D1.4
Prototipi Analisi Visuale

Lucia Vadicamo, Claudio Vairo, Luca Ciampi,
Claudio Gennaro, Donato Cafarelli, Fabrizio Falchi

INDEX

Introduzione **2**

Dataset **3**

2.1 Raccolta Dati 3

2.2 Annotazione dei Dati **7**

Sperimentazione riconoscimento primo prototipo **11**

Sperimentazione riconoscimento secondo prototipo **16**

Architettura Hardware **21**

Architettura Software **25**

Bibliografia **27**

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 A1.4 "Realizzazione seconda versione dei prototipi Analisi Visuale"
Tale attività ha avuto per scopo la realizzazione della seconda versione del 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 e annotazione di un dataset di riprese da drone di persone in mare realizzato appositamente per il progetto e necessario per 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 il secondo prototipo.
- Architettura Hardware: vengono descritte le specifiche del drone (inclusi radio controller e modulo camera), del dispositivo mobile collegato al drone dove viene installata l'applicazione sviluppata (in questo caso un tablet Android) e della piattaforma dove è installata la rete neurale e che esegue il task di analisi del video per la detection della persona in mare (NVIDIA Jetson).
- Architettura Software: viene descritto in dettaglio lo schema di funzionamento dell'applicazione mobile sviluppata.

2. Dataset

Questa sezione descrive l'attività di realizzazione del dataset ***MOBDrone***, contenente oltre 125.000 immagini acquisite da droni aerei in un ambiente marino sotto diverse condizioni, come varie altitudini, angoli di ripresa e illuminazione. Le immagini sono state annotate manualmente con più di 180K oggetti, di cui circa 113K uomini in mare, dove ciascun oggetto è stato localizzato con precisione con una bounding box.
Il Dataset è stato presentato alla 21esima edizione della International Conference on Image Analysis and Processing (ICIAP) [Cafarelli et al 2022] ed è stato rilasciato pubblicamente su Zenodo (https://doi.org/10.5281/zenodo.5996890).

La creazione di tale dataset è stata necessaria per lo svolgimento delle attività A1.2 ed A1.4 del progetto che prevedono 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 stato di fondamentale importanza avere a disposizione dati per la 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.

2.1 Raccolta Dati

Questa sezione descrive la campagna di raccolta dei dati utilizzati per la creazione del MOBDrone dataset. 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.

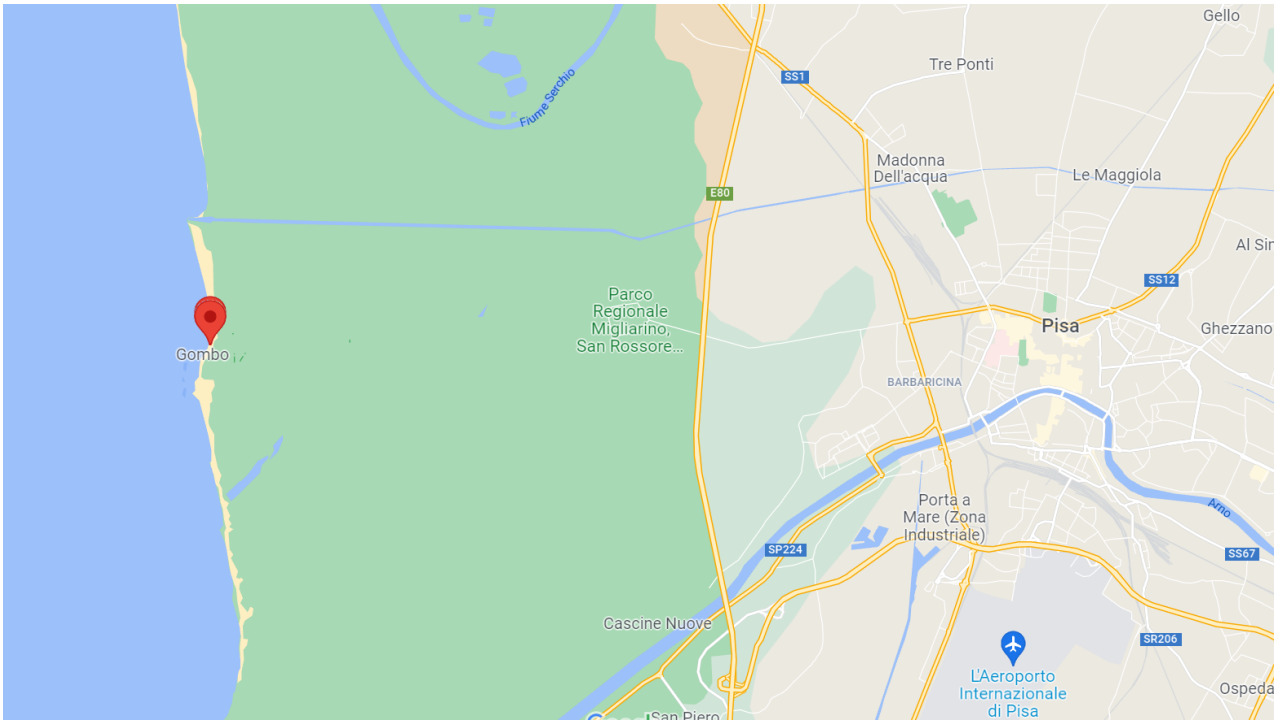


Fig. 1: Luogo in cui sono state effettuate le attività di riprese da drone aereo e di immersione in acqua (coordinate 43.7212778,10.2787972).

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 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 è previsto di poter 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)

In totale sono state effettuate 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 sì è provveduto ad eliminare alcune porzioni di video in cui le persone erano identificabili, per cui alcuni dei video originali sono stati divisi in più videoclip.

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.

Il set di dati finale contiene **66 videoclip.** I dati sono stati poi post-elaborati come descritto nella sezione seguente.

2.2 Annotazione dei Dati

Per prima cosa, abbiamo convertito i 66 videoclip dalla risoluzione 4K a 1080p e da essi abbiamo estratto i fotogrammi a una velocità di 30 FPS, ottenendo un totale di 126,170 immagini (si veda la Tabella 1 per le statistiche riassuntive).

Altezza Drone		**# Immagini**	**# Videoclips**
10 m		958	1
20 m		10, 053	6
30 m		29, 404	15
40 m		33, 046	13
50 m		29, 183	16
60 m		23, 526	15
	tot	**126,170**	**66**

Tabella 1. Dettagli Dataset MobDrone.

I dati ottenuti, per poter essere utilizzati ai fini delle attività del progetto, devono essere annotati, ovvero bisogna identificare nei keyframe dei video se appare un oggetto o una persona ed 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 Cafarelli (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 che consente sia di annotare il dataset manualmente e sia di sfruttare uno dei modelli di Deep Learning presenti per un'annotazione automatica.

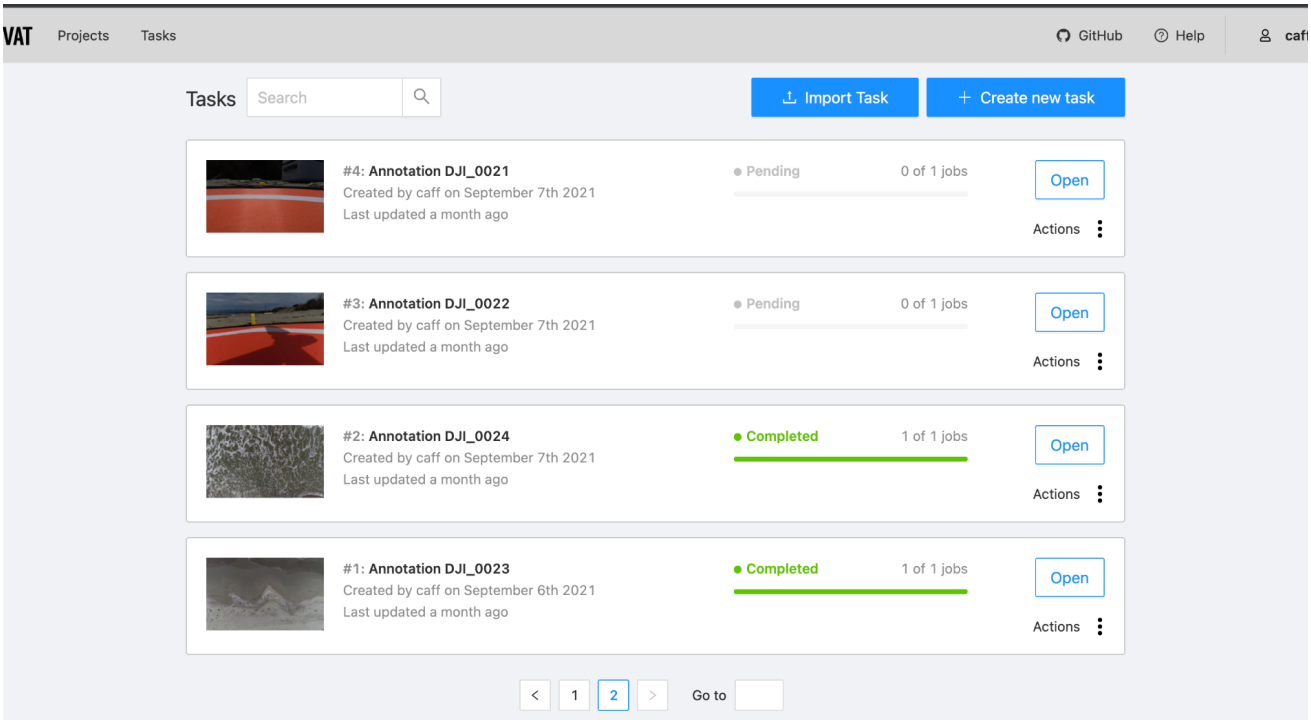


Fig. 4: CVAT Master Task page

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.

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 sono state impiegate circa 60 ore di lavoro dell'annotatore.

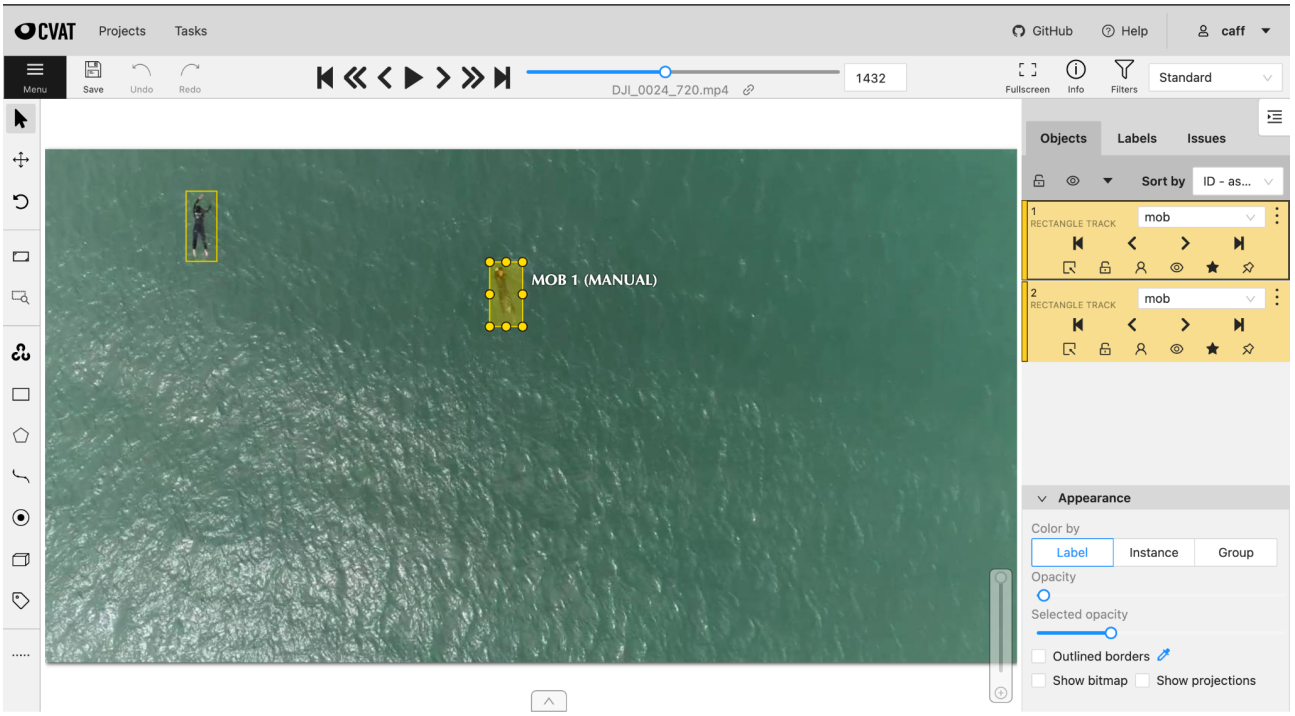


Fig. 5: Esempio processo di annotazione mediante il tool CVAT

Sebbene il nostro lavoro si concentri principalmente sulla localizzazione e sul riconoscimento di persone, abbiamo annotato anche altri oggetti presenti nelle scene. In particolare, abbiamo considerato un totale di 5 classi (*person, boat, surfboard, wood, life buoy*) ed abbiamo individuato una bounding box che localizza con precisione ogni istanza degli oggetti di interesse.
In totale sono state annotate 181,689 istanze, di cui 113,408 relative alla classe *person*, che è di interesse primario nello scenario MOB. Si noti tuttavia che circa il 27.72% delle immagini non contiene oggetti (ad esempio, immagini di acqua limpida).
Riportiamo dall'articolo [Cafarelli et al 2022] la Tabella 2 dove sono riassunte alcune statistiche relative alle annotazioni e la Figura 6 dove vengono mostrati alcuni campioni del nostro dataset

Class	#Annotations	#Images	Samples
person	113,408	77,365	
boat	39,967	31,238	
wood	15,980	9,040	
life buoy	10,401	10,386	
surfboard	1,933	1,933	
no object		34,976	
total	181,689		

Tabella 2: Statistiche relative all'annotazione

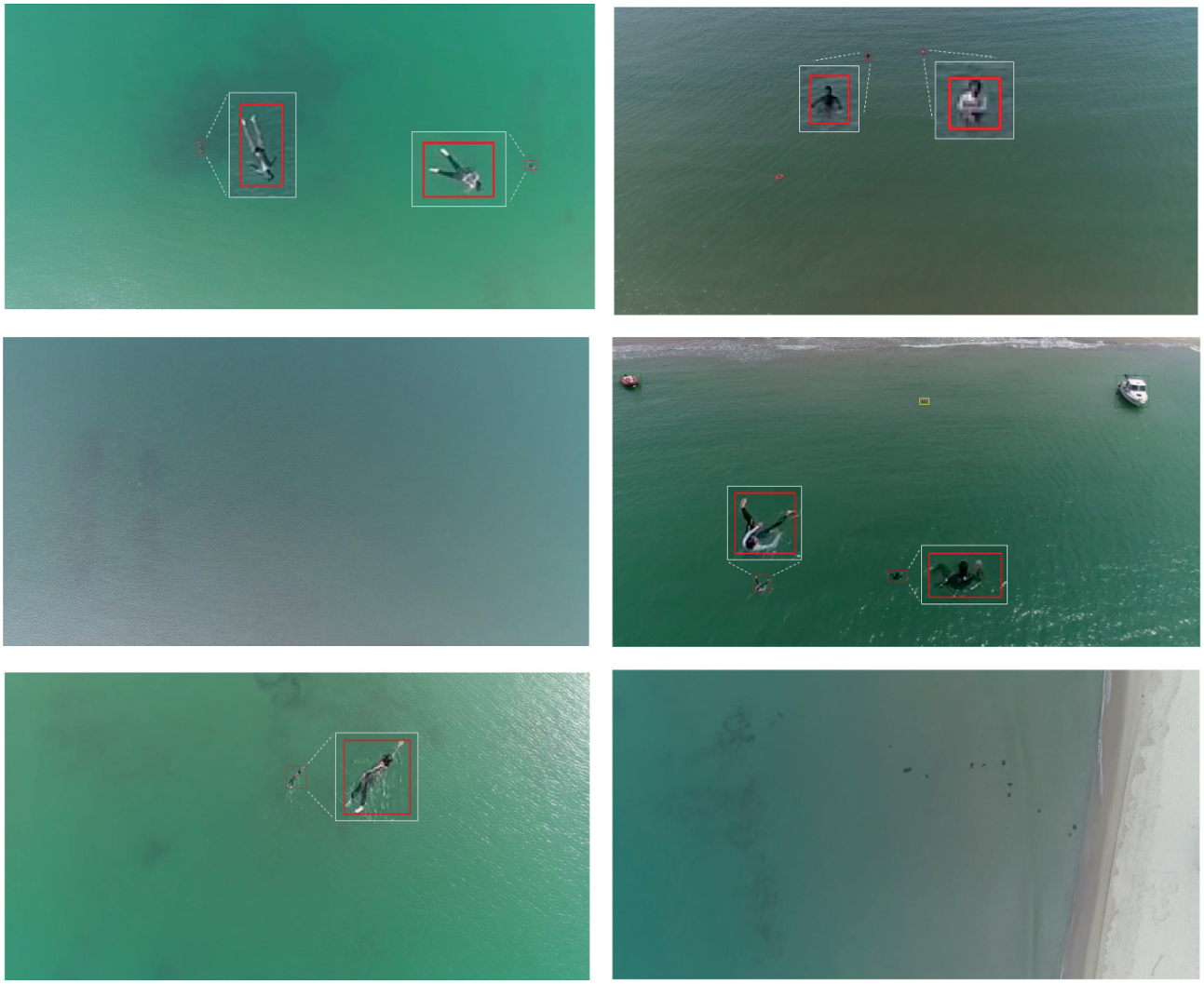


Fig.6: Esempi di immagini acquisite a diverse altitudini, condizioni di luce e direzioni della telecamera. Sono mostrate anche le annotazioni con le bounding box che localizzano gli oggetti etichettati. Gli oggetti appartenenti alla categoria *person*, che è quella di maggiore interesse negli scenari MOB, sono delineati con bounding box rosse ed ingranditi con uno zoom. Si noti che il 27,72% delle immagini non contengono oggetti (ad esempio, immagini di acqua limpida) e che gli oggetti interferenti sullo sfondo, come le rocce, spesso innescano rilevamenti di falsi positivi.

3. Sperimentazione riconoscimento primo prototipo

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.

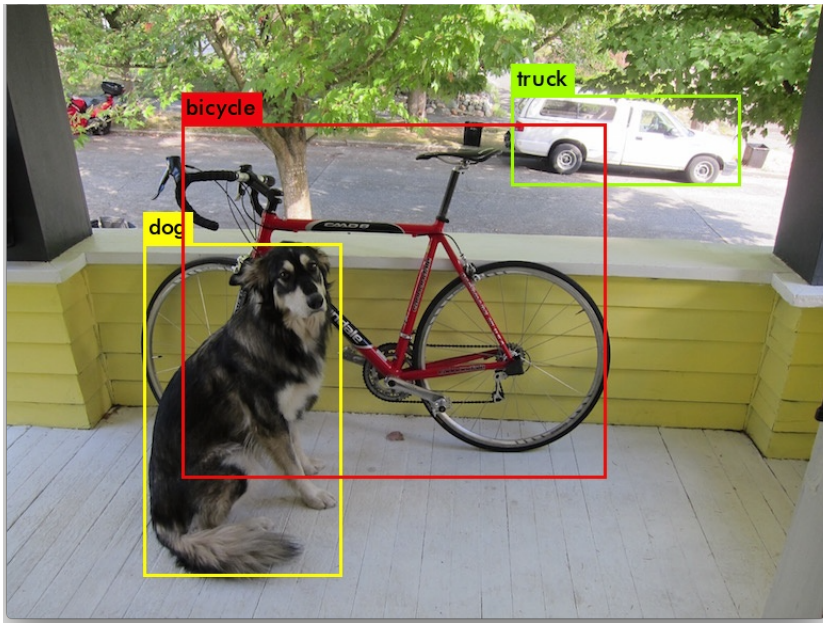


Fig. 7: Esempio di detection e riconoscimento mediante l'uso di YOLOv3 (fonte: https://pjreddie.com/darknet/yolo/)

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 pre-allenata 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 è 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 bounding box 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.

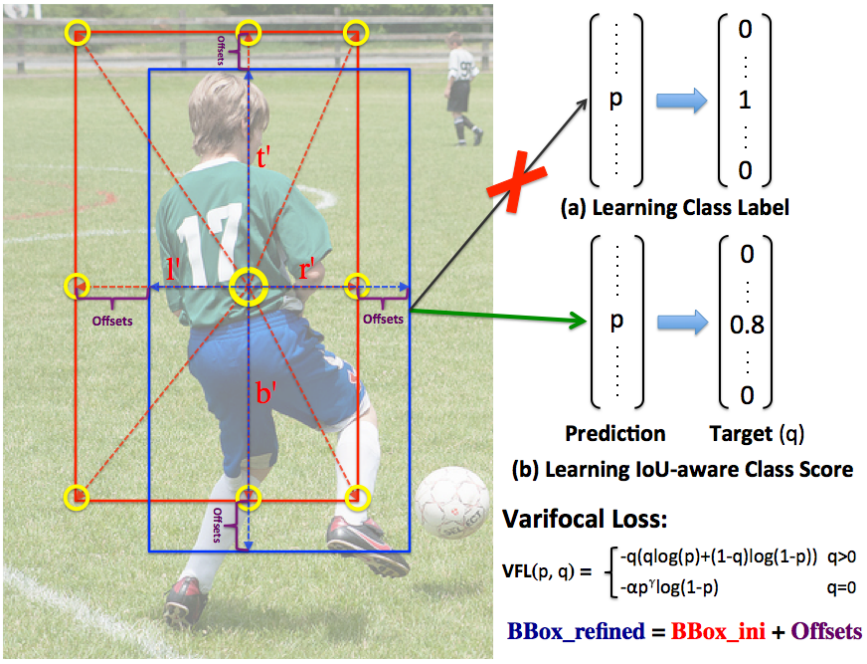


Fig. 8: illustrazione del metodo usato in VarifolcalNet (fonte [Zhang et al 2021]). Invece di imparare a prevedere l'etichetta di una classe per un determinato un bounding box, come mostrato in (a), il metodo proposto impara a predire un punteggio di classificazione della localizzazione (ovvero la misura 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* (composta 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) ai tempi della realizzazione del primo prototipo era ancora in corso, in prima istanza non era stato 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. 9-13). 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. È stato osservato che la rete YOLOv3 riesce a identificare e riconoscere l'uomo in mare per quote di volo basse (20 m), a identificare l'uomo ma spesso classificandolo erroneamente come uccello, 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 a 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 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 più 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.

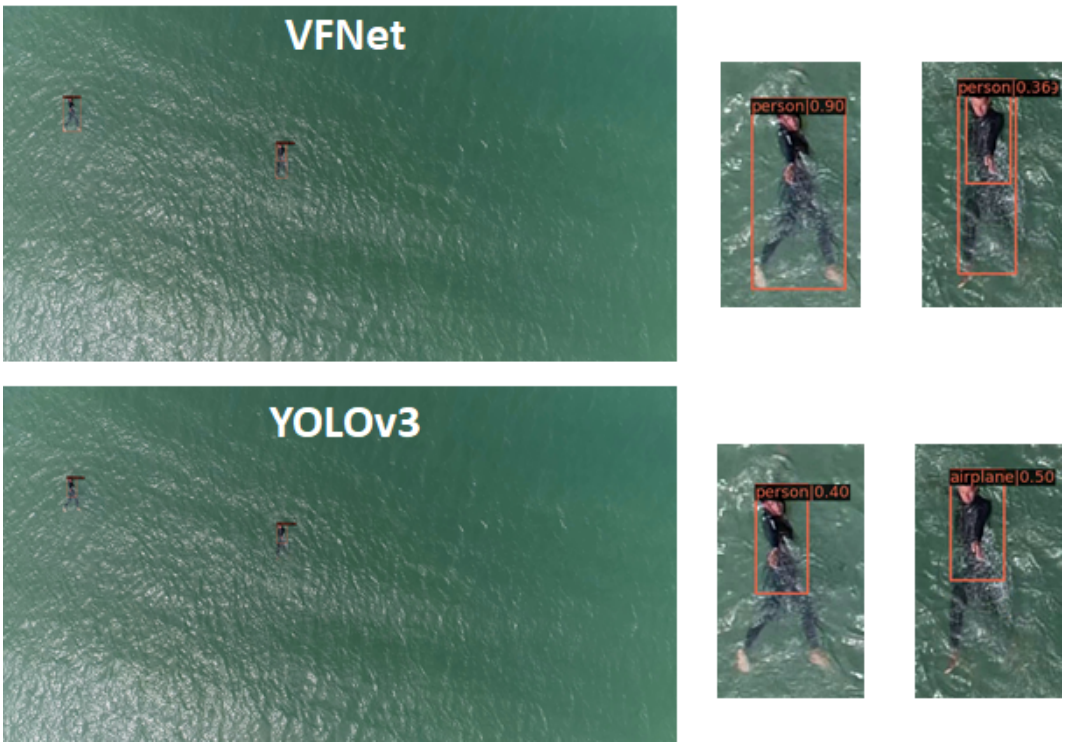


Fig. 9: Risultati qualitativi - video ID 4 (altezza 20m), in alto un esempio di frame processato con VFNet, in basso con YOLOv3. A destra dei frame analizzati vi è uno zoom sui due sub. In generale, sia YOLO che VFNet hanno identificato in quasi tutti i frame del video entrambi i sub che stanno nuotando. 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".

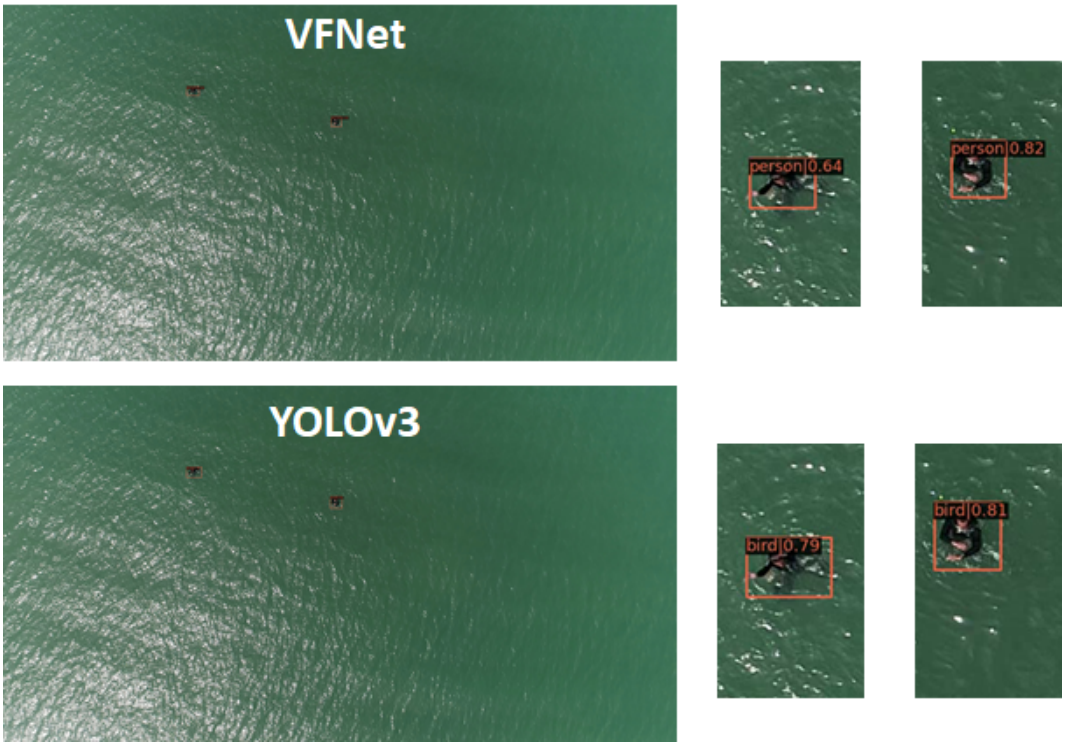


Fig. 10: Risultati qualitativi video ID 5 (altezza 30m), in alto un esempio di frame processato con VFNet, in basso con YOLOv3. A destra dei frame analizzati vi è uno zoom sui due sub. Sia YOLO che VFNet hanno identificato in quasi tutti i frame del video entrambi i sub che stanno galleggiando e sbracciando. 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"

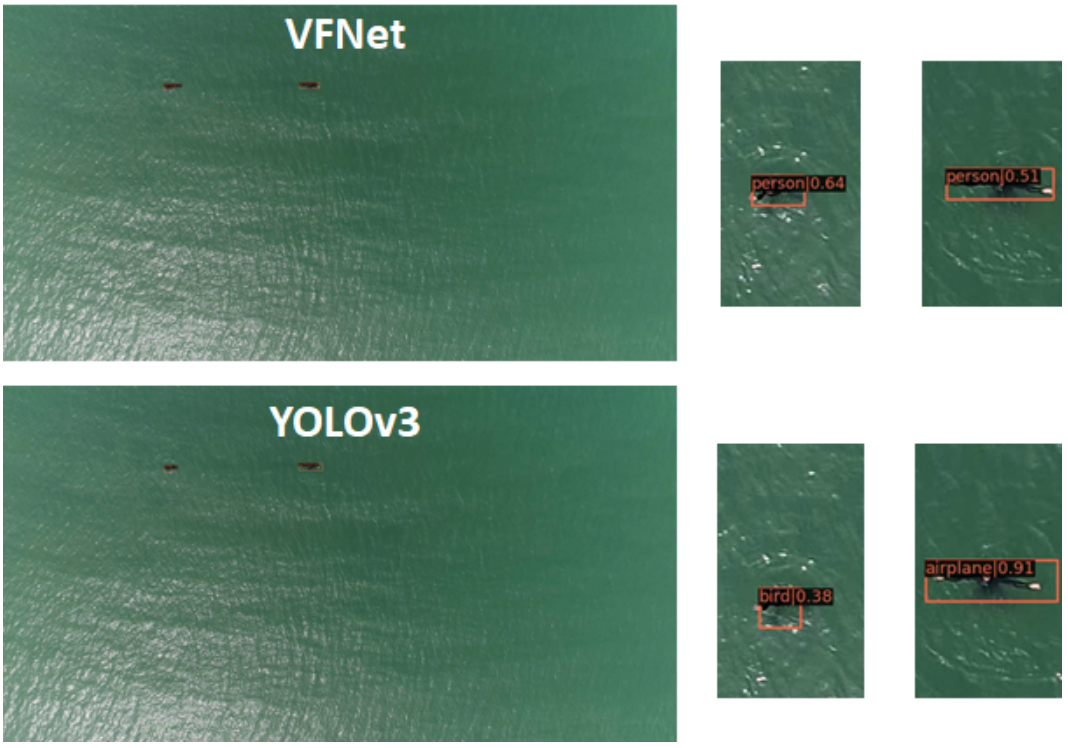


Fig. 11: Risultati qualitativi video ID 6 (altezza 40m), in alto un esempio di frame processato con VFNet, in basso con YOLOv3. A destra dei frame analizzati vi è uno zoom sui due sub. Sia YOLO che VFNet hanno identificato in quasi tutti i frame del video entrambi i sub che stanno galleggiando e sbracciando. 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".

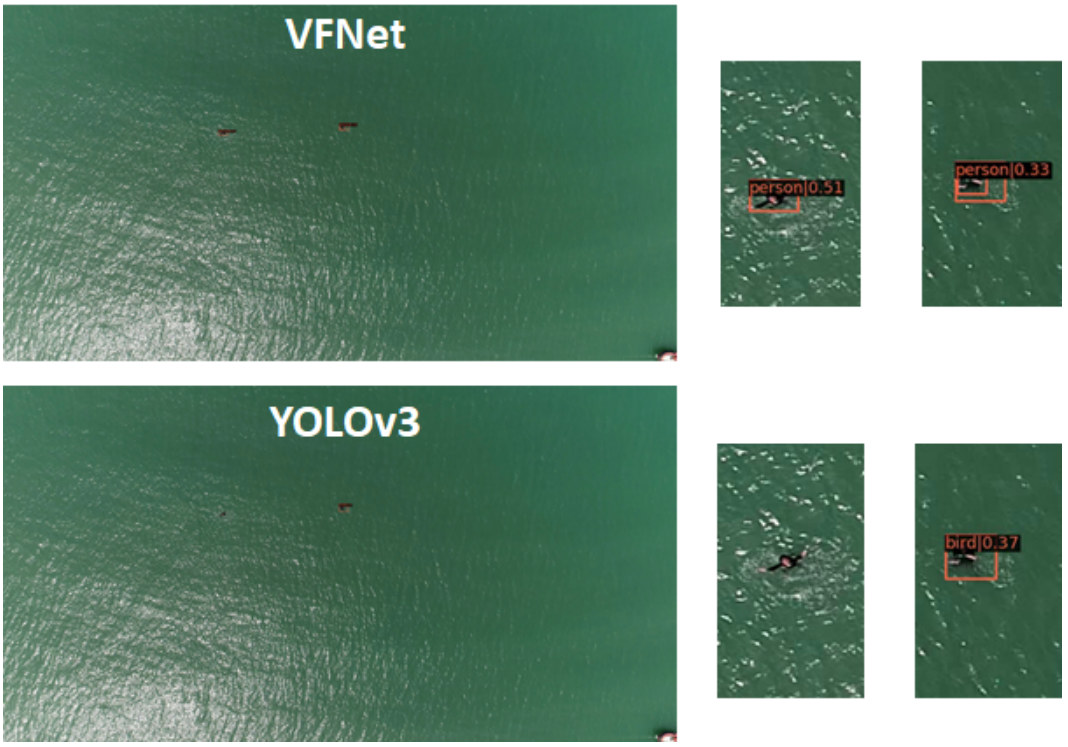


Fig. 12: Risultati qualitativi video ID 7 (altezza 50m), in alto un esempio di frame processato con VFNet, in basso con YOLOv3. A destra dei frame analizzati vi è uno zoom sui due sub. In questo caso VFNet ha quasi sempre fatto correttamente la detection di entrambi i sub, che stanno galleggiando e sbracciando, riconoscendoli principalmente come "person" o "bird". YOLO invece 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").

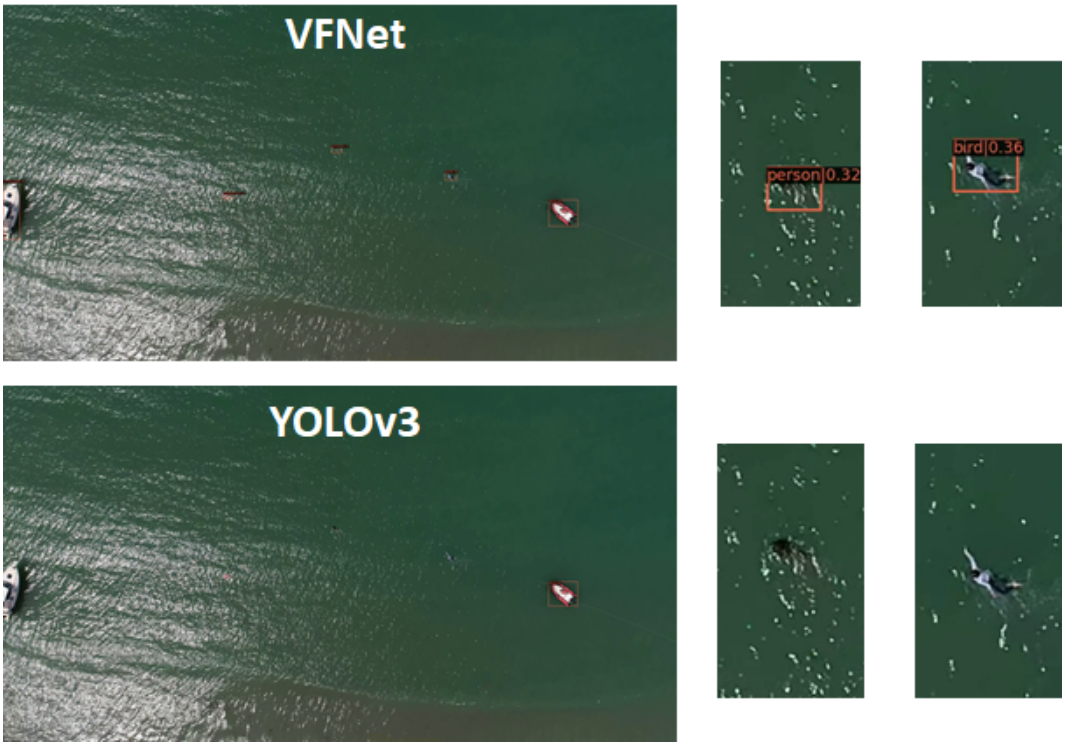


Fig. 13: Risultati qualitativi video ID 10 (altezza 60m), in alto un esempio di frame processato con VFNet, in basso con YOLOv3. A destra dei frame analizzati vi è uno zoom sui due sub. In questo caso VFNet ha quasi sempre fatto la detection di entrambi i sub,

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. Sperimentazione riconoscimento secondo prototipo

Durante l'attività di realizzazione del secondo prototipo è stata ultimata l'annotazione del dataset ed è stato quindi possibile effettuare una sperimentazione quantitativa sui dati raccolti. I risultati di tale attività sono stati presentati su un articolo a conferenza internazionale [Cafarelli et al 2022].

Abbiamo valutato diversi object detector allo stato dell'arte sul nostro dataset MOBDrone concentrandoci sul rilevamento delle persone in mare, cioè sulla localizzazione delle istanze di oggetti appartenenti alla classe *person*.

Nella prima parte della nostra analisi delle prestazioni, abbiamo confrontato 9 dei più popolari e performanti *detector* di oggetti presenti in letteratura. In seguito, abbiamo esaminato i tre migliori, effettuando un'analisi più approfondita dei risultati ottenuti.

I detector considerati nella nostra analisi possono essere approssimativamente raggruppati in tre categorie: i metodi basati su reti neurali convoluzionali (CNN) con ancoraggio, i metodi CNN senza ancoraggio e i metodi basati su transformer. Li riassumiamo brevemente di seguito (per maggiori dettagli rimandiamo il lettore ai relativi articoli).

- ***Anchor-based CNN methods*:** metodi che calcolano le posizioni delle bounding box e le classi degli oggetti sfruttando architetture di tipo CNN che si basano su ancore, cioè su bounding box preliminari con varie scale e aspect ratio. Si possono dividere in due gruppi: i) l'approccio a due stadi, in cui un primo modulo è responsabile della generazione di un insieme sparso di proposte di oggetti e un secondo modulo è incaricato di raffinare queste previsioni e classificare gli oggetti; e ii) l'approccio a uno stadio che fa un'inferenza direttamente delle bounding box campionando su griglie regolari e dense, saltando la fase di proposta delle regioni.
  Di questa classe di metodi abbiamo testato *Faster R-CNN* [Ren et al. 2017] e *Mask R-CNN* [He et al 2017] per quanto riguarda il primo gruppo, e YOLOv3 [Redmon and Farhadi 2018], TOOD [Feng et al 2021] e VarifocalNet (VfNet) [Zhang et al 2021] per il secondo.
- ***Anchor-free CNN methods*:** metodi che si basano sulla previsione di key-points, come i punti d'angolo o centrali, per la detection degli oggetti, invece di utilizzare le bounding box di ancoraggio e le loro limitazioni intrinseche. Di questa classe abbiamo considerato CenterNet [Zhou et al 2019] e YOLOX [Ge et al 2021].
- ***Transformer-based methods:*** metodi che utilizzano moduli di attenzione di tipo *Transformer*, introdotti di recente, per l'elaborazione di image feature maps, eliminando la necessità di componenti progettati a mano, come procedure di non-maximum suppression o la generazione di ancore. Di questa classe abbiamo considerato il DEtection TRansformer (DETR) [Carion et al 2020] e una sua evoluzione, il Deformable DETR [Zhu et al. 2021].

Abbiamo valutato e confrontato i detector sopra descritti sul nostro dataset MOBDrone utilizzando la *Average Precision* (AP), ovvero il valore medio di precisione per valori di *recall* compresi tra 0 e 1. In particolare, per la valutazione dell'efficacia abbiamo utilizzato le seguenti metriche

- la **MS COCO mAP@[0.50:0.95**] (definita come in [Lin et al 2014]), ovvero, l'AP mediato su 10 soglie di *Intersection-over-Union* (IoU) nell'intervallo [0.50 , 0.95] con un passo di 0.05
- l'**AP50**, cioè l'AP calcolato al singolo valore di soglia IoU di 0.50.

Rimandiamo il lettore a [Lin et al 2014] per maggiori dettagli su tali metriche standard di valutazione.

Tutte le reti che abbiamo utilizzato sono state pre-addestrate[1] sul dataset COCO [Lin et al 2014], una popolare benchmark di immagini in contesti quotidiani che comprendono oggetti appartenenti a 80 categorie diverse, tra cui la classe *person*. Per valutare le prestazioni dei vari detector, abbiamo filtrato le rilevazioni ottenute considerando solo quelle classificate come "person". I risultati ottenuti sono riportati nella Tabella 3.

[1] I modelli pre-addrestati sono disponibili su https://modelzoo.co/model/mmdetection

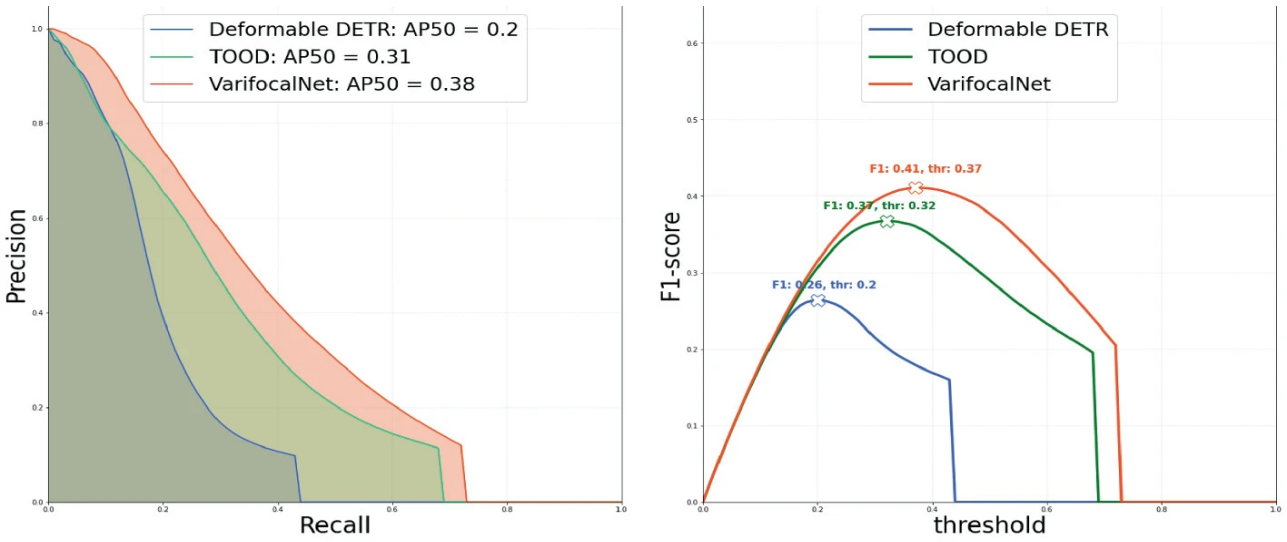
Metodo	AP50 ↑	mAP@[0.50:0.95] ↑
VarifocalNet	**0.378**	**0.144**
TOOD	0.314	0.116
Deformable DETR	0.199	0.075
YOLOX	0.126	0.049
Faster R-CNN	0.126	0.041
CenterNet	0.124	0.041
DETR	0.128	0.040
Mask R-CNN	0.109	0.033
YOLOv3	0.011	0.009

Tabella 3: Confronto dei detector considerati. mAP@[0.50:0.95] è l'AP mediato su 10 soglie IoU nell'intervallo [0.50 : 0.95] con una dimensione di passo di 0.05, mentre AP50 è l'AP calcolato al singolo valore di soglia IoU di 0.50.

Il modello che risulta più performante è VarifocalNet, considerando entrambe le metriche, seguito da TOOD e Deformable DETR. Tuttavia, in generale, tutti i detector mostrano prestazioni moderate, indicando le difficoltà del problema di localizzazione delle persone in questo scenario difficile. Riteniamo che la metrica più significativa nel nostro scenario sia l'AP50, poiché 1) il dataset è etichettato manualmente dagli esseri umani e quindi è accurato in termini di classificazione e impreciso in termini di confini delle bounding box, 2) è fondamentale localizzare con precisione le istanze, cioè è fondamentale rilevare le persone in mare ma la qualità della localizzazione è meno importante.

Tenendo conto di ciò, di seguito mostriamo un'analisi approfondita dei tre migliori modelli AP50, ovvero VarifocalNet, TOOD e Deformable DETR.

Nella Figura 14a, sono riportate le curve di *precision-recall*, ovvero i valori di *precision* e *recall* per diverse soglie di confidenza del rilevamento, di VarifocalNet, TOOD e Deformable DETR, impostando la soglia IoU a 0.50. Le aree sotto queste curve corrispondono ai valori AP50. Come si può notare, la rete VarifocalNet mostra le migliori prestazioni a tutte le soglie di confidenza. La stessa tendenza è confermata nella Figura 14b, dove vengono mostrati i valori di F_1 -score (dove $F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$) a diverse soglie di confidenza del rilevamento, sempre impostando la soglia IoU a 0.50. VarifocalNet mostra comunque prestazioni superiori rispetto agli altri due detector. Si noti che i valori massimi di queste curve indicano la soglia di confidenza del rilevamento che può essere utilizzato dai potenziali utenti, racchiudendo un compromesso tra i valori di *precision* e *recall* risultanti. La figura 15 mostra alcuni risultati qualitativi prodotti da VarifocalNet utilizzando questo la soglia di confidenza.



(a) Precision vs. Recall curves (IoU=0.5). Areas under curves correspond to AP50.

(b) F_1-score vs. detection threshold curves (IoU=0.5).

Fig. 14: Confronto tra i tre migliori detector. Riportiamo le curve di precision-recall (a) e di soglia di rilevamento F1 (b) dei tre migliori modelli (VfNet, TOOD e Deformable DETR). VfNet mostra le migliori prestazioni.

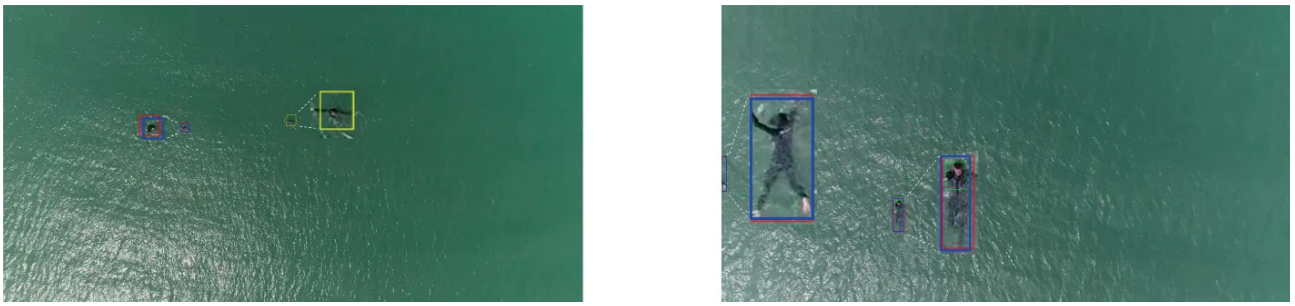


Fig. 15: Esempio di detection prodotti da VarifocalNet: in verde i falsi positivi, in blu i true positive in rosso le bounding box del ground truth

Nella Tabella 4, riportiamo il confronto tra i migliori tre detector a diverse altitudini in termini di AP50 e F_1-score. Come previsto, in generale le prestazioni diminuiscono con l'aumentare dell'altitudine. Tuttavia, è interessante notare che TOOD e Deformable DETR faticano particolarmente a rilevare oggetti di piccole dimensioni, cioè quando l'altitudine è superiore ai 40 metri, mentre ottengono risultati comparabili o addirittura migliori di VarifocalNet ad altitudini inferiori ai 30 metri.

	VarifocalNet		**TOOD**		**Deformable DETR**	
Altitudine	**AP50 ↑**	**F_1↑**	**AP50 ↑**	**F_1↑**	**AP50 ↑**	**F_1↑**
10 m	0.973	0.444	0.989	0.363	0.959	0.636
20 m	0.771	0.318	0.681	0.308	0.514	0.279
30 m	0.400	0.199	0.407	0.223	0.240	0.210
40 m	0.540	0.226	0.406	0.203	0.314	0.209
50 m	0.241	0.161	0.187	0.140	0.107	0.08
60 m	0.205	0.223	0.171	0.196	0.063	0.131

Tabella 4: Confronto dei tre migliori detector su immagini a diverse altitudini

Infine, nella Tabella 5 riportiamo un'analisi per classi delle detection ottenute, cioè consideriamo le detection appartenenti a tutte le 80 classi e non solo quelle classificate come "*person*". In particolare, si tiene conto anche degli errori dovuti agli oggetti mal classificati, cioè agli oggetti rilevati che corrispondono alle annotazioni "*person*" nel ground-truth ma che sono stati classificati come oggetti appartenenti a un'altra categoria. A tale scopo definiamo il *True Positive Rate* (TPR) come il rapporto tra il numero di istanze di persona correttamente rilevate e classificate (TP) e il numero totale di istanze di persona nel ground-truth (P). Inoltre, definiamo $dTPR(c) = \frac{dTP(c)}{P}$ come il tasso di rilevamento dei True positive per la classe di output *c* rispetto alla classe target *person* data dal numero dTP(c) di istanze di persona rilevate correttamente (cioè, considerando solo le bounding box) ma classificate con la categoria *c* diviso per il numero totale di istanze di persona nel ground-truth. In altre parole, $dTPR(c)$ indica la percentuale di istanze di persone rilevate correttamente ma classificate erroneamente con la categoria *c*. La somma del TPR e di tutti i $dTPR(c)$dà la *overall detection Recall* (dR), cioè la percentuale di istanze di persone rilevate correttamente senza considerare la classificazione in output. Allo stesso modo, la *detection Miss Rate*, definito come $dMR = 1 - dR$, è la porzione di istanze di persone che non sono state rilevate affatto. Ad esempio, dalla Tabella 5, si può osservare che la VarifocalNet pre-addestrata ha rilevato correttamente l'81.8% delle persone del ground-truth, anche se, nella maggior parte dei casi, le ha classificate male. Ciò può suggerire che lo stesso modello, messo a punto sui dati MOB, può avere un margine di crescita nella localizzazione e riconoscimento delle persone.

Il codice di valutazione e tutte le altre risorse per riprodurre i risultati sono disponibili al seguente link: http://aimh.isti.cnr.it/dataset/MOBDrone

Nella realizzazione del secondo prototipo dell'attività 1.4 è stata quindi scelta la rete VarifocalNet visto che è stata la rete che ha dimostrato le migliori performance in tutte le analisi riportate in questa sezione.

	Person	Bird	Airplane	Kite	Other	Overall	
Method	**TPR ↑**	**dTPR ↑**	**dTPR ↑**	**dTPR ↑**	**dTPR ↑**	**dR↑**	**dMR ↓**
VfNet	0.285	0.266	0.190	0.067	0.012	**0.818**	**0.182**
TOOD	0.326	0.118	0.212	0.125	0.017	0.799	0.201
Def. DETR]	0.206	0.311	0.072	0.041	0.026	0.657	0.343

Tabella 5: Analisi per classi. Sono state considerate le detection di tutte le 80 classi COCO, tenendo conto degli errori dovuti agli oggetti mal classificati, ossia alle persone che sono state rilevate ma classificate come oggetti di un'altra categoria. TPR è la percentuale dei True Positive rispetto alla classe “person”. dTPR è il rapporto tra le istanze di persone rilevate correttamente ma classificate in modo errato. La overall detection Recall (dR) è la proporzione di istanze di persona rilevate considerando anche gli oggetti classificati in modo errato; il Miss Rate complessivo del rilevamento è la proporzione di istanze di persona che non sono state rilevate affatto. È stata usata un a IoU uguale a 0.5.

5. Architettura Hardware

In questa sezione vengono illustrate in dettaglio le specifiche tecniche delle componenti hardware utilizzate 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.

Samsung Galaxy A7 Lite (cambiato rispetto a D1.2)

Specifiche Tecniche	
Display	**8,7" WXGA+ / 800 X 1340 PX**
Fotocamera	**8 MPX**
Frontale	**2 MPX**
CPU	**OCTA 2.3 GHZ**
RAM	**3 GB**
Batteria	**5100 MAH**
Android	**11**

Il dispositivo 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.

DJI MAVIC 2 ZOOM (cambiato rispetto a D1.2)

- ## Aeromobile

Specifiche Tecniche aeromobile	
Peso (con batteria ed eliche)	1202 g
Massima velocità di salita	Modalità S: 5 m/s Modalità P: 4 m/s
Massima velocità di discesa	Modalità S: 3 m/s Modalità P: 3 m/s
Velocità massima	Modalità S: 72 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	-10 – 40 °C
Sistemi di posizionamento satellitare	GPS/GLONASS

- ## Fotocamera

Specifiche Tecniche fotocamera	
Sensore	CMOS 1” Pixel effettivi: 20 M
Obiettivo	FOV 77° 28 mm (formato 35mm equivalente) f/2.8 - f/11 messa a fuoco automatica 1 m – ∞
Intervallo ISO	Video: 100-6400 (manuale) Foto: 100-3200 (automatico) 100-12800 (manuale)
Velocità dell’otturatore elettronico	8-1/8000 s

Dimensione foto	**3:2 rapporto d'aspetto: 5472 × 3648**
Modalità di registrazione video	**4K: 3840×2160 24/25/30p** **2.7K: 2688x1512 24/25/30/48/50/60p** **FHD: 1920×1080 24/25/30/48/50/60/120p**
Bit-rate del video (max.)	**100 Mbps**
Formato Foto	**JPEG / DNG (RAW)**
Formato Video	**MP4 / MOV (MPEG-4 AVC/H.264, HEVC/H.265)**

Il DJI Mavic 2 Zoom 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 Mavic 2 Zoom garantisce riprese video in 4K/30fps o 2.7k/60fps. Inoltre, è dotato di un sistema di rilevamento ostacoli omnidirezionale che protegge completamente il velivolo a 360 gradi.

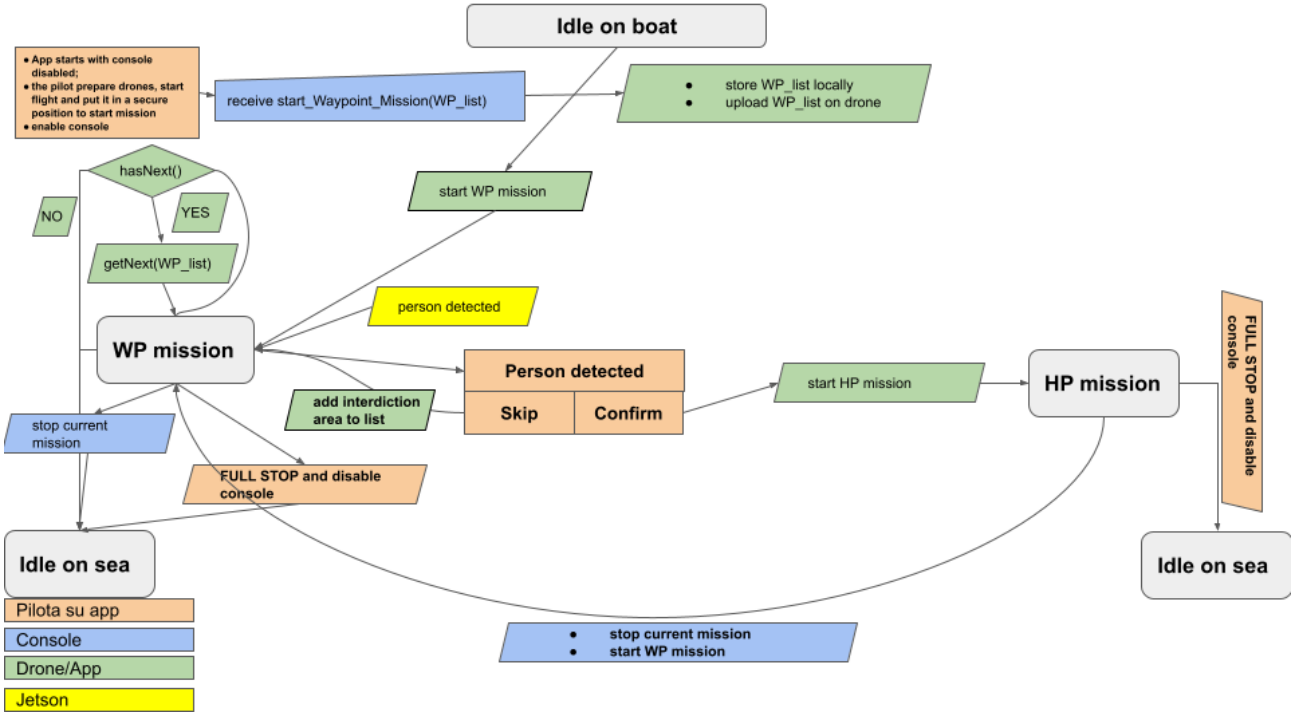
- ## Radiocomando

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: 8000 m** **CE: 5000m** **SRRC: 5000m** **MIC: 5000m**
Intervallo di temperatura operativa	**0 – 40 °C**
Batteria	**3950 mAh**
Potenza del trasmettitore (EIRP)	**2.400-2.483 GHz** **FCC: ≤26 dBm** **CE: ≤20 dBm** **SRRC: ≤20 dBm** **MIC: ≤20 dBm** **5.725-5.850 GHz** **FCC: ≤26 dBm** **CE: ≤14 dBm** **SRRC: ≤26 dBm**
Tensione/Corrente operativa	**1,8 A a 3,83 V**
Dimensioni dei dispositivi mobili supportati (max.)	**Lunghezza: 160 mm** **Spessore: 6,5 – 8,5 mm**

Porte USB supportate	**Lightning, Micro-USB (Tipo B), USB-C**

6. Architettura Software

Nella figura seguente è riportato lo schema di funzionamento di uno scenario tipico in cui il drone avvia la ricerca di una persona e ne individua una, iniziando il volo circolare sul punto dove è stata rilevata la persona in mare.



Nello stato iniziale il drone è fermo sulla nave; quando accade un evento di uomo in mare ed è necessario l'intervento del drone per cercarlo, innanzitutto il pilota del drone avvia l'app MOBDrone e fa decollare il drone mettendolo in una posizione di sicurezza da cui si può avviare la WaypointMission per la ricerca della persona. Fatto questo, il pilota deve abilitare i comandi della plancia premendo l'apposito tasto sull'app mobile. Di default all'avvio dell'app i comandi della plancia sono disabilitati per ragioni di sicurezza.
L'app MOBDrone una volta avviata fa partire in automatico lo stream video verso la NVIDIA Jetson e allo stesso tempo crea una connessione parallela in cui invia le coordinate GPS di ogni frame, utile per capire a che coordinate viene eventualmente rilevata una persona. La Jetson analizza il flusso video ed esegue l'algoritmo di detection della persona.
A questo punto la plancia può caricare sul drone la lista dei waypoint da seguire (una chiamata NADPopulateListCoordinate() per ogni waypoint) e avviare la missione di ricerca (WaypointMission) tramite il comando NADUploadAndStartWaypointMission().
Il drone, quindi, inizia a seguire i waypoints nella lista finché una delle seguenti condizioni occorre:

1. i waypoint nella lista sono finiti;
2. dalla console arriva un comando di pausa/stop missione;
3. il pilota preme sull'app il tasto di sicurezza "disable console" che stoppa la missione in corso e disabilita la console, assumendo quindi il controllo manuale del drone;
4. una notifica di persona rilevata arriva dalla jetson.

Nei primi 3 casi il drone rimane in hovering nella posizione corrente in attesa di ulteriori comandi da parte della plancia o del pilota. Il caso 4 invece è quello in cui dall'analisi del flusso video viene notificato che è stata rilevata una persona in mare. In questo caso il drone si ferma e interrompe la WaypointMission e sull'app mobile viene visualizzato un popup con due opzioni: "skip" e "confirm". Se il pilota si accorge che è una detection di falso allarme, allora preme "skip" e il drone riprende l'esecuzione della WaypointMission aggiungendo una zona d'interdizione nella posizione corrente del drone per evitare ulteriori notifiche di persona in mare. Altrimenti, se il pilota preme "confirm", viene

interrotta l'esecuzione dell'attuale WaypointMission e viene avviata l'esecuzione della HotpointMission in cui il drone inizia ad eseguire un volo circolare intorno al punto in cui è stata rilevata la persona.
Questa prosegue finché dalla plancia non arriva un comando pausa/stop missione o il comando NADSearchNextTarget(), che di fatto interrompe l'esecuzione dell'attuale HotpointMission e riprende l'esecuzione della WaypointMission precedentemente in esecuzione (questo è utile quando ci si rende conto che la persona rilevata precedentemente non fosse quella cercata, nonostante il pilota avesse confermato l'avvio della HotpointMission).
Infine, per far tornare il drone nei pressi della nave, la console può invocare il comando GoToShip().

7. Bibliografia

[Cafarelli et al 2022] Cafarelli, D., Ciampi, L., Vadicamo, L., Gennaro, C., Berton, A., Paterni, M., ... & Falchi, F. (2022). MOBDrone: A Drone Video Dataset for Man OverBoard Rescue. In International Conference on Image Analysis and Processing (pp. 633-644). Springer, Cham.

[Carion et al 2020] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13Return to ref 4 in article

[Feng et al 2021] Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W.: TOOD: task-aligned one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3510–3519, October 2021

[Ge et al 2021] Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430 (2021)

[He et al 2017] He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: IEEE International Conference on Computer Vision, ICCV 2017, pp. 2980–2988. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.322

[Lin et al 2014] Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48

[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.

[Ren et al. 2017] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137–1149 (2017). https://doi.org/10.1109/tpami.2016.2577031

[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).

[Zhou et al. 2019] Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)

[Zhu et al. 2021] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: 9th International Conference on Learning Representations, ICLR 2021. OpenReview.net (2021)

[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).