D1.2
Prototipi Analisi Visuale

Lucia Vadicamo, Claudio Gennaro, Donato Cafarelli, Fabrizio Falchi

INDEX

1. Introduzione

In questo documento vengono descritte le principali attività svolte nell'ambito dell'Obiettivo Operativo n. 1 (OO1) "Progettazione dei sistemi di Intelligenza Artificiale e di Visione Artificiale per la sicurezza dell'imbarcazione" e in particolare dell'Attività A1.2 "Realizzazione prima versione prototipi Analisi Visuale".
Tale attività ha avuto per scopo la realizzazione di un primo prototipo per il riconoscimento e il tracking automatico di persone in mare e oggetti all'interno di flussi video provenienti da fonti eterogenee.
L'attività svolte sono complessivamente riconducibili a tre parti ad ognuna delle quali è dedicato un capitolo di questo documento:

- Dataset: creazione di un dataset di riprese da drone di persone in mare realizzato appositamente per il progetto e necessario per il training e il test del prototipo da realizzare.
- Sperimentazione riconoscimento: è stata analizzata la letteratura scientifica allo scopo di selezionare le tecnologie più promettenti che sono state testate sul dataset sviluppato per il progetto in modo da comprendere performance e realizzare un primo prototipo.
- Architettura Hardware: ha preso in considerazione le problematiche hardware relative al drone aereo e al controllo da remoto dello stesso.

2. Dataset

Questa sezione descrive una campagna di raccolta dati utili allo svolgimento dell'attività A1.2 del progetto, il quale prevede la realizzazione di un prototipo per il riconoscimento e il tracking automatico di persone ed oggetti in mare all'interno di flussi video provenienti da fonti eterogenee, tra cui telecamere poste su droni aerei. Infatti, per la realizzazione di tale prototipo è di fondamentale importanza avere a disposizione dati per l'addestramento e/o validazione di un'intelligenza artificiale in grado di riconoscere in tempo reale persone e oggetti in mare nelle riprese effettuata da un drone, per poi poter guidare il drone stesso in modo che rimanga sopra di esse fino all'arrivo di mezzi di salvataggio. Sebbene pubblicamente siano disponibili molti dataset annotati contenenti persone ed oggetti in scenari di vita quotidiana, non si può dire lo stesso per la situazione di persone ed oggetti in mare ripresi dall'alto come previsto nel progetto NAUSICAA. Vista la scarsità di una tale tipologia di dati accessibili pubblicamente, l'ISTI-CNR ha lavorato alla realizzazione di un dataset di riprese in ambiente marino mediante l'uso di droni aerei ed in particolare riprese che coinvolgano persone/oggetti che, trovandosi in acqua, simulino di aver bisogno di essere soccorsi/identificati. La raccolta dati è stata possibile grazie ad una collaborazione che l'ISTI-CNR ha avviato con il Servizio Fly&Sense dell'Area territoriale del CNR di Pisa per le operazioni di volo di Sistemi Aeromobili a Pilotaggio Remoto, il cui responsabile è stato Dr. Marco Paterni, e con l'Istituto di Fisiologia Clinica (IFC) del CNR per le operazioni di immersione in acqua di due subacquei professionisti abilitati alle attività scientifiche, il cui responsabile è stato il Dr. Mirko Passera. La Dr.ssa Lucia Vadicamo (ISTI-CNR) ha coordinato le attività di riprese da drone e le attività in acqua per la realizzazione del dataset.

Le attività sono state svolte nella spiaggia del Gombo del Parco di Migliarino, San Rossore e Massaciuccoli, poiché tale area permette di svolgere le azioni previste per la realizzazione delle riprese in piena sicurezza in quanto l'area identificata è segregata dai regolamenti vigenti del Parco. Infatti, per l'accesso alla spiaggia del Gombo (Figura 1), raggiungibile via terra dall'interno del Parco, è stato necessario richiedere ed ottenere una speciale autorizzazione dall'Ente Parco.

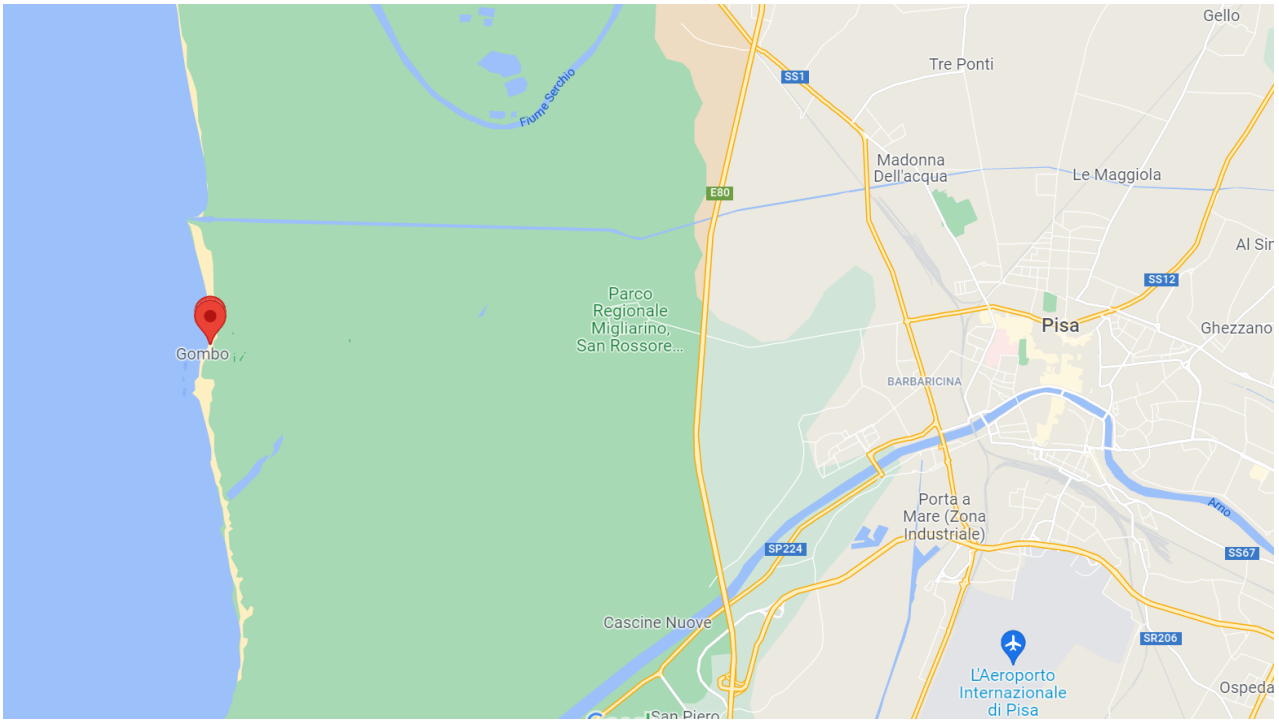

Fig. 1: Luogo in cui sono state effettuate le attività di riprese da drone aereo e di di immersione in acqua (coordinate 43.7212778,10.2787972).

Prima delle attività in mare sono stati individuati diversi scenari di interesse per la realizzazione di un insieme di riprese che fossero quanto più possibile varie e mirate al contesto del soccorso uomo in mare previsto dal progetto. In particolare sono state individuate dieci dimensioni di interesse:

1. *Soggetti/oggetti da riprendere in mare*: persone, boe, salvagente, vestiti, scogli, pezzi di legno, parti di terra ferma e quant'altro ci sia naturalmente;

2. *Scenari per la simulazione soccorso persona in mare*: persona vigile (nuota, galleggia, o sbraccia per attirare l'attenzione), persona semi-cosciente (galleggia o si muove limitatamente), persona non cosciente (corpo galleggiante in posizione supina o prona, oppure parzialmente galleggiante, ossia parte del corpo è sotto la superficie dell'acqua);
3. *Abbigliamento persona in mare*: persona vestita (colori chiari, scuri o misti), persona in costume, persona che indossa accessori (per esempio un cappello);
4. *Orari di ripresa (cambiamenti di luce)*: mattina, intorno a mezzogiorno, pomeriggio, sera;
5. *Altezza di volo*: provare a volare sia su quote alte per vedere una porzione di mare più ampia (per es. 40-60 metri) sia a volare ad altezze inferiori per vedere bene gli oggetti e quindi anche l'uomo in mare (per es. 10-30 metri).
6. *Angolo di ripresa (pitch):* direzione camera a perpendicolo sul mare e altre direzioni di ripresa (per es. 45° e 60°)
7. *Direzione di volo:* direzioni di volo diverse (per es. sud-nord, est-ovest, o zig-zag) per acquisire immagini con diversi angoli di illuminazione;
8. *Condizioni meteo:* cielo sereno, parzialmente nuvoloso, o coperto così da avere diversi effetti di luce ed ombre sul mare -- il caso di pioggia, maltempo, o vento forte non sono di interesse poiché il progetto non prevede il volo del drone in tali situazioni.
9. *Condizioni del mare:* mare calmo, mosso o molto mosso
10. *Specifiche Video:* video in alta, media, o bassa risoluzione.

Avendo dieci dimensioni di interesse, e molte opzioni per ogni dimensione, non è possibile realizzare tutte le combinazioni possibili (per comprendere quelle menzionate sopra servirebbero più di 260 mila riprese diverse). Inoltre alcuni scenari sono di difficile realizzazione (per es. mare molto mosso o riprese serali) o difficilmente programmabili (condizioni meteo e mare). Si è deciso quindi di procedere con una una modalità "best-effort", variando principalmente le altezze di volo, gli scenari di soccorso uomo in mare, l'abbigliamento e le condizioni di luminosità, limitatamente alle condizioni circostanziali relative alle date in cui è stato possibile effettuare le riprese e le attività in acqua. Da notare, infatti, che le date disponibili per la realizzazione delle riprese vere proprie sono state limitate dalle condizioni meteo, dalle procedure necessarie per l'accesso al Parco (l'autorizzazione per l'accesso al parco è stata concessa in data 09/07/2021 ed è valida fino al 31/12/2022, ma ogni singolo accesso va autorizzato nuovamente dall'ente Parco qualche giorno prima delle attività), ed anche dalle disponibilità dei vari operatori coinvolti nelle attività in spiaggia/mare. Ad oggi, è stato possibile realizzare due set di riprese nelle date 04/08/2021 e 15/09/2021, in orari diversi (mattina e primo pomeriggio) e con diverse condizioni meteo (cielo parzialmente nuvoloso e cielo molto nuvoloso). In entrambi i casi le immagini sono state acquisite con la camera DJI FC6310 del drone Phantom 4 Pro V2. Le riprese sono state fatte ad alta risoluzione (4K) in quanto si prevede di generare anche immagini a bassa risoluzione a partire dai dati raccolti. In entrambe le occasioni le attività in acqua sono state effettuate da due subacquei professionisti (un uomo ed una donna), che hanno simulato vari scenari per il soccorso in mare.

Fig. 2: Alcune foto rappresentative delle attività svolte il 4 Agosto (in alto) ed il 15 Settembre (in basso)

ID Video	Data e ora Riprese	Altezza Drone (m)	Scenario				Abbigliamento Sub					Altre persone e/o oggetti in acqua
			Vigile		Non cosciente							
			Gallegia e/o sbraccia	*Nuota*	*Supino*	*Prono*	*Maglia scura e pantaloni scuri*	*Maglia chiara e pantaloni scuri*	*Maglia colorata e pantaloni scuri*	*Maglia scura e pantaloni chiari*	*Maglia chiara e pantaloni chiari*	
1	04/08/21, 12:02	30										Barche, persone, ed altro
2	04/08/21, 12:15	30	✔				✔					Barche, persone, ed altro
3	04/08/21, 12:20	10	✔				✔					
4	04/08/21, 12:21	20		✔			✔					
5	04/08/21, 12:25	30	✔				✔					
6	04/08/21, 12:27	40	✔				✔					Barca
7	04/08/21, 12:28	50	✔				✔					Barca, gommone
8	04/08/21, 12:30	20	✔				✔					
9	04/08/21, 13:15	40				✔	✔	✔				Gommone, salvagente
10	04/08/21, 13:17	60	✔	✔		✔	✔	✔				Barca, gommone, salvagente
11	04/08/21, 13:20	30				✔	✔	✔				Salvagente
12	04/08/21, 13:23	50	✔				✔	✔				Barche, persone, ed altro
13	04/08/21, 13:24	40*	✔			✔	✔	✔				Barche, persone, ed altro
14	04/08/21, 13:38	50*	✔				✔	✔				Barche, persone, ed altro
15	04/08/21, 13:40	30*	✔				✔	✔				Barche, salvagente
16	04/08/21, 13:43	20*	✔				✔	✔				Barche, salvagente
17	04/08/21, 13:45	40	✔				✔	✔				Barche, salvagente
18	04/08/21, 13:46	40		✔			✔	✔				Barche, salvagente
19	15/09/21, 10:33	40	✔						✔	✔		Barca
20	15/09/21, 10:35	40		✔					✔	✔		Barca
21	15/09/21, 10:37	50	✔						✔	✔		Barca
22	15/09/21, 10:40	50		✔					✔	✔		Barca
23	15/09/21, 10:42	60	✔						✔	✔		Barca
24	15/09/21, 10:44	60		✔					✔	✔		Barca ed altro
25	15/09/21, 10:52	40				✔			✔	✔		
26	15/09/21, 10:54	50				✔			✔	✔		
27	15/09/21, 10:56	60				✔			✔	✔		Barca ed altro
28	15/09/21, 10:58	60				✔			✔	✔		Barca
29	15/09/21, 11:03	30				✔			✔	✔		Barca
30	15/09/21, 11:05	30	✔	✔					✔	✔		
31	15/09/21, 11:22	40	✔						✔		✔	Barca e legno
32	15/09/21, 11:25	40		✔					✔		✔	Barca e legno
33	15/09/21, 11:29	50	✔						✔		✔	Barca e legno
34	15/09/21, 11:30	50		✔					✔		✔	Barca
35	15/09/21, 11:32	60		✔					✔		✔	Barca e legno
36	15/09/21, 11:34	60	✔						✔		✔	Barca e legno
37	15/09/21, 11:35	60		✔					✔		✔	Barca e legno
38	15/09/21, 11:37	30		✔					✔		✔	Barca
39	15/09/21, 11:38	30	✔						✔		✔	Barca
40	15/09/21, 11:49	40			✔	✔			✔		✔	Barca
41	15/09/21, 11:50	50			✔	✔			✔		✔	Barca
42	15/09/21, 11:52	60			✔	✔			✔		✔	Barca
43	15/09/21, 11:54	30			✔	✔			✔		✔	Barca
44	15/09/21, 11:56	40			✔	✔			✔		✔	Barca
45	15/09/21, 11:57	50			✔	✔			✔		✔	Barca
46	15/09/21, 11:59	60			✔	✔			✔		✔	Barca
47	15/09/21, 12:00	30			✔	✔			✔		✔	Barca
48	15/09/21, 12:02	20	✔						✔		✔	Barca
49	15/09/21, 12:04	20		✔					✔		✔	Barca
		* Angolo di ripresa 45°										

Fig. 3: Dettagli dei video acquisiti e scenari simulati. L'angolo di ripresa è di 90° (perpendicolo sull'acqua) eccetto i quattro casi segnati con l'asterisco.

In totale sono state fatte 49 riprese video, le cui caratteristiche sono riportate in Figura 3, e riassunte qui di seguito. Le altezze di volo sono state variate da 10 m a 60 m sul livello del mare, con una maggiore frequenza per le quote medio-alte. In particolare sono state utilizzate le seguenti quote: 10 m (1 ripresa), 20 m (5 riprese), 30m (11 riprese), 40 m (12 riprese), 50 m (10 riprese), 60 m (10 riprese). La velocità di ripresa è stata in media di 3 metri al secondo. L'angolo di ripresa è stato fissato a perpendicolo sull'acqua (90°), ma nelle attività del primo giorno sono state effettuate anche quattro riprese usando un'angolazione di 45°. Dato che in base ad alcuni esperimenti preliminari l'angolazione di 90° è risultata efficace per l'analisi dei video, questa è stata fissata per le successive attività di riprese così da semplificare una dimensione del problema. I due sub hanno simulato sia scenari in cui la persona è vigile (nuota, galleggia o sbraccia) che non cosciente (corpo galleggiante supino o prono). L'abbigliamento dei sub è stato variato quanto più possibile, tenendo comunque presente la necessità di indossare una muta in quanto le immersioni in acqua sono durate circa due ore per ogni sessione di riprese. In alcuni casi è stato possibile riprendere anche oggetti che si trovavano in acqua o che sono stati posizionati appositamente (barche, pezzi di legno, salvagente, tavola da sup, gommoni, rocce). Nelle inquadrature di alcuni video sono stati incidentalmente ripresi anche dei bagnanti che si trovavano nelle vicinanze della porzione di mare in cui si stazionavano i nostri sub. Per questioni di privacy alcune porzioni di video verranno quindi rimosse dal dataset qualora il viso o le persone siano identificabili.

Al momento non è stato possibile acquisire video con situazioni di mare mosso, ma si prevede nella primavera del 2022, o quando le condizioni meteo saranno adatte alle immersioni, di ampliare ulteriormente il dataset.

I dati raccolti, per poter essere utilizzati ai fini delle attività del progetto, devono essere annotati, ovvero bisogna identificare nei keyframe dei video se appare un oggetto od una persona e in quale posizione dell'immagine. Tale processo è dispendioso in quanto coinvolge attivamente un annotatore umano che deve visionare molte ore di video. Nel progetto l'attività di annotazione è stata svolta da Donato Caffarelli (ISTI-CNR) che ha utilizzato un approccio semiautomatico mediante l'uso del *Computer Vision Annotation Tool (CVAT)*, disponibile a link https://github.com/openvinotoolkit/cvat.
Il tool permette una più agevole annotazione del dataset grazie all'utilizzo di una interfaccia semplice ed intuitiva.

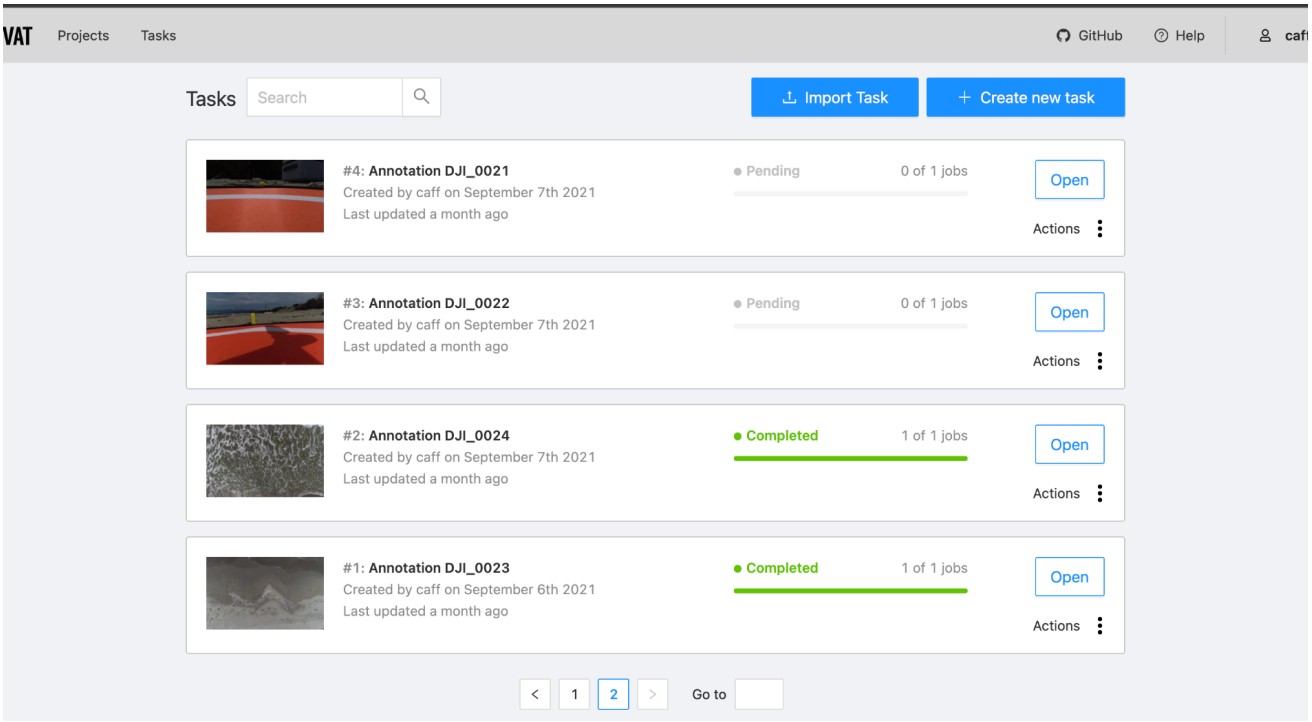

Fig 4: CVAT Master Task page

Consente sia di annotare il dataset manualmente e sia di sfruttare uno dei modelli di Deep Learning presenti per un'annotazione automatica.

Il suo funzionamento prevede il caricamento di un gruppo di immagini o video; da questi ultimi, in particolare, verranno estratti tutti i frame e sarà possibile effettuare un'annotazione frame per frame, con la possibilità quindi di saltare dei frame o di selezionarne alcuni specifici da annotare.
L'annotazione avviene mediante la scelta del nome di una o più "*label*" e si procede a disegnare una bounding box (ossia un riquadro) intorno all'oggetto da etichettare. Il tool prevede, inoltre, una funzione "*tracking*" che consente, una volta creata la bounding box, solamente di riposizionarla sull'oggetto, evitando di doverla disegnare per ogni frame.
Infine, è possibile scaricare sia i frame del video che le annotazioni. Quest'ultime, grazie all'interfaccia ottimizzata per i task di computer vision, possono essere esportate in base al modello (se presente) su cui verranno effettuati i test.

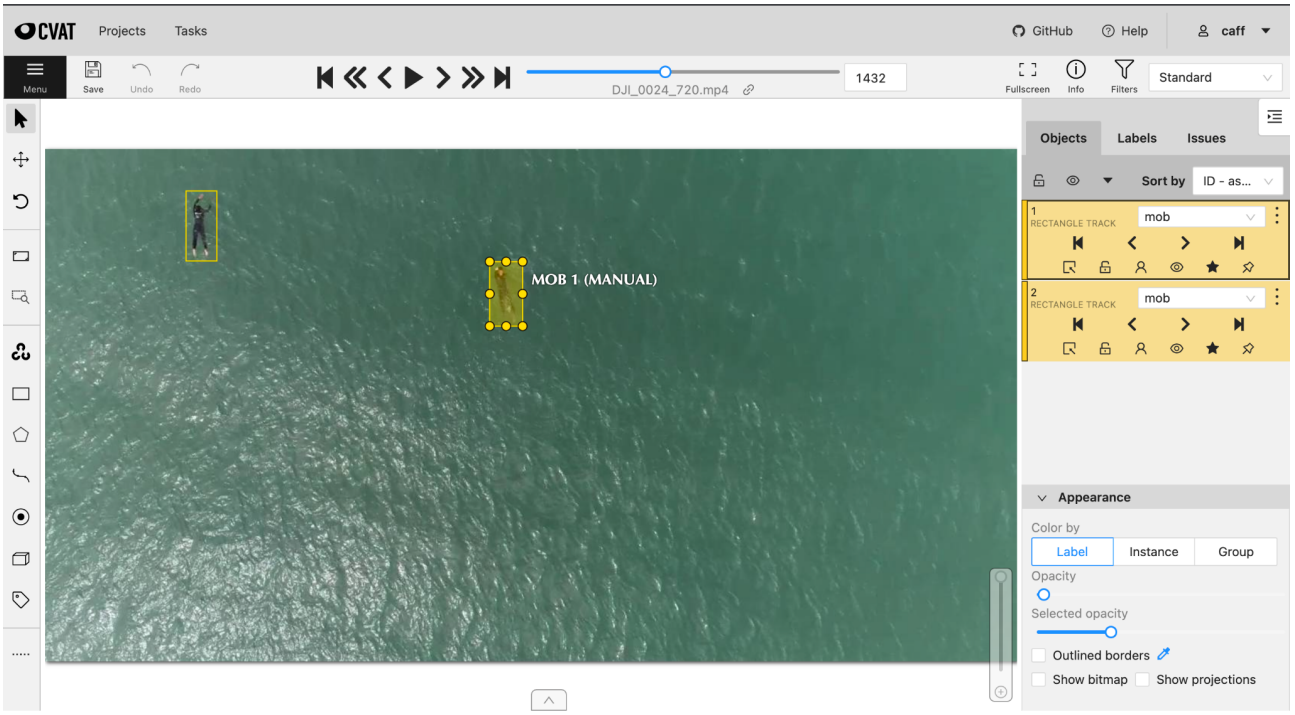

Fig 5: Esempio processo di annotazione mediante il tool CVAT

Nella Figura 5 si può notare il processo di annotazione di uno dei video ottenuti dalle riprese effettuate con il drone. Le bounding box (di colore giallo) sono state disegnate intorno all'oggetto di interesse rappresentato dai due sub intenti a simulare una tipica situazione di uomo in mare.
Nonostante l'uso del tool CVAT agevoli il processo di annotazione bisogna notare che tale processo è comunque dispendioso in termini di risorse umane: per l'annotazione (comprensiva dell'inserimento di tutte le bounding box) di tutti i frame di un video di circa 1 minuto è stato impiegata quasi un'ora di lavoro dell'annotatore. Tenendo conto che i 49 video acquisiti hanno una durata variabile tra uno e quattro minuti, non è stato ancora possibile ultimare il processo di annotazione che proseguirà nei prossimi mesi.

3. Sperimentazione riconoscimento

Le prime sperimentazioni sul riconoscimento di persone/oggetti in mare in riprese aeree sono state svolte utilizzando due reti neurali che sono allo stato dell'arte per il riconoscimento di oggetti: *YOLOv3* [Redmon and Farhadi 2018] e *VarifocalNet* [Zhang et al 2021]. Sul Web sono disponibili numerosi modelli pre-allenati di tali reti. Per una prima analisi abbiamo utilizzato i modelli YOLOv3 - backbone: Darknet53 - Input size: 608x608 e VarifocalNet - backbone: X101-64x4d, entrambi scaricabili dal "Model Zoo" di *MMDetection (*https://github.com/open-mmlab/mmdetection*)* che è un tool open source per il riconoscimento di oggetti basato sulla libreria PyTorch (https://pytorch.org/).

YOLO (*You only look once*) è uno degli algoritmi più veloci per il riconoscimento di oggetti e YOLOv3 è una sua versione migliorata e pubblicata nel 2018 [Redmon and Farhadi 2018] . Seppure nel panorama della letteratura scientifica degli ultimi tre anni, YOLO non detiene più il primato in termini di accuratezza nell'identificazione e riconoscimento di oggetti, esso risulta ancora oggi uno degli algoritmi più utilizzati grazie al suo ottimo rapporto tra accuratezza ed efficienza, il che lo rende particolarmente adatto ad applicazioni che debbano funzionare in tempo reale.
YOLOv3 si basa sull'utilizzo di una singola rete neurale sull'intera immagine, analizzando tutte le parti dell'immagine in parallelo (non usa quindi il paradigma della sliding-window), ed effettua la detection su tre differenti scale dell'immagine (l'immagine viene ridimensionata rispettivamente di un fattore 32, 16 ed 8 per ottenere una maggiore accuratezza su scale piccole) . Ad alto livello, l'architettura della rete è suddivisa in due componenti principali: *Feature Extractor* e *Feature Detector* (detector multiscala). L'immagine viene prima processata dal feature extractor che estrae delle feature (descrittori numerici) e poi dal detector della rete che restituisce l'immagine processata con dei bounding box attorno alle classi rilevate. In altre parole, la rete divide l'immagine in regioni e predice dei bounding-boxes e delle probabilità di presenza di oggetti per ogni regione. Le probabilità vengono usate per predire le classi degli oggetti che appaiono in un'immagine, le bounding-boxes per localizzare le posizioni spaziali di tali oggetti.

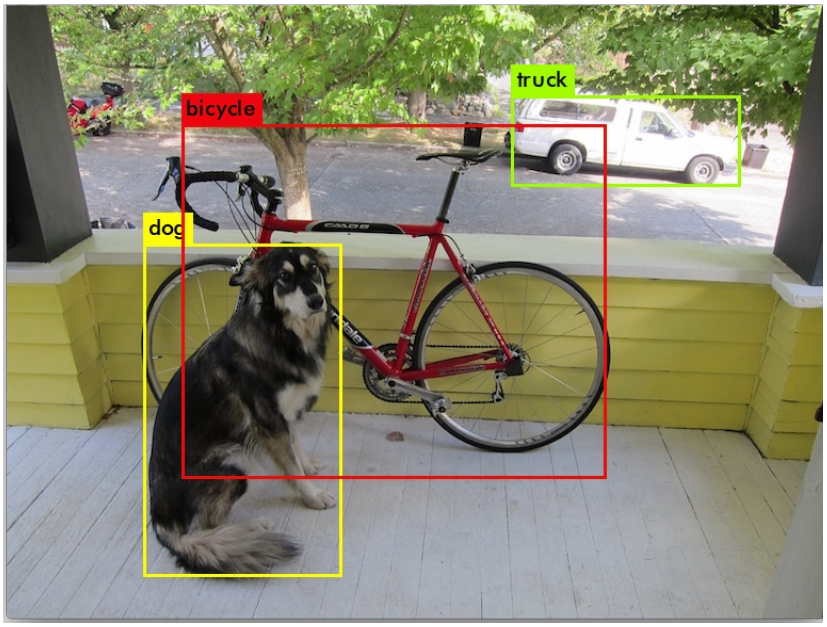

Fig 6: Esempio di detection e riconoscimento mediante l'uso di YOLOv3 (fonte:https://pjreddie.com/darknet/yolo/)

L'architettura di YOLOv3 si chiama Darknet-53 poiché è composta da 53 layer convoluzionali per la parte di Feature Extraction ed è stata sviluppata a partire dalla rete Darknet-19 usata in YOLOv2. I 53 layer della Darknet sono affiancati da ulteriori 53 layer per la parte di detection, per cui l'intera architettura è costituita da 106 layer. La rete preallenata su COCO, come quella utilizzata nel progetto, permette il riconoscimento di 80 classi di oggetti, tra cui "persona", "cane", "barca", etc..

VarifocalNet è un object detector "denso" (ossia si basa sul paradigma "sliding-window" su una griglia dell'immagine) ed stato presentato quest'anno a CVPR 2021 (International Conference on Computer Vision and Pattern Recognition) che è una delle conferenze più importanti nel campo della computer vision e del pattern recognition. VarifocalNet, o

VFNet in breve, è un metodo che mira a classificare accuratamente un enorme numero di detection candidate per l'identificazione di oggetti. Esso utilizza una nuova funzione di loss, chiamata Varifocal Loss, per l'addestramento di un detector "denso" di oggetti che predica lo "IoU-Aware Classification Score (IACS)" che misura simultaneamente la fiducia nella presenza di un oggetto di una determinata classe e la precisione nella localizzazione della boundbox generata, ed impiega una nuova efficiente rappresentazione "a stella" delle bounding box sia per la stima dello score IACS che per il raffinamento di bounding box grossolane.

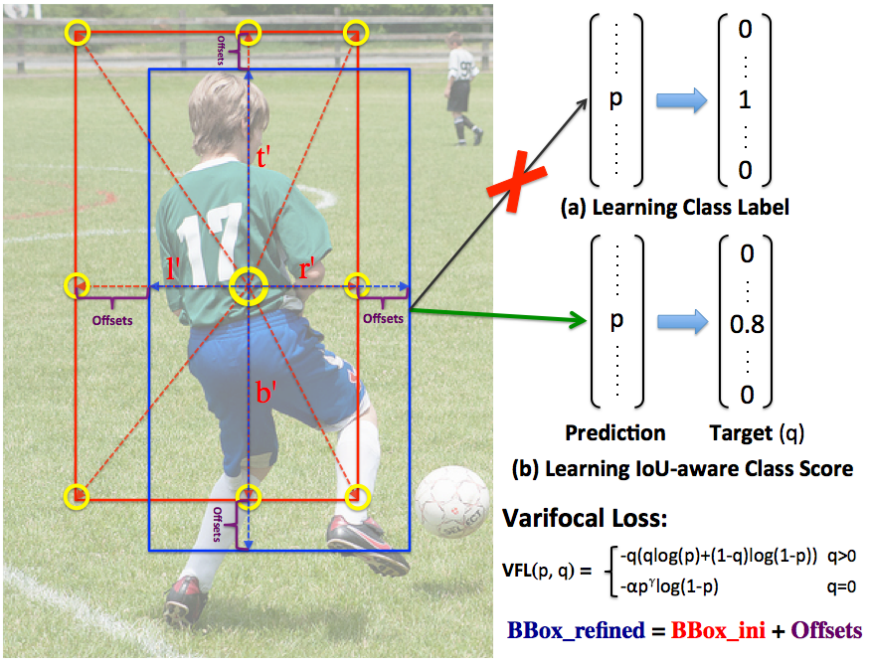

Fig 7: illustrazione del metodo usato in VarifolcalNet (fonte [Zhang et al 2021]). Invece di imparare a prevedere l'etichetta di una classe per un determinato un bounding box, come mostrato in (a), il metodo proposto impara a predire un punteggio di classificazione della localizzazione (ovvero la misura IACS) come una misura che unisce la fiducia della presenza dell'oggetto all'accuratezza della localizzazione (b). La Varifocal Loss ed una nuova rappresentazione a forma di stella delle bounding box (usa feature estratte da nove punti, come quelli gialli mostrati in figura) sono usate durante l'addestramento dell'object detector per la stima dello IACS. Le bounding box inizialmente stimate (in rosso) vengono poi raffinate in riquadri più accurati (in blu).

L'architettura della VFNet è costruita da un modello di base (*backbone*), una *Feature pyramid networks* (FPN) [Lin et al 2017] e dalla *VarifocalNet Head* (cosposta da due sottoreti, una per la localizzazione delle bounding box ed una per il loro raffinamento). Le reti backbone ed FPN usate da VFNet sono le stesse di quelle usate dal *Fully convolutional one-stage object detection* (FCOS) [Tian et al 2019]

Siccome l'attività di annotazione dei video raccolti nel progetto NAUSICAA (Sezione 2) è ancora in corso, non è stato ancora possibile effettuare dei test quantitativi sulle performance delle due reti considerate per lo scenario del riconoscimento uomo in mare. Pertanto sono stati effettuati dei test qualitativi. Alcuni risultati sono riportati di seguito (Fig 8-12). Le reti YOLOv3 e VFNet sono state testate su un insieme rappresentativo di 12 video (caratterizzati da diverse quote di volo, angoli di ripresa e scenari uomo in mare) al fine di valutare eventuali potenzialità o criticità nel riconoscimento e detection delle persone/oggetti in mare. E' stato osservato che la rete YOLOv3 riesce ad identificare e riconoscere l'uomo in mare per quote di volo basse (20 m), ad identificare l'uomo ma spesso classificandolo erroneamente come ucello, areoplano od altro su quote medio basse (30-40 m) mentre presenta criticità (anche solo per la detection) per quote superiori ai 40m. Il modello VFNet si è invece dimostrato più robusto nella detection sulle varie quote di volo, riuscendo quasi sempre ad identificare i sub anche nelle riprese fatte a 60m sul livello del mare, Tuttavia per quote di volo alte presenta alcune criticità nel riconoscimento in quanto l'uomo in mare spesso viene riconosciuto come "uccello". Questa imprecisione, così come quelle di YOLO su quote di volo basse, possono ricondursi al fatto che l'addestramento delle reti usate è stata fatta su COCO (https://cocodataset.org/) che è un dataset che contiene varie foto di persone, animali ed oggetti ritratti in situazioni di vita quotidiana. Nel dataset esistono alcune immagini di uomo in mare, come ad esempio persone che fanno sport in acqua, ma questi sono ripresi da angolazioni e distanze completamente diverse da quelle usate nel progetto (principalmente sono foto frontali e quasi mai riprese

dall'alto, non confrontabili quindi con le riprese aeree da quote di volo elevate). D'altro canto esistono molte immagini nel dataset che ritraggono uccelli che galleggiano sul pelo dell'acqua o aerei su uno sfondo uniforme del cielo, da cui potrebbe derivare il possibile bias riscontrato nelle reti testate che anche se riescono a fare la detection dell'uomo in acqua a volte lo classificano in maniera errata. Da questa analisi qualitativa si è evidenziato quindi che la rete YOLO potrebbe essere utilizzata per l'analisi di riprese da drone con quote di volo inferiori a 30 metri. Tuttavia, la rete Per VFNet si dimostrata più adatta allo scopo del progetto perché riesce a fare la detection di oggetti anche molto piccoli in un'immagine (come accade quando le riprese sono fatte da quote di volo alte), tuttavia la parte di riconoscimento dell'oggetto identificato dovrà essere migliorata per i fini del progetto per eliminare classificazioni non corrette.
Un altro aspetto da considerare ai fini del progetto è la velocità di analisi: YOLO è una rete molto veloce di VFNet. Ad esempio per l'analisi di un video 4K, YOLOv3 ha impiegato circa 1 secondo per frame, mentre VFNet ha richiesto circa 1.5 secondi per frame.
Si prevede quindi, nel corso del progetto, di continuare l'analisi delle performance (efficienza ed efficacia) usando varie risoluzioni di input delle immagini ed eventualmente modificando le reti utilizzate per renderle più performanti. Risultati quantitativi saranno disponibili non appena sarà ultimata l'annotazione del dataset.

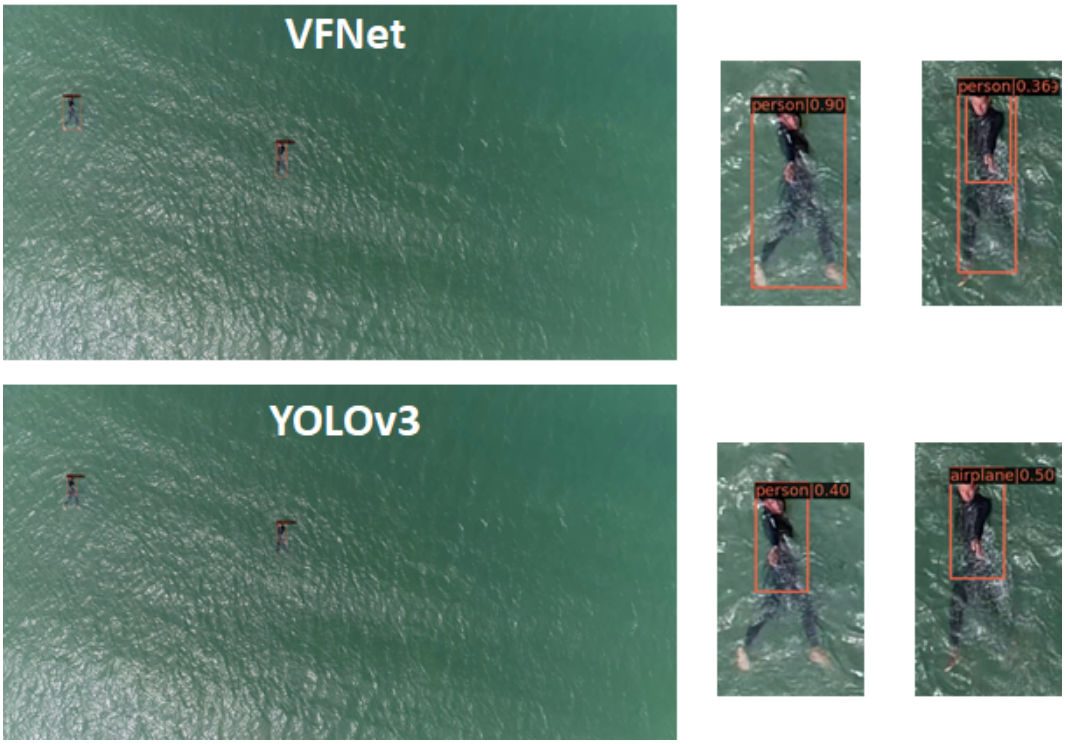

Fig 8: Risultati qualitativi - video ID 4 (altezza 20m), in alto un esempio di frame processato con VFNet, in basso con YOLOv3. A destra dei frame analizzati vi è uno zoom sui due sub. In generale, sia YOLO che VFNet hanno identificato in quasi tutti i frame del video entrambi i sub che stanno nuotando. VFNet li ha riconosciuti correttamente, classificandoli quasi sempre come "person". YOLO, invece, li ha riconosciuti spesso come "person" ma alcune volte anche come "kite", "airplane" e "bird".

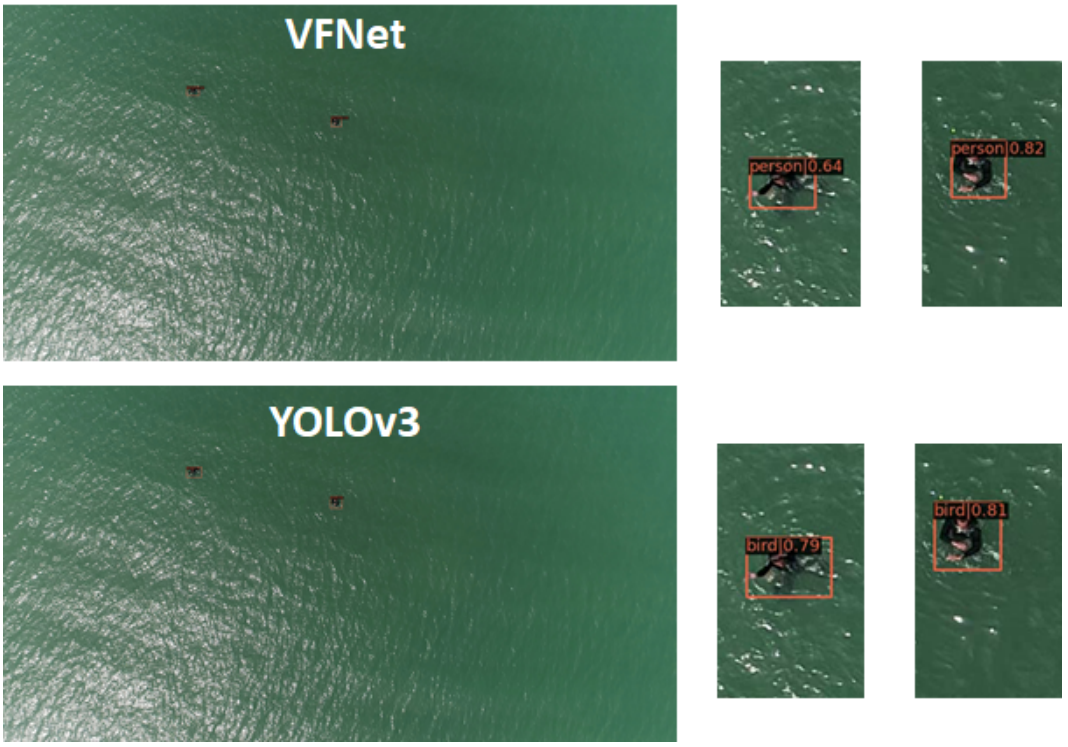

Fig 9: Risultati qualitativi video ID 5 (altezza 30m), in alto un esempio di frame processato con VFNet, in basso con YOLOv3. A destra dei frame analizzati vi è uno zoom sui due sub. Sia YOLO che VFNet hanno identificato in quasi tutti i frame del video entrambi i sub che stanno galleggiando e sbracciando. VFNet li ha riconosciuti quasi sempre correttamente come "person" e a volte come "bird" e raramente come "kite". La classificazione fatta da YOLO sembra meno accurata in quanto i sub sono stati riconosciuti a volte come "person", ma molto spesso anche come "bird" o "airplane"

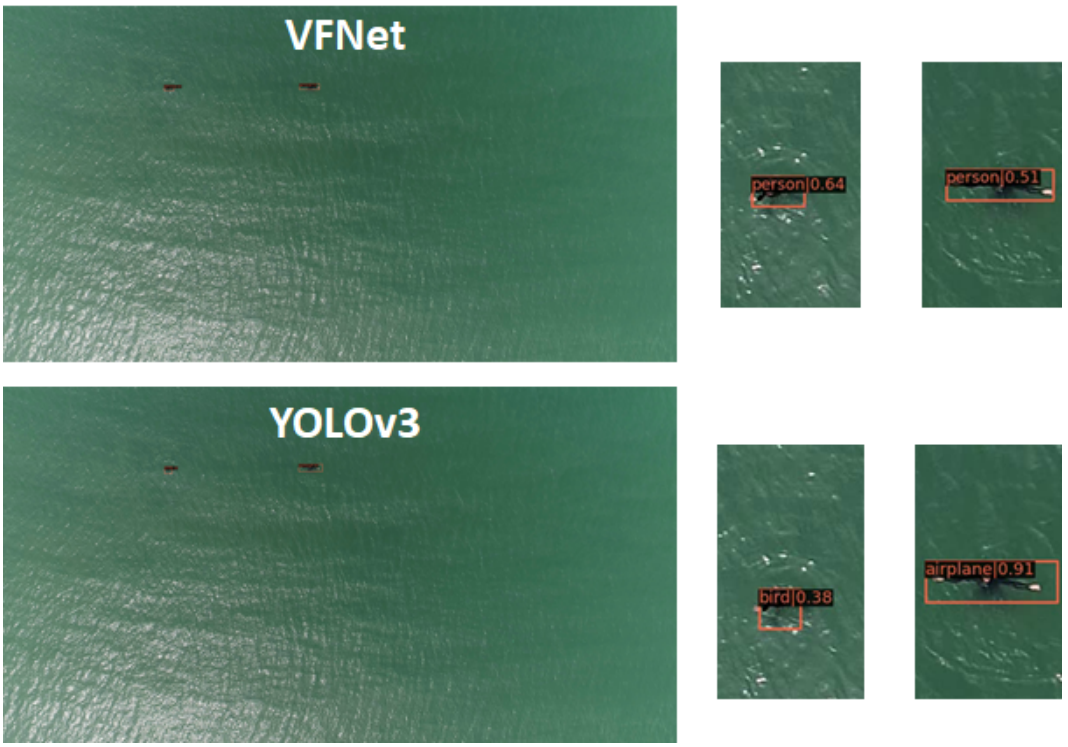

Fig 10: Risultati qualitativi video ID 6 (altezza 40m), in alto un esempio di frame processato con VFNet, in basso con YOLOv3. A destra dei frame analizzati vi è uno zoom sui due sub. Sia YOLO che VFNet hanno identificato in quasi tutti i frame del video entrambi i sub che stanno galleggiando e sbracciando. VFNet li ha riconosciuti spesso correttamente come "person" o come "bird" ed a volte come "surfboard". La classificazione fatta da YOLO sembra meno accurata in quanto i sub sono stati riconosciuti quasi sempre come "bird" o "airplane", e qualche volta come "kite", ma quasi mai come "person".

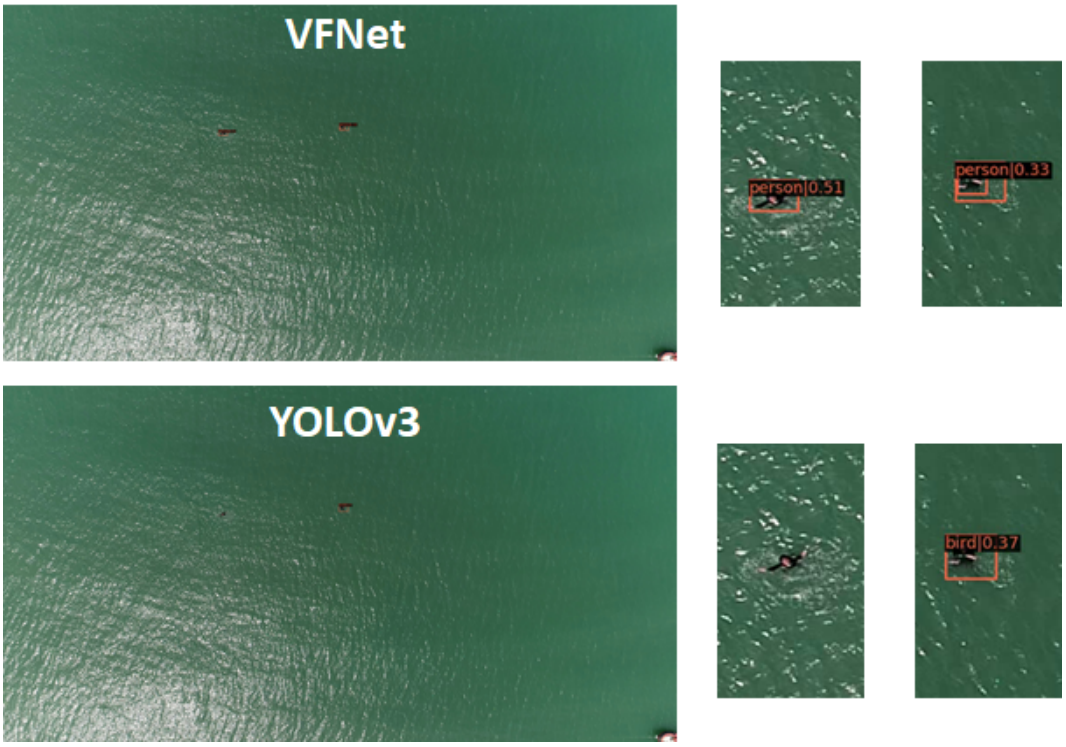

Fig 11: Risultati qualitativi video ID 7 (altezza 50m), in alto un esempio di frame processato con VFNet, in basso con YOLOv3. A destra dei frame analizzati vi è uno zoom sui due sub. In questo caso VFNet ha quasi sempre fatto correttamente la detection di entrambi i sub,che stanno galleggiando e sbracciando, riconoscendoli principalmente come “person” o “bird”. YOLOinvece quasi sempre non è riuscita a fare la detection di uno od entrambi i sub, e quando la detection è andata a buon fine la classificazione era quasi sempre errata (“bird” o “airplane”).

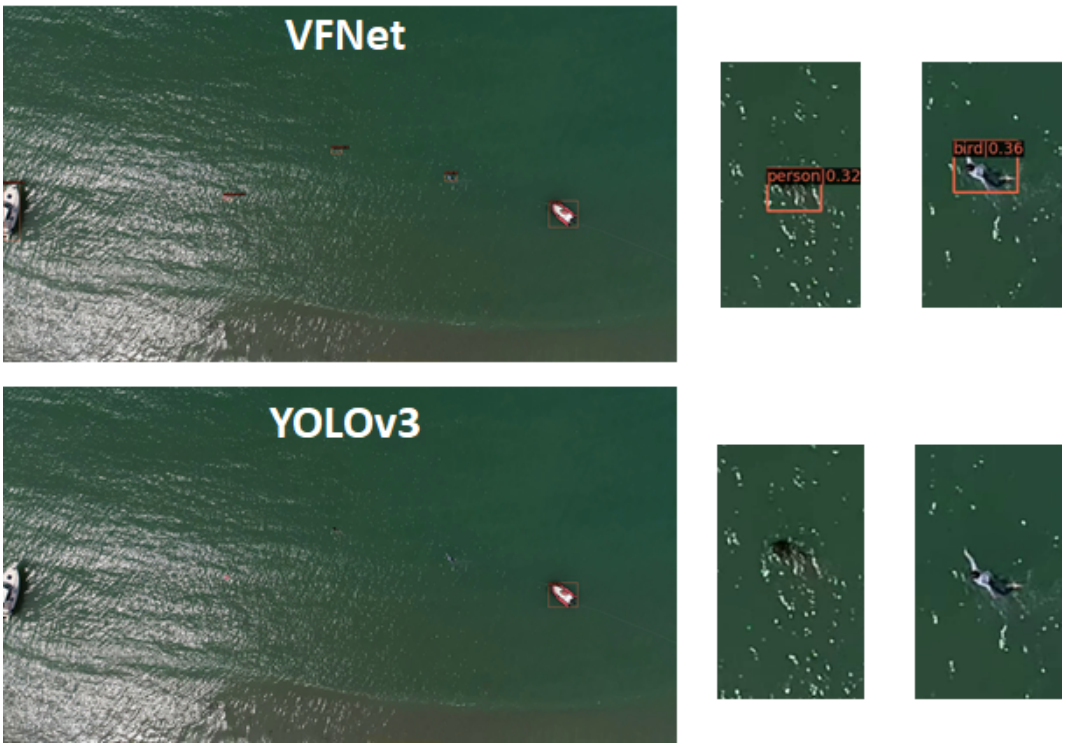

Fig 12: Risultati qualitativi video ID 10 (altezza 60m), in alto un esempio di frame processato con VFNet, in basso con YOLOv3. A destra dei frame analizzati vi è uno zoom sui due sub. In questo caso VFNet ha quasi sempre fatto la detection di entrambi i sub, riconoscendoli principalmente come “person” o “bird, e più raramente con altri oggetti. VFNet non è riuscito quasi mai a fare la detection dei due sub.

4. Architettura hardware

In questa sezione vengono illustrate in dettaglio le specifiche tecniche delle componenti hardware utili per la realizzazione ed il funzionamento della parte software del progetto volta alla realizzazione di un sistema di comunicazione tra il drone ed il sistema supervisore e di un algoritmo per il tracking automatico di persone ed oggetti in mare.
Di seguito vengono evidenziate, in dettaglio le specifiche tecniche dell'hardware utilizzato.

Huawei Honor10

Specifiche Tecniche	
Display	5,84" FHD+ / 1080 X 2280 PX
Fotocamera	16 MPX ƒ/1.8
Frontale	4 MPX ƒ/2.0
CPU	OCTA 2.36 GHZ
RAM	4 GB
Batteria	3400 MAH
Android	10 Q

Il device Android è utilizzato per testare l'applicazione (MoB App) sviluppata per garantire lo streaming del flusso video proveniente dal drone verso la Jetson ed il sistema supervisore e per gestire le richieste provenienti da quest'ultimo (es: volare verso una o più posizioni specifiche) verso il drone stesso.

Jetson Xavier NX

Specifiche Tecniche	
GPU	**NVIDIA Volta architecture con 384 NVIDIA CUDA® cores e 48 Tensor cores**
CPU	**6-core NVIDIA Carmel ARM®v8.2 64-bit CPU 6 MB L2 + 4 MB L3**
DL Accelerator	**2x NVDLA Engines**
Vision Accelerator	**7-Way VLIW Vision Processor**
Memory	**8 GB 128-bit LPDDR4x @ 51.2GB/s**
USB	**4x USB 3.1, USB 2.0 Micro-B**
Camera	**2x MIPI CSI-2 DPHY lanes**
Connectivity	**Gigabit Ethernet, M.2 Key E (WiFi/BT included), M.2 Key M (NVMe)**

Nvidia Jetson Xavier NX è un “embedded system-on-module” (SoM) utile, grazie alle sue componenti tra cui i Deep Learning Accelerators (DLAs), per lo sviluppo e l’esecuzione di algoritmi di deep learning e computer vision. Su questa scheda verrà eseguito l’algoritmo di rilevamento e di tracking di uomini in mare.

Phantom 4 Pro V 2.0

- **Aeromobile**

Specifiche Tecniche aeromobile	
Peso (con batteria ed eliche)	1375 g
Massima velocità di salita	Modalità S: 6 m/s Modalità P: 5 m/s
Massima velocità di discesa	Modalità S: 4 m/s Modalità P: 3 m/s
Velocità massima	Modalità S: 72 km/h Modalità A: 58 km/h Modalità P: 50 km/h
Quota massima di tangenza sopra il livello del mare	19685 piedi (6000 m)
Autonomia di volo	Circa 30 minuti
Intervallo di temperatura operativa	0 – 40 °C
Sistemi di posizionamento satellitare	GPS/GLONASS

- **Fotocamera**

Specifiche Tecniche fotocamera	
Sensore	**CMOS 1” Pixel effettivi: 20 M**
Obiettivo	**FOV 84° 8,8 mm/24 mm (formato 35mm equivalente) f/2.8 - f/11 messa a fuoco automatica 1 m – ∞**
Intervallo ISO	**Video: 100-3200 (automatico) 100-6400 (manuale) Foto: 100-3200 (automatico) 100-12800 (manuale)**
Velocità dell’otturatore meccanico	**8-1/2000 s**
Velocità dell’otturatore elettronico	**8-1/8000 s**
Dimensione dell’immagine	**3:2 rapporto d’aspetto: 5472 × 3648 4:3 rapporto d’aspetto: 4864 × 3648 16:9 rapporto d’aspetto: 5472 × 3078**
Dimensione immagine PIV	**4096×2160 (4096×2160 24/25/30/48/50p) 3840×2160 (3840×2160 24/25/30/48/50/60p) 2720×1530 (2720×1530 24/25/30/48/50/60p) 1920×1080 (1920×108024/25/30/48/50/60/120p) 1280×720 (1280×720 24/25/30/48/50/60/120p)**
Modalità di registrazione video	**H.265 C4K:4096×2160 24/25/30p a 100Mbps 4K:3840×2160 24/25/30p a 100Mbps 2.7K:2720×1530 24/25/30p a 65Mbps 2.7K:2720×1530 48/50/60p a 80Mbps FHD:1920×1080 24/25/30p a 50Mbps FHD:1920×1080 48/50/60p a 65Mbps FHD:1920×1080 120p a 100Mbps HD:1280×720 24/25/30p a 25Mbps HD:1280×720 48/50/60p a 35Mbps HD:1280×720 120p a 60Mbps H.264 C4K:4096×2160 24/25/30/48/50/60p a 100Mbps 4K:3840×2160 24/25/30/48/50/60p a 100Mbps 2.7K:2720×1530 24/25/30p a 80Mbps 2.7K:2720×1530 48/50/60p a 100Mbps FHD:1920×1080 24/25/30p a 60Mbps FHD:1920×1080 48/50/60 a 80Mbps FHD:1920×1080 120p a 100Mbps HD:1280×720 24/25/30p a 30Mbps HD:1280×720 48/50/60p a 45Mbps HD:1280×720 120p a 80Mbps**

Bit-rate del video (max.)	**100 Mbps**
Video	**MP4/MOV (AVC/H.264; HEVC/H.265)**

Il DJI Phantom 4 Pro V2.0 è dotato di un sensore CMOS da 20 Megapixel da 1 pollice e un otturatore meccanico che elimina le distorsioni da rolling shutter. Con obiettivo f/2.8 grandangolare ottimizzato, la fotocamera di Phantom 4 Pro V2.0 garantisce riprese video in 4K/60fps e immagini in modalità Burst a 14 fps. Inoltre, il sistema FlightAutonomy comprende doppi sensori di visione posteriore e sistemi di rilevamento a infrarossi per un totale di 5 direzioni di rilevamento degli ostacoli e 4 direzioni di evitamento degli ostacoli.

- ## Radiocomando

Specifiche Tecniche radiocomando	
Frequenza operativa	**2.400-2.483 GHz e 5.725-5.850 GHz**
Distanza massima di trasmissione	**2.400 – 2.483 GHz, 5.725 – 5.850 GHz (senza ostacoli né interferenze)FCC: 10000 m CE: 6000m SRRC: 6000m MIC: 6000m**
Intervallo di temperatura operativa	**0 – 40 °C**
Batteria	**6000 mAh LiPo 2S**
Potenza del trasmettitore (EIRP)	**2.400 – 2.483 GHz FCC: 26 dBm CE: 20 dBm SRRC: 20 dBm MIC: 17 dBm 5.725-5.850 GHz FCC: 26 dBm CE: 14 dBm SRRC: 20 dBm MIC: -**
Tensione/Corrente operativa	**1,2 A a 7,4 V**
Porta di uscita video	**GL300K: HDMI GL300L: USB**
Supporto per dispositivi mobili	**GL300K: Dispositivo con display incorporato (schermo 5.5-pollici, 1920×1080, 1000 cd/m2, sistema Android, 4 GB RAM＋16 GB ROM) GL300L: Tablet e smartphone**

- **Bibliografia**

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

[Redmon and Farhadi 2018] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[Tian et al 2019] Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9627-9636).

[Zhang et al 2021] Zhang, H., Wang, Y., Dayoub, F., & Sunderhauf, N. (2021). VarifocalNet: An iou-aware dense object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8514-8523).