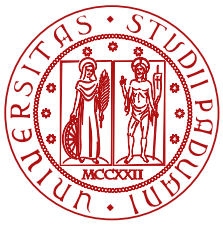


UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

Rilevamento di oggetti 3D da immagini 2D: metodi e applicazioni

Relatore
Prof. Stefano Ghidoni

Laureando
Zlatko Kovachev

Correlatore
Prof. Matteo Terreran

ANNO ACCADEMICO 2022-2023

Data di laurea 16/11/2023

Vorrei iniziare facendo i miei sentiti ringraziamenti a tutte le persone che mi hanno aiutato e che sono state vicino a me in tutti questi anni.

Desidero ringraziare soprattutto la mia famiglia, che grazie al loro amore e al loro sostegno hanno reso tutto questo possibile. Grazie papà per tutti gli sforzi che fai per me e per la nostra famiglia, così che possiamo essere tutti felici. Grazie mamma per tutti i consigli che mi hai dato e per essere sempre accanto a me. Grazie Miki perché hai sempre creduto in me e ti sei comportato come il fratello di cui sono sempre andato fiero. Grazie Cristina per tutte le risate che ci siamo fatti, che grazie al tuo scherzare mi hai sempre reso felice. Grazie Gessica, grazie di essere arrivata nella mia vita e di avermi sostenuto sempre, nei momenti belli e nei momenti più brutti. Senza di voi, senza la mia famiglia non sarei mai arrivato fino a questo punto.

Ringrazio di cuore il Prof. Stefano Ghidoni e il Prof. Matteo Terreran per la loro guida esperta e il loro incoraggiamento costante durante tutto il mio lavoro di tesi.

Desidero ringraziare i miei amici e colleghi per le preziose discussioni, il supporto morale e le risate condivise durante questo percorso. Grazie di aver passato questi ultimi 3 anni insieme, tra risate, scherzi e studio. Avete reso questa esperienza più divertente e preziosa e grazie a voi ho creato dei ricordi che rimarranno per sempre.

Infine vorrei ringraziare le persone che ci sono venute a mancare negli ultimi anni e che adesso sarebbero state orgogliose della persona che sono diventato. Nonno Vane e nonna Zlata, mi dispiace che oggi non potete festeggiare insieme a tutti noi però spero di avervi comunque resi fieri di me. Grazie Jack, grazie di essere sempre stato accanto a me negli ultimi 14 anni, eri il mio migliore amico e sapevo che non ero mai da solo perché tu eri sempre accanto a me. Tutto questo è anche grazie a te.

Grazie a tutti di aver sempre creduto in me, tutto questo non sarebbe stato possibile senza di voi.

Zlatko Kovachev

Abstract

La rilevazione di oggetti 3D da immagini 2D rappresenta una sfida cruciale nell'abito della visione artificiale e nella robotica, che si pone come obbiettivo quello di individuare la rotazione 3D e la traslazione 3D di oggetti, stimandone così la posa 6D. Questo campo di ricerca mira a determinare con precisione la posizione e l'orientamento degli oggetti in un ambiente tridimensionale. Metteremo in luce l'importanza della stima della posa 6D, un campo in rapida evoluzione, che va dai metodi tradizionali all'apprendimento avanzato, aprendo nuove prospettive di applicazione e offrendo soluzioni innovative negli ambiti industriali, robotici e della realtà aumentata. Questa tesi esplora diverse categorie di metodi utilizzati per affrontare questa sfida, suddividendoli in tre approcci principali: i metodi Template-Based, i metodi Feature-Based e i metodi Learning-Based. Discuteremo le problematiche principali di questo compito difficile, come la presenza di occlusioni, la simmetria degli oggetti e le variazioni di illuminazione. Allo stato dell'arte molti approcci riescono ad affrontare questa sfida con ottimi risultati, riuscendo a calcolare la stima della posa 6D in tempi relativamente brevi ($\approx$ 100ms) e con una precisione notevole, permettendo di stimare la posa anche in real-time. La velocità e l'efficacia della stima della posa 6D è aumentata negli ultimi anni grazie all'utilizzo di metodi di deep-learning e di CNN, che hanno portato questa sfida a un livello successivo. Rimangono comunque ancora tante questioni da migliorare e da risolvere: le CNN devono essere allenate e ciò richiede molto tempo e molte risorse, le occlusioni, gli oggetti senza texture e gli oggetti simmetrici rimangono ancora problemi chiave da affrontare. Esploreremo diversi approcci presenti in letteratura, analizzandone le relative sfide e opportunità, discutendone i metodi utilizzati e individuando possibili direzioni di ricerca future.

Indice

1 Introduzione **1**

1.1 La posa 6D 1

1.1.1 Applicazioni e utilità 3

1.1.2 Problematiche principali 5

1.2 Struttura della tesi 7

2 Approcci al problema **9**

2.1 Approcci Non-Learning-Based 10

2.2 Approcci Learning-Based 12

2.3 Dataset 13

3 Analisi dei metodi **17**

3.1 Metodi Template-Based 18

3.1.1 Esempio di metodi Template-Based 19

3.2 Metodi Feature-Based 21

3.2.1 Esempio di metodi Feature-Based 22

3.3 Metodi Learning-Based 25

3.3.1 Esempi di metodi basati sulla previsione di un bounding box utilizzando e sull'algoritmo PnP 27

3.3.2 Esempi di metodi basati sulla classificazione 28

3.3.3 Esempi di metodi basati sulla regressione 29

4 Conclusioni **33**

Bibliografia **35**

Elenco delle figure

1.1 La trasformazione della posa dal sistema di coordinare dell'oggetto al sistema di coordinate della telecamera è determinata dalla matrice di rotazione 3D R e dal vettore di traslazione 3D t [2]. 2

1.2 Convenzione comune per gli angoli yaw, pitch e roll [3]. 2

1.4 Esempio di realtà aumentata [9]. 4

1.5 Questa immagine contiene tutte le problematiche di cui discuteremo: (a) il trapano è coperto da vari oggetti, (b) ci sono molti oggetti nello sfondo, (c) la papera è monocolore e (d) la tazza blu è simmetrica [11]. 5

2.1 Esempio di modelli CAD 3D [12]. 11

2.2 Rappresentazione schematica della metodologia usata da [13]. Possiamo vedere come avviene la corrispondenza dei punti chiave 2D-3D nei punti (d) e (e) . . . 13

2.4 Immagini esempio dal dataset YCB-Video [12]. 15

2.5 Immagini esempio dal dataset T-LESS [16]. 15

3.1 Rappresentazione schematica dei metodi Template-Based. Una prima fase offline costruisce un database di template e in una seconda fase online, l'immagine di input viene confrontata con i template per calcolare la posa 6D [17]. 19

3.2 Rappresentazione finale della metodologia presentata Konishi et al. [19]. Possiamo notare come il contorno dell'oggetto viene definito anche in presenza di occlusioni e di sfondo disordinato. 20

3.3 Rappresentazione della metodologia usata da Massa et al. [21]. 21

3.4 Rappresentazione schematica dei metodi Feature-Based. Dall'immagine di input vengono estratte le informazioni chiave, poi vengono confrontate con le informazioni del modello 3D per ottenere la posa 6D dell'oggetto [17]. 22

3.5 Rappresentazione schematica della metodologia usata in PVNet [13]. 23

3.6 Rappresentazione schematica della metodologia usata da Kundu et al. [24]. . . 24

3.7 Rappresentazione schematica della metodologia usata da Chen et al. [25]. . . . 24

3.8 Rappresentazione schematica di metodi Learning-Based (a) monostadio e (b) bi stadio [17]. 25
3.9 Rappresentazione schematica della metodologia usata da Hu et al. [27] 28
3.10 Rappresentazione schematica della metodologia usata da Su et al. [29] 28
3.11 L'approccio di Mousavian et al. [27] rileva un bounding box 2D tramite classificazione per poi rilevare un bounding box 3D. 29
3.12 PoseCnn proposto in [12]. 30
3.13 Rappresentazione schematica della metodologia usata da [34]. 30

Capitolo 1

Introduzione

Lo sviluppo tecnologico degli ultimi decenni ci ha permesso di automatizzare alcuni processi lavorativi tramite l'utilizzo di macchinari, permettendoci di migliorarne l'efficienza, la precisione e la sicurezza in una vasta gamma di settori. Una buona stima della posa 6D è importante in molti settori dell'automazione, in quanto fornisce un'informazione chiave che consente a macchine, robot e applicazioni di interagire in modo più intelligente con l'ambiente circostante. Prendiamo ad esempio un braccio robotico, se volessimo automatizzare il processo di raccolta di un oggetto, dovremmo innanzitutto riuscire a capire la posizione e la rotazione dell'oggetto interessato rispetto a una telecamera per poi dare questa informazione al nostro braccio per poter eseguire dei comandi inerenti. Il bin-picking utilizza proprio questa ideologia: questo sistema è molto utilizzato nell'industria per poter riconoscere, raccogliere ed estrarre oggetti da un recipiente. Questo è solo un esempio di come la stima della posa 6D di un oggetto può essere utilizzata nel mondo reale e in questo capitolo ne approfondiremo le caratteristiche, parlando delle sue applicazioni e dei principali problemi che bisogna affrontare per definirla.

1.1 La posa 6D

Rilevare un oggetto 3D da un'immagine 2D è un modo per definire il problema della stima della posa 6D di un oggetto tramite un'immagine. La stima della posa 6D di un oggetto è una tecnica di computer visione che ci permette di definire la posizione e l'orientamento di un oggetto in relazione a un sistema di riferimento. La parola 6D sta a indicare i "6 gradi di libertà" di cui abbiamo bisogno: utilizziamo 3 gradi di libertà per definire la traslazione x, y, z e altri 3 gradi per definire la rotazione di ogni asse [1].

Se guardiamo la Figura 1.1 possiamo notare che la traslazione 3D dell'oggetto può essere definita tramite un vettore t che definisce la distanza dal centro delle coordinate x, y, z dell'oggetto in relazione al centro delle coordinate x', y', z' della telecamera. Invece, la rotazione

3D dell'oggetto viene definita tramite una matrice *R*, che definisce gli angoli della rotazione delle assi dell'oggetto rispetto alla rotazione delle assi della telecamera. Conoscendo il vettore traslazione *t* e la matrice di rotazione *R* possiamo definire la posa 6D di un oggetto [2].

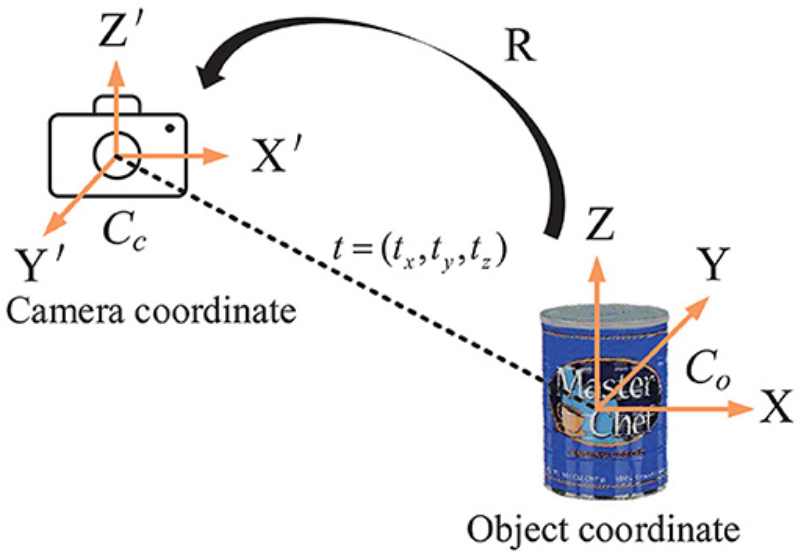


Figura 1.1: La trasformazione della posa dal sistema di coordinare dell'oggetto al sistema di coordinate della telecamera è determinata dalla matrice di rotazione 3D R e dal vettore di traslazione 3D t [2].

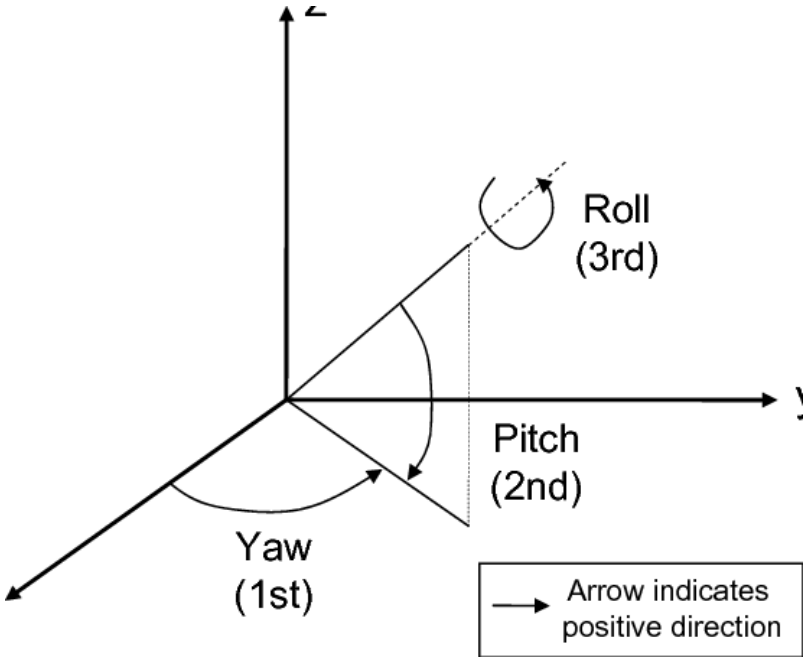


Figura 1.2: Convenzione comune per gli angoli yaw, pitch e roll [3].

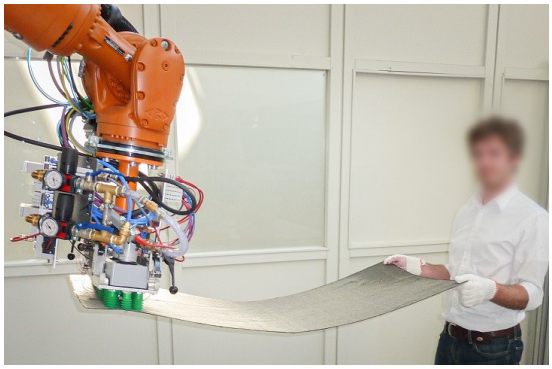
Alle rotazioni di ogni asse, come mostrato nella Figura 1.2, sono stati dati dei nomi convenzionali: la rotazione dell'asse x viene chiamata *yaw*, quella dell'asse y viene chiamata *pitch* e quella dell'asse z viene chiamata *roll* [3].

La posa 6D di un oggetto ci permette di conoscerne la traslazione e rotazione in relazione alla posizione della telecamera. Tuttavia, la stima della posa 6D di un oggetto da un'immagine si dimostra una complessa sfida a causa di diversi fattori. Le difficoltà principali includono la presenza di ombre e variazioni di illuminazione nell'immagine, occlusioni parziali o complete dell'oggetto, la simmetria dell'oggetto che può portare ad ambiguità nella stima della rotazione, e la necessità di disporre di dati di addestramento significativi e rappresentativi per approcci basati sull'apprendimento (Capitolo 2.2). Affrontare queste sfide richiede l'implementazione di metodi avanzati e l'integrazione di diverse strategie per ottenere stime precise della posa 6D. Nel prosieguo di questa tesi, esploreremo le diverse soluzioni proposte in letteratura e le relative sfide.

1.1.1 Applicazioni e utilità

Nel paragrafo iniziale abbiamo introdotto l'argomento del bin-picking, un sistema che utilizza la stima della posa 6D per poter individuare oggetti, afferrarli e muoverli. In particolare questo sistema da la capacità a un braccio robotico di riconoscere un singolo oggetto da un contenitore e decidere i movimenti necessari da compiere per poterlo estrarre e spostare nella zona interessata (Figura 1.3a): la competizione di *"Amazon Picking Challenge"* [4] ha portato molta attenzione a questo settore, favorendone lo sviluppo. Un altro campo della robotica che necessita della posa 6D è l'interazione uomo-robot, che permette all'operatore umano di poter lavorare in sicurezza con la presenza di un robot (Figura 1.3b) [5]. Per garantire un'interazione sicura ed efficace con un operatore umano, è essenziale poter realizzare sistemi robotici che possano analizzare l'area di lavoro, riconoscere e definire la posa 6D degli oggetti circostanti a esso e identificare le azioni umane.

(a) Esempio di bin-picking [6].



(b) Esempio di interazione umano-robot [5].

La stima della posa 6D è altrettanto rilevante e cruciale nel contesto della realtà aumentata. In questo settore, la posa 6D gioca un ruolo fondamentale nel posizionare e sovrapporre in modo accurato gli oggetti virtuali all'ambiente fisico. Questo processo, noto come "ancoraggio spaziale", richiede una comprensione precisa della posizione e dell'orientamento della telecamera rispetto agli oggetti del mondo reale. La realtà aumentata può offrire esperienze coinvolgenti e informative, ad esempio, nella manutenzione industriale, i tecnici possono utilizzare occhiali o dispositivi AR per visualizzare istruzioni dettagliate direttamente sulle apparecchiature complesse, sfruttando la stima accurata della posa 6D per garantire che le istruzioni siano sovrapposte con precisione agli oggetti da riparare o ispezionare (Figura 1.4) [7].

La guida autonoma utilizza la tecnologia della stima della posa 6D per poter riconoscere le strade, gli ostacoli da evitare, i veicoli circostanti e gli essere umani (pedoni e ciclisti). In questo settore è necessaria una stima accurata per poter permettere alle vetture autonome di raggiungere ottimi livelli di sicurezza e autonomia. Lo sviluppo di questa tecnologia ha portato alla creazione di una vasta raccolta di dataset RGB-D, poiché è stato necessario mappare un grande numero di strade al fine di raccogliere informazioni fondamentali per l'addestramento dei sistemi di visione che permettono alle auto autonome di comprendere l'ambiente che li circonda. [8].

Figura 1.4: Esempio di realtà aumentata [9].

1.1.2 Problematiche principali

Nel paragrafo precedente abbiamo parlato di cos'è la stima della posa 6D di un oggetto e in questa sezione parleremo delle problematiche principali che troviamo per ottenere questa informazione. Innanzitutto, per ottenere la posa 6D di un oggetto, abbiamo bisogno di un'immagine di partenza: questa immagine deve raffigurare una scena in cui il nostro oggetto selezionato è presente. Spesso il nostro oggetto non compare da solo nella scena, ma potrebbero esserci presenti anche altri elementi di disturbo. La presenza di altri oggetti nella scena possono portare a occlusioni dell'oggetto in studio, ma possono anche rendere difficile la lettura dell'immagine a causa di uno sfondo disordinato che contiene molti oggetti superflui. L'illuminazione è un altro componente importante per avere una buona stima della posa: se l'immagine è troppo scura potrebbe risultare difficile distinguere i vari oggetti e se l'immagine è troppo luminosa, alcuni oggetti potrebbero riflettere la luce e causare rumore nell'immagine. Quando ci sono difficoltà nella lettura di un'immagine è spesso utile aggiungere una fase preliminare di rilevamento o localizzazione dell'oggetto per distinguere l'area dell'immagine che contiene l'oggetto per migliorare l'accuratezza e la robustezza della stima della posa 6D [10].

Figura 1.5: Questa immagine contiene tutte le problematiche di cui discuteremo: (a) il trapano è coperto da vari oggetti, (b) ci sono molti oggetti nello sfondo, (c) la papera è monocolore e (d) la tazza blu è simmetrica [11].

In questa sezione parleremo di quattro problematiche principali:

- **Occlusioni (Foreground Occlusion):**

  La presenza di occlusioni della scena può comportare a un impreciso calcolo della posa, poiché il bersaglio è oscurato da altri oggetti e una o più parti di esso possono risultare nascoste od oscurate (Figura 1.5 (a)). Questo fenomeno si può verificare quando la scena è in disordine oppure quando ci sono tanti oggetti raggruppati insieme, come in molte scene industriali. Per affrontare questo problema molti metodi si affidano alla parte visibile dell'oggetto per poter descriverne le parti nascoste, così da calcolarne la posa 6D. Utilizzando informazioni 3D si possono ottenere risultati migliori rispetto al solo utilizzo di un'immagine RGB, visto che alcune parti dell'oggetto sono oscurate è difficile utilizzare informazioni relative alla texture e si devono utilizzare informazioni come il contorno del soggetto.

- **Sfondo in disordine (Background Clutter):**

  Abbiamo appena discusso che la presenza di oggetti davanti al nostro soggetto rende difficile la stima della posa 6D, ma anche il disordine dello sfondo rappresenta una sfida. La presenza di oggetti in sottofondo rende l'immagine più complessa, aumentando il numero di informazioni inutili che rendono difficile misurare la posa 6D (Figura 1.5 (b)). Nel complesso è molto raro poter avere un scenario un cui nello sfondo non ci sia disordine: segmentando l'immagine possiamo separare i vari oggetti e selezionando il soggetto desiderato possiamo fare la stima della posa 6D con più facilità.

- **Oggetti senza texture (Textureless objects):**

  Una sfida molto importante nella stima della posa 6D sono gli oggetti senza texture (Figura 1.5 (c)). La mancanza di texture sulla superficie di tali oggetti rende arduo l'estrazione di punti chiave, e molti approcci si basano sull'informazione di texture per effettuare una stima accurata. Nell'ambito industriale gli oggetti privi di texture sono comuni, rendendo la stima della posa 6D di tali oggetti cruciale. In assenza di informazioni di texture affidabili, l'utilizzo delle caratteristiche geometriche e di diverse tipologie di informazioni (contorno, colore, forma e altro ancora) è possibile ottenere risultati precisi e affidabili per il calcolo della posa 6D.

- **Simmetria (Symmetrical objects):**

  La stima della posa 6D di oggetti simmetrici costituisce una sfida significativa, poiché quando questi oggetti vengono ruotati di 180 gradi, generano osservazioni identiche, il che implica che la posizione e l'orientamento non cambiano rispetto a un punto di vista

fisso Figura (1.5 (d)). Gli oggetti simmetrici sono caratterizzati da forme e caratteristiche che possono essere specularmente identiche da diverse prospettive, il che rende difficile distinguerli. Quando si cerca di stimare la posa di tali oggetti, il sistema deve affrontare l'ambiguità intrinseca causata dalla simmetria, portando a una maggiore incertezza nei risultati. Per affrontare questa sfida si potrebbe incorporare alcune conoscenze pregresse sugli oggetti per migliorare la precisione nella stima della posa 6D di oggetti simmetrici.

Come vedremo nei capitoli successivi, le diverse metodologie alla stima della posa 6D affrontano queste problematiche in diversi modi, ma possiamo già affermare che gli approcci Learning-Based (Capitolo 2.2) sono relativamente più robusti nell'affrontare tali sfide, mentre gli approcci Non-Learning-Based (Capitolo 2.1) ne riscontrano maggiori difficoltà.

1.2 Struttura della tesi

L'obiettivo di questa tesi è fornire una panoramica generale sullo stato dell'arte della stima della posa 6D utilizzando immagini RGB come input. Nel secondo capitolo fornirò un'analisi dettagliata sugli approcci Non-Learning-Based e sugli approcci Learning-Based, per fornire una solida base sull'argomento. Nel terzo capitolo esamineremo e discuteremo i principali metodi proposti in letteratura: metodi Template-Based, metodi Feature-Based e metodi Learning-Based. Infine, nell'ultimo capitolo, dopo un riassunto dei risultati principali, esporrò i miei commenti sul possibile futuro di questo campo, discutendo delle possibili aree di miglioramento e delle sfide che rimangono aperte nella stima della posa 6D tramite informazioni 2D.

Capitolo 2

Approcci al problema

Quando si è iniziato a studiare il problema della stima della posa 6D, le reti neurali convoluzionali (CNN) e metodi di deep-learning non erano ancora disponibili o ampiamente utilizzate. In quel periodo, la ricerca si basava principalmente sull'utilizzo di informazioni geometriche e sull'aspetto degli oggetti (colore, texture), realizzando approcci che utilizzavano metodologie e algoritmi tradizionali per affrontare il problema della stima della posa (approcci Non-Learning-Based). Successivamente, con l'emergere delle CNN e delle tecniche di deep-learning, si è assistito a un cambiamento significativo nella progettazione di modelli e approcci per affrontare la stima della posa 6D (approcci Learnig-Based) [1].

In questo capitolo daremo una panoramica generale e teorica degli approcci sviluppati per affrontare il problema della stima della posa, classificandoli in approcci Learning-Based, se si basano sull'apprendimento, e in approcci Non-Learning-Based se non si basano sull'apprendimento. Invece, nel prossimo capitolo (Capitolo 3), entreremo più nel dettaglio e analizzeremo i vari metodi realizzati per la stima della posa 6D combinando tecniche di entrambi gli approcci.

- **Approcci Non-Learning-Based:**

  Gli approcci che non utilizzano CNN o tecniche di deep-learning rientrano in una delle due categorie seguenti. Ci sono gli approcci che si basano solamente sull'informazione 2D, come le immagini RGB, e ci sono gli approcci che si basano su informazioni 3D, come immagini RGB-D o point cloud, però escluderemo quest'ultimi perché non fanno parte del nostro campo di studio. Entrambi questi approcci utilizzano l'immagine di input per poter calcolare dei punti chiave dell'oggetto per poi trovare l'immagine più simile nel dataset, convertendo la stima della posa 6D in una ricerca di immagini. Gli approcci basati sull'informazione 2D possono essere divisi in approcci basati su immagini reali e approcci basati su immagini CAD in base al tipo di modello utilizzato.

- **Approcci Learning-Based:**

  Gli approcci basati sull'apprendimento utilizzano principalmente CNN, regressione o altri metodi di deep-learning per addestrare un modello di apprendimento con dati di addestramento adeguati e quindi calcolare la stima della posa 6D di un oggetto in una situazione sconosciuta in base ai dati di addestramento. Vedremo due categorie principali di approcci learning-based, gli approcci basati su punti chiave (keypoint-based) e gli approcci olistici che mirano ad addestrare una rete end-to-end per misurare la posa 6D di un oggetto.

2.1 Approcci Non-Learning-Based

Gli approcci basati solamente su informazioni 2D hanno come obiettivo principali il trovare una correlazione tra l'immagine di input e una delle immagini nel dataset attraverso le informazioni contenute nell'immagine. Per stimare la posa 6D degli oggetti possiamo utilizzare molte informazioni ricavabili da un'immagine, ad esempio abbiamo informazioni geometriche (contorno), informazioni sulla texture e sul colore. Questi approcci convertono la stima della posa 6D in un problema di corrispondenza tra immagini. Dobbiamo dire che utilizzare solamente informazioni 2D ha i propri svantaggi, sono meno robusti di approcci che utilizzando informazioni 3D, perché contengono meno dati con cui lavorare. Scene complesse, con forti variazioni di luminosità e prive di texture influiscono negativamente sulle prestazioni di questi approcci.

Possiamo classificare gli approcci basati sull'informazione 2D in due categorie: gli approcci che utilizzano immagini CAD e gli approcci che utilizzano immagini reali [1]. La prima categoria richiede un modello CAD dell'oggetto, invece la seconda categoria utilizza immagini reali come modelli base.

1. **Metodi basati su immagini CAD:**

   Questi approcci hanno bisogno del modello CAD 3D dell'oggetto interessato per poterne stimare la posa 6D: da questi modelli vengono generati delle immagini dell'oggetto da diversi punti di vista, per poter essere utilizzate come template di base. Le immagini virtuali (Figura 2.1) generate dai modelli CAD sono più accurate rispetto alle immagini reali, in quanto il processo di rendering non è influenzato dall'illuminazione o dalla sfocatura. Questi approcci sono molto adatti nelle applicazioni industriali, poiché è molto facile creare un modello CAD dei prodotti industriali.

Figura 2.1: Esempio di modelli CAD 3D [12].

2. **Metodi basati su immagini reali:**

   Sebbene sia possibile ottenere risultati precisi utilizzando modelli CAD 3D, a volte non è possibile ottenere il modello CAD di un oggetto. Pertanto, vengono utilizzate immagini reali come template. Per oggetti comuni (non pezzi/elementi industriali), oggetti unici o personalizzati e per oggetti con superfici complesse (difficili da modellare in 3D) è più facile realizzare immagini dal vivo in confronto a dover create o reperire il modello 3D degli oggetti stessi. Gli approcci basati su immagini reali utilizzano immagini reali per la stima della posa 6D, ma questa loro versatilità ha un problema fondamentale: se ci sono elementi di disturbo nelle immagini, le informazioni possono essere estratte in modo non corretto e la posa 6D può risultare incorretta o addirittura impossibile da ottenere se il dataset di immagini non è adeguato.

In generale, gli approcci basati su immagini CAD sono più accurati rispetto agli approcci basati su immagini reali, perché le immagini generate dai modelli CAD contengono pochi rumori. Tuttavia, quando non c'è la possibilità di ottenere facilmente i modelli CAD, si utilizzano immagini reali come template. L'accuratezza della stima della posa 6D è influenzata dal numero di template che si utilizza. Più template ci sono, più accurata sarà la stima della posa. Tuttavia, un gran numero di template richiede molto spazio di archiviazione e tempo di ricerca.

2.2 Approcci Learning-Based

Con l'emergere dei concetti di reti neurali convoluzionali (CNN) e metodi di deep-learning, negli ultimi anni molti studiosi hanno applicato questi concetti alla stima della posa 6D, ottenendo ottimi risultati. Gli approcci Learning-Based hanno rivoluzionato il modo in cui affrontiamo la stima della posa 6D, offrendo vantaggi distinti rispetto ai metodi tradizionali basati su caratteristiche geometriche o modelli matematici. Questi approcci si basano sulla potenza delle CNN, che sono in grado di apprendere modelli complessi dai dati, e sfruttano un'ampia quantità di dati di addestramento per ottenere risultati precisi e generalizzabili [1]. Gli approcci learning-based per la stima della posa 6D che vedremo, possono essere divisi in due categorie: gli approcci basati su punti chiave e gli approcci olistici.

1. **Approcci basati su punti chiave:**

   Gli approcci basati su punti chiave stabiliscono corrispondenze 2D-3D tra le immagini e misurano la posa in base a queste corrispondenze. Questi metodi seguono generalmente una procedura a due fasi: prima estraggono i punti chiave 2D dall'immagine di input e poi stimano la posa 6D utilizzando un algoritmo PnP (Perspective-n-Point). Questi approcci sono spesso più semplici da implementare rispetto agli approcci olistici e possono funzionare bene quando è possibile identificare punti chiave affidabili nell'immagine. Ad esempio, nella Figura 2.2 viene rappresentata la metodologia di un metodo che utilizza punti chiave. Partendo da (a) l'immagine di input, (b) per ogni pixel viene calcolato un vettore che punta a un keypoint e tramite un algoritmo basato su RANSAC (c) viene realizzato uno schema di votazioni. La corrispondenza tra i punti chiave 2D-3D nei punti (d) e (e) ci permette di ottenere la stima della posa 6D dell'oggetto (tramite un algoritmo PnP).

2. **Approcci olistici:**

   A differenza degli approcci basati sui punti chiave, gli approcci olistici adottano una struttura end-to-end per misurare la posa 6D, il che può renderli più veloci rispetto agli approcci basati su punti chiave. Questi metodi considerano l'immagine nel suo insieme e cercano di prevedere la posizione e l'orientamento dell'oggetto in un singolo passaggio, discretizzando lo spazio della posa e trasformando il compito di stima della posa in un compito di classificazione. Tuttavia, gli approcci olistici possono essere più complessi e richiedere più tempo rispetto agli approcci basati su punti chiave.

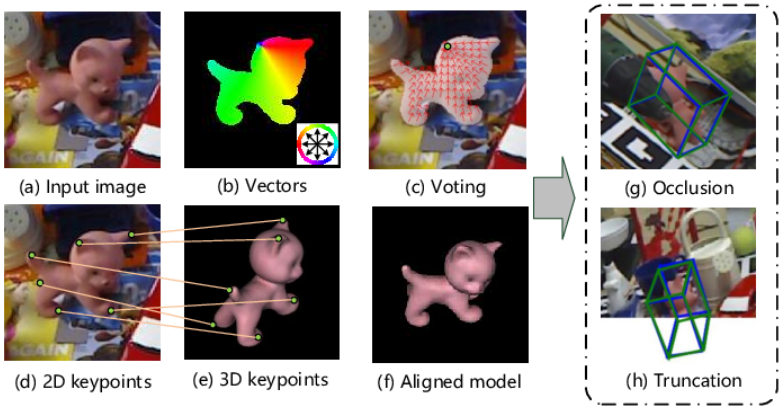


Figura 2.2: Rappresentazione schematica della metodologia usata da [13]. Possiamo vedere come avviene la corrispondenza dei punti chiave 2D-3D nei punti (d) e (e)

Entrambe queste categorie di approcci hanno i loro punti di forza e le loro debolezze, e la scelta tra di esse dipende spesso dalla natura specifica del problema e dai requisiti dell'applicazione. Gli approcci learning-based presentano ancora alcune debolezze, esse richiedono una notevole capacità di archiviazione per conservare i dati di addestramento e un tempo sufficiente per addestrare il modello. Per alcuni oggetti complessi che richiedono molti dati di addestramento, ciò potrebbe richiedere diverse ore o addirittura giorni per addestrare il modello, limitando quindi le possibili applicazioni di tali approcci. Infine, questi approcci non riescono a performare bene nella stima della posa 6D di oggetti che non esistono nel dataset.

2.3 Dataset

Un dataset è una collezione di dati e, nell'ambito della stima della posa 6D, può comprende immagini 2D, ottenute da oggetti reali o da render di un modello CAD 3D, oppure contengono modelli 3D degli oggetti necessari. Grazie ai dataset possiamo ottenere le informazioni dell'oggetto interessato, necessarie per realizzare una stima della posa 6D. Negli approcci Non-Learning-Based i dataset possono essere la collezione delle immagini reali o virtuali dell'oggetto e, grazie a essi, possiamo calcolare la posa dell'oggetto utilizzando un'immagine di input e il dataset che abbiamo realizzato. Invece, per gli approcci Learning-Based, i dataset possono essere immagini dell'oggetto in cui vengono annotate le informazioni sulla posa a priori. Questi dataset possono essere utilizzati per allenare una CNN o algoritmi per la stima della posa.

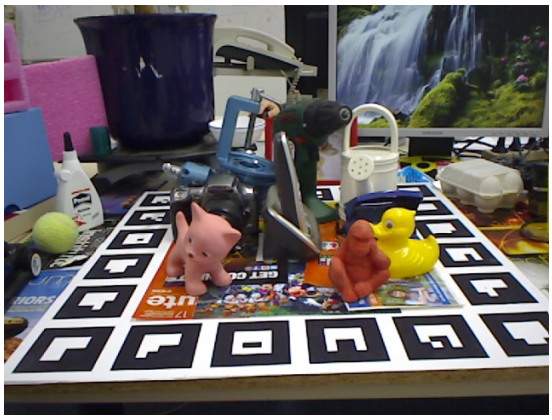
Oltre all'allenamento, i dataset possono essere usati anche per la validazione e il test delle CNN già allenate su altri dataset, così da poterne valutare le prestazioni. Per avere dei punti di riferimento tra tutta la comunità scientifica, si possono utilizzare dei dataset affidabili e noti a tutti

per poter confrontare le prestazioni dei propri metodi. Questa azione di chiama ”benchmark”: favoriscono la creazione di standard nell'ambito.

Alcuni dei dataset più utilizzati nell'ambito della stima della posa 6D:

- **LINEMOD [14]:**

  LINEMOD è un benchmark standard per la stima della posa 6D degli oggetti ed è strutturato con 18000 immagini RGBD di scenari con occlusioni e sfondi disordinanti. Contiene annotazioni manuali sulle vere pose 6D degli oggetti e 15 oggetti senza texture (Figura 2.3a).



(a) Dataset LINEMOD [14].

(b) Dataset Occlusion LINEMOD [15].

- **Occlusion LINEMOD [15]:**

  Questo dataset è una variazione del LINEMOD che è stato preso come base di partenza: contiene 10000 immagini di 20 oggetti con e senza texture, catturati sotto 3 condizioni di luminosità diversi. La peculiarità di questo dataset è che ogni immagine contiene oggetti che sono altamente occlusi, rendendo difficile la loro stima della posa 6D (Figura 2.3b).

- **YCB-Video [12]:**

  Questo dataset contiene 92 video, con annotazioni manuali, ripresi da una telecamera RGBD e contenenti 21 oggetti differenti presi dal dataset YCB. Gli oggetti sono disposti in varie pose e configurazioni, e alcuni fotogrammi contengono anche occlusioni significative tra gli oggetti. Questo dataset rappresenta una sfida a causa delle condizioni di illuminazione variabili, del notevole rumore nelle immagini e delle occlusioni (Figura 2.4).

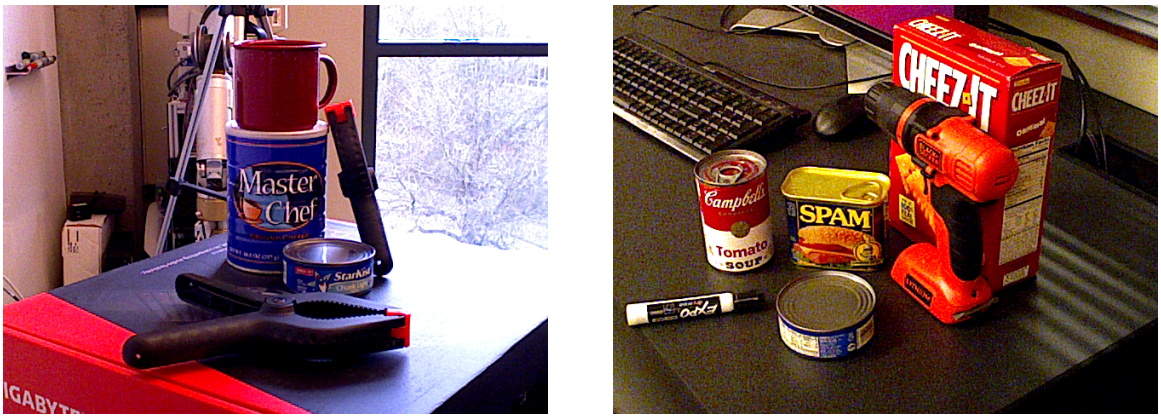


Figura 2.4: Immagini esempio dal dataset YCB-Video [12].

- **T-LESS [16]:**

  Il dataset T-LESS contiene 20 differenti scene industriali, enfatizzando la presenza di oggetti senza texture e oggetti rigidi. Contiene 30 oggetti industriali, senza texture significanti, colori discriminanti o distinguibili proprietà riflettenti. Le forme e/o le dimensioni degli oggetti spesso presentano somiglianze, e alcuni oggetti sono costituiti da composizioni di altri oggetti (Figura 2.5).

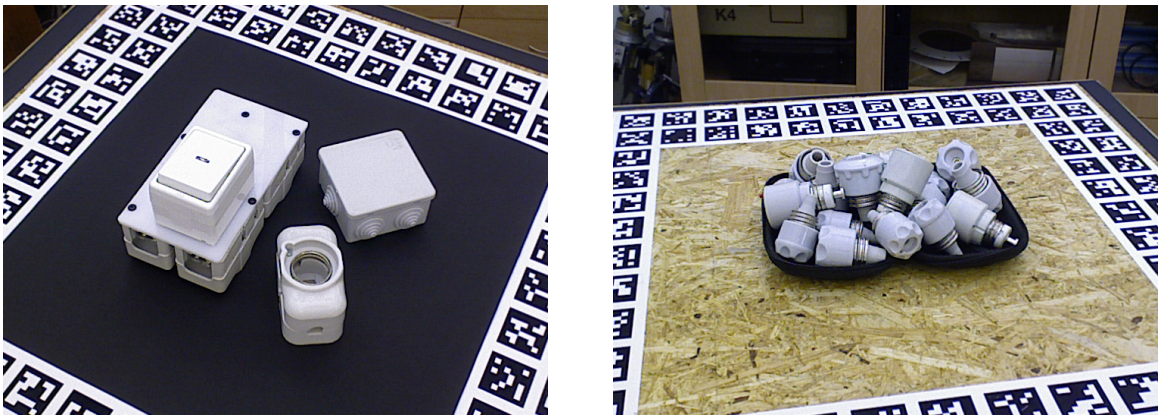


Figura 2.5: Immagini esempio dal dataset T-LESS [16].

L'utilizzo di questi dataset ha portato a una standardizzazione nel mondo della posa 6D, facilitandone lo sviluppo della ricerca e il confronto tra i diversi approcci. I dataset forniscono una base per creare e valutare nuovi metodi per la stima della posa 6D ed è fondamentale scegliere un dataset appropriato per permettere agli algoritmi di affrontare sfide reali.

Capitolo 3

Analisi dei metodi

Come menzionato nel capitolo precedente (Capitolo 2), lo studio della stima della posa 6D può essere divisa in due tipologie di approcci, gli approcci Non-Learning-Based e gli approcci Learning-Based. Gli approcci Non-Learning-Based utilizzano principalmente le informazioni geometriche dell'oggetto ricavate dall'immagine però, grazie alla diffusione del Deep-Learning e all'uso di CNN, hanno visto un'evoluzione negli ultimi anni. Non parleremo più di approcci Non-Learning-Based e Learning-Based, perché i metodi che andremmo a vedere combinano tecniche di entrambe le categorie, utilizzando sia le informazioni geometriche, sia utilizzando i benefici che possono fornire le CNN.

In questa sezione parleremo dei tre metodi principalmente utilizzati, dividendoli in metodi basati su template (metodi Template-Based), metodi basati su determinate caratteristiche dell'oggetto (metodi Feature-Based) e metodi basati sull'apprendimento, quindi sull'utilizzo di CNN e tecniche di deep-learning (metodi Learning-Based)[17]. I primi due metodi menzionati possono essere definiti tradizionali, poiché la loro stima della posa 6D si basa su approcci geometrici. I metodi Tempalte-Based utilizzano rappresentazioni 2D degli oggetti da diversi punti di vista per fare una comparazione con l'immagine originale per poter stabilire la posizione e l'orientamento dell'oggetto. Questi metodi possono gestire oggetti privi di texture ma sono molto sensibili alle variazioni di illuminazione e alle occlusioni. I metodi Feature-Based si basano sull'estrazione e la corrispondenza di caratteristiche locali tra modelli 3D e immagini 2D per stimare la posa 6D di un oggetto. Invece, i metodi Learning-Based, sono metodi concentrati principalmente sull'apprendimento, si addestra una rete neurale su misura per una specifica attività. Visto l'elevato interesse negli ultimi anni per questa categoria di metodi, possiamo suddividerla ulteriormente in altre tre sotto categorie: abbiamo metodi che si basano sulla predizione di un bounding box utilizzando poi algoritmi PnP, metodi che risolvono un problema di classificazione e metodi che risolvono un problema di regressione. Questi metodi possono raggiungere livelli molto elevati di precisione ma hanno bisogno di molti dati per addestrare

la rete con precisione e per essere efficaci nei casi reali. Come detto in precedenza, le CNN possono essere utilizzate anche per eseguire le fasi più critiche dei metodi tradizionali, cos' da poter unire i vantaggi delle varie strategie nella soluzione finale [17].

I metodi che vedremo utilizzano una singola immagine 3D per calcolare una stima della posa 6D degli oggetti. Vedremo come ogni metodo affronta le problematiche principali di occlusione, simmetria, oggetti senza texture e cattiva illuminazione.

3.1 Metodi Template-Based

I metodi Template-Based, noti anche come metodi basati su template, costituiscono una delle categorie tradizionali per la stima della posa 6D. Questi approcci si basano sulla comparazione tra immagini di template, che rappresentano oggetti da diversi punti di vista, e l'immagine di input per determinare la posizione e l'orientamento dell'oggetto. I metodi Template-Based seguono una procedura chiara per stimare la posa 6D di un oggetto, come è possibile vedere nella Figura 3.1. In una prima fase offline avviene la creazione di un database, composto da una serie di template 2D dell'oggetto da diverse prospettive. Questi template rappresentano l'oggetto da rilevare e possono essere ottenuti da modelli CAD 3D o da immagini reali dell'oggetto da diverse angolazioni. La seconda fase avviene online ed è una fase di test. Una volta ottenuti i template, vengono confrontati con l'immagine di input per cercare una corrispondenza. Questo processo può coinvolgere l'individuazione dei punti chiave o delle caratteristiche salienti dell'oggetto nell'immagine. Se viene trovata una corrispondenza tra un template e l'immagine di input, è possibile stimare la posa 6D dell'oggetto in base alla posizione e all'orientamento del template corrispondente.

Vantaggi:

- Sono relativamente facili da implementare e se il database è esaustivo possono raggiungere un'alta precisione.
- Possono gestire oggetti che possono non avere texture o caratteristiche distintive, a condizione che i template siano accurati.

Svantaggi:

- Sono suscettibili alle variazioni di illuminazione, poiché le immagini di template potrebbero non corrispondere esattamente alle condizioni di illuminazione reali.
- Sono sensibili alla presenza di occlusioni, poiché possono ostacolare il rilevamento e la corrispondenza tra template e immagine.

- La velocità di esecuzione è inversamente proporzionale al numero di elementi appartenenti al modello. Bisogna bilanciare l'accuratezza del modello con il tempo di esecuzione.

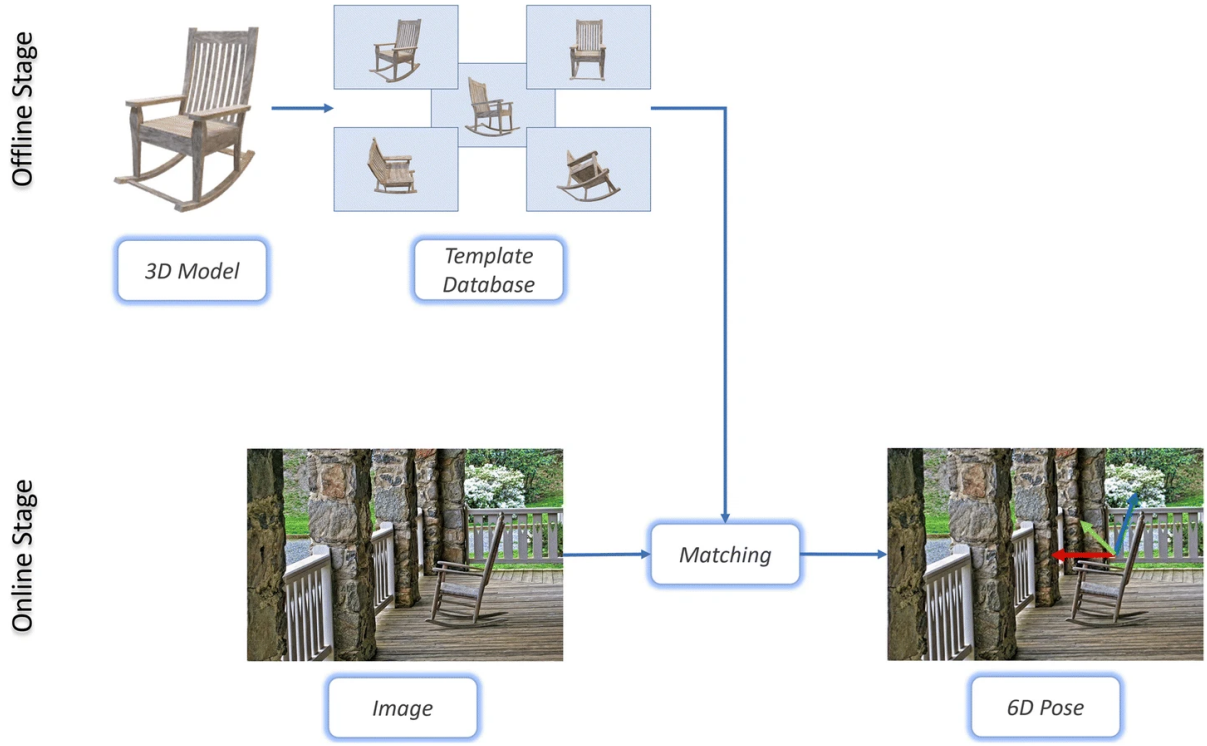


Figura 3.1: Rappresentazione schematica dei metodi Template-Based. Una prima fase offline costruisce un database di template e in una seconda fase online, l'immagine di input viene confrontata con i template per calcolare la posa 6D [17].

3.1.1 Esempio di metodi Template-Based

Come già discusso in precedenza, i primi approcci alla stima della posa 6D utilizzavano approcci geometrici. In questo contesto, Ulrich et al. [18], ha sviluppato un sistema che genera una struttura gerarchica utilizzando solamente le informazioni geometriche prese da un modello CAD 3D di un oggetto. Questo modello viene utilizzato per effettuare una ricerca gerarchica sull'immagine di input poter localizzare l'oggetto e definirne la sua posa 6D in modo efficace e in tempi veloci: il tempo di riconoscimento non dipende dalla complessità dell'oggetto ma da quante direzioni diverse potrebbe essere visto l'oggetto. Questa metodologia affronta con successo problemi comuni, come le occlusioni, ambienti disordinanti ed è molto robusto con oggetti senza texture e riflettenti e, grazie al solo utilizzo di informazioni geometriche, è particolarmente utile in contesti industriali e robotici (es. bin-picking).

Sappiamo che la velocità di esecuzione dei metodi Tempalte-Based è inversamente proporzionale al numero di template, quindi l'aumento del numero di template comporta un rallentamento nella stima della posa. Per affrontare questo problema Konishi et al. [19] ha introdotto "Perspectively Cumulated Orientation Feature" (PCOF), estratto da immagini realizzate da un modello CAD 3D. Il modello che gestisce PCOF gestisce una determinata gamma di pose dell'oggetto 3D, diminuendo il numero di template necessari. Per la stima della posa 6D è stato introdotto "Hierarchical Pose Trees" (HPT), una struttura gerarchica costruita ragruppando le pose degli oggetti 3D e riducendone la risoluzione. La combinazione dell'utilizzo di PCOF e HPT rende efficiente la stima della posa 6D di oggetti senza texture, riflettenti e in scene con occlusioni e sfondi disordinanti (Figura 3.2).

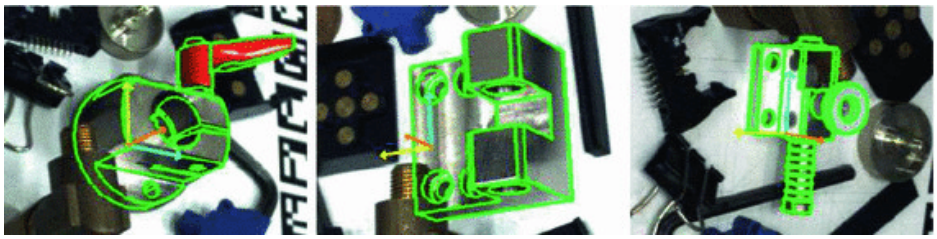


Figura 3.2: Rappresentazione finale della metodologia presentata Konishi et al. [19]. Possiamo notare come il contorno dell'oggetto viene definito anche in presenza di occlusioni e di sfondo disordinato.

Payet e Todorovic in [20] hanno realizzato un modello che utilizza i contorni dell'oggetto rilevati nell'immagine come caratteristica base. Utilizzando degli esempi di pose arbitrarie dell'oggetto, realizzano un modello di forma in termini di pochi template che dipendono dalla vista dell'oggetto (raccolta dei contorni dell'oggetto da diversi punti di vista). Questi template di viste vengono utilizzati per poter stimare la posa 6D attraverso un nuovo attributo, "Bag of Boundaries" (BOB), creato per confrontare il contorno dell'oggetto nell'immagine di input con i template di vista. Il metodo si dimostra efficace per gestire oggetti senza texture, sfondi disordinati e occlusioni.

Alcuni studi hanno implementato l'utilizzo di CNN per poter migliorare i propri risultati, ad esempio Massa et al. in [21] ha incorporato una CNN end-to-end, CaffeNet, come passaggio preliminare per generare un vettore di caratteristiche dell'oggetto, sia nella fase offline che nella fase online (Figura 3.3). Nella parte offline viene realizzata una libreria di templates partendo da dei render CAD filtrati da un modello CaffeNet. Anche l'immagine di input viene filtrata, prima da un modello CaffeNet e poi da una CNN che la combina con l'immagine originale. Infine viene fatta una comparazione tra questa immagine e i template realizzati per ottenere la posa 6D. Questo approccio unisce tradizionali tecniche utilizzate nei metodi Tempalte-Based con i vantaggi di precisione e velocità di una CNN.

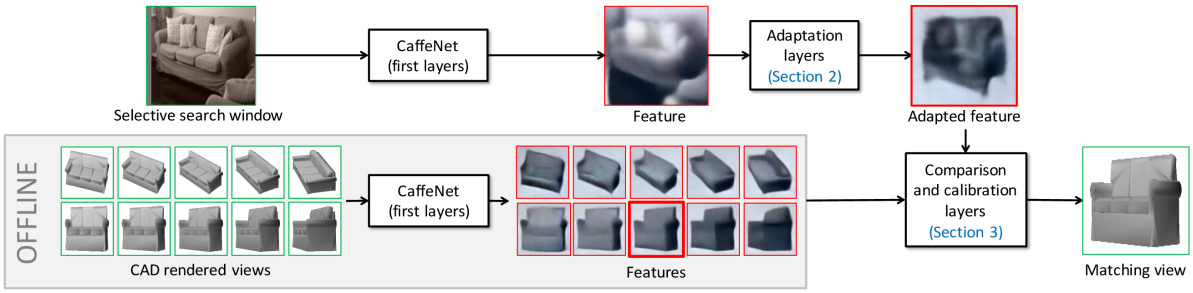


Figura 3.3: Rappresentazione della metodologia usata da Massa et al. [21].

Anche Sundermeyer et al. in [22] ha utilizzato una CNN chiamata Augmented Autoencoder per stimare la posa 6D dell'oggetto. Questa neural network è addestrata sui template dell'oggetto generati da render del modello 3D dell'oggetto per poterne imparare la rappresentazione da diverse angolazioni. Questo cosiddetto Augmented Autoencoder riesce a gestire intrinsecamente simmetrie degli oggetti e occlusioni.

3.2 Metodi Feature-Based

I metodi Feature-Based si concentrano sull'estrazione e sulla corrispondenza di caratteristiche locali (punti chiave, valori di grigio, bordi o intersezioni di linee rette) tra modelli 3D e immagini 2D per stimare la posa 6D di un oggetto. Questa procedura combinano approcci tradizionali con CNN, ed è basata sull'identificazione e sulla corrispondenza di punti chiave o caratteristiche distintive tra il modello 3D noto e l'immagine di input per stabilire corrispondenze 2D-3D. I metodi Feature-Based offrono un'alternativa robusta ai metodi Template-Based, affrontando alcune delle limitazioni legate all'illuminazione e alla texture. La scelta tra questi due approcci dipende spesso dalla natura specifica del problema e dai requisiti dell'applicazione. Questi metodi hanno la necessità di avere il modello 3D dell'oggetto, per poterlo utilizzare come riferimento per la stima della posa 6D. Nella prima fase l'immagine di input viene analizzata per l'estrazione di punti chiave o caratteristiche distintive. Le caratteristiche estratte dall'immagine di input vengono confrontate con quelle presenti nel modello 3D risolvendo un problema geometrico di corrispondenze 2D-3D. Una volta ottenuta una corrispondenza affidabile tra le caratteristiche dell'immagine di input e il modello 3D, è possibile stimare la posa 6D dell'oggetto, tenendo conto della posizione e dell'orientamento dell'oggetto rispetto alla telecamera. Questa procedura può essere vista nella Figura 3.4.

Vantaggi:

- Robusti alle variazioni di illuminazione e alla presenza di occlusioni.

Svantaggi:

- Non funzionano bene con gli oggetti simmetrici.
- Oggetti senza texture o con texture poco ricche rendono difficile se non impossibile il calcolo dei punti chiave.
- L'analisi e l'elaborazione delle caratteristiche richiedono risorse computazionali, il che può influire sulla velocità di esecuzione.
- L'oggetto deve avere caratteristiche distintive facili da individuare, poiché la qualità dei punti chiave estratti influisce direttamente sull'accuratezza della stima della posizione.

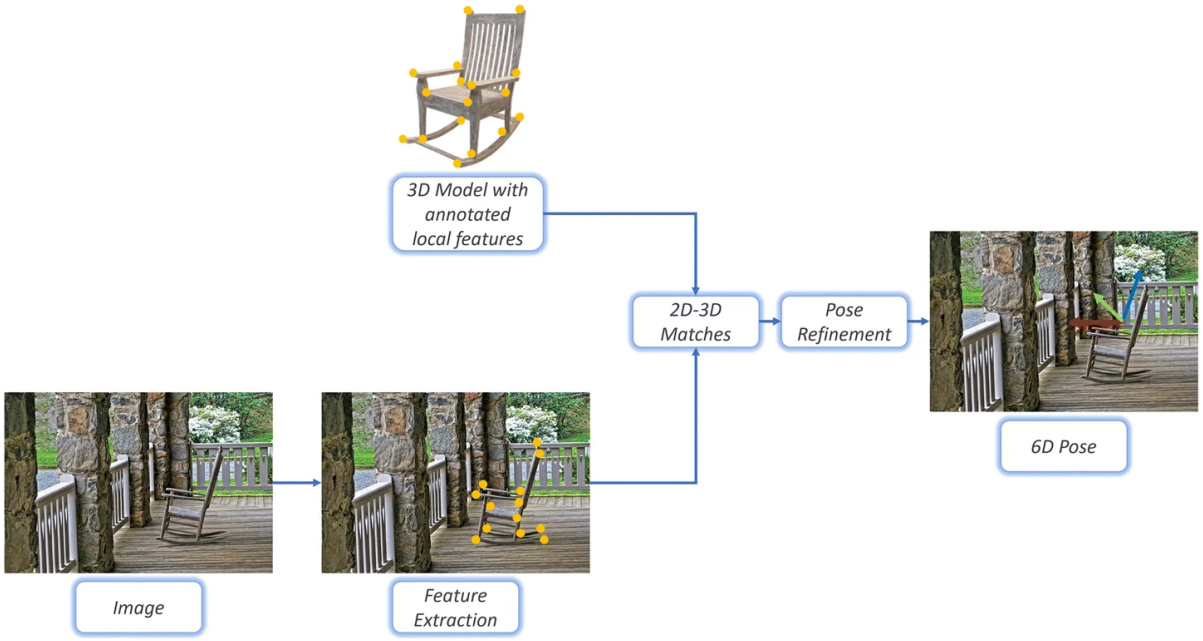


Figura 3.4: Rappresentazione schematica dei metodi Feature-Based. Dall'immagine di input vengono estratte le informazioni chiave, poi vengono confrontate con le informazioni del modello 3D per ottenere la posa 6D dell'oggetto [17].

3.2.1 Esempio di metodi Feature-Based

Il sistema realizzato da Peng et al. [13], è un ottimo esempio di metodo Feature-Based per la stima della posa 6D, concentrandosi specificamente su scene che presentano grandi occlusioni (g)

e oggetti troncati (h) (Figura 3.5). Partendo da (a) l'immagine di input, (b) per ogni pixel viene calcolato un vettore che punta a un keypoint. Per fare ciò è stata realizzata una CNN chiamata "Pixel-wise Votino Network" (PVNet) per predire la corrispondenza 2D-3D, regredendo i pixel in vettori per farli puntare verso la posizione dei punti chiave. Il risultato è una distribuzione di probabilità per ogni keypoint e tramite un algoritmo basato su RANSAC (c) viene realizzato uno schema di votazioni. Infine, per ottenere la posa 6D viene utilizzato un algoritmo PnP tra i ponti (d) 2D e (e) 3D. Questo processo riesce a ottenere ottimi risultati con oggetti troncati e occlusi anche a real-time.

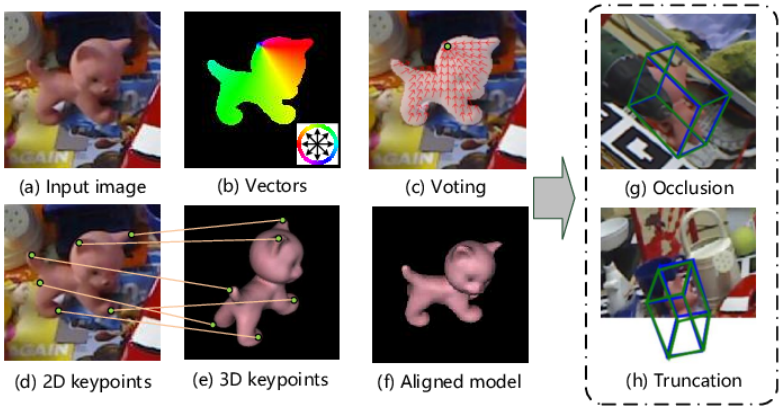


Figura 3.5: Rappresentazione schematica della metodologia usata in PVNet [13].

I metodi Feature-Based utilizzano una strategia di votazione per stimare i punti caratteristici in uno spazio vettoriale al fine di migliorare l'accuratezza della posa stimata, tuttavia la funzione di perdita utilizzata successivamente, tiene conto solo della direzione del vettore che, portando alla comparsa di errori. Basandosi su PVNet, You et al. [23] ha realizzo un sistema che aggiunge un nuovo passaggio di raffinatura della posa 6D, ottenendo ottimi risultati. Hanno considerato una funzione di perdita che gestisce l'errore del campo vettoriale e incorpora una rete di affinamento per rivedere la posa prevista al fine di ottenere un buon risultato finale.

Kundu et al. [24] ha utilizzato la regolarità della geometria degli oggetti, fondendo informazioni ottenute da punti di vista differenti per migliorare la comprensione geometrica del sistema, il che a sua volta, ha potenziato le prestazioni nella stima della posa 6D. Il processo avviene in due fasi (Figura 3.6): l'immagine di input viene associata a multiple viste 2D del modello 3D per poter generare una mappa di corrispondenza tra i punti 2D-3D attraverso una prima CNN allenata per individuare le posizioni invarianti locali dei descrittori per ottenere i corrispondenti punti chiave. Una seconda CNN, unendo le informazioni delle diverse mappe di corrispondenza, fornisce la posa finale.

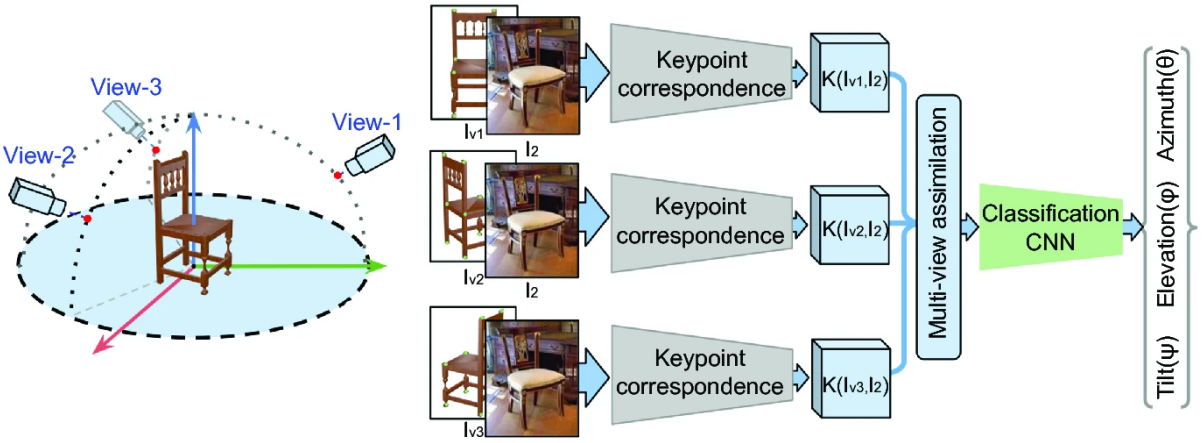


Figura 3.6: Rappresentazione schematica della metodologia usata da Kundu et al. [24].

Nell'ambito industriale molti oggetti sono privi di texture e con materiali lucidi, rendendo questa tipologia di oggetti solitamente difficili da trattare. In questo contesto, Chen et al. [25] propone un sistema incentrato su bersagli metallici. Le problematiche di questi oggetti derivano dall'impraticabilità di individuare punti chiave o altre informazioni partendo dalle texture, poiché la maggior parte di essi non rappresenta effettive caratteristiche del bersaglio, bensì riflessi dell'ambiente circostante. Il processo proposto è definito da tre fasi: rilevamento dell'oggetto, rilevamento delle caratteristiche e stima della posa (Figura 3.7). Per rilevare gli oggetti e definirne i punti chiave vengono utilizzate le informazioni sui contorni, definiti utilizzando la densità limite dei punti discreti sulle superfici metalliche per poter realizzare una segmentazione degli oggetti, per poi poterne rilevare i contorni. Successivamente vengono sfruttate sia le informazioni sui punti chiave che il modello CAD per poter calcolare la posa 6D di ciascun oggetto. Questo lavoro ha proposto anche una risoluzione a un problema degli approcci Learning-Based, cioè la necessità di una grande quantità di dati di addestramento e la necessità di lavoro umano per l'etichettatura degli oggetti di tali dataset: viene proposto un approccio per generare dataset partendo da modelli CAD ed etichettare gli oggetti presenti nelle immagini automaticamente.

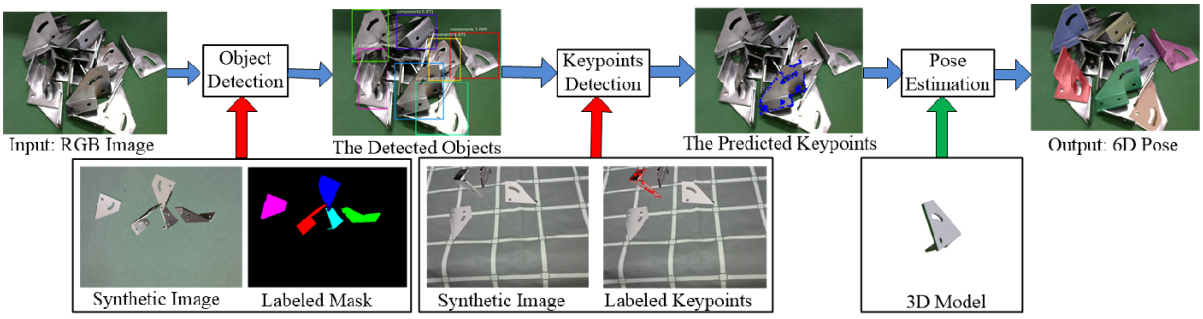


Figura 3.7: Rappresentazione schematica della metodologia usata da Chen et al. [25].

3.3 Metodi Learning-Based

I metodi Learning-Based rappresentano una categoria avanzata di approcci per la stima della posa 6D che si basano sull'uso di reti neurali convoluzionali (CNN) e tecniche di deep-learning. Questi metodi differiscono dai tradizionali metodi Template-Based e Feature-Based in quanto si concentrano principalmente sull'apprendimento da dati di addestramento per prevedere la posa 6D di un oggetto in situazioni sconosciute. Questi metodi possono essere a uno o due stadi, come mostrato nella Figura 3.8. I modelli end-to-end monostadio (Figura 3.8 (a)) possono essere utilizzati quando si vuole avere sistema unificato, in cui stima la posa 6D dell'oggetto al primo colpo, senza dover passare attraverso diversi passaggi. Appena il modello monostadio riceve l'immagine di input, la CNN definisce l'oggetto e ne calcola la posa 6D direttamente. Invece, se si vuole utilizzare un algoritmo Perspective-n-Point (PnP) per perfezionare i parametri della posa si può optare per un modello bi stadio (Figura 3.8 (b)). Si comporta come un modello monostadio all'inizio, ma dopo il primo calcolo della posa 6D, esegue un ulteriore passaggio per poterne migliorare la stima. In generale, le CNN a due stadi sono più precise di quelle a end-to-end, in particolare su oggetti di piccole dimensioni e su oggetti multipli[17]. I metodi Learning-Based rappresentano un campo di ricerca in crescita nell'ambito della stima della posa 6D e promettono di affrontare alcune delle sfide dei metodi tradizionali.

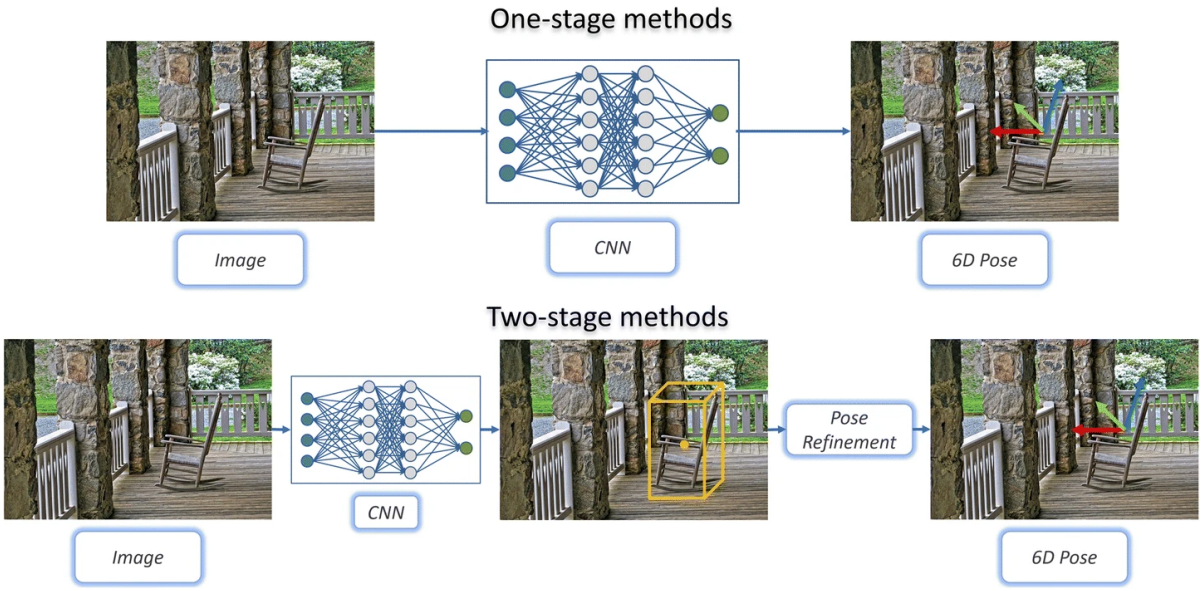


Figura 3.8: Rappresentazione schematica di metodi Learning-Based (a) monostadio e (b) bi stadio [17].

Possiamo classificare questi metodi in tre categorie:

- **Metodi basati sulla previsione di un bounding box utilizzando e sull'algoritmo PnP:**

  Questi metodi per la previsione della posa 6D sono composti da due fasi principali, prima si prevede un bounding box attorno all'oggetto e poi viene calcolata la posa 6D utilizzando l'algoritmo PnP. Per poter calcolare un bounding box attorno all'oggetto nell'immagine di input bisogna innanzitutto allenare una CNN. Le CNN vengono addestrate utilizzando dei dataset che contengono delle immagini dell'oggetto con le annotazioni sul bounding box di quella posa, così da poter rilevare e prevedere il bounding box di una posa che non ha mai visto. Nella seconda fase viene calcolata la posa 6D effettiva utilizzando l'algoritmo PnP, utilizzato per stimare la posa 6D basandosi sulla relazione tra i punti 2D rilevati nel bounding box e i corrispondenti punti 3D noti, ottenuti da un modelli CAD 3D dell'oggetto o da altri metodi. L'algoritmo PnP, sfruttando la geometria della prospettiva, stima la posa 6D dell'oggetto rispetto alla telecamera.

- **Metodi basati sulla classificazione:**

  In questi metodi, il problema della stima della posa 6D viene trasformato in un problema di classificazione. La rete neurale end-to-end viene addestrata per classificare l'oggetto in diverse categorie di posa 6D. Sfruttano le CNN per ottenere una distribuzione di probabilità nello spazio della posa e la associano alle informazioni del modello 3D per acquisire la posizione e la rotazione 3D. Questo richiede un'ampia quantità di dati di addestramento contenenti esempi di immagini con etichette di posa 6D.

- **Metodi basati sulla regressione:**

  Questi sistemi risolvono la posa 6D come un problema di regressione e utilizzano le CNN per stimare la posizione. Effettuano direttamente una regressione dei parametri della posa 6D dell'oggetto di destinazione dall'immagine di input. Di solito, c'è una fase preliminare di rilevamento dell'oggetto per semplificare il processo di stima della posizione. Questi sistemi appartengono alla categoria dei metodi end-to-end, cioè progettano una rete neurale addestrata per prevedere direttamente la posa 6D dell'oggetto senza passare per la fase di classificazione.

Vantaggi:

- Alte prestazioni con oggetti occlusi o in presenza di sfondi affollati.
- Robusti in presenza di oggetti con poche texture e con variazioni di illuminazione.
- Se addestrati bene i modelli, possono fornire elevati livelli di precisione nella stima della posa 6D nella maggior parte delle situazioni.

Svantaggi:

- L'addestramento del modello richiede molto tempo e richiedono una grande quantità di dati di addestramento.
- Sensibilità agli oggetti assenti dal dataset di attestamento.
- La loro capacità di generalizzare è ancora un problema in alcuni casi.

3.3.1 Esempi di metodi basati sulla previsione di un bounding box utilizzando e sull'algoritmo PnP

Rad e Lepetit in [26] hanno proposto BB8, un approccio olistico alla stima della posa 6D definita da multiple CNN in serie. Una prima CNN si occupa della segmentazione semantica degli oggetti, capace di localizzarli anche con la presenza di occlusioni e sfondi disordinati. Una seconda CNN cerca di predire gli angoli delle bounding box 3D attorno agli oggetti e una terza CNN utilizza questi dati, dopo essere stati processati con un algoritmo PnP, per rifinire la posa 6D. Per risolvere il problema degli oggetti simmetrici, durante la fase di allenamento, hanno ristretto la gamma delle possibili pose che gli oggetti possono assumere: dato che un oggetto simmetrico ha un angolo di simmetria di 180°, la restrizione è stata posta per avere una gamma di pose che vanno da 0° a 90°.

Per superare il problema delle occlusioni, Hu et al. [27] ha introdotto un metodo basato sulla segmentazione dell'immagine, in cui ogni parte dell'oggetto contribuisce alla stima della posizione della propria bounding box 3D. Questo sistema è composto da due principali flussi (Figura 3.9): da una parte una CNN è allenata per poter etichettare e segmentare gli oggetti presenti nell'immagine, e dall'altra parte si cerca di predire gli angoli delle bounding box 3D regredendo i pixel dell'immagine. L'unione di questi due flussi permette di calcolare la stima della posa 6F anche con la presenza di occlusioni.

Li et al. [28] ha realizzato un sistema che cerca di predire separatamente la rotazione e la traslazione degli oggetti, introducendo "Coordinates-based Disentangled Pose Network" (CDPN) per poter ottenere questa divisione. Questa separazione viene fatta per ottenere una stima della posa estremamente accurata e robusta e per poter gestire oggetti senza texture e occlusi. Per risolvere la rotazione viene utilizzata una funzione Masked Coordinate-Confidence Loss (MCC loss) per il calcolo delle coordinate della bounding box, che verranno utilizzate in un algoritmo PnP. Invece la traslazione viene calcolata direttamente dall'immagine usando Scale-Invariant Translation Estimation (SITE).

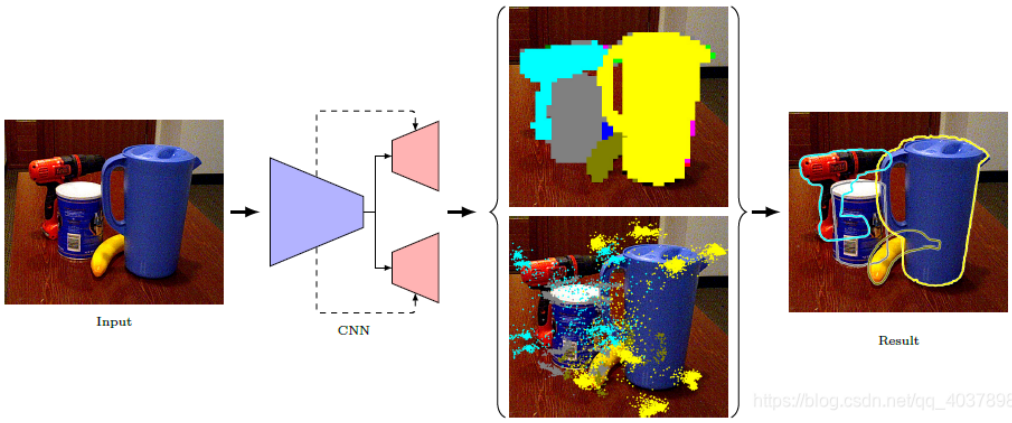


Figura 3.9: Rappresentazione schematica della metodologia usata da Hu et al. [27]

3.3.2 Esempi di metodi basati sulla classificazione

Data la scarsità di dati di addestramento con annotazioni di punti di vista e l'abbondanza di modelli 3D, Su et al. [29] ha realizzato una CNN allenata con immagini sintetiche di oggetti 3D su sfondi reali. L'approccio proposto permette di stimare i punti di vista di oggetti in situazioni reali, per poter stimare la posa 6D risolvendo un problema di classificazione.

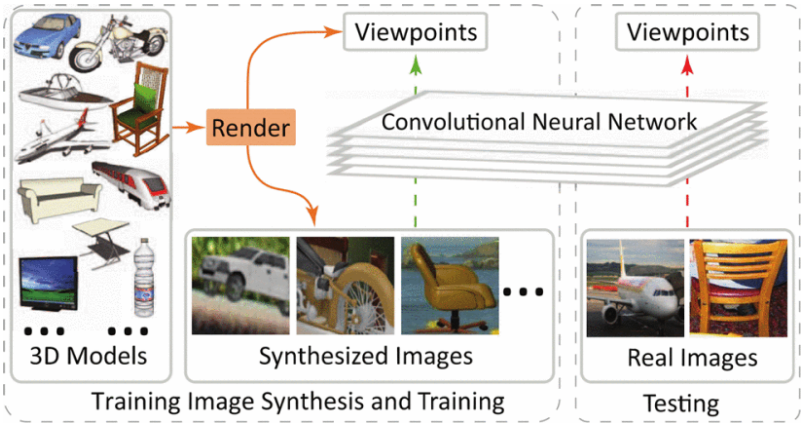


Figura 3.10: Rappresentazione schematica della metodologia usata da Su et al. [29]

Come mostrato nella Figura 3.10, il sistema sintetizza immagini di addestramento sovrapponendo immagini, generate da ampie raccolte di modelli 3D, su immagini reali. La CNN, addestrata utilizzando una combinazione di immagini reali e immagini sintetizzate, è utilizzata per mappare le immagini alle visualizzazioni reali degli oggetti, per poi stimarne la posa 6D.

Kehlt et al. [30] ha esteso le capacità del paradigma SSD (Single Shot MultiBox Detector [31]) per poter definire la stima della posa 6D realizzando un sistema end-to-end. Una CNN

è allenata per poter riconoscere i bounding box 2D degli oggetti da un'immagine RGB e per fornire ogni box con un set delle possibili pose 6D per quella determinata istanza. Lo spazio di rotazione 3D viene scomposto in punti di vista discreti e rotazioni nel piano per poter trattare la stima della rotazione come un problema di classificazione.

Mousavian et al. [32] cerca di stimale la posizione e la grandezza della bounding box 3D di un oggetto partendo da una bounding box 2d e utilizzando i pixel vicini. Il primo passo della rete è la stima dell'orientamento 3D e poi si regrediscono le dimensioni dell'oggetto utilizzando una CNN (Figura 3.11). Queste stime, combinate con i vincoli geometrici sulla traslazione imposti dal bounding box 2D, ci consentono di stimare una posa 3D stabile e accurata dell'oggetto. Questo metodo semplice ed efficiente è adatto a molte applicazioni del mondo reale, comprese le vetture con guida autonoma.

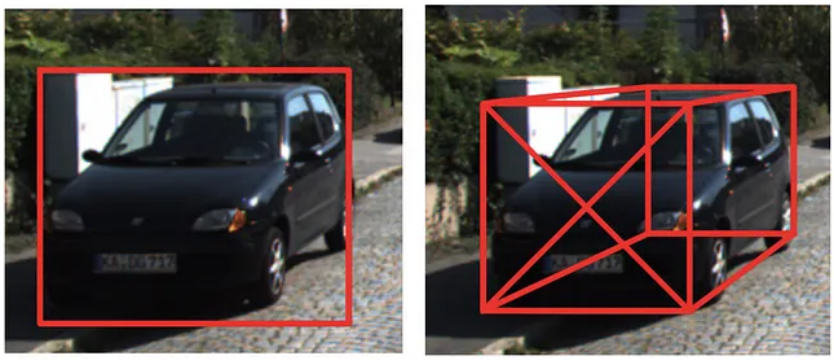


Figura 3.11: L'approccio di Mousavian et al. [27] rileva un bounding box 2D tramite classificazione per poi rilevare un bounding box 3D.

3.3.3 Esempi di metodi basati sulla regressione

Xiang et al. [12] ha realizzato PoseCNN, considerato uno dei migliori metodi per performare la stima della posa 6D utilizzando immagini RGB. La CNN esegue tre compiti principali: etichettatura semantica, stima della traslazione 3D e regressione della rotazione 3D (Figura 3.12). Una volta segmentati gli oggetti, per stimare la loro traslazione 3D viene individuato il centro di ognuno di essi e poi viene calcolata la loro distanza dalla telecamera. La rotazione 3D viene stimata tramite regressione a una rappresentazione quaternionica. Le prime due fasi estraggono e integrano mappe di caratteristiche, dall'immagine di input, con diverse risoluzioni. PoseCNN è robusto alle occlusioni e con l'introduzione di una funzione di perdita, Shape-Match-Loss, riesce a stimare la posa anche di oggetti simmetrici, però trova complicazioni quando nelle scene ci sono oggetti identici.

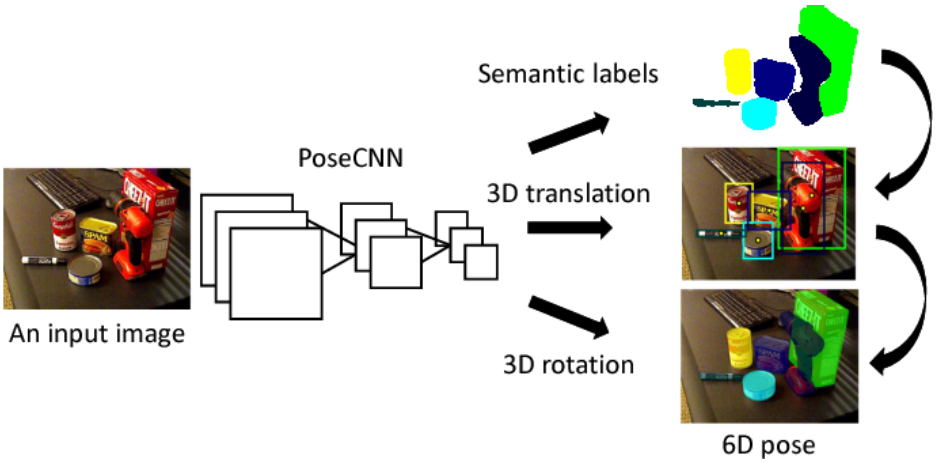


Figura 3.12: PoseCnn proposto in [12].

Il sistema realizzato da Hu et al. [33] effettua le pose 6D regredendo gruppi di corrispondenze 2D-3D associate a punti chiave. Le corrispondenze 3D-2D vengono inizialmente stabilite tramite una CNN guidata dalla segmentazione. Il sistema utilizza 3 moduli per stimare la posa: modulo per l'estrazione delle caratteristiche locali, modulo per aggregare le caratteristiche e un modulo di inferenza globale. Nel primo modulo vengono estratte le caratteristiche locali dell'immagine utilizzando una CNN con parametri condivisi che verranno poi aggregate all'interno di differenti cluster di corrispondenze 3D-2D nel secondo modulo. Infine viene effettuate una inferenza globale per poter ottenere la posa 6D.

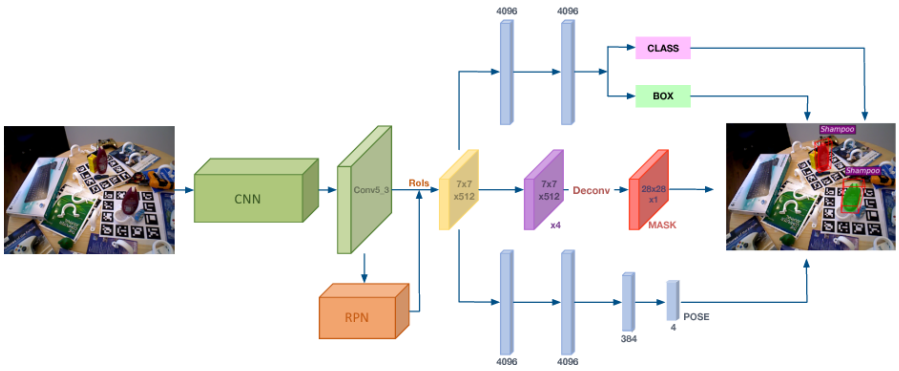


Figura 3.13: Rappresentazione schematica della metodologia usata da [34].

Infine Do et al. [34] ha realizzato un sistema end-to-end chiamato Deep-6DPose che simultaneamente riesce a individuare, segmentare e stimare la posa degli oggetti. Il framework utilizza una CNN, in questo caso la VGG (Figura 3.13 $Conv5_3$), per estrarre le caratteristiche principali dall'immagine. Un network di proposte di regione (RPN) è collegato all'ultimo strato convoluzionale della VGG e produce Region of Interest (RoIs). Per ogni RoI, vengono estratte

e raggruppate le caratteristiche corrispondenti dalla mappa delle caratteristiche convoluzionali. Queste caratteristiche raggruppate vengono utilizzate come input per quattro rami principali: regressione della bounding box, classificazione, segmentazione e posa 6D. L'output finale del modello comprende le istanze rilevate con le relative classi, le maschere di segmentazione previste e le pose 6D previste per le istanze rilevate, mostrate con bounding box 3D.

Capitolo 4

Conclusioni

In questo capitolo finale, trarremo delle conclusioni riguardo alla stima della posa 6D degli oggetti e riepilogheremo i punti chiave discussi in tutta la tesi. Esploreremo le principali scoperte dalla nostra analisi delle diverse categorie di metodi e delle sfide inerenti a questo campo di ricerca in continua evoluzione.

Abbiamo visto come la stima della posa 6D abbia un ampio spettro di applicazioni pratiche, che vanno dalla robotica alla realtà aumentata, dalla visione artificiale alle vetture autonome. Queste tecnologie stanno rapidamente cambiando il modo in cui interagiamo con il mondo fisico e hanno il potenziale per trasformare settori chiave.

Durante la nostra analisi, abbiamo identificato diverse sfide chiave nel campo della stima della posa 6D: la presenza di occlusioni e uno sfondo in disordine rendono difficile l'individuazione dell'oggetto, oggetti simmetrici e oggetti senza texture sono difficili da gestire in molti metodi ma estremamente importanti, le variazioni di illuminazione rendono difficile la lettura dell'immagine e la necessità di grandi quantità di dati di addestramento comportano la necessità di adottare dei compromessi.

Abbiamo suddiviso l'analisi dei metodi di stima della posa 6D in tre categorie principali: i metodi Template-Based, che si basano sulla comparazione tra immagini di template e l'immagine di input per determinare la posizione e l'orientamento dell'oggetto. Questi metodi sono relativamente facili da implementare e possono raggiungere un'alta precisione quando il database dei template è esaustivo. Tuttavia, sono suscettibili alle variazioni di illuminazione e alle occlusioni, e la velocità di esecuzione può dipendere dal numero di elementi nel modello. I metodi Feature-Based, che si concentrano sull'estrazione e la corrispondenza di caratteristiche locali tra modelli 3D e immagini 2D. Questi metodi affrontano alcune delle sfide dei metodi Template-Based ma possono essere complessi da implementare e richiedere un'accurata estrazione delle caratteristiche. I metodi Learning-Based, una categoria avanzata di approcci che si basano sull'apprendimento da dati di addestramento utilizzando reti neurali convoluzionali

(CNN) e tecniche di deep-learning. Questi metodi possono raggiungere elevati livelli di precisione e possono essere più robusti alle variazioni di illuminazione e alle occlusioni. Tuttavia, richiedono una grande quantità di dati di addestramento e risorse computazionali significative. Si potrebbe pensare che i metodi Template-Based e Feature-Based siano obsoleti, visto l'elevato numero di metodi che utilizzano deep-learning negli ultimi anni. Però possono essere utilizzati come supporto per i metodi Learning-Based, usati come singoli passaggi di un sistema più ampio basato su CNN. In generale, la scelta del metodo dipende dalla natura specifica del problema e dai requisiti dell'applicazione, ciascuna categoria presenta vantaggi e svantaggi specifici che possono essere adatti a diversi contesti di applicazione.

Il campo di studio continua a evolversi, e vi sono promettenti opportunità di ricerca future: sviluppare metodi ibridi che combinano le diverse categorie di approcci potrebbe portare a risultati migliori e più robusti, bisognerebbe ridurre la dipendenza da grandi dataset di addestramento e bisognerebbe migliorare l'efficienza computazionale dei metodi Learning-Based per renderli più accessibili in applicazioni reali. Prevedo che in futuro, gli approcci Learning-Based saranno ulteriormente sviluppate, mantenendo da un lato la loro robusta performance e dall'altro lato, il tempo di addestramento dovrebbe essere ridotto. Questi approcci dovrebbero anche essere combinati con approcci Non-Learning-Based per ottenere risultati più accurati. In conclusione, la stima della posa 6D è un campo di ricerca entusiasmante che offre soluzioni innovative per problemi complessi e che con ulteriori ricerche e sviluppi, continuerà a migliorare la nostra capacità di percepire e interagire con il mondo circostante.

Bibliografia

[1] Z. He, W. Feng, X. Zhao e Y. Lv, «6D Pose Estimation of Objects: Recent Technologies and Challenges,» *Applied Sciences*, vol. 11, n. 1, 2021, issn: 2076-3417. doi: 10.3390/app11010228. indirizzo: https://www.mdpi.com/2076-3417/11/1/228.

[2] G. Liang, F. Chen, Y. Liang, Y. Feng, C. Wang e X. Wu, «A Manufacturing-Oriented Intelligent Vision System Based on Deep Neural Network for Object Recognition and 6D Pose Estimation,» *Frontiers in Neurorobotics*, vol. 14, 2021, issn: 1662-5218. doi: 10.3389/fnbot.2020.616775. indirizzo: https://www.frontiersin.org/articles/10.3389/fnbot.2020.616775.

[3] J. L. Blanco, «A tutorial on SE(3) transformation parameterizations and on-manifold optimization,» set. 2010.

[4] N. Correll, K. E. Bekris, D. Berenson et al., «Analysis and Observations From the First Amazon Picking Challenge,» *IEEE Transactions on Automation Science and Engineering*, vol. 15, n. 1, pp. 172–188, 2018. doi: 10.1109/TASE.2016.2600527.

[5] S. Ghidoni, M. Terreran, D. Evangelista et al., «From Human Perception and Action Recognition to Causal Understanding of Human-Robot Interaction in Industrial Environments,» 2022.

[6] H. E. Team, «Automating Your Processes Through Bin Picking – Here's What to Look For,» *HowToRobot*, indirizzo: https://howtorobot.com/expert-insight/automating-your-processes-through-bin-picking-heres-what-look.

[7] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier e B. MacIntyre, «Recent advances in augmented reality,» *IEEE Computer Graphics and Applications*, vol. 21, n. 6, pp. 34–47, 2001. doi: 10.1109/38.963459.

[8] L. Wang, X. Zhang, Z. Song et al., «Multi-Modal 3D Object Detection in Autonomous Driving: A Survey and Taxonomy,» *IEEE Transactions on Intelligent Vehicles*, vol. 8, n. 7, pp. 3781–3798, 2023. doi: 10.1109/TIV.2023.3264658.

[9] M. R. Nichols, «How Augmented Reality Will Disrupt The Manufacturing Industry,» *Thomasnet*, indirizzo: https://blog.thomasnet.com/augmented-reality-manufacturing.

[10] S. Hoque, M. Y. Arafat, S. Xu, A. Maiti e Y. Wei, «A Comprehensive Review on 3D Object Detection and 6D Pose Estimation With Deep Learning,» *IEEE Access*, vol. 9, pp. 143 746–143 770, 2021. doi: 10.1109/ACCESS.2021.3114399.

[11] T. Hodaň, M. Sundermeyer, B. Drost et al., «BOP Challenge 2020 on 6D Object Localization,» *European Conference on Computer Vision Workshops (ECCVW)*, 2020. indirizzo: https://bop.felk.cvut.cz/datasets/.

[12] Y. Xiang, T. Schmidt, V. Narayanan e D. Fox, «PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes,» *CoRR*, vol. abs/1711.00199, 2017. arXiv: 1711.00199. indirizzo: http://arxiv.org/abs/1711.00199.

[13] S. Peng, Y. Liu, Q. Huang, X. Zhou e H. Bao, «PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation,» in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] S. Hinterstoisser, V. Lepetit, S. Ilic et al., «Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes,» in *Computer Vision – ACCV 2012*, K. M. Lee, Y. Matsushita, J. M. Rehg e Z. Hu, cur., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 548–562, isbn: 978-3-642-37331-2.

[15] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton e C. Rother, «Learning 6D Object Pose Estimation Using 3D Object Coordinates,» in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele e T. Tuytelaars, cur., Cham: Springer International Publishing, 2014, pp. 536–551, isbn: 978-3-319-10605-2.

[16] T. Hodan, P. Haluza, S. Obdrzalek, J. Matas, M. Lourakis e X. Zabulis, *T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects*, 2017. arXiv: 1701.05498 [cs.CV].

[17] G. Marullo, L. Tanzi, P. Piazzolla e E. Vezzetti, «6D object position estimation from 2D images: a literature review,» *Multimedia Tools and Applications*, vol. 82, n. 16, pp. 24 605–24 643, 2023, issn: 1573-7721. doi: 10.1007/s11042-022-14213-z. indirizzo: https://doi.org/10.1007/s11042-022-14213-z.

[18] M. Ulrich, C. Wiedemann e C. Steger, «Combining Scale-Space and Similarity-Based Aspect Graphs for Fast 3D Object Recognition,» *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, n. 10, pp. 1902–1914, 2012. doi: 10.1109/TPAMI.2011.266.

[19] Y. Konishi, Y. Hanzawa, M. Kawade e M. Hashimoto, «Fast 6D Pose Estimation from a Monocular Image Using Hierarchical Pose Trees,» in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe e M. Welling, cur., Cham: Springer International Publishing, 2016, pp. 398–413, isbn: 978-3-319-46448-0.

[20] N. Payet e S. Todorovic, «From contours to 3D object detection and pose estimation,» in *2011 International Conference on Computer Vision*, 2011, pp. 983–990. doi: 10.1109/ICCV.2011.6126342.

[21] F. Massa, B. C. Russell e M. Aubry, «Deep Exemplar 2D-3D Detection by Adapting from Real to Rendered Views,» in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 6024–6033. doi: 10.1109/CVPR.2016.648.

[22] M. Sundermeyer, Z.-C. Marton, M. Durner e R. Triebel, «Augmented Autoencoders: Implicit 3D Orientation Learning for 6D Object Detection,» *International Journal of Computer Vision*, vol. 128, n. 3, pp. 714–729, 2020, issn: 1573-1405. doi: 10.1007/s11263-019-01243-8. indirizzo: https://doi.org/10.1007/s11263-019-01243-8.

[23] J.-K. You, C.-C. J. Hsu, W.-Y. Wang e S.-K. Huang, «Object Pose Estimation Incorporating Projection Loss and Discriminative Refinement,» *IEEE Access*, vol. 9, pp. 18 597–18 606, 2021. doi: 10.1109/ACCESS.2021.3054493.

[24] J. N. Kundu, M. V. Rahul, A. Ganeshan e R. V. Babu, «Object Pose Estimation from Monocular Image Using Multi-view Keypoint Correspondence,» in *Computer Vision – ECCV 2018 Workshops*, L. Leal-Taixé e S. Roth, cur., Cham: Springer International Publishing, 2019, pp. 298–313, isbn: 978-3-030-11015-4.

[25] C. Chen, X. Jiang, W. Zhou e Y.-H. Liu, *Pose Estimation for Texture-less Shiny Objects in a Single RGB Image Using Synthetic Training Data*, 2019. arXiv: 1909.10270 [cs.RO].

[26] M. Rad e V. Lepetit, «BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth,» in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3848–3856. doi: 10.1109/ICCV.2017.413.

[27] Y. Hu, J. Hugonot, P. Fua e M. Salzmann, «Segmentation-Driven 6D Object Pose Estimation,» in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[28] Z. Li, G. Wang e X. Ji, «CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation,» in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7677–7686. doi: 10.1109/ICCV.2019.00777.

[29] H. Su, C. R. Qi, Y. Li e L. J. Guibas, «Render for CNN: Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views,» in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2686–2694. doi: 10.1109/ICCV.2015.308.

[30] W. Kehl, F. Manhardt, F. Tombari, S. Ilic e N. Navab, «SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again,» in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1530–1538. doi: 10.1109/ICCV.2017.169.

[31] W. Liu, D. Anguelov, D. Erhan et al., «SSD: Single Shot MultiBox Detector,» in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe e M. Welling, cur., Cham: Springer International Publishing, 2016, pp. 21–37, isbn: 978-3-319-46448-0.

[32] A. Mousavian, D. Anguelov, J. Flynn e J. Košecká, «3D Bounding Box Estimation Using Deep Learning and Geometry,» in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5632–5640. doi: 10.1109/CVPR.2017.597.

[33] Y. Hu, P. Fua, W. Wang e M. Salzmann, «Single-Stage 6D Object Pose Estimation,» in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[34] T. Do, M. Cai, T. Pham e I. D. Reid, «Deep-6DPose: Recovering 6D Object Pose from a Single RGB Image,» *CoRR*, vol. abs/1802.10367, 2018. arXiv: 1802.10367. indirizzo: http://arxiv.org/abs/1802.10367.