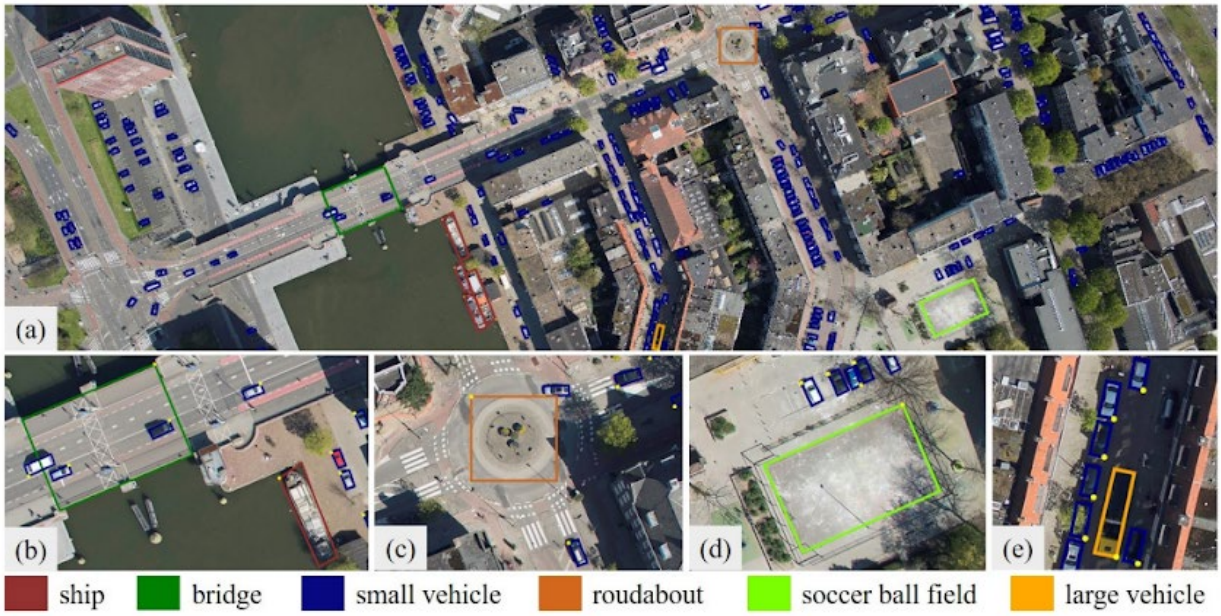

Figura 3-4: Esempi di immagini e annotazioni dal dataset DOTA.

Storicamente, il primo dataset costruito per testare degli algoritmi di object detection è il PASCAL Visual Object Classes (VOC) (Everingham, 2010). Esso ha consentito la realizzazione delle prime grandi challenge su object detection, tenute dal 2005 al 2015 (passando da 4 a 20 classi di oggetti, con dataset di train/val con 11530 immagini per un totale di 27450 oggetti annotati).

Successivamente fu realizzato ImageNet (Deng, 2009), un dataset molto più ricco di PASCAL VOC, che contiene 200 classi di oggetti, e circa 500 mila bounding box annotati. ImageNet, oltre ad essere estremamente ricco (negli ultimi aggiornamenti sia arrivato a oltre 14 milioni di immagini per 1000 classi), è attualmente il dataset più popolare per il benchmarking di algoritmi di detection e classification su larga scala (ad es. nella Large Scale Visual Recognition Challenge).

Un altro dataset interessante è MS COCO (Lin T. Y., 2014), costituito da circa 328 mila immagini, per un totale di 91 classi e 2,5 milioni di oggetti segmentati. Ciò che lo differenzia dagli altri dataset è che MS COCO ha in media più istanze e classi presenti per immagine e contiene anche delle annotazioni ulteriori di

contesto associate ad ogni oggetto, che consentono di impostare delle challenge più complicate che negli altri dataset.

In Tabella -1 inseriamo le statistiche relative alle varie versioni di DOTA, a confronto con i classici PASCAL VOC, MS COCO e ImageNet.

Tabella 3-1

Dataset	Classes	Image quantity	Megapixel area	BBox quantity	Avg. BBox quantity
PASCAL VOC (07++12)	20	21503	5133	52090	2.42
MS COCO (2014 trainval)	80	123287	32639	886266	7.19
ImageNet (2014 train)	200	456567	82820	478807	1.05
DOTA-v1.0	15	2806	19173	188282	67.10
DOTA-v1.5	16	2806	19173	402089	143.73
DOTA-v2.0	18	11268	107133	1793658	159.18

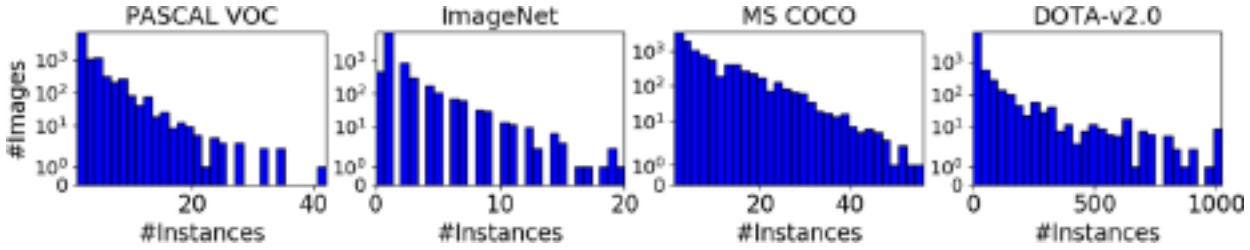

Figura 3-5: Numero di istanze per immagine a confronto tra i dataset generali di Object detection e DOTA. Per PASCAL VOC, ImageNet e MS COCO, contiamo le statistiche su 10.000 immagini casuali. Poiché le immagini in DOTA sono molto grandi (20.000x20.000), per un confronto equo, contiamo le statistiche di [10:1000] patch di immagini di dimensione 1024x1024. DOTA ha una gamma più ampia di istanze per immagine.

HRSC2016 (Liu Z. W., 2016)

High Resolution Ship Collection 2016 (HRSC2016) contiene immagini ottiche di scenari portuali caratterizzati dalla presenza contemporanea di una o più imbarcazioni. Per ogni immagine viene fornita una descrizione delle navi presenti nella scena. Ogni nave viene localizzata tramite bounding box, sia allineati agli assi coordinati sia ruotati opportunamente per tener conto della generica orientazione della nave. Il dataset, reperibile nella sua versione più recente da (Feng, 2016) – vedi anche http://www.escience.cn/people/liuzikun/DataSet.html – contiene immagini a risoluzione variabile tra 0.4 m e 2 m per pixel, con dimensioni da 300x300 fino a 1500x900 (la maggior parte superiore a 1000x600). Dati ausiliari riguardano l'identificativo della struttura portuale, il sensore da cui provengono i dati, informazioni sulla data di acquisizione, coordinate geografiche, lo strato di risoluzione di Google Earth e la scala dell'immagine.

HRSC2016 è arricchito da informazioni e strumenti rilevanti:

- **La classificazione nel dataset è organizzata mediante struttura ad albero, con accuratezza progressiva man mano che si avanza nei tre livelli previsti (Classe dell'oggetto → Categoria della nave → Tipo di nave).**
- **La localizzazione è riportata i) tramite bounding box orientato secondo gli assi coordinati, ii) mediante bounding box ruotato secondo gli assi principali della nave e iii) tramite segmentazione fine dell'area occupata dalla nave.**
- **La suddivisione nei sottoinsiemi di Training (436 immagini), Validation (181 immagini) e Test (444 immagini) è stata realizzata mediante un algoritmo che garantisce una distribuzione delle istanze di oggetti il più possibile omogenea e non polarizzata.**
- **Il dataset è corredato di applicazioni dedicate per l'annotazione e la gestione dei dati.**

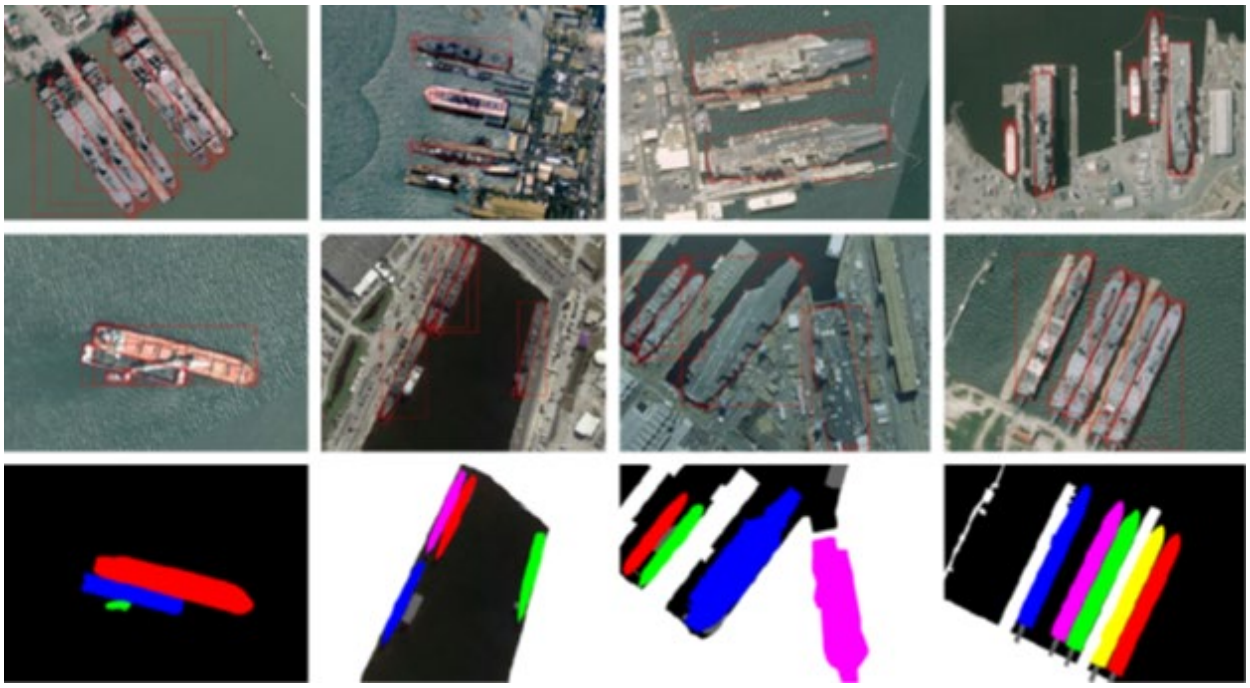

Figura 3-6: Esempi di immagini e annotazioni del dataset HRSC2016 (Liu Z. W., 2016)**.**

Nella seguente Tabella 3 è riportato un confronto tra vari dataset menzionati specificamente per le immagini aeree, rispetto a: sorgente del dato, tipologia di annotazione di riferimento (HBB, OBB, o CP - center point), il numero di categorie principali e totali, il numero di istanze totali presenti, il range di dimensioni delle immagini e anno di riferimento.

Compaiono anche altri dataset, attualmente non descritti, perché non usati per la valutazione delle performance degli algoritmi illustrati. Tuttavia, vengono riportati perché hanno caratteristiche che li rendono interessanti e/o utili nell'immediato futuro rispetto allo sviluppo/fine tuning di algoritmi di object detection in RTOD.

Tabella 3-2

Dataset	Source	Annotation	# of main categories	Total # of categories	# of instances	# of images	Image width	Year
NWPU VHR-10	Multi-source	HBB	10	10	3651	800	~1000	2014
VEDAI	satellite	OBB	3	9	2950	1268	512~1024	2015
COWC	aerial	CP	1	1	32716	53	2000~19000	2016
HRSC2016	GoogleEarth	OBB	1	26	2976	1061	~1100	2016
RSOD (Long, 2017)	GoogleEarth	HBB	4	4	6950	976	~1000	2017
CARPPK (Hsieh, 2017)	drone	HBB	1	1	89777	1448	1280	2017
ITCVD (Yang M. Y., 2018)	aerial	HBB	1	1	228	23543	5616	2018
HRSD (Zhang Y. Y., 2019)	Multi-source	HBB	13	13	55740	21761	152~10569	2019
DIOR (Li, 2020)	GoogleEarth	HBB	20	20	190288	23463	800	2019
FGSD (Chen K. W., 2020)	GoogleEarth	OBB	1	43	5634	2612	930	2020
DOTA-v1.0	Multi-source	OBB	14	15	188282	2806	800~13000	2018
DOTA-v1.5	Multi-source	OBB	15	16	402089	2806	800~13000	2019
DOTA-v2.0	Multi-source	OBB	17	18	1793658	11268	800~20000	2021

3.2 Performance delle architetture

Nella valutazione delle performance delle architetture, si fa presente che alcune di esse sono già state precedentemente valutate nel documento RTOD-TN-ST-011-INT (Sezione 3.2); queste sono ad esempio: LeNet e AlexNet, VGG, GoogleNet (Inception), Resnet (50, 101, …), DenseNet, SE block (squeeze and excitation), YoloV3 e CenterNet. Si nota che le performance riportate nel documento non sono state valutate sul task di localizzazione (object detection), ma solo sulla classificazione (image classification), in cui per la maggior parte dei casi l'input è rappresentato da un crop (generalmente quello centrale) estratto dall'immagine di input (ImageNet).

Le performance che riportiamo qui di seguito invece sono relative alle reti descritte in Sezione e ad altre usate come termine di paragone in letteratura, nella **Tabella** l'elenco di quelle che saranno menzionate:

Tabella 3-3

Elenco reti considerate nella valutazione delle performance	
SSD	**RefineDet**
YOLO-V2	**RefineDet + PIoU**
R-FCN	**CenterNet**
FR-H	**CenterNet + PIoU**
FR-O	**S2 A-Net**
R-DFPN	**FR-O + RT**
R2CNN	**FR-O + RT Augmented**
RRPN	

Come già citato, DOTA è uno dei dataset di riferimento maggiormente utilizzati come benchmark per una larga parte di algoritmi di l'object detection, in particolare per l'object detection da immagini aeree.

Si è quindi provveduto a presentare una valutazione comprensiva e comparativa delle performance dei vari algoritmi presentati nelle precedenti sezioni al fine basandosi principalmente sulle prestazioni ottenute su tale dataset di riferimento.

In modo da rendere consistente e più omogeneo il confronto, le performance principali sono state confrontate a parità di versione del dataset DOTA, utilizzando quindi la versione DOTA-v1.0.

Le altre differenziazioni necessarie per effettuare un'analisi comparativa riguardano il backbone (tipologia dell'architettura) utilizzato dall'algoritmo e la categoria di oggetti su cui l'algoritmo è specializzato, le tipologie di immagini su cui vengono testati gli algoritmi, ed eventualmente la dimensione delle immagini.

Un'ulteriore distinzione che può esser fatta, e che risulta di interesse per l'analisi riguardante il progetto RTOD, è quella tra la possibilità di effettuare un riconoscimento basato su Horizontal Bounding Box (HBB) o anche su Oriented Bounding Box (OBB).

Le categorie su cui sono state effettuate le valutazioni, previste in DOTA-v1.0, sono: plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer ball field and swimming pool. Per consentire una valutazione più idonea al progetto RTOD, si sono riportate le precisioni media ed aggregata su categorie di oggetti di maggior interesse per il progetto. Tali categorie scelte sono: plane, ship, storage tank, e harbor.

Nella Tabella si mostrano le precisioni medie (mean Avg. Prec. - all categories) su tutte le categorie per ogni metodo, associato con un relativo backbone, e la precisione sulle categorie selezionate (Selected categories). Inoltre, si fornisce la dimensione delle immagini di training e testing per ognuno dei metodi valutati (Size), dimensioni che sono solitamente ottenute riducendosi tramite l'utilizzo di patch dell'immagine originale. In questo scenario sono trattati dati con il formato OBB.

Tabella 3-4

Method	Backbone	Size	mean Avg. Prec. (all categories)	mean Avg. Prec. (selected categories)
SSD	VGG16	512	10.6	**8.4**
YOLOV2	DarkNet19	416	21.4	**19.8**
R-FCN	R-101	800	26.8	**19.1**
FR-H	R-101	800	32.3	**19.9**
R-DFPN	R-101	800	57.9	**60.7**
R2CNN	R-101	800	60.7	**61.8**
RRPN	R-101	800	61	**62.9**
RefineDet	VGG16	512	50.9	**73.6**
RefineDet+PIoU	VGG16	512	52.5	**73.7**
RefineDet	R-101	512	55.1	**75.3**
RefineDet+PIoU	R-101	512	56.5	**75.2**
CenterNet	DLA-34	512	59.1	**67.1**
CenterNet+PIoU	DLA-34	512	60.5	**67.0**
FR-O	R-101	1024	54.1	**48.1**
S^2 A-Net	R-50-FPN	1024	79.4	**87.4**
FR-O* + RT	R-50-FPN	1024	73.8	**85.1**
FR-O* + RT (Aug.)	R-50-FPN	1024	79.8	**86.5**

Nella Tabella 3-5 sottostante invece si riportano i risultati ottenuti in maniera più specifica su scenario con dati che utilizzano dati HBB.

Tabella 3-5

Method	Backbone	mean Avg. Prec. (all categories)	mean Avg. Prec. (selected categories)
ICN	DR-101-FPN	72.5	**80.3**
SCR-Det	R-101-SF-MDA	75.4	**80.7**
CenterMap	R-101-FPN	77.3	**86.0**
Li et al.	R-101	78.8	**85.1**
FR-O* + RT	R-50-FPN	74.6	**85.5**
FR-O* + RT (Aug.)	R-50-FPN	80.8	**87.4**

Una ulteriore valutazione viene fatta, seppur in maniera più ristretta, per diverse reti testate sul dataset HRSC2016 menzionato nella precedente sezione, per queste, riportate in Tabella 3-6, è stato possibile

rilevare soltanto la mean Average Precision generale e non ristretta alle categorie selezionate, poiché il dataset è diverso.

Tabella 3-6

Method	Backbone	Size	mean Avg. Prec.
R2CNN	ResNet101	800	**73.0**
RC1&RC2	VGG-16	800	**75.7**
RRPN	ResNet101	800	**79.1**
R2PN	VGG-16	800	**79.6**
RetinaNet-H	R-101	800	**82.9**
RetinaNet-R	R-101	800	**89.2**
RoI-Transformer	R-101	512	**86.2**
R3Det	R-101	300	**87.1**
R3Det	R-101	600	**89.0**
R3Det	R-101	800	**89.3**
CenterNet-OBB	ResNet18	512	**67.7**
CenterNet-OBB+PIoU	ResNet18	512	**78.5**
CenterNet-OBB	R-101	512	**77.4**
CenterNet-OBB+PIoU	R-101	512	**80.3**
CenterNet-OBB	DLA-34	512	**88.0**
CenterNet-OBB+PIoU	DLA-34	512	**89.2**

Come si può notare facilmente dalle tabelle sopra, spesso le performance ottenute utilizzando dati di tipo HBB sono leggermente più elevate rispetto a quelle utilizzando gli OBB. Tale risultato ha una motivazione collegata alla metrica utilizzata nella valutazione del risultato, questa è basata sul concetto di Intersection over Union (IoU), che provoca una migliore performance globale di copertura delle aree classificate con un bounding box di tipo allineato con gli assi (HBB), rispetto ad un bounding box orientato (OBB) che è necessariamente più piccolo del corrispettivo horizontal, essendo esattamente allineato con l'oggetto da classificare.

Per mostrare meglio questa discrepanza, si mostra il seguente esempio in Figura 3-7. Come si vede, la IoU del classificatore basato su OBB (a destra) è di solito leggermente inferiore alla IoU per oggetti HBB (a

sinistra), utilizzando lo stesso algoritmo poiché la detection tramite OBB necessita di un posizionamento più preciso rispetto al corrispettivo tramite HBB: un piccolo spostamento della classificazione HBB porta ad una perdita di precisione minore rispetto al caso in cui lo stesso spostamento sia applicato al caso orientato.

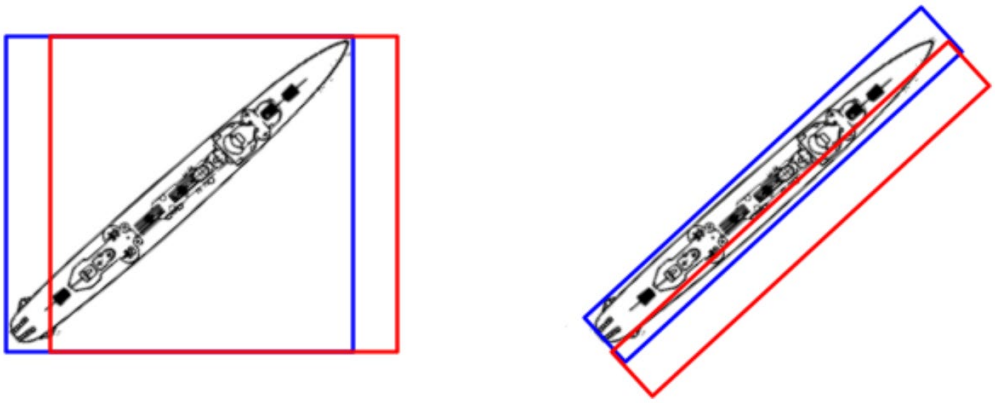

Figura 3-7: Esempio di differenza nelle classificazioni con HBB e OBB.

In Figura 3-8, è possibile confrontare la detection di Faster R-CNN (seconda colonna) che localizza tramite HBB rispetto agli altri algoritmi, che usano OBB, mentre nella prima colonna sono riportate le rispettive ground truth di partenza.

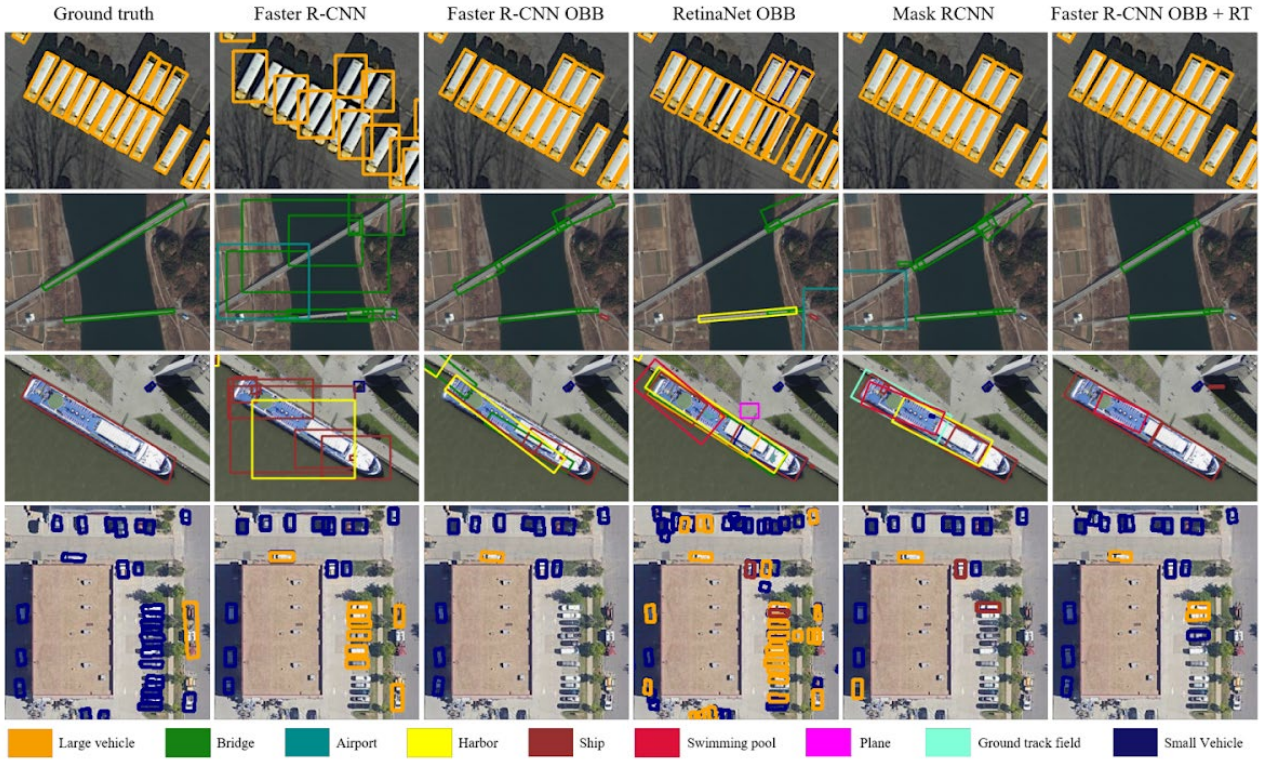

Figura 3-8: Confronto di detection tra alcune delle principali reti citate secondo varia tipologia di target annotati da identificare (Ding, 2021)**.**

Oltre ad una valutazione relativa ai valori di accuratezza riportati in **Tabella** e **Tabella** , in **Figura -8** è possibile apprezzare un esempio degli output di detection di cinque algoritmi (Faster R-CNN, faster R-CNN

OBB, Retinanet OBB, Mask RCNN e Faster R-CNN OBB+ RoI Transformer) sullo stesso set di 4 immagini (contenute nel test-dev di DOTA-v2.0). La scelta delle immagini vuole mettere in evidenza le differenze tra i risultati rispetto alla presenza di oggetti variamente ruotati, con affollamento, con aspect ratio grande o piccolo.

Si ritiene comunque necessario, soprattutto approcciandosi alla fase di test, un tuning tra i risultati presentati in questa sezione (i.e. sui dataset di riferimento) e il dataset reale previsto nello studio del progetto RTOD.

3.3 Controllo sui vincoli fissati dai requisiti di sistema RTOD

Nella Tabella 3-7 si riportano le prestazioni in termini di velocità di inferenza (riferite per una singola NVIDIA Tesla V100 GPU) e media della average precision (mAP) come precedentemente menzionata, per alcune delle architetture presentate. Per ottenere una base comune di confronto, tutti questi algoritmi sono stati valutati con un backbone "Resnet-50" con FPN. Le immagini di riferimento sono tutte riportate alla dimensione 1024x1024. Le mAP sono invece riportate come medie su tutte le versioni DOTA su cui sono state testate (v1.0, v1.5 e v2.0) e sono invece suddivise per HBB ed OBB.

Tabella 3-7

Method	Speed (fps)	HBB mAP (average on DOTA-v*.0)	OBB mAP (average on DOTA-v*.0)
RetinaNet-H	**16.7**	59.5	-
RetinaNet-OBB	**12.1**	60.3	57.4
Mask R-CNN	**9.7**	62.4	61.0
Cascade Mask R-CNN	**7.2**	62.2	61.5
Hybrid Task Cascade	**7.9**	62.6	61.7
Faster R-CNN	**14.3**	61.9	-
Faster R-CNN OBB	**14.3**	61.7	59.6
Faster R-CNN OBB + Dpool	**12.1**	61.9	60.4
Faster R-CNN H-OBB	**13.7**	61.7	60.5
Faster R-CNN OBB + RoI Transformer	**12.4**	64.7	63.9
S2A-Net	**16.0**	-	74.1

Come si nota anche nel peggiore dei casi sopra citati, la performance è sempre superiore ai minimi fissati (i.e. 5fps), e volendosi riportare ad una casistica più ampia che ricopra immagini di dimensioni superiori (e.g. fino a 2048x2048) si stima che si possa raggiungere una velocità di inferenza pari al minimo richiesto con un minimo impatto sulla precisione.

Uno schema di confronto rispetto ad accuracy (mAP) e velocità di inferenza (FPS) a parità di setting è quindi riportato nella Figura 3-9. I backbone usati sono: ResNet50 (simboli piccoli) o ResNet101 (simboli grandi); le immagini di input sono di dimensione 1024×1024.

Le architetture confrontate sono: Faster R-CNN (FR-CNN), Mask R-CNN, RetinaNet, Hybrid Task Cascade (HTC), RoI Transformer (RoITrans) e S2A-Net. I calcoli sono stati effetttuati usando una V100 GPU.

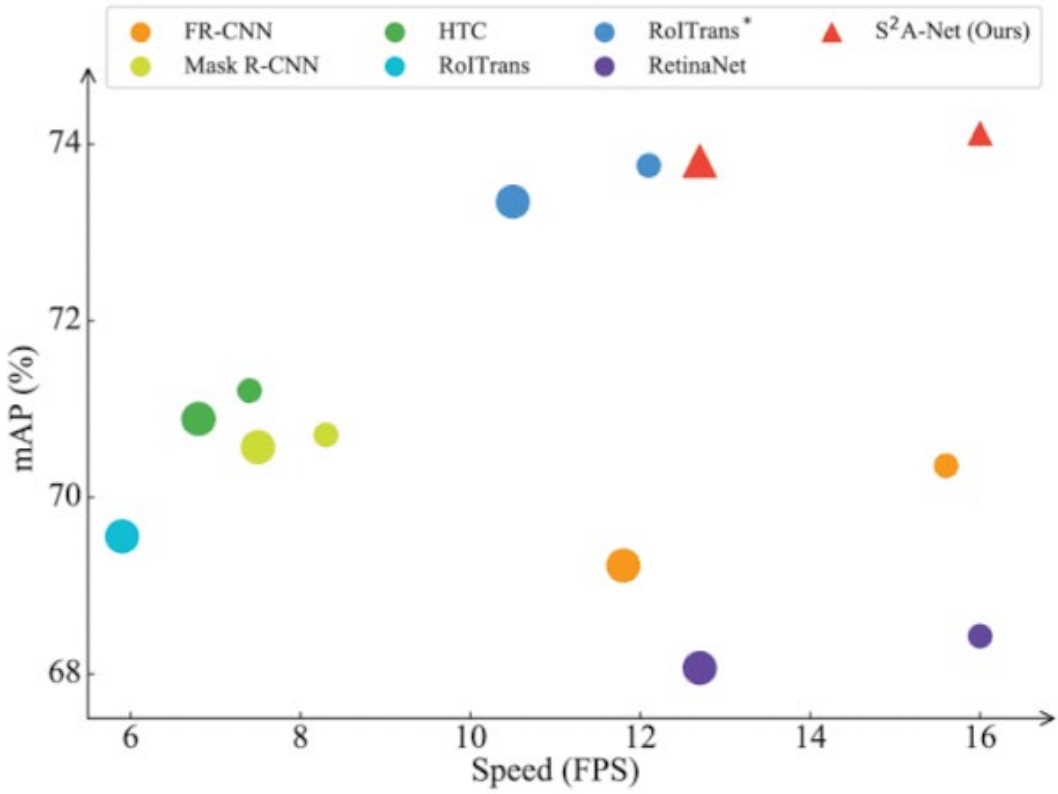

Figura 3-9: Performance delle reti presentate rispetto alla loro accuratezza, da (Han, 2021)**.**

4 Bibliografia

Chen, K. W. (2020). *Fgsd: A dataset for fine-grained ship detection in high resolution satellite images.* arXiv preprint. doi:arXiv:2003.06832

Chen, Z. C. (2020). PIoU loss: Towards accurate oriented object detection in complex environments. *European Conference on Computer Vision* (p. 195-211). Springer, Cham. doi:doi.org/10.1007/978-3-030-58558-7_12

Cheng, G. H. (2014). Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing, 98*, 119-132.

Dai, J. L. (2016). R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*(29).

Deng, J. D.-F. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition* (p. 248-255). IEEE.

Ding, J. X. (2021). Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Duan, K. B. (2019). Centernet: Keypoint triplets for object detection. *Proceedings of the IEEE/CVF international conference on computer vision* (p. 6569-6578). IEEE.

Everingham, M. V. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision, 88*(2), 303-338.

Feng, G. (2016). *HRSC2016*. Tratto da https://www.kaggle.com/guofeng/hrsc2016

Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE international conference on computer vision* (p. 1440-1448). IEEE.

Girshick, R. D. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE.

Han, J. D. (2021). Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1-11.

Hsieh, M. R. (2017). Drone-based object counting by spatially regularized regional proposal network. *Proceedings of the IEEE international conference on computer vision* (p. 4145-4153). IEEE.

Jiang, Y. Z. (2017). *R2CNN: Rotational region CNN for orientation robust scene text detection.* arXiv preprint arXiv:1706.09579. doi:10.48550/arXiv.1706.09579

Law, H. a. (2018). Cornernet: Detecting objects as paired keypoints. *Proceedings of the European conference on computer vision* (p. 734-750). Springer.

Li, K. W. (2020). Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing, 159*, 296-307.

Lin, T. Y. (2014). Microsoft COCO: Common objects in context. *European conference on computer vision* (p. 740-755). Springer, Cham.

Lin, T. Y. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* (p. 2117-2125). IEEE.

Lin, T. Y. (2017). Focal loss for dense object detection. *IEEE international conference on computer vision* (p. 2980-2988). IEEE.

Liu, W. A. (2016). Ssd: Single shot multibox detector. *European conference on computer vision* (p. 21-37). Springer Cham.

Liu, Z. W. (2016). Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE Geoscience and Remote Sensing Letters, 13*(8), 1074-1078.

Long, Y. G. (2017). Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing, 55*(5), 2486-2498.

Ma, J. S. (2018). Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia, 20*(11), 3111-3122.

Mundhenk, T. N. (2016). A large contextual dataset for classification, detection and counting of cars with deep learning. *European conference on computer vision* (p. 785-800). Springer, Cham.

Newell, A. Y. (2016). Stacked hourglass networks for human pose estimation. *European conference on computer vision* (p. 483-499). Springer.

Razakarivony, S. &. (2016). Vehicle detection in aerial imagery: A small target detection benchmark. *ournal of Visual Communication and Image Representation, 34*, 187-203.

Ren, S. H. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*(28).

Viola, P. a. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. 1*, p. I-I. IEEE.

Xia, G. S. (2018). DOTA: A large-scale dataset for object detection in aerial images. *Proceedings of the IEEE conference on computer vision and pattern recognition* (p. 3974-3983). IEEE.

Yang, M. Y. (2018). Deep learning for vehicle detection in aerial images. *2018 25th IEEE International Conference on Image Processing* (p. 3079-3083). IEEE.

Yang, R. P. (2021). A novel CNN-based detector for ship detection based on rotatable bounding box in SAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*(14), 1938-1958. doi:doi:10.1109/JSTARS.2021.3049851

Yang, X. L. (2021). R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35 (4)*, p. 3163-3171.

Yang, X. S. (2018). Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sensing, 10*(1), 132.

Zhang, S. W. (2018). Single-shot refinement neural network for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* (p. 4203-4212). IEEE.

Zhang, Y. Y. (2019). Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing, 57*(8), 5535-5548.

Zhou X., K. V. (2020). Tracking Objects as Points. In B. H. Vedaldi A. (A cura di), *European Conference on Computer Vision. Lecture Notes in Computer Science vol. 12349.* Springer, Cham. doi:doi.org/10.1007/978-3-030-58548-8_28

Zhou, X. W. (2019). *Objects as points.* arXiv:1904.07850. doi:https://doi.org/10.48550/arXiv.1904.07850

Zhou, Y. Y. (2017). Oriented response networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (p. 519-528). IEEE.