VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

DETEKCE OBJEKTU S VYUŽITÍM HLOUBKOVÝCH DAT

OBJECT DETECTION USING DEPTH DATA

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **MAREK VALKO**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. PETR MUSIL, Ph.D.**
SUPERVISOR

BRNO 2024

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Zadání bakalářské práce

155964

Ústav:	Ústav počítačové grafiky a multimédií (UPGM)
Student:	**Valko Marek**
Program:	Informační technologie
Název:	**Detekce objektu s využitím hloubkových dat**
Kategorie:	Počítačové vidění
Akademický rok:	2023/24

Zadání:

1. Seznamte se s metodami detekce objektů v obraze a hloubkových datech (RGBD).
2. Vyberte, případně vytvořte, vhodnou datovou sadu pro detekci objektu v RGBD.
3. Vyhledejte metody hlubokého učení zaměřující se na problematiku detekce objektů v RGBD, případně navrhněte úpravy standardních metod tak, aby mohly pracovat s RGBD daty.
4. Vyberte vhodné metody a experimentálně je otestujte.
5. Vhodným způsobem vyhodnoťte výsledky experimentů a diskutujte je.
6. Vytvořte stručné video prezentující cíle a výsledky vaší práce.

Literatura:

- S. Song, S. Lichtenberg, and J. Xiao; SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite; CVPR2015
- Xiao, Z; Xue, J; Xie, P; Wang, G; FETNet: Feature Exchange Transformer Network for RGB-D Object Detection; BMVC2021

Podrobné závazné pokyny pro vypracování práce viz https://www.fit.vut.cz/study/theses/

Vedoucí práce:	**Musil Petr, Ing., Ph.D.**
Vedoucí ústavu:	Černocký Jan, prof. Dr. Ing.
Datum zadání:	1.11.2023
Termín pro odevzdání:	9.5.2024
Datum schválení:	9.11.2023

Abstrakt

Táto bakalárska práca sa zaoberá detekciou objektov v obraze s využitím hĺbkových dát. Cieľom bolo zvoliť vhodné metódy hlbokého učenia a experimentálne ich overiť na relevantných dátových sadách. Práca začína prehľadom základných techník detekcie objektov v obraze a hĺbkových dátach. V rámci riešenia boli vybrané dátové sady NYU Depth v2 a Washington RGB-D, na ktorých sa testovali upravené modely YOLOv5 a YOLOv8. Experimenty skúmali rôzne reprezentácie hĺbkových informácií a analyzovali, ako integrácia hĺbkových dát zlepšuje výkon týchto modelov. Výsledky ukázali výrazné zlepšenie metrík mAP pri porovnaní s klasickými modelmi využívajúcimi iba RGB dáta. Integrácia hĺbkových dát tak umožnila dosiahnuť presnejšie a spoľahlivejšie výsledky pri detekcii objektov.

Abstract

This bachelor thesis addresses the detection of objects in images using depth data. The goal was to select appropriate deep learning methods and experimentally verify them on relevant datasets. The thesis begins with an overview of basic techniques for detecting objects in images and depth data, utilizing selected datasets NYU Depth v2 and Washington RGB-D to test modified YOLOv5 and YOLOv8 models, adapted for effective processing of RGB-D data. The experiments explored various representations of depth information and analyzed how the integration of depth data enhances the performance of these models. The results demonstrated significant improvements in mAP metrics compared to traditional models that use only RGB data. The integration of depth data thus allowed for more accurate and reliable object detection results.

Kľúčové slová

detekcia objektov, RGBD, hĺbkové dáta, YOLOv5, YOLOv8, reprezentácia hĺbkových dát, HHA, jet, skoré a neskoré zlúčenie

Keywords

object detection, RGBD, depth data, YOLOv5, YOLOv8, depth data representation, HHA, jet, early and late fusion

Citácia

VALKO, Marek. *Detekce objektu s využitím hloubkových dat.* Brno, 2024. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Petr Musil, Ph.D.

Detekce objektu s využitím hloubkových dat

Prehlásenie

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Ing. Petra Musila, Ph.D.. Uviedol som všetky literárne pramene, publikácie a dalšie zdroje, z ktorých som čerpal.

.....
Marek Valko
8. mája 2024

Poďakovanie

Rád by som poďakoval svojmu vedúcemu práce Ing. Petrovi Musilovi, Ph.D. za pomoc, odborné vedenie a cenné rady.

Obsah

1 Úvod **2**

2 Hĺbkové dáta a ich spracovanie **3**
2.1 Počítačové videnie a spracovanie obrazu 3
2.2 Hĺbkové dáta 6
2.3 Metódy spracovania hĺbkových dát 10
2.4 Detektory objektov 16

3 Návrh riešenia **20**
3.1 Dátové sady 20
3.2 Zvolené metódy integrácie hĺbkových dát do detekčných algoritmov 24
3.3 Metriky pre zrovnávanie 25
3.4 Trénovacie prostredie 27

4 Experimenty **28**
4.1 Experimenty s upravenými modelmi YOLOv5 a rôznymi reprezentáciami hĺbky na dátovej sade NYU Depth v2 28
4.2 Experiment s upravenými modelmi YOLOv5 na dátovej sade Washington RGB-D 48
4.3 Experiment s pridaním hĺbkových dát do modelu YOLOv8 50

5 Záver **54**

Literatúra **57**

A Obsah priloženej SD karty **62**

Kapitola 1

Úvod

V súčasnosti sme svedkami prudkého vývoja a zlepšovania metód počítačového videnia, ktoré sú čoraz účinnejšie aj vďaka rýchlemu pokroku v oblasti strojového učenia. Tieto technológie prenikajú do mnohých aspektov každodenného života, priemyslu a vedeckého výskumu. Vďaka schopnosti počítačových systémov učiť sa z veľkého množstva dát a neustále sa zlepšovať, sú dnes metódy strojového učenia neoddeliteľnou súčasťou pokročilých výskumov a vývoja v počítačovom videní. Integrácia hĺbkových dát do procesov detekcie a rozpoznávania objektov ponúka nové perspektívy a umožňuje presnejšie a spoľahlivejšie výsledky než kedykoľvek predtým.

Hĺbkové dáta predstavujú ďalšiu dimenziu informácií, ktorá sa pridáva k tradičným RGB dátam. Táto dimenzia poskytuje cenné informácie o vzdialenostiach objektov od senzora a ich vzájomných polohách v priestore. Význam pridávania hĺbkových dát k RGB dátam spočíva v schopnosti modelov počítačového videnia lepšie rozumieť a interpretovať scény zložitých a preplnených prostredí, kde môžu byť objekty čiastočne zakryté alebo inak vizuálne nepostrehnuteľné. Integrácia hĺbkových a RGB dát vedeckými tímami z celého sveta ukazuje, že tieto prístupy môžu dramaticky zlepšiť presnosť a efektivitu detekčných algoritmov.

Dostupnosť senzorov na získavanie hĺbkových dát, akými sú LiDAR alebo RGB-D kamery, je v súčasnosti vyššia ako kedykoľvek predtým, čo umožňuje ich širšie využitie a integráciu do komerčných a výskumných projektov. Tieto zariadenia, ktoré boli kedysi drahé a bolo obtiažne ich získať, sú teraz cenovo oveľa dostupnejšie, čo otvára dvere pre inovácie vo viacerých sektoroch.

Táto práca sa systematicky zaoberá viacerými metódami detekcie objektov s využitím hĺbkových dát. Začiatok práce predstavuje základné metódy detekcie objektov v obraze a hĺbkových dátach. Postupne sa prechádza k výberu konkrétnych metód a príprave vhodných dátových sád, na ktorých je možné tieto metódy otestovať. V práci sú tiež analyzované a upravované existujúce metódy strojového učenia tak, aby efektívne pracovali s kombinovanými RGBD dátami. Práca pokračuje sériou experimentov, ktoré testujú účinnosť týchto metód na vybraných dátových sadách NYU Depth Dataset V2 a Washington RGB-D. Cieľom experimentov je objektívne posúdiť účinnosť navrhovaných riešení a možnosť ich aplikovania v praxi.

Kapitola 2

Hĺbkové dáta a ich spracovanie

Nasledujúca kapitola sa zaoberá spracovaním hĺbkových dát a poskytuje teoretický základ spolu s prehľadom doterajšieho výskumu v oblasti detekcie objektov s využitím hĺbkových dát. Kapitola začína krátkym úvodom do spracovania obrazu a počítačového videnia, v ktorom sú predstavené niektoré základné techniky a pojmy. Jej ďalšia časť sa venuje samotným hĺbkovým dátam, ich reprezentáciou v podobe point cloudých dát alebo hĺbkových máp a spôsobom získavania takýchto dát pomocou špecializovaných senzorov. Následne sú predstavené rôzne metódy detekcie objektov v point cloudových dátach a RGB-D obrazoch. Opísané sú viaceré prístupy k zlúčeniu farebného obrazu a hĺbky, akými sú metódy skorého, neskorého, hlbokého a doplnkového zlúčenia. Pozornosť je venovaná aj rôznym metódam reprezentácie hĺbky v hĺbkových mapách, akými sú odtiene šedi, farebná paleta jet, povrchové normály a pokročilé kódovanie hĺbky HHA. Táto časť končí analýzou výziev, ktoré so sebou spracovanie hĺbkových dát prináša. Kapitolu uzatvára prehľad o jednokrokových detektoroch objektov, konkrétne o detektoroch série YOLO vo verziách YOLOv5 a YOLOv8.

2.1 Počítačové videnie a spracovanie obrazu

Človek dokáže vnímať, rozpoznávať a následne reagovať na svoje okolie pomocou zmyslových orgánov. Jedným z najdôležitejších zmyslových orgánov je ľudský zrak, vďaka ktorému človek získava až 80% všetkých informácií zo svojho okolia [55]. Je preto pochopiteľné, že existuje stále väčšia snaha implementovať a vyvíjať túto vlastnosť aj v oblasti počítačovej techniky. Schopnosť vnímať a následne reagovať na svoje okolie predstavuje v počítačovej technike veľmi dôležitú úlohu.

Počítačové videnie je odvetvie, ktoré sa zaoberá získavaním dôležitých informácií zo zachyteného obrazu. Pod pojmom získať informácie je myslené správne rozoznanie skupiny objektov či javov a nasledovné využitie týchto informácií pre ďalšie spracovanie. Obrazové dáta môžu mať rôzne formy ako napríklad fotografie, video sekvencie, pohľady z viacerých kamier alebo viacrozmerné dáta z rôznych skenerov. Tieto dáta môžu byť uložené v pamäti počítača, alebo získané z pripojenej kamery. Spracovanie obrazu predstavuje jednu z oblastí počítačového videnia, ktorá má široké uplatnenie vo výrobnom priemysle, medicíne, vojenskom priemysle, bezpečnostných systémoch alebo dopravnom priemysle.

Prvým krokom v procese počítačového videnia sú techniky predspracovania obrazu, ktoré slúžia na optimalizáciu snímok pre následné analytické účely. Tento krok predstavuje úpravy, ktoré zlepšujú kvalitu obrazu tým, že eliminujú šum, korigujú skreslenia a zlep-

šujú celkovú čitateľnosť snímok. Tieto úpravy sú nevyhnutné pre efektívnejšiu extrakciu a klasifikáciu charakteristických príznakov, čo vedie k presnejšej detekcii a rozpoznávaniu objektov.

V tejto práci sa používajú techniky predspracovania, ktoré sú podrobne rozobraté v nasledujúcej časti.

Filtrácia

Hlavným účelom filtrácie je zvýšiť kvalitu obrazu prostredníctvom odstránenia šumu a nechcených artefaktov, zatiaľ čo sa zachovajú kľúčové atribúty ako sú hrany a textúry [15]. V tejto štúdii sa používajú filtre na zlepšenie kvality hĺbkových máp. Nižšie sú uvedené niektoré z najčastejšie používaných typov filtrov.

- **Mediánový filter** je nelineárna filtračná metóda, kde sa hodnota každého pixelu nahradí mediánom hodnôt zo svojho okolia. Táto metóda je veľmi efektívna pri eliminácii impulzného šumu a zároveň zachováva hrany a ostré prechody. Mediánové filtrovanie je často vyberané pre jeho schopnosť udržať štruktúru obrazu a účinne redukovať šum.

- **Gaussovský filter** predstavuje lineárnu filtráciu, kde sa používa Gaussovo jadro pre konvolúciu s vstupným obrazom. Tento prístup efektívne znižuje vysokofrekvenčný šum a zároveň udržiava nízkofrekvenčné charakteristiky. Gaussovské filtrovanie je obľúbené pre jeho schopnosť vyhladiť obrazy bez značného skreslenia alebo straty detailov.

- **Bilaterálny filter** je sofistikovanejší nástroj, ktorý kombinuje priestorové a intenzitné súradnice pre filtráciu obrazu. Tento filter poskytuje vyhladenie, pričom súčasne zachováva hrany a jemné detaily tým, že zohľadňuje nielen blízkosť pixelov, ale aj podobnosť ich intenzít. Táto dualita umožňuje bilaterálnemu filtru vykonať vyhladenie oblastí v obraze, kde sú farby a intenzity podobné, pričom súčasne zabraňuje rozmazaniu cez hrany, kde sú tieto hodnoty odlišné. Bilaterálny filter sa najčastejšie využíva v pokročilých grafických aplikáciách a video post-produkcii, kde je potrebné dosiahnuť vysokú úroveň detailu a vizuálnej kvality.

Normalizácia

Metóda normalizácie v oblasti predspracovania obrazu slúži na zjednotenie hodnôt intenzity v rámci zvoleného rozsahu, ako je napríklad 0 až 255. Tento postup zaisťuje konzistentnejšiu a jednotnejšiu reprezentáciu obrazových dát pre ďalšie spracovanie, ako je extrakcia príznakov a ich klasifikácia [15].

- **Min-max normalizácia** je proces, ktorý upravuje intenzity obrazu tak, aby spadali do preddefinovaného rozsahu, napríklad od 0 do 1 alebo od 0 do 255. Tento postup sa realizuje použitím lineárnej transformácie, ktorá premapuje existujúci rozsah hodnôt na požadovaný rozsah na základe najnižších a najvyšších hodnôt zaznamenaných v obraze.

- **Vyrovnávanie histogramov** je technika, ktorá optimalizuje distribúciu intenzity obrazu. Využíva kumulatívnu distribučnú funkciu (CDF) príslušného histogramu na

redistribúciu intenzít, čo vedie k zvýšeniu kontrastu a lepšej viditeľnosti detailov v obraze. Tento prístup výrazne prispieva k zlepšeniu vizuálnej kvality obrazu, čo je dôležité pre precíznejšie spracovanie a analýzu obrazových dát.

Rozpoznávanie obrazov

Proces identifikácie objektov v digitálnych obrazoch, známy ako rozpoznávanie obrazov, slúži k identifikácii a kategorizácii objektov alebo vzorov. Táto analýza umožňuje počítačovým systémom rozlišovať objekty na základe ich jedinečných atribútov. Rozpoznávanie obrazov je dôležitou súčasťou oblasti počítačového videnia a má široké praktické uplatnenie, ako napríklad v analýze medicínskych obrazov (napr. röntgenových snímok), detekcii ciest a prekážok pre autonómne vozidlá a v detekcii defektov pri kontrolách kvality. Medzi najčastejšie používané techniky patrí klasifikácia obrazov, detekcia objektov, sémantická segmentácia a segmentácia inštancií.

Klasifikácia

Klasifikácia obrázkov je technika určená na rozdeľovanie obrázkov do viacerých kategórií alebo skupín [26]. Táto metóda je široko využívaná na rozpoznávanie a identifikáciu konkrétnych objektov, ako napríklad druhy zvierat, druhy rastlín, alebo typy vozidiel. Klasifikačnému procesu predchádza tréningová fáza, počas ktorej je každý obrázok v dátovej sade označený príslušnou kategóriou a na základe tohto tréningu je model schopný klasifikovať nové obrázky do príslušných kategórií.

Detekcia objektov

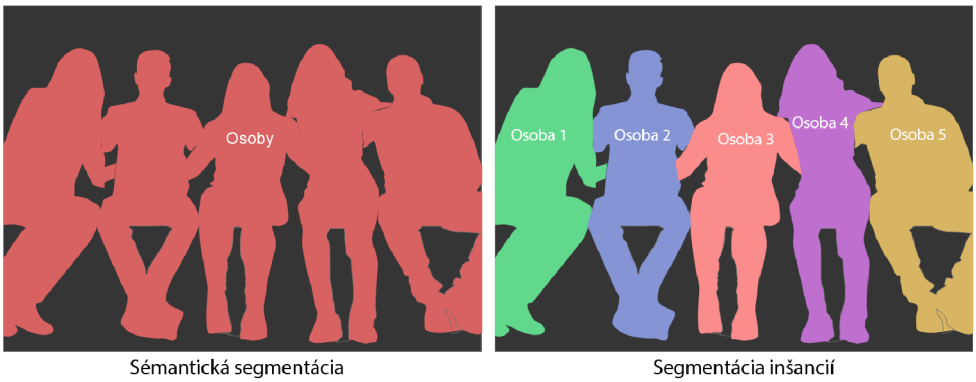
Detekcia objektov je pokročilejšia technika než klasifikácia obrázkov, pretože umožňuje nielen rozpoznávanie objektov, ale aj určenie ich polohy na obrázku alebo vo videu [66]. Detekcia objektov je využívaná v aplikáciách, kde je dôležité poznať presnú polohu objektov, ako napríklad v autonómnych vozidlách alebo v systémoch pre monitorovanie bezpečnosti. Tréningové dáta pre detekciu objektov obsahujú anotácie (označenia) určujúce presné umiestnenie objektov v rámci obrázku.

Ilustrácia rozdielov medzi klasifikáciou obrazu a detekciou objektov je zobrazená na obrázku 2.1.

Obr. 2.1: Porovnanie klasifikácie obrazu a detekcie objektov, upravené z [25].

Sémantická segmentácia a segmentácia inštancií

Sémantická segmentácia je proces, ktorý rozdeľuje obraz do rôznych oblastí na základe významu alebo kontextu pixelov, pričom každému pixelu priraďuje vopred definovanú triedu, ako je napríklad budova alebo strom, a nerozlišuje medzi rôznymi výskytmi tej istej triedy [12]. Na druhej strane, segmentácia inštancií predstavuje zložitejšiu techniku, ktorá kombinuje detekciu objektov a sémantickú segmentáciu. Táto metóda nielen identifikuje a lokalizuje objekty v obraze, ale aj ich segmentuje na jednotlivé inštancie, čím každý výskyt objektu segmentuje samostatne, bez ohľadu na počet inštancií rovnakej triedy v obraze [53]. Rozdiely medzi týmito dvoma technikami ilustruje obrázok 2.2.



Obr. 2.2: Porovnanie sémantickej segmentácie a segmentácie inštancií, upravené z [8].

2.2 Hĺbkové dáta

Formát RGB-D (Red Green Blue – Depth) predstavuje kombináciu farebných a hĺbkových informácií. Hoci farebná zložka, ktorá obsahuje informácie o farbách, je všeobecne známa, hĺbková zložka, ktorá nesie informácie o vzdialenosti objektov od kamery, je medzi laickou verejnosťou menej rozšírená. Táto hĺbková zložka, reprezentovaná point cloudom alebo hĺbkovými mapami, rozširuje možnosti vizuálneho obsahu o nový rozmer a umožňuje nové využitia siahajúce za hranice tradičného spracovania obrazu.

Point cloud

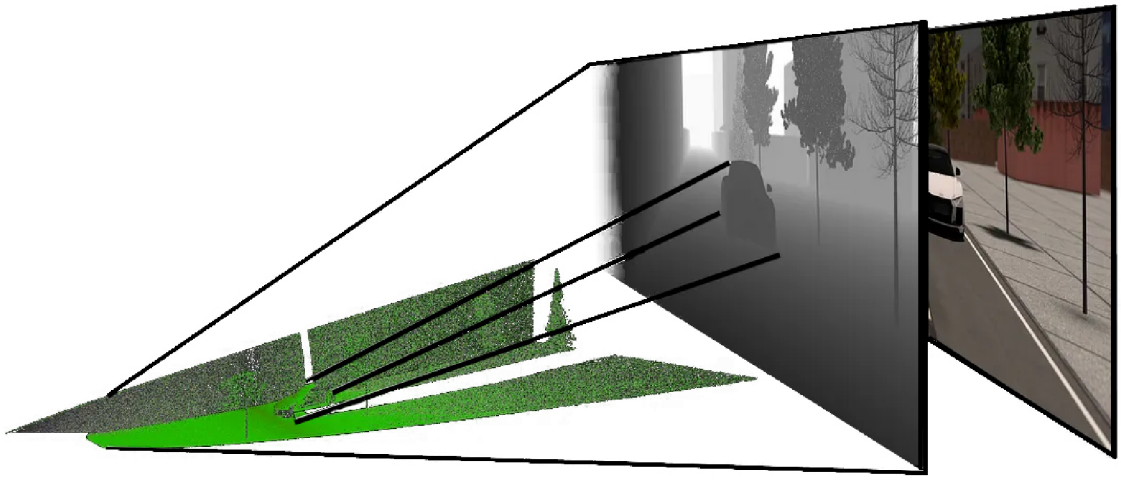
Point cloud (slovensky doslovne oblak bodov) predstavuje spôsob reprezentácie trojrozmerných objektov pomocou súboru bodov v priestore. Každý bod v takomto súbore má priradené tri súradnice - X, Y a Z, ktoré určujú jeho pozíciu v kartézskej súradnicovej sústave. Často sú k týmto bodom pridávané aj dodatočné informácie, napríklad farebné hodnoty v RGB formáte, ktoré poskytujú ďalšie podrobnosti o povrchu skenovaných objektov. Pri získavaní point cloudových dát sa často využívajú 3D skenery alebo fotogrammetrické techniky. Výhodou týchto metód oproti tradičným 2D obrazom je ich schopnosť zaznamenať presné detaily objektov bez ohľadu na vonkajšie svetelné podmienky [39]. Tieto dáta sú cenné hlavne v oblastiach, kde je potrebné poznať objem rozoznávaných objektov alebo sa kladie dôraz na ich povrchové štruktúry. Point cloudové dáta sa spravidla skladajú z obrovského množstva bodov, často v miliónoch, ktoré sú ukladané bez špecifického poriadku. Pre spracovanie a analýzu týchto dát sú preto potrebné pokročilé metódy. S ich pomocou

je možné výrazne zjednodušiť a zefektívniť prácu s 3D objektami. Point cloud najčastejšie nachádza svoje využitie pri 3D modelovaní, 3D vykresľovaní alebo dokonca metrológií [43].

Hĺbkové mapy

Zatiaľ čo point cloudové dáta poskytujú bohatý, trojrozmerný obraz objektov s vysokou mierou detailu, hĺbkové mapy umožňujú zobraziť tieto trojrozmerné informácie na dvojrozmernom obrázku, čo je často pri spracovaní obrazu praktickejšie. Hĺbkové mapy sú zvyčajne dvojrozmerné obrazy v odtieňoch šedej, ktoré majú rovnaké rozlíšenie ako pridružené RGB obrázky. Intenzita šedi jednotlivých pixelov na hĺbkovej mape indikuje vzdialenosť daného bodu od kamery, čím každý pixel na hĺbkovej mape definuje polohu svojho RGB ekvivalentu v osi Z. Hodnota 255 na hĺbkovej mape zvyčajne označuje body scény najvzdialenejšie od kamery, zatiaľ čo hodnota 0 signalizuje body nachádzajúce sa najbližšie k objektívu.

Na obrázku 2.3 sú znázornené tri rôzne pohľady na rovnakú scénu. Na ľavo je point cloud, kde zelené body zobrazujú trojrozmerné súradnice objektov. V strede je hĺbková mapa, ktorá pomocou odtieňov šedej ukazuje vzdialenosť jednotlivých objektov od kamery. Napravo je zarovnaný RGB náprotivok, ktorý poskytuje reálny farebný pohľad na scénu. Každá z týchto vizualizácií ponúka iný druh informácií a je užitočná pre rôzne účely pri spracovaní obrazu.



Obr. 2.3: Porovnanie point cloudu a hĺbkových máp.

RGB-D senzory

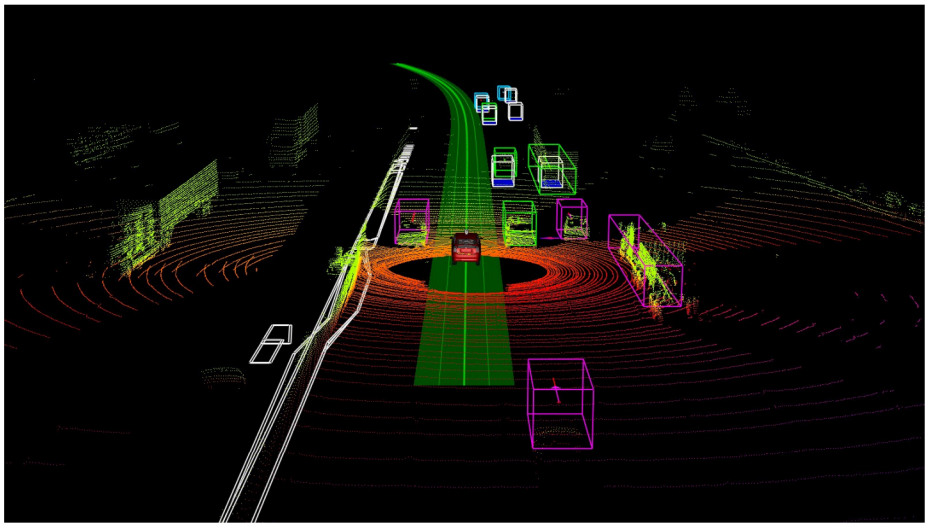
RGB-D senzory sú zariadenia navrhnuté na zachytávanie dát o hĺbke. Hlavnými typmi aktívnych RGB-D senzorov sú senzory doby letu (ToF z anglického time-of-flight) a senzory štruktúrovaného svetla. Senzory doby letu, ako je Azure Kinect, určujú hĺbku meraním času, ktorý svetlo potrebuje na cestu od senzora k objektu a naspäť. Senzory štruktúrovaného svetla, napríklad Microsoft Kinect, premietajú na scénu známy vzor a na základe analýzy jeho deformácie vypočítavajú údaje o hĺbke.

V posledných rokoch sa začali senzory schopné zachytávať hĺbku integrovať aj do smartfónov, ako napríklad kamery TrueDepth a LiDAR skenery na niektorých modeloch iPhone, čím sa táto technológia dostáva k širšiemu okruhu používateľov. Podobné rozšírenie tejto

technológie umožňuje širokému publiku ľahšie využívať možnosti rozšírenej reality, zdokonaliť portrétne fotografie a zlepšiť bezpečnostné funkcie ako napríklad rozpoznávanie tváre.

LIDAR

Lidar, skratka pre Light Detection and Ranging, predstavuje technológiu, ktorá operuje na princípe využitia elektromagnetických vĺn s optickým alebo infračerveným rozsahom. V podstate funguje podobne ako radar, avšak namiesto rádiových vĺn využíva elektromagnetické vĺny s kratším rozsahom. Jednou z najväčších výhod lidaru je jeho schopnosť operovať aj v nepriaznivých poveternostných podmienkach, ako je hmla alebo oblačnosť [41]. Vďaka tejto vlastnosti sa z lidaru stal dôležitý nástroj pre snímanie okolia v rôznych podmienkach a prostrediach. Lidary sú schopné generovať 2D aj 3D obrazy tým, že pre každý pixel vypočíta jeho vzdialenosť od senzoru. Senzor tiež poskytuje údaje o miere schopnosti materiálov odrážať svetlo a dokáže snímať aj farbu povrchu materiálu. Ďalšou výhodou je ich schopnosť merať rýchlosť pohybu objektov vďaka zaznamenaným zmenám vo frekvencii. Lidar nachádza široké využitie vo vojenskom priemysle, najmä vzhľadom na jeho schopnosť fungovať v nízkom osvetlení. Okrem toho sa lidar využíva aj v civilnom sektore, kde poskytuje informácie pre identifikáciu objektov a orientáciu strojov v priestore. Využitie lidaru sa čoraz viac rozširuje do rôznych odvetví. Najčastejšie sa používa na vytváranie detailných máp a mapovania prostredia, ako aj na orientáciu vozidiel v priestore. Získané dáta sú pre človeka veľmi intuitívne a podobné tomu, čo by videl svojimi vlastnými očami. Príklad konkrétneho využitia dát získaných pomocou lidaru je zobrazený na obrázku 2.4.



Obr. 2.4: Detekované 3D objekty v point cloudových dátach získaných LiDARom pomocou modelu YOLO [63].

Odhadovanie hĺbkových máp

Ak nie je k dispozícii špecializovaný hardvér, je možné odhadnúť mapy hĺbky priamo z RGB obrázkov využitím pokročilých algoritmov spracovania obrazu. Tieto metódy často spočívajú v aplikácii techník stereo videnia, ktoré odhadujú hĺbku z obrázkov získaných z viacerých kamier alebo z dvojíc obrázkov. Stereo videnie funguje na princípe porovnávania rozdielov v zobrazení objektov na dvoch snímkach, ktoré boli vyfotografované z rôznych

uhlov. Tieto rozdiely umožňujú algoritmom určiť vzdialenosť objektov, podobne ako ľudské oči vnímajú hĺbku porovnávaním obrazu získaného z každého oka. Tento princíp sa využíva aj pri konverzii starších filmov z 2D na 3D formát, ako napríklad pri filmoch Jurský park (1993) alebo Leví kráľ (1994).

Využitie hĺbkových dát

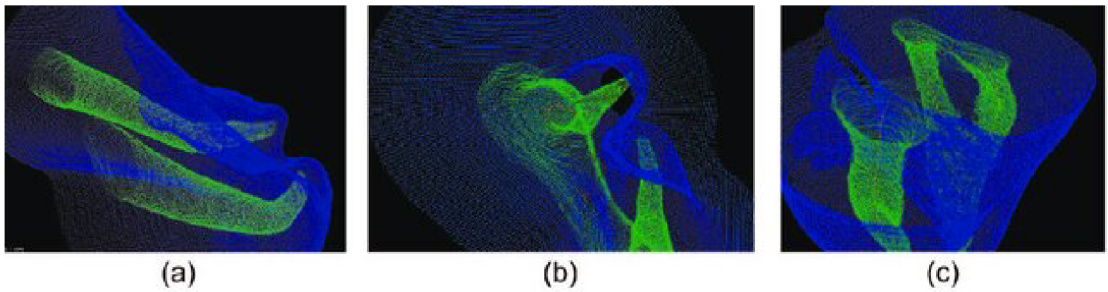
Hĺbkové dáta vzbudzujú značný záujem vďaka geometrickým informáciám, ktoré poskytujú, čo umožňuje podrobné porozumenie zachytenej scény. Na rozdiel od RGB dát, hĺbkové informácie nie sú ovplyvnené uhľom pohľadu, slabým osvetlením, ani textúrou a farbami objektov, čo je čo je výhodou pri využití v mnohých oblastiach, ako sú virtuálna/rozšírená realita a počítačové videnie (rozpoznávanie gest, sledovanie a rozpoznávanie činností, detekcia osôb v dave, navigácia robotov atď.). Aj v oblasti Internetu vecí (IoT) sa budúcnosť uberá smerom k 3D videniu a hĺbkovým technológiám IoT, ktoré umožňujú strojom, ako sú roboty, autonómne vozidlá a drony, dosiahnuť hlboké vnímanie podobné ľudskému. Niektoré z praktických využití hĺbkových dát sú viac predstavené v nasledujúcej časti.

Autonómne systémy: Hĺbkové mapy sú dôležité pre zlepšenie interakcií a navigácie v reálnom svete v reálnom čase, čo pomáha autonómnym technológiám v rôznych oblastiach. S ich pomocou môžu autonómne zariadenia presnejšie interpretovať svoje okolie. Integrácia hĺbkových dát teda v tejto oblasti zvyšuje bezpečnosť v samo-riadiacich autách, pomáha dronom pri kontrolovaní plodín a zlepšuje schopnosť robotov kontrolovať montážne procesy a zisťovať chyby.

Zdravotníctvo: V zdravotníctve hĺbkové mapy prinášajú presné trojrozmerné zobrazenia častí tela. Toto pomáha zvyšovať presnosť chirurgických zákrokov a pomáha to tiež pri tvorbe prispôsobených zdravotníckych pomôcok. Hĺbkové dáta pomáhajú tiež pri monitorovaní dýchacích pohybov v reálnom čase a pri presnej lokalizácií nádorov pri radiačnej terapii, čo znižuje riziko poškodenia zdravých buniek. Príklad využitia hĺbkových dát v medicíne je možné vidieť na obrázku 2.5.

Kamerové systémy a bezpečnosť: Integrácia hĺbkového vnímania do videonahrávok zvyšuje efektivitu bezpečnostného monitorovania a poskytuje lepšie chápanie prostredia. Tento prístup pomáha zlepšiť bezpečnostné opatrenia na preplnených miestach ako sú napríklad letiská. S použitím hĺbkových snímačov je možné presnejšie detekovať nezvyčajné správania alebo predmety.

Virtuálna a rozšírená realita: Hĺbkové mapy umožňujú presné umiestnenie objektov a zabezpečujú plynulú interakciu s virtuálnym prostredím. Pomáhajú predchádzať kolíziám s reálnymi objektami a zlepšujú užívateľský zážitok tým, že umožňujú prirodzenú manipuláciu s virtuálnymi objektmi a tiež zlepšujú simuláciu realistických svetelných podmienok.



Obr. 2.5: Vizualizácia point cloudových 3D údajov predstavujúcich zub a segmentovaný koreňový kanálik. Tri rôzne pohľady (a), (b) a (c) na zub s dvoma koreňovými kanálikmi [10].

2.3 Metódy spracovania hĺbkových dát

Detekcia objektov v point cloudových dátach

Jeden z prístupov k spracovaniu hĺbkových dát sa špecificky zameriava na detekciu 3D objektov z point cloudových dát, ako sú napríklad LiDARové skeny.

Pan a kolgeovia v štúdií [46] predstavujú architektúru Pointformer, navrhnutú špecificky pre point cloudové dáta s využitím transformera na efektívne učenie sa príznakov. Ich lokálny transformerový modul umožňuje modelovať interakcie medzi bodmi v lokálnej oblasti a učiť sa príznaky závislé od kontextu na úrovni objektov. Tento lokálny modul však dopĺňa predstavený globálny transformer, ktorý sa sústredí na učenie reprezentácií citlivých na kontext na úrovni scény. Autori štúdie tiež predstavujú efektívny modul na korekciu súradníc, ktorý výrazne zlepšuje generovanie návrhov objektov, čo aj demonštrujú na dátových sadách z interiérov a exteriérov.

Štúdia od Qi a jeho kolegov [48] sa zameriava na nevyužité možnosti 3D detekcie objektov mimo reálneho času. Konkrétne sa venujú automatickému generovaniu kvalitných 3D anotácií. Ich navrhovaný detekčný systém využíva sekvenciu point cloudových dát a kombinuje viacero snímok pre zachytenie doplnkových (komplementárnych) pohľadov na objekty. S využitím takéhoto prístupu dokázali dosiahnuť výsledky zrovnateľné s ľudským označovaním, čo podnecuje vývoj nových využití pre strojové učenie s čiastočným dohľadom.

Tian a jeho kolegovia v štúdií [59] predstavili prvú praktickú metódu detekcie objektov s využitím učenia bez učiteľa. Ich prístup využíva dáta z LiDARových senzorov na riešenie problémov spojených s nejednoznačnými a nespoľahlivými hranicami objektov odvodenými z 2D obrazu. V prvom kroku tejto metódy sa najskôr z 3D point cloudových dát vygenerujú kandidátske segmenty objektov a následne sa začne iteratívny proces označovania takýchto segmentov. Cieľom tohto procesu je natrénovať sieť na označovanie segmentov, ktorá pracuje s príznakmi z 2D obrázkov a zároveň aj z 3D point cloudových dát. Výsledky ich experimentov na rozsiahlej dátovej sade Waymo Open naznačujú, že aj nesupervizované metódy detekcie objektov môžu dosahovať uspokojivé výsledky. Ich zistenia poskytli nové perspektívy pre výskum algoritmov zameraných na analýzu a spracovanie 3D prostredí, najmä v prípadoch, kde je použitie tradičných učebných techník pod dohľadom náročné.

Senzory LiDAR na generovanie point cloudových dát sú však drahé a produkujú riedky a komplikovaný výstup, ktorý vyžaduje veľa ďalšieho predspracovania. V posledých rokoch došlo k rýchlemu zlepšeniu a zvýšeniu dostupnosti cenovo dostupných komerčných hĺbkových senzorov. Ich výstupom sú často hĺbkové mapy, ktorých spracovanie je jednoduchšie.

Detekcia objektov v RGB-D obrazoch

Druhý prístup spracovania hĺbkových dát sa zameriava na použitie ľahšie spracovateľných hĺbkových máp namiesto zložitých point cloudových dát. Hĺbkové mapy dopĺňajú detekciu objektov založenú iba na RGB viacerými spôsobmi. Hĺbkové obrazy lepšie vykresľujú hranice objektov, čo uľahčuje lokalizáciu objektov a ich správne pokrytie ohraničujúcimi rámčekmi. Toto je dôležité najmä v prípadoch, keď hranice objektov nie sú na farebných snímkach jasné v dôsledku slabého osvetlenia alebo silných tieňov. Hĺbkové mapy dokážu tiež vyriešiť skreslenie mierky, ktoré sa často objavuje vo farebných obrazoch v dôsledku perspektívnych projekcií.

Aj keď existuje mnoho štúdií zameraných na využitie hĺbky pri detekcii 3D objektov [51], [37], [7], [54], len niekoľko sa pokúsilo integrovať hĺbku do hlbokých detektorov objektov [3], [18], [22] pre detekciu 2D objektov. Práca Gupta et al. [18] bola jednou z prvých, ktorá sa o to pokúsila. Využili informácie o hĺbke na zlepšenie procesu návrhu regiónov, ako aj procesu klasifikácie. Konkrétne použili RGB-D obrázky na výpočet 2,5D kontúr, z ktorých identifikovali oblasti, kde pravdepodobne sa nachádzajú objekty. Pre klasifikáciu objektov natrénovali dva modely konvolučnej siete na extrakciu príznakov. Jeden model získaval príznaky z RGB informácií a druhý z hĺbkových informácií. Následne trénovali lineárny SVM na klasifikáciu týchto príznakov do objektov. Ďalším významným prínosom bolo použitie kódovania HHA ako vstupné údaje do modelu konvolučnej siete pre extrakciu príznakov. Ukázali, že tento prístup je účinnejší ako priamy vstup hĺbky.

Cao a jeho kolegovia [3] demonštrovali, že hĺbka odhadnutá z RGB obrazu môže výrazne zlepšiť detekciu objektov bez potreby hĺbkového senzora. Na odhad hĺbky scény z RGB obrazu použili podmienené náhodné pole (angl. Conditional Random Field) a následne trénovali dve nezávislé konvolučné siete na klasifikáciu oblastí pomocou RGB vstupov a odhadnutej hĺbky. Na kódovanie informácií o hĺbke priamo použili logaritmus odhadnutej hĺbky a nevyužili povrchové normály alebo kódovanie HHA, využité v práci [18], pričom argumentovali, že tieto signály nie sú dostatočne informatívne kvôli približnej a zašumenej povahe odhadnutej hĺbky a potrebe informácií o kamere.

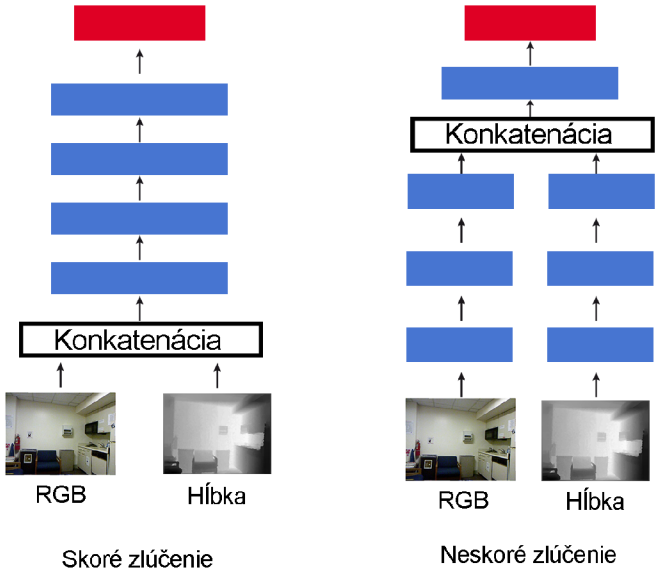
Hou jeho kolegovia. [22] poskytli podrobnú analýzu rôznych spôsobov integrácie informácií o hĺbke do hlbokých detektorov objektov. Skúmali dôležitosť rôznych vizuálnych vstupov (RGB, hĺbka, uhol, výška, kontúra atď.) a rôzne úrovne (úroveň vstupu vs. úroveň príznaku) pre kombináciu RGB a hĺbkových informácií pri detekcii objektov. Argumentovali a preukázali, že kombinácia farby a hĺbky na úrovni vstupu nie je efektívna kvôli odlišnej povahe informácií a spracovanie HHA dát samostatne pomocou oddelených sietí funguje lepšie ako spoločné spracovanie týchto informácií jednou sieťou.

Zlúčenie farebného obrazu a hĺbky

Pri detekcií objektov v RGB-D priestore je nutné efektívne využívanie všetkých druhov dát, ktoré tieto obrázky poskytujú - konkrétne farbu a hĺbku. Tieto informácie sa navzájom dopĺňajú, avšak ich správne využitie vyžaduje sofistikované metódy ich zlúčenia/fúzie.

Skoré zlúčenie (angl. early fusion alebo zlúčenie na vstupnej vrstve) predstavuje techniku, pri ktorej sa RGB obrázok a hĺbková mapa spoja do jedného štvorkanálového obrazu už v úvodnej fáze spracovania dát neurónovou sieťou, teda ešte pred akoukoľvek základnou výpočtovou operáciou [19]. Tento zlúčený obraz sa následne spracováva pomocou 2D alebo 3D konvolučných filtrov. Metóda skorej fúzie bola využitá napríklad pri detekcií najvýraznejšieho objektu v obraze v štúdií [21].

Neskoré zlúčenie (angl. late fusion alebo zlúčenie na koncovej vrstve) sa líši tým, že RGB obrázok a hĺbková mapa sa spracovávajú nezávisle, často s využitím dvoch rozdielnych neurónových sietí. Výsledné príznaky z oboch neurónových sietí sa neskôr kombinujú buď spojením, alebo ďalším spracovaním prostredníctvom konvolučných sietí. Eitel a spol. [13] využili práve tento prístup, čím dosiahli zvýšenú presnosť na ich skúmanej dátovej sade.



Obr. 2.6: Porovnanie metód skorého a neskorého zlúčenia.

Hlboké zlúčenie (angl. deep fusion) prekonáva obmedzenia skorej a neskorej fúzie tým, že umožňuje finálnym predpovediam využívať buď skoré alebo hlboké reprezentácie. Zabráni sa tým strate dôležitých informácií. Jedná sa o metódu, ktorá kombinuje príznaky získané z viacerých reprezentácií vstupu, pričom namiesto jednoduchého spojenia (konkatenácie) využíva tzv. priemerovanie po prvkoch (angl. element-wise mean pooling), čím dosahuje lepšie výsledky [7].

Doplnkové zlúčenie (angl. complementarity-aware fusion) predstavuje najsofistikovanejší spôsob fúzie. Metódy zlučovania popísané vyššie sa buď učia vlastnosti z farebných a hĺbkových dát samostatne, alebo jednoducho spracúvajú RGB-D ako štvorkanálové dáta. Wang a kolegovia [60] predpokladajú, že rôzne modality by mali obsahovať nielen niektoré dátovo špecifické vzory, ale aj nejaké spoločné vzory. Navrhujú multimodálny rámec učenia vlastností pre rozpoznávanie objektov v RGB-D. Najprv sa vytvoria dve konvolučné neurónové siete, jedna pre farbu a druhá pre hĺbku. Následne sú tieto siete prepojené tzv. multimodálnymi vrstvami, ktoré spájajú informácie o farbe a hĺbke vynútením spoločnej časti, ktorú majú zdieľať príznaky rôznych modalít. Takýto prístup vytvára príznaky odrážajúce spoločné vlastnosti, ako aj vlastnosti špecifické pre rôzne modality.

Cheng a spol. [5] navrhli mechanizmus zlučovania nazvaný zohľadnenie komplementárnosti (CA), ktorý podporuje určovanie doplňujúcich informácií z rôznych modalít na rôznych úrovniach abstrakcie. Zaviedli modul CAFuse, ktorý umožňuje krížové prepojenia medzi modalitami a úrovňami a modálnymi/úrovňovými dohľadmi, explicitne podporujúce zachytenie doplňujúcich informácií od partnera, čím sa znižuje nejednoznačnosť zlučovania a zvyšuje sa účinnosť zlučovania.

Metódy reprezentácie hĺbky v hĺbkových mapách

Existuje mnoho metód, ako najlepšie reprezentovať hĺbku v RGB priestore, každá s vlastnými výhodami a nevýhodami. Efektívne metódy reprezentácie hĺbky sú dôležité, pretože umožňujú neurónovým sieťam efektívnejšie spracovávať a interpretovať priestorové informácie. Nasledujúci text sa preto zameriava na niektoré z týchto metód.

Odtiene šedi

Kódovanie šedou škálou je jednoduchá metóda, pri ktorej sa hodnoty hĺbky premieňajú na intenzity šedej škály. Tento proces je definovaný vzorcom 2.1, kde $d_{\min}$ a $d_{\max}$ sú najnižšia a najvyššia hodnota hĺbky pozorovaná vo všetkých scénach dátovej sady. Aj keď táto metóda zjednodušuje údaje na jeden kanál, zachováva dôležité prvky ako sú hrany a rohy. Tieto prvky sú dôležité pre identifikáciu obrysov objektov a ďalších priestorových charakteristík a umožňujú neurónovým sieťam efektívnejšie spracovanie údajov.

$$g(d) = 255\frac{d - d_{\min}}{d_{\max} - d_{\min}} \tag{2.1}$$

Farebná paleta jet

Pri kódovaní farebným priestorom Jet sa šedé hodnoty mapujú na farebnú mapu Jet, ktorá sa rozkladá od červenej (pre bližšie objekty) cez zelenú až po modrú (pre vzdialenejšie objekty). Tento prístup je veľmi efektívny, pretože farebný gradient robí rozdiely v hĺbke výraznejšími, čo zlepšuje schopnosť siete rozlišovať jemné variácie v hĺbke. Výskumy [13], [57] ukázali, že kódovanie v tomto farebnom priestore môže zlepšiť výsledky pri rozpoznávaní 3D objektov, pretože poskytuje bohatšie a detailnejšie vizuálne znázornenie údajov o hĺbke.

Povrchové normály

Pri spracovaní RGB-D obrazu sa na reprezentáciu hĺbkových dát využívajú aj normálové vektory. Tieto lokálne deskriptory povrchu poskytujú dôležité informácie o orientácii a tvare objektov v scéne a často sa využívajú ako vstupy pre viaceré úlohy počítačového videnia, vrátane rekonštrukcie povrchu [32], geometrickej registrácie [47] a segmentácie [16]. Sada povrchových normál obsahuje podstatné informácie o tvare objektu, pričom poskytuje lepší popis tvaru ako len samotné gradienty hĺbky. Štúdia [1] tvrdí, že použitie farebne kódovaných povrchových normál lepšie zachytáva štrukturálne informácie a jemné detaily objektov ako metóda farebnej palety Jet. Získanie týchto normál môže byť realizované priamo z údajov získaných hĺbkovými senzormi (napr. Kinect) alebo ich možno odhadnúť pomocou pokročilých algoritmov spracovania obrazu a hlbokého učenia. Podľa existujúcich štúdií [64], [18], [50] integrácia normál povrchu do učebných procesov modelov môže zvýšiť ich výkonnosť, čo je spôsobené lepším rozpoznávaním geometrických detailov a zlepšenou robustnosťou voči šumom a variabilitám v dátach.

Príkladom využitia normál povrchu v praxi je štúdia od Ľubora Ladického a jeho kolegov [33], ktorá bola realizovaná na dátovej sade NYUv2. Táto štúdia demonštrovala, ako spracovanie normál povrchu môže prispieť k lepšiemu porozumeniu trojrozmerných scén a zvýšeniu presnosti detekcie objektov v rôznych prostrediach.

Kódovanie HHA

HHA alebo aj HDHA kódovanie [18] je komplexnejší prístup, ktorý do kódovania hĺbky prináša tri rozdielne aspekty: horizontálny rozdiel (HD), výšku od zeme (H) a uhol voči smeru gravitácie (A). Táto metóda nielenže uchováva pôvodné informácie o hĺbke, ale pridáva aj dôležité kontextové detaily, ktoré pomáhajú pochopiť trojrozmernú štruktúru scény.

Horizontálny rozdiel (HD) sa formálne vypočíta ako:

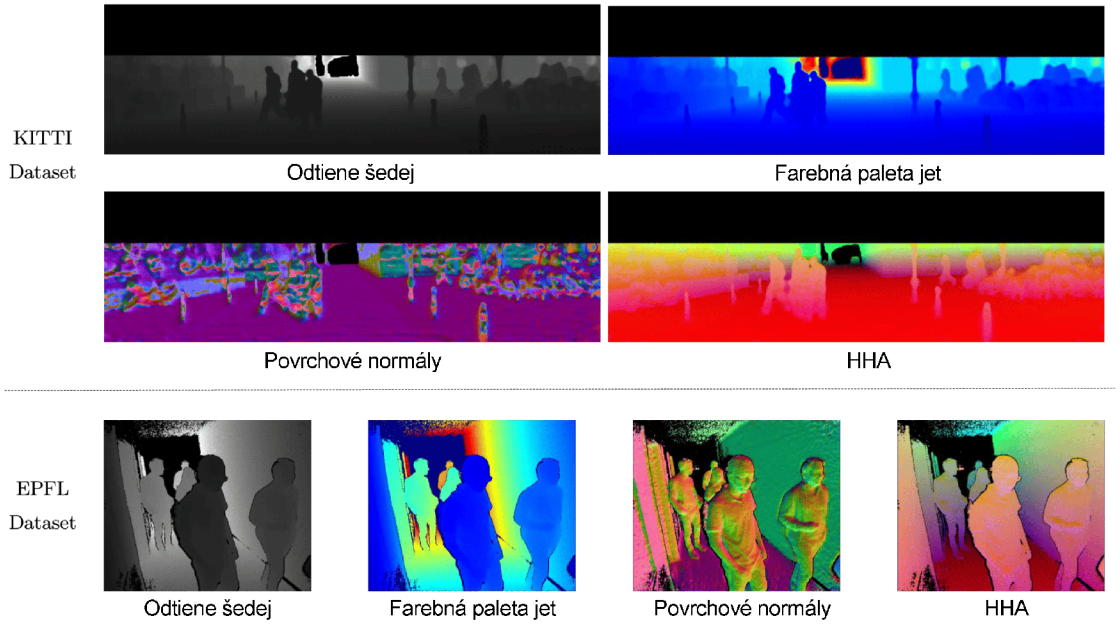
$$HD = p_x - p_x^c, \tag{2.2}$$

kde p_x je súradnica x pixelu p a p^c je zodpovedajúci pixel v pohľade z kamery c. Tento parameter predstavuje relatívnu vzdialenosť rôznych častí objektu alebo scény od kamery.

Výška (H) je kvôli zložitosti získania presných informácií o polohe zeme často meraná od najnižšieho bodu v scéne namiesto priameho merania od zeme [17]. Tento tento parameter umožňuje relatívne určenie výšky objektov v scéne.

Na výpočet uhla s gravitáciou (A) sa používa iteratívny postup, podobný tomu v [17]. Za počiatočný odhad smeru gravitácie sa považuje vertikálna os, vzhľadom na ktorú sa všetky normály povrchu zoskupia do zhlukov, ktoré sú na túto os buď približne rovnobežné alebo približne kolmé (ortogonálne). Po zhlukovaní sa nový smer gravitácie odhadne vzhľadom na tieto rovnobežné a ortogonálne zhluky. Tieto kroky sa opakujú s cieľom minimalizovať odchýlku medzi predpokladaným smerom gravitácie a skutočnou orientáciou povrchov.

Takýto viacdimenzionálny prístup umožňuje neurónovým sieťam lepšie chápať a využívať priestorové štruktúry podobne ako ľudské vizuálne vnímanie, čo výrazne zlepšuje detekciu príznakov v hĺbkových mapách.



Obr. 2.7: Porovnanie metód reprezentácie hĺbky v hĺbkových mapách.

Výzvy pri spracovaní hĺbkových dát

Detekcia objektov s využitím hĺbkových dát prináša so sebou rad výziev, ktoré sú dôležité pre správne porozumenie a spracovanie scény. Vzhľadom na rôznorodosť prostredí a pod-

mienok je potrebné vyvinúť robustné metódy, ktoré dokážu účinne detekovať objekty bez ohľadu na okolité faktory. Nasledujúca časť preto približuje niektoré z týchto výziev.

Technologické obmedzenia senzorov ako napríklad rozsah hĺbkového senzora môžu zásadne obmedziť efektivitu detekčných systémov. Senzory ako Microsoft Kinect majú presnosť merania do vzdialenosti približne 4,5 metra [57], čo predstavuje problém pri získavaní dát vo vonkajšom prostredí, kde sú často potrebné drahšie a sofistikovanejšie zariadenia ako sú napríklad laserové skenery [42].

Neúplnosť dát predstavuje ďalšiu výzvu. Snímače RGB-D môžu generovať dáta, ktoré sú v niektorých oblastiach, najmä v hĺbkových mapách, neúplné alebo riedke. Na doplnenie neúplných dát o hĺbke sa často využívajú rôzne druhy filtrovania. Detekčné systémy by mali byť schopné detekovať celý objekt aj keď nie sú dostupné všetky jeho časti.

Prekrývanie objektov je častý problém hlavne pri dynamických scénach vo vnútorných ale aj vonkajších prostrediach. Zrakové systémy, či už ľudské alebo zvieracie, sa s týmito situáciami vyrovnávajú veľmi dobre. Detekčné algoritmy často musia objekty nielen lokalizovať v trojrozmernom priestore, ale tiež odhadnúť ich polohu alebo veľkosť, aj keď sú viditeľné len čiastočne. To predstavuje problém, pretože nedostatočne viditeľné časti objektov a rušivé elementy na pozadí môžu často takéto algoritmy zmiasť a viesť k nesprávnym výsledkom.

Osvetlenie tiež predstavuje jednu z výziev pri detekcií objektov v RGB-D, najmä v prostrediach, kde dochádza k jeho častej zmene. Napríklad autonómne riadené drony a domáce interiérové roboty musia fungovať počas dňa aj noci a preto je pravdepodobné, že sa stretnú s extrémnymi hodnotami osvetlenia prostredia. Vzhľad objektov môže byť výrazne ovplyvnený svetelnými podmienkami nielen v RGB obraze, ale aj v hĺbkovej mape, v závislosti od typu 3D senzorov použitých na získavanie dát. Detekčné systémy by mali byť voči týmto zmenám robustné, aby dokázali správne interpretovať dáta bez ohľadu na svetelné podmienky.

Nároky na výpočet RGB-D algoritmov sú v porovnaní s tradičnými 2D metódami výrazne vyššie, nakoľko pridanie ďalšej priestorovej dimenzie podstatne zväčšuje objem dát, ktoré sa musia spracovať. Mnohé praktické využitia, napríklad autonómne riadenie, vyžadujú detekciu objektov v reálnom čase. Napriek hardvérovej akcelerácii s použitím GPU sú algoritmy detekcie založené na RGB-D pomalšie v porovnaní s ich 2D náprotivkami. Dokonca aj techniky, ktoré sú veľmi efektívne pri RGB obrázkoch ako sú posuvné okná a konvolučné operácie sa stávajú výrazne nákladnejšími z hľadiska výpočtového času a pamäťového priestoru.

Trénovacie dáta obsahujúce hĺbkovú zložku je v porovnaní s čisto RGB dátovými sadami náročnejšie získať a to aj napriek čoraz rozšírenejšej dostupnosti RGB-D senzorov. Hoci sú v súčasnosti dostupné aj nízkonákladové senzory, ako napríklad Microsoft Kinect, tie sú zvyčajne efektívnejšie pri použití vo vnútorných prostrediach. V dôsledku týchto obmedzení je možné pozorovať trend, že dátové sady z exteriérového prostredia sú menej časté a často tiež majú menší rozsah.

Okrem týchto výziev musia byť detekčné algoritmy navrhnuté tak, aby zvládli rozpoznať aj rôzne deformované objekty. Je tiež dôležité aby algoritmy vedeli rozlíšiť objekty, ktoré majú veľké variácie v rámci jednej triedy. Takéto variácie sú znázornené na obrázku. Zároveň musia rozlišovať aj medzi triedami, ktoré sa môžu navzájom výrazne podobať. Vysoká podobnosť medzi rôznymi triedami môže negatívne ovplyvniť efektivitu detekčných algoritmov, najmä ak je na tréning použitých len málo RGB-D obrázkov. Pár príkladov podobností medzi triedami, variácií v rámci jednej triedy a niektorých ďalších výziev pri detekcií objektov je možné vidieť na obrázku 2.8.

Obr. 2.8: Niektoré výzvy pri detekcií objektov.

2.4 Detektory objektov

Jednokrokové detektory

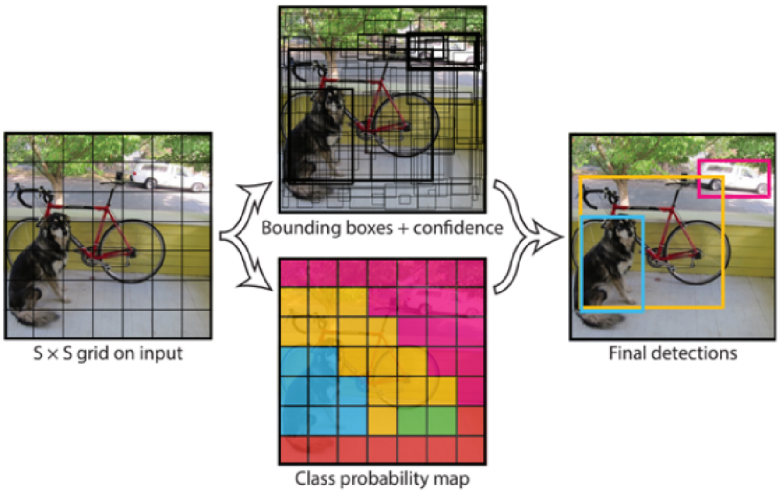
Jednokrokové detektory sú výkonné nástroje na detekciu a klasifikáciu objektov v obrazových dátach. Tieto systémy fungujú tak, že pri jednom prechode obrazom dokážu identifikovať a lokalizovať objekty pomocou priamo mapovaných pixelov na súradnice obmedzujúcich rámcov a pridelením pravdepodobností k príslušným kategóriám. Vďaka tomu, že sú založené na regresnej analýze, umožňujú efektívne spracovanie bez potreby viacstupňového trénovania, ako je tomu u regiónovo orientovaných detektorov.

Oproti viackrokovým detektorom, ktoré vyžadujú oddelené tréningové fázy pre rôzne časti modelu, jednokrokové detektory umožňujú trénovanie modelu naraz. Tento prístup značne zjednodušuje proces vývoja a zvyšuje efektivitu trénovania. S ich schopnosťou rýchlo spracovávať obrazové dáta sú jednokrokové detektory ideálne pre použitie v situáciách, kde je dôležitá rýchlosť, napríklad pri monitorovaní a rozpoznávaní v reálnom čase. Aj keď sú jednokrokové detektory výpočtovo menej náročné a poskytujú porovnateľnú presnosť so systémami založenými na regiónoch, majú tendenciu byť menej presné pri detekcii malých objektov.

YOLO – You Only Look Once

Architektúra YOLO [52] je založená na princípe získania všetkých potrebných informácií (trieda objektu a jeho umiestnenie) v rámci jedného vyhodnotenia, čo je základom jej názvu. Vznikla v roku 2016 s cieľom presnej detekcie objektov v reálnom čase. Veľkou výhodou tejto architektúry je, že sa učí veľmi všeobecné reprezentácie objektov, takže je schopná relatívne presnej detekcie aj v prípade, keď testovacie obrázky sa výrazne líšia od trénovacích (napríklad obrázky z prírody a umelecké diela). Architektúra YOLO modelov využíva pre predikciu každého ohraničujúceho rámu príznaky z celého obrázka a zároveň ich predikuje pre všetky triedy objektov súčasne.

Proces tejto detekcie začína rozdelením vstupného obrázka podľa mriežky veľkosti $S \times S$. Každá bunka mriežky sa potom stará o detekciu objektov, ktorých stred leží v danej bunke. Pre každú bunku je potom predpovedaných B ohraničujúcich rámov obsahujúcich polohu stredov objektov, ich výšku a šírku vzhľadom na celý obrázok, a pravdepodobnosť, že sa tam nachádza nejaký objekt. Súčasne každá bunka vypočíta vektor C pravdepodobností, že daná bunka obsahuje daný objekt, pričom hodnota C zodpovedá počtu tried, pre ktoré je daný model trénovaný. Tento vektor je vždy jeden nezávisle na hodnote B. Všetky predikcie sú potom uložené ako tenzor veľkosti $S \times S \times (B * 5 + C)$ [52]. Zjednodušený princíp detekcie je znázornený na obrázku 2.9.



Obr. 2.9: Zjednodušený princíp detekcie objektov architektúrou YOLO. Prevzaté z [52].

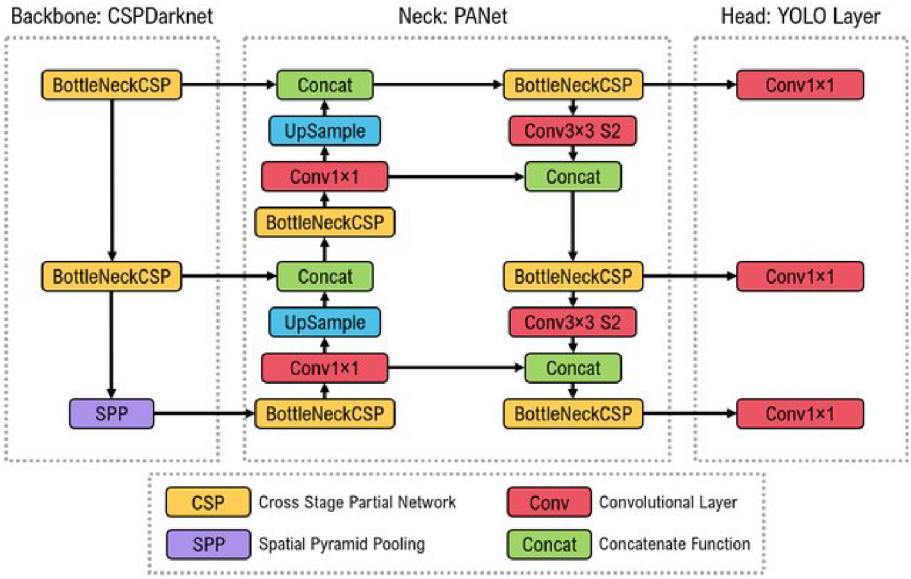
Nevýhodou tohto prístupu je obmedzenie počtu objektov, ktoré je táto architektúra schopná detekovať, najmä ak sa vyskytujú blízko seba. Okrem toho model veľmi zle detekuje malé objekty. Z týchto dôvodov nie je model vhodný pre detekciu niektorých typov objektov, ako sú napríklad hejna vtákov. Ďalšou nevýhodou je, že táto architektúra má problémy s detekciou objektov, ak sa vyskytujú v odlišných pomeroch strán.

Architektúra siete sama osebe je založená na jednej konvolučnej sieti, nie je teda rozdelená na viacero podsietí. Je inšpirovaná architektúrou GoogLeNet pre klasifikáciu objektov.

YOLOv5

YOLOv5 je moderná (tzv. state-of-the-art) sieť na detekciu objektov. Je open source, vyvíjaná a spravovaná spoločnosťou Ultralytics, pričom väčšina príspevkov je pripísaná Glennovi Jocherovi. Sieť dosahuje výborné výsledky na Microsoft COCO datasete, čo sa týka presnosti aj rýchlosti inferencie [28]. Architektúra tejto siete v piatej iterácii je podobná architektúre použitej v YOLOv4 [2], avšak obsahuje aj niektoré zmeny, ktoré zlepšujú a výraznejšie zjednodušujú proces trénovania. Ďalšou významnou odlišnosťou je, že YOLOv5 je implementované v Pythone pomocou PyTorch, zatiaľ čo YOLOv4 je implementované v jazyku C s použitím rámca Darknet.

Architektúra YOLOv5 sa skladá z troch hlavných častí. Základnú kostru (angl. backbone) tvorí CSPDarknet53 [2], ktorý je známy svojou rýchlosťou a presnosťou. Kostra zabezpečuje získavanie základných príznakov ako sú hrany, farby a podobne z obrázkov. Navyše sa využíva Spatial Pyramid Pooling (SPP), ktorý zdôrazňuje najdôležitejšie vlastnosti takmer bez dodatočnej spotreby času. Časť nazývaná krk (angl. neck) obsahuje sieť Path Aggregation Network (PANet) [2], kde sa ďalej spracovávajú informácie z detekovaných príznakov. Medzi kostrou a krkom sú tiež tzv. skokové cesty, konkrétne Cross Stage Partial connections (CSP). Vrcholová časť, alebo hlava (angl. head), využíva architektúru založenú na kotvách, ako bola použitá už v YOLOv3 [14]. V tejto časti sú detekované objekty v obrázku reprezentované kombináciou označení tried, ohraničujúcich rámčekov a skóre dôveryhodností. Aktuálne je vyvinutých hneď päť variantov YOLOv5, ktoré sa líšia podľa veľkosti a s tým súvisiacej rýchlosti a presnosti.



Obr. 2.10: Architektúra YOLOv5. Prevzaté z [31].

YOLOv8

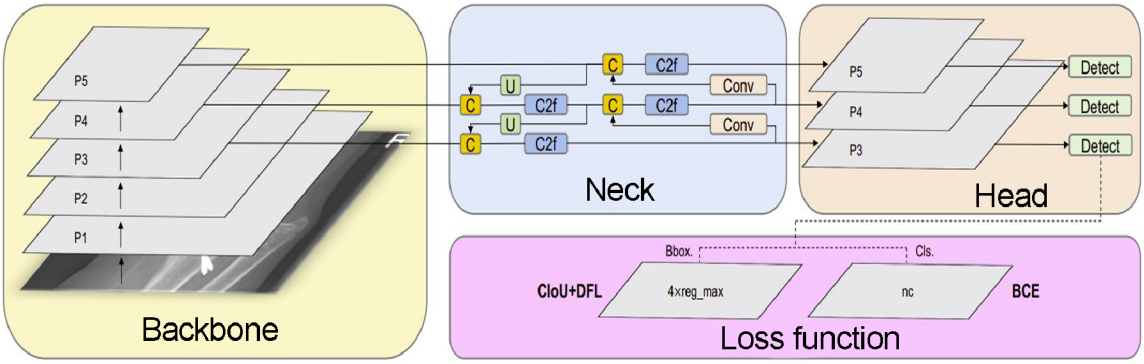
V januári 2023 sa spoločnosť Ultralytics, ktorá stála za vývojom úspešného YOLOv5 rozhodla vydať novú verziu svojho modelu pre detekciu objektov v obraze YOLOv8 [29]. YOLOv8 sa drží súčasného trendu a je bez kotiev (angl. anchors), čo znižuje počet potrebných predpovedí rámčekov a tým zrýchľuje proces post-processingu, konkrétne zrýchľuje algoritmus NMS (Non-Maximum Suppression). Model YOLOv8 je možné spustiť pomocou

rozhrania príkazového riadka (CLI) alebo môže byť nainštalovaný ako balík PIP. Navyše, model YOLOv8 je dodávaný s viacerými integráciami na označovanie, trénovanie a nasadenie, čo uľahčuje jeho použitie v praxi. V nasledujúcej časti je opísaná architektúra tohto modelu, ktorá je zobrazená na obrázku 2.11.

Model YOLOv8 používa techniku CSP [62] na rozdelenie mapy príznakov na dve časti, pričom jedna časť využíva konvolučné operácie a druhá je spojená s výstupom konvolučných operácií predchádzajúcej časti. Tým sa zlepšuje schopnosť modelu učiť sa a znižuje sa jeho výpočtová náročnosť. Namiesto modulu C3 použitého v YOLOv5 používa YOLOv8 modul C2f, čo umožňuje modelu získať bohatšie informácie o zmene gradientov. YOLOv8 tiež znižuje počet blokov v každej časti a na zvýšenie rýchlosti inferencie modelu využíva modul SPPF [20].

Vo všeobecnosti hlbšie neurónové siete poskytujú viac informácií a vedú k lepším výsledkom. V prípade malých objektov však môže príliš veľa konvolučných operácií viesť k strate informácií. Na vyriešenie tohto problému je potrebná tzv. viacškálová fúzia príznakov [23] pomocou architektúr FPN [38] a PAN [40].

Na rozdiel od jeho predchodcov, ktorí používajú spojenú hlavovú časť, používa YOLOv8 tzv. oddelenú hlavu (angl. decoupled head), ktorá oddeľuje klasifikačnú a detekčnú časť. YOLOv8 nahrádza prístup založený na kotvách tzv. Anchor-Free prístupom, ktorý tieto kotvy nepotrebuje a objekt lokalizuje podľa jeho stredu tým, že predpovedá vzdialenosť od jeho stredu k ohraničujúcemu rámčeku.



Obr. 2.11: Architektúra modelu YOLOv8, upravené z [30].

Kapitola 3

Návrh riešenia

Táto kapitola sa zaoberá metódami pre detekciu objektov v RGBD obrázkoch využitými v experimentálnej časti. Zameriava sa na dátové sady, metriky, implementácie algoritmov a upravené modely, ktoré sú základom pre naše experimenty a vývoj v oblasti detekcie objektov v RGBD obrazoch. Konkrétne sú predstavené dve dátové sady, NYU Depth Dataset V2 a Washington RGB-D, ktoré sú zdrojom hĺbkových dát potrebných pre výskum. Popísané sú zvolené upravené modely YOLOv5 a tiež vykonané úpravy modelu YOLOv8, ktoré boli potrebné na to aby bol schopný pracovať s hĺbkovými dátami. Predstavené sú metriky na porovnávanie efektívnosti detekčných metód, ktoré umožňujú objektívne vyhodnotiť účinnosť navrhnutých riešení. V závere kapitoly je priblížené aj tréningové prostredie, v ktorom sa uskutočňuje vývoj a validácia modelov.

3.1 Dátové sady

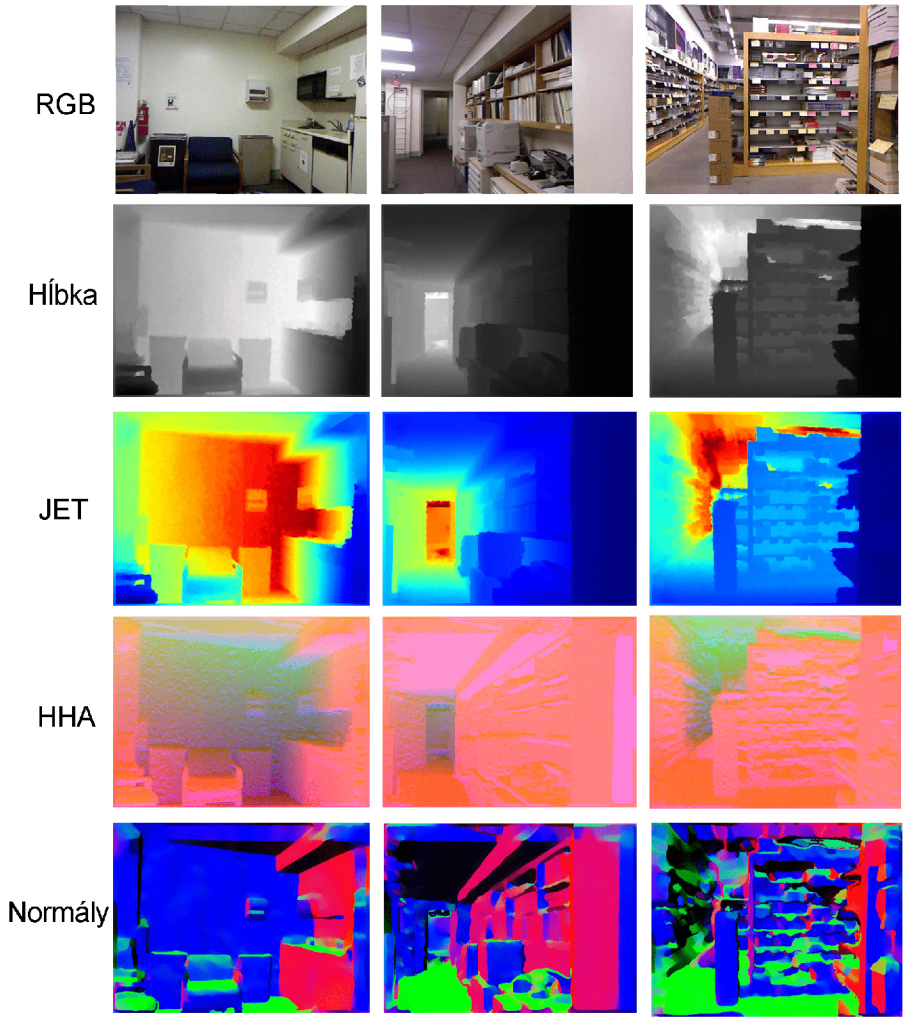
Nasledujúca časť sa zameriava na dôležitosť dátových sád NYU Depth Dataset V2 a Washington RGB-D v kontexte počítačového videnia. Obidve sady poskytujú cenné údaje pre vývoj a validáciu algoritmov zameraných na spracovanie interiérových scén s využitím hĺbkových dát. Podrobne predstavená je ich štruktúra, obsah a význam pre pokrok v technológiách spracovania hĺbkových údajov.

NYU Depth Dataset V2

Dátová sada NYU Depth Dataset V2 [56] bola vytvorená laboratóriom Vision Learning Graphics (VLG) na Newyorskej univerzite a priniesla významný pokrok v oblasti počítačového videnia, konkrétne pri spracovaní a analýze interiérových scén s využitím hĺbkových dát. Táto dátová sada je rozšírením predchádzajúcej sady NYU Depth V1, ktorá bola pôvodne vytvorená na segmentáciu objektov vo vnútorných scénach. Dátová sada V2 bola publikovaná v roku 2012 a významne rozšírila rozsah dát svojho predchodcu. Nová dátová sada obsahuje približne 408 000 RGB obrázkov na ktorých sa nachádza 894 rôznych tried a približne 35 064 objektov. Jej súčasťou je aj označená podmnožina, ktorá obsahuje 1449 zarovnaných RGB-D obrázkov s podrobnými anotáciami z 464 vnútorných scén v 26 rôznych prostrediach. Ukážku tejto označenej podmnožiny v rôznych hĺbkových reprezentáciách je možné nájsť na obrázku 3.1. RGB-D obrázky boli zozbierané z mnohých budov v troch mestách USA. Rozmanitosť zachytených scén umožnila vývoj robustných algoritmov, ktoré sú prispôsobiteľné rôznym kontextom v interiéri. Každý obrázok je anotovaný presnými označeniami viacerých tried, ktoré rozlišujú rôzne objekty a prvky v rámci scény.

V porovnaní s RGB-D dátovými sadami, ktoré boli vytvorené pred vynálezom nízkonákladových RGB-D kamier (napr. Kinect), mali NYU Depth V1 a V2 výrazne viac kategórií a dát, čo pomohlo a stále pomáha výskumníkom preskúmať sofistikovanejšie a efektívnejšie algoritmy. Označená podmnožina uľahčuje výskum algoritmov strojového učenia aj tým, že poskytuje predspracované obrázky, na ktorých boli doplnené chýbajúce hodnoty hĺbky, ktoré sa inak často objavujú v hĺbkových obrázkoch ako čierne body. Vďaka takémuto predspracovaniu je možné túto dátovú sadu využiť aj na trénovanie modelov, ktoré vyžadujú konzistentnú kvalitu vstupných údajov.

NYU Depth Dataset V2 sa aj vďaka významným štúdiám rýchlo stal referenčným štandardom pre hodnotenie metód založených na dátach RGB-D scén [61], [27]. Podporuje širokú škálu úloh počítačového videnia, vrátane odhadu hĺbky, sémantickej segmentácie a 3D rekonštrukcie. Rozmanitosť a komplexnosť tejto dátovej sady vyzýva existujúce metodiky a podnecuje vývoj sofistikovanejších algoritmov schopných pochopiť a interpretovať zložité vnútorné prostredia s využitím dát o hĺbke. Viaceré významné štúdie [18], [4], [9], [58] využívajú na porovnanie výsledkov 19 najčastejšie vyskytujúcich sa tried tejto dátovej sady a preto sú presne tieto triedy použité aj v tejto práci. Pri experimentoch s touto dátovou sadou boli použité dáta obsahujúce povrchové normály, ktoré boli súčasťou práce [33].



Obr. 3.1: Ukážka dátovej sady NYU Depth Dataset V2 v rôznych reprezentáciách hĺbky.

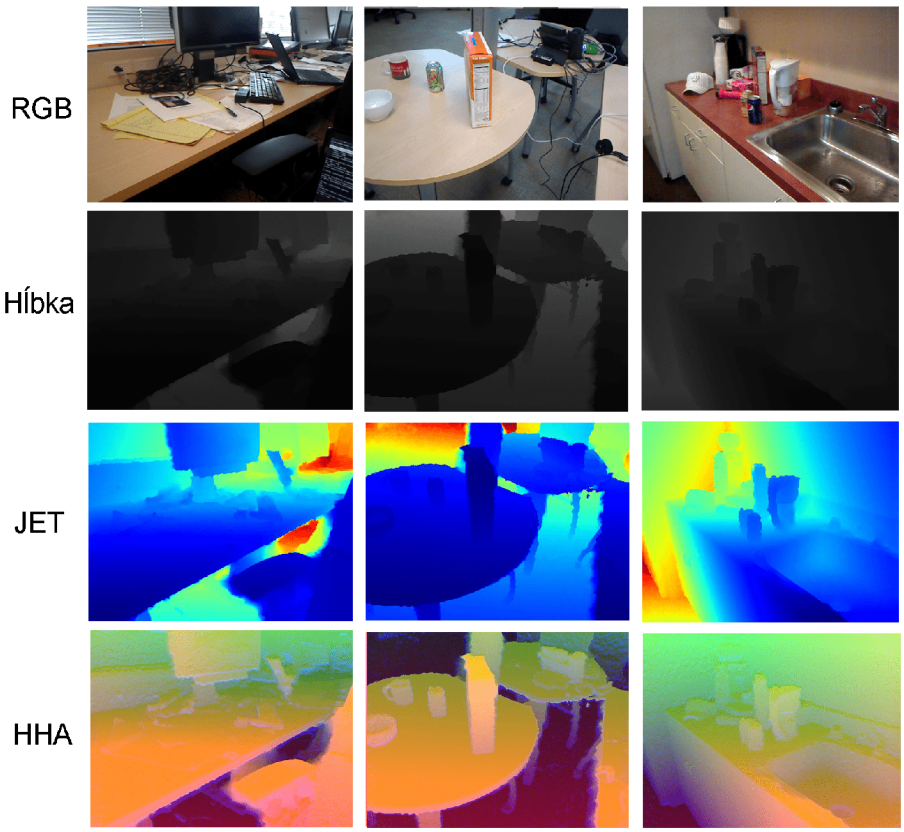
Washington RGB-D

Dátová sada Washington RGB-D [35], bola predstavená verejnosti v roku 2011 Univerzitou vo Washingtone, konkrétne jej laboratóriom Intel Labs. Dátová sada Washington RGB-D je rozdelená na dve časti. Prvá časť má názov RGB-D Objects a obsahuje orezané obrázky 300 objektov zaradených do 51 kategórií. Druhú časť predstavuje sada RGB-D Scenes, ktorá je využitá v tejto práci a obsahuje osem videosekvencií zobrazujúcich viaceré scény. Ukážku tejto časti dátovej sady je možné vidieť na obrázku 3.2. V týchto scénach sa nachádzajú objekty uvedené v prvej časti, pričom sú zachytené z rozličných perspektív a často sú čiastočne alebo úplne zakryté.

Scény v dátovej sade pokrývajú kancelárske priestory, zasadacie miestnosti, obývaciu izbu, kaviareň, kuchyňu a obsahujú objekty v 6 kategóriách. Videosekvencie boli nahrávané držaním RGB-D kamery približne na úrovni ľudského oka a prechádzaním po jednotlivých scénach. Každá scéna je reprezentovaná point cloudom, vytvoreným zarovnaním sady snímok z Kinectu pomocou pokročilých techník RGB-D mapovania.

Hoci táto dátová sada poskytuje podrobné anotácie, zaznamenané boli aj niektoré jej obmedzenia. V rámci označených obrázkov scén nie sú označené všetky objekty, ktoré sa tam vyskytujú a ohraničujúce boxy občas orezávajú niektoré objekty. Svoje nedostatky majú aj dostupné hĺbkové dáta, najmä ide o ich občasnú neúplnosť a nesúvislosti na povrchoch objektov, čo môže v niektorých prípadoch viesť k nadmernej segmentácií. Niektoré z týchto problémov je však možné zmierniť rôznymi technikami počítačového videnia akou je napríklad filtrovanie.

Dátová sada bola významnou súčasťou výskumu [11] na tému RGB-D mapovania prezentovaného na konferencii o robotike a automatizácií ICRA 2012, ktorý sa zameriaval na využitie hĺbkových kamier typu Kinect pre detailné 3D modelovanie vnútorných prostredí. Okrem toho bola využitá vo veľa rôznych štúdiách [34], [49], [65] na porovnávanie a zdokonaľovanie algoritmov súvisiacich s hĺbkovým vnímaním, detekciou objektov a mapovaním prostredia, čo preukazuje jej účinnosť a relevanciu pri vývoji technológií robotického vnímania.



Obr. 3.2: Ukážka dátovej sady RGB-D Objects v rôznych reprezentáciách hĺbky.

S cieľom zlepšiť kvalitu dostupných hĺbkových máp dátovej sady Washington RGB-D bola v tejto práci použitá a upravená technika na doplnenie chýbajúcich hĺbkových údajov, ktorá bola pôvodne vyvinutá pre NYU Depth Dataset V2. Tento postup, ktorý navrhol Levin a jeho kolegovia [36], využíva ako podklad na doplnenie nedostatkov hĺbkových máp ich zodpovedajúce RGB obrazy, konkrétne ich zobrazenie v odtieňoch šedej. Metóda najskôr v takomto šedotónovom obraze analyzuje lokálne podobnosti okolo každého pixelu a na

ich základe následne interpoluje a vyhladí chýbajúce hodnoty hĺbky. Implementácia tejto metódy je voľne dostupná vo viacerých verejných repozitároch[1].

3.2 Zvolené metódy integrácie hĺbkových dát do detekčných algoritmov

Upravený model YOLOv8

Pre účely tejto práce bol prispôsobený model YOLOv8 pre prácu so štvorkanálovými obrazmi, konkrétne s RGBD formátom, ktorý kombinuje tri farebné kanály s jedným hĺbkovým kanálom. Táto modifikácia spočívala predovšetkým v zmenách spôsobu načítania dát. Využila sa funkcia `imread` z knižnice OpenCV, do ktorej bol pridaný parameter `cv2.IMREAD_UNCHANGED`. Tento parameter zabezpečil, že obrazy sú načítané so všetkými štyrmi kanálmi, čo je kľúčové pre správnu funkčnosť modelu. Ďalej bolo potrebné upraviť konfiguračný `.yaml` súbor modelu pridaním parametru `ch: 4`.

Ďalšou zásadnou úpravou bolo obmedzenie alebo úplné vypnutie augmentačných techník, ktoré štandardne nepodporujú spracovanie štvrtého kanálu. Pred trénovaním takéhoto modelu bolo potrebné predspracovať dátovú sadu tak, aby sa tri RGB kanály zlúčili s hĺbkovým kanálom do jediného štvorkanálového obrazu. Tento zlúčený obraz bol uložený vo formáte TIFF, ktorý bol vybraný ako najvhodnejší pre tento účel, pretože pri použití formátu PNG sa vyskytlo viacero rôznych technických problémov. Takto upravený model predstavuje metódu skorého zlúčenia a umožňuje využitie dodatočných hĺbkových informácií spolu s tradičnými RGB dátami.

Upravené modely YOLOv5

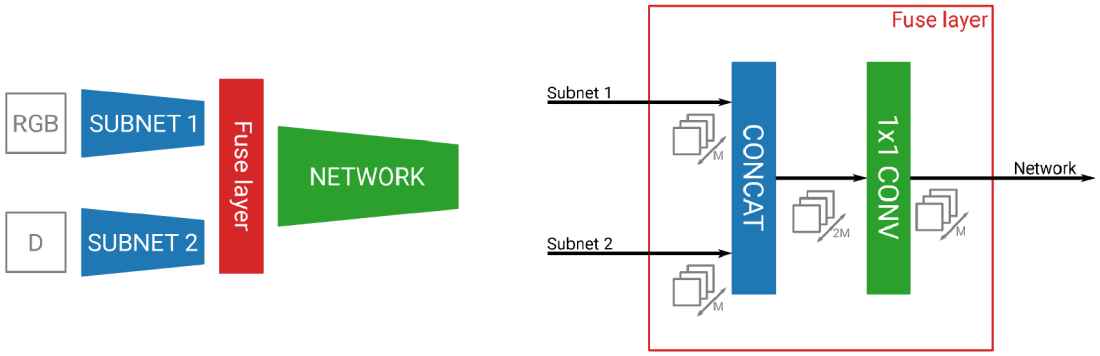
Výskum v tejto záverečnej práci je čiastočne založený aj na štúdii [44], ktorá využíva upravené verzie modelu YOLOv5 na detekciu objektov pomocou RGB a infračervených obrazov zachytených dronom. V tejto štúdií boli použité dve rôzne metódy integrácie infračervených snímkov do procesu detekcie, ktoré si vyžadovali špecifické úpravy v architektúre modelu. Práve tieto upravené modely sú využité v experimentálnej časti.

Prvá metóda, označovaná ako neskorá fúzia, využíva dve samostatné chrbticové siete pre RGB a hĺbkové obrazy. Tieto obrazy sa v rámci svojich chrbticových sietí spracovávajú paralelne, pričom každá sieť má svoje vlastné váhy a učí sa rozpoznávať príznaky špecifické pre svoju dátovú doménu. Po spracovaní obrazov v chrbticovej časti dochádza k zlúčeniu výsledkov, ktoré sú potom poslané do spoločného krku a hlavy modelu. Takéto zlúčenie vstupov je rovnako ako v [45] realizované konkatenáciou, po ktorej nasleduje 1×1 konvolúcia, ktorá redukuje rozmerovosť dát a tým zabezpečuje správne spracovanie informácií v krčnej časti siete. Zjednodušené zobrazenie takejto upravenej architektúry je na obrázku 3.3.

Druhá metóda predstavuje prístup skorej fúzie, kde sa RGB a hĺbkové obrazy spájajú už na úrovni vstupu. Model je upravený tak, aby prvá vrstva prijímala 4-kanálový vstup, ktorý kombinuje tri RGB kanály s jedným kanálom IR/hĺbkového obrazu prevedeného do odtieňov šedej. Po tejto fúzii na úrovni vstupného obrazu sa sieť správa rovnako ako pôvodný YOLOv5 model. Tento prístup je priamočiarejší a nevyžaduje paralelné spracovanie dvoch chrbticových sietí, čo je výhodnejšie z hľadiska výpočtovej efektívnosti.

[1] https://gist.github.com/ialhashim/be6235489a9c43c6d240e8331836586a

Oba spomenuté prístupy si vyžiadali špecifické úpravy v architektúre a inferenčnom procese modelu YOLOv5. Napríklad na pridanie infračervených/hĺbkových obrázkov do tréningu bolo potrebné upraviť modul zodpovedný za načítavanie dát (angl. dataloader). Takéto úpravy umožnili efektívne načítavať a spracovávať snímky a ukladať ich do vyrovnávacej pamäte, čo znížilo čas potrebný na trénovanie. Okrem toho bolo nutné prispôsobiť aj metódy augmentácie aby zachovali zarovnanie hĺbkových obrazov s RGB obrázkami.



Obr. 3.3: Architektúra fúznej siete použitej v metóde skorej fúzie (a) a štruktúra fúznej vrstvy (b) z [45].

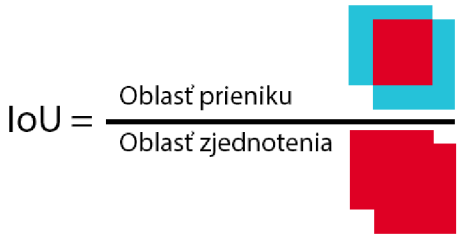
Implementácia kódovania HHA

V experimentálnej časti tejto práce je na reprezentáciu hĺbkových máp aplikovaná aj metóda HHA. Táto metóda, detailnejšie predstavená v podkapitole 2.3, umožňuje efektívnejšie využitie informácií dostupných v hĺbkových obrazoch. Pri implementácií tejto metódy som využil verejne dostupný repozitár [6], ktorý poskytuje implementáciu tohto algoritmu v programovacom jazyku Python 3. Pôvodná verzia tohto algoritmu bola vyvinutá v MATLABe, avšak pre účely tejto práce som uprednostnil Pythonovú verziu, ktorá dosahuje identické výsledky ako MATLABový kód.

3.3 Metriky pre zrovnávanie

V oblasti počítačového videnia je dôležité porozumieť tomu, ako presne sú objekty na obrázkoch detekované. Na to slúži viacero metrík, ktoré pomáhajú vyhodnotiť a porovnať účinnosť použitých metód. Medzi základné metriky patria Intersection over Union (IoU), presnosť (precision), citlivosť (recall), a priemerná presnosť (average precision).

Intersection over Union (IoU) je základná metrika, ktorá meria prekrytie medzi predpovedaným a skutočným ohraničujúcim rámčekom. IoU je vypočítané ako pomer prieniku (intersection) medzi predpovedanou a skutočnou oblasťou k zjednoteniu (union) týchto oblastí. Hodnota IoU môže byť od 0 (žiadne prekrytie) do 1 (úplné prekrytie).



Obr. 3.4: Ilustrácia metriky IoU

Pre konkrétny prah IoU môžeme predpovede rozdeliť do štyroch kategórií:

- **Skutočne pozitívna** (TP) - objekt je správne detekovaný a IoU je nad stanoveným prahom.
- **Falošne pozitívna** (FP) - objekt je nesprávne detekovaný a IoU je pod stanovým prahom.
- **Skutočne negatívna** (TN) - objekt správne nebol detekovaný.
- **Falošne negatívna** (FN) - objekt bol prehliadnutý napriek tomu, že mal byť detekovaný.

Z týchto hodnôt potom je možné odvodiť ďalšie metriky [24]:
Precision je metrika, ktorá vyjadruje, koľko z detekovaných objektov bolo správne identifikovaných. Je definovaná ako pomer skutočne pozitívnych predikcií k celkovému počtu pozitívnych predikcií (TP + FP).

$$Precision = \frac{TP}{TP + FP} \tag{3.1}$$

Recall (alebo citlivosť) ukazuje, aký podiel zo všetkých skutočne pozitívnych prípadov bol správne identifikovaný. Je to pomer skutočne pozitívnych k súčtu skutočne pozitívnych a falošne negatívnych prípadov.

$$Recall = \frac{TP}{TP + FN} \tag{3.2}$$

Average precision (AP) je ďalšia dôležitá metrika, ktorá hodnotí presnosť detekcie pre každú triedu objektov. AP sa počíta ako priemer maximálnych hodnôt precision pre všetky úrovne recall, čo znamená, že vyjadruje plochu pod krivkou precision-recall.
Mean average precision (mAP) poskytuje celkovú mieru úspešnosti detektoru objektov, pričom priemeruje AP pre všetky triedy objektov. Bežne sa používa napríklad mAP pri IoU prahoch 0,50 a 0,95.

$$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k \tag{3.3}$$

Ohraničujúce rámčeky

Pri každej predpovedi pre ohraničujúce rámčeky (angl. bounding boxes) sa berú do úvahy predpovedané hodnoty x, y, w, h, c, kde x, y sú súradnice stredu detekovaného objektu, w, h sú šírka a výška ohraničujúceho rámčeka vzhľadom na veľkosť obrázka a c je skóre dôveryhodnosti, ktoré odhaduje *Intersection over Union* (IOU) medzi skutočným a predpovedaným ohraničujúcim rámčekom.

$$c = P(Object) \times IOU(0, x) \tag{3.4}$$

3.4 Trénovacie prostredie

Pre trénovanie modelov na detekciu objektov bolo v tejto práci využité prostredie Google Colab. Toto online prostredie poskytuje prístup k výkonnej výpočtovej technike bez nutnosti osobného investovania do drahého hardvéru. Hlavnou výhodou bezplatnej verzie Google Colab je dostupnosť grafickej karty NVIDIA Tesla T4, ktorá je vhodná pre strojové učenie a umožňuje efektívne vykonávanie náročných výpočtových operácií potrebných pri trénovaní neurónových sietí. Nevýhodou tohto prostredia je, že každý užívateľ môže mať spustený maximálne jeden virtuálny stroj, musí byť aktívny, maximálna doba spustenia každého virtuálneho stroja je obmedzená a dáta na disku nie sú perzistentné. Pri väčších modeloch a vyššom rozlíšení občas nestačila bezplatná verzia, a preto boli niektoré modely trénované na výkonnejších grafických kartách NVIDIA L4 s využitím zakúpených výpočtových jednotiek v rámci platenej verzie Colab.

Kapitola 4

Experimenty

Táto kapitola sa zameriava na tri experimenty, ktoré hodnotia účinnosť integrácie hĺbkových dát do objektových detektorov YOLO v rôznych kontextoch. Prvý a zároveň najobsiahlejší experiment obsahuje viacero podexperimentov, ktoré skúmajú rozličné metódy fúzie a reprezentácie hĺbkových dát v rámci dátovej sady NYU Depth v2. Druhý experiment rozširuje skúmanie na inú dátovej sadu, Washington RGB-D, a testuje, či pozitívne výsledky z prvého experimentu možno replikovať v iných podmienkach. Záverečný experiment testuje efektivitu integrácie hĺbkových dát do modelu YOLOv8, s cieľom potvrdiť ich použiteľnosť a výkonnosť v modernejších architektúrach. Tieto experimenty spolu poskytujú komplexný pohľad na prínos využitia hĺbkových dát pre detekciu objektov.

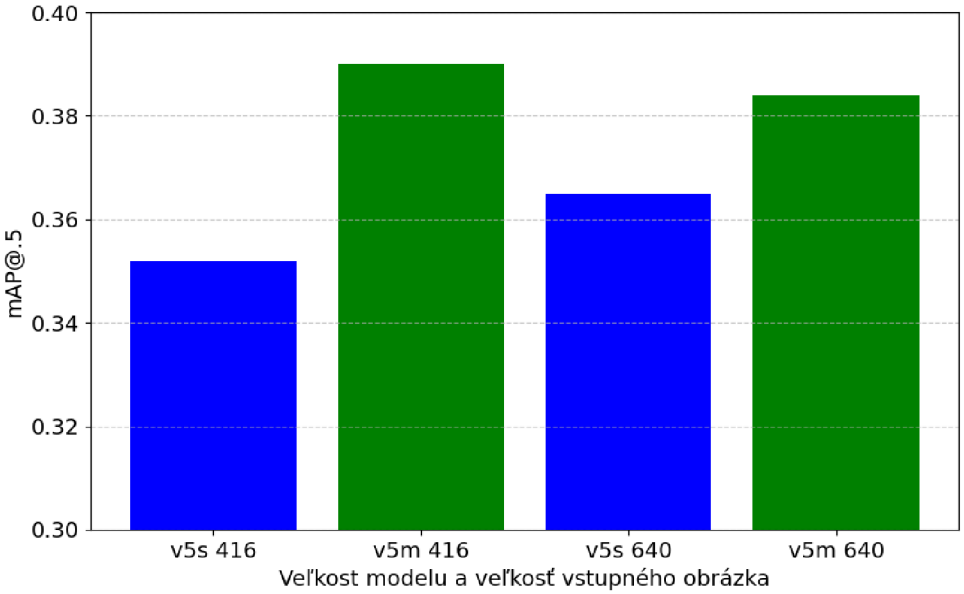
4.1 Experimenty s upravenými modelmi YOLOv5 a rôznymi reprezentáciami hĺbky na dátovej sade NYU Depth v2

Nasledujúca časť popisuje experimenty s upravenými modelmi YOLOv5 a rôznymi reprezentáciami hĺbky na dátovej sade NYU Depth v2, ktoré prebiehali v štyroch fázach. Najprv sa hodnotila základná výkonnosť s použitím štandardných RGB vstupov. Následne sme skúmali efektivitu skorej a neskorej fúzie hĺbkových dát s RGB vstupmi. Záverečná časť porovnáva tieto prístupy a analyzuje ich vplyv na presnosť detekcie objektov. Každý model bol testovaný v dvoch veľkostiach na rôznych rozlíšeniach, pričom kľúčové boli metriky AP@.5 a mAP@.5:.95. Pre podrobné porovnanie medzi jednotlivými triedami bola zvolená metrika AP@.5, zatiaľ čo metrika mAP@.5:.95 je uvedená vždy len pre všetky triedy dohromady.

Výkonnosť základného modelu

Úvodná fáza bola venovaná zisteniu základnej výkonnosti s použitím štandardných RGB vstupov, bez pridanej hĺbkovej zložky V úvodnej fáze experimentov bolo mojím cieľom získať základné výkonnostné metriky s použitím štandardných RGB vstupov, bez pridanej hĺbkovej zložky. Získané metriky budú slúžiť ako referenčný rámec pre ďalšie pokročilé testy. Na tento účel som využil modely YOLOv5 v konfiguráciách YOLOv5s a YOLOv5m testované na obrazoch s rozlíšením 416×416 pixelov. Okrem toho boli modely YOLOv5s testované aj pri rozlíšení 640×640, aby som získal predstavu o vplyve väčšieho rozlíšenia na výkonnosť modelu. Tieto testy poskytli základné porozumenie schopností modelov spracovávať len RGB dáta bez hĺbkových informácií. Vizuálne porovnanie metrík týchto

konfigurácií je znázornené na obrázku 4.1. V tabuľke 4.1 sú výsledky AP@.5 vyjadrené číselne pre každú triedu. Výsledky mAP@.5:.95 je možné nájsť v tabuľke 4.2.



Obr. 4.1: Porovnanie mAP@.5 a mAP@.5:.95

Class	YOLOv5s 416 × 416	YOLOv5m 416 × 416	YOLOv5s 640 × 640	YOLOv5m 640 × 640
All	0.352	0.390	0.365	0.384
Bathtub	0.350	0.316	0.243	0.381
Bed	0.510	0.540	0.449	0.512
Bookshelf	0.403	0.473	0.372	0.440
Box	0.0634	0.089	0.0928	0.114
Chair	0.416	0.475	0.423	0.413
Counter	0.358	0.414	0.366	0.315
Desk	0.0757	0.0938	0.136	0.116
Door	0.290	0.367	0.351	0.318
Dresser	0.0564	0.136	0.0655	0.100
Garbage Bin	0.376	0.641	0.489	0.547
Lamp	0.451	0.462	0.486	0.552
Monitor	0.520	0.479	0.567	0.510
Night Stand	0.608	0.672	0.690	0.710
Pillow	0.247	0.303	0.326	0.288
Sink	0.274	0.354	0.218	0.343
Sofa	0.451	0.434	0.400	0.435
Table	0.147	0.155	0.158	0.139
Television	0.521	0.460	0.576	0.483
Toilet	0.575	0.556	0.523	0.573

Tabuľka 4.1: Porovnanie celkovej metriky mAP@.5 a metrík AP@.5 pre všetky skúmané triedy naprieč rôznymi veľkosťami modelu YOLOv5 pri rozlíšeniach 416 × 416 a 640 × 640.

	YOLOv5s 416 × 416	YOLOv5m 416 × 416	YOLOv5s 640 × 640	YOLOv5m 640 × 640
mAP@0.5:.95	0.170	0.209	0.178	0.202

Tabuľka 4.2: Porovnanie celkovej metriky mAP@.5:.95 modelov YOLOv5s a YOLOv5m pri rozlíšeniach 416 × 416 a 640 × 640.

Metóda skorého zlúčenia

Ako prvú som skúmal efektivitu metódy skorej fúzie, kde sa priama integrácia hĺbkových dát do vstupnej vrstvy modelu vykonáva pridaním štvrtého kanálu k RGB kanálom. Konkrétne bol k RGB kanálom pridaný štvrtý kanál, ktorý reprezentuje hĺbku ako bolo bližšie popísané v časti 3.2. Experimenty začali s modelom YOLOv5s a postupne som pridal aj modely YOLOv5m, čím som chcel zistiť, ako veľkosť modelu ovplyvňuje schopnosť spracovať pridané hĺbkové informácie. Napriek tomu, že interpretácie hĺbkových dát ako farebná paleta Jet, kódovanie HHA a normály povrchu sú farebné, v prvej vrstve modelu sú všetky tieto dáta konvertované do odtieňov sivej. Aj po odstránení farebnosti si však každá reprezentácia zachováva niektoré svoje charakteristické príznaky, ktoré ovplyvňujú presnosť detekcie. Všetky tieto interpretácie hĺbky boli testované najskôr pri rozlíšení vstupu 416 × 416 a po-

tom pri rozlíšení 640×640, čo mi umožnilo porovnať efektívnosť štvorkanálového vstupu pri rôznych úrovniach detailov v hĺbkových mapách.

Výsledky experimentov, ktoré porovnávajú vplyv pridania hĺbkových dát ako štvrtého kanálu do modelov YOLOv5s a YOLOv5m pri rozlíšení 416×416 sú prezentované v tabuľkách 4.3 a 4.4, kde sú uvedené hodnoty metriky AP@.5 pre rôzne hĺbkové reprezentácie a triedy objektov. V tabuľke 4.5 je možné nájsť celkovú metriku mAP@.5:.95 naprieč rôznymi reprezentáciami hĺbky.

Class	**Jet**	**HHA**	**Normals**	**RGB**
All	**0.379**	0.376	0.373	0.352
Bathtub	**0.482**	0.317	0.426	0.350
Bed	0.510	**0.571**	0.510	0.510
Bookshelf	0.517	**0.452**	0.329	0.403
Box	0.0679	**0.0948**	0.0741	0.0634
Chair	0.476	0.445	**0.479**	0.416
Counter	0.327	**0.419**	0.307	0.358
Desk	**0.141**	0.0836	0.101	0.0757
Door	**0.303**	0.299	0.256	0.290
Dresser	0.0553	0.185	**0.217**	0.0564
Garbage Bin	0.29	0.403	**0.516**	0.376
Lamp	0.443	**0.490**	0.429	0.451
Monitor	**0.612**	0.441	0.528	0.520
Night Stand	0.600	**0.624**	0.548	0.608
Pillow	0.388	0.352	**0.460**	0.247
Sink	0.273	**0.349**	0.294	0.274
Sofa	**0.472**	0.465	0.420	0.451
Table	0.175	0.182	**0.196**	0.147
Television	**0.564**	0.389	0.428	0.521
Toilet	0.501	**0.581**	0.568	0.575

Tabuľka 4.3: Metriky AP@.5 upraveného YOLOv5s s využitím skorého zlúčenia pri rozlíšení 416×416 v rôznych hĺbkových reprezentáciách.

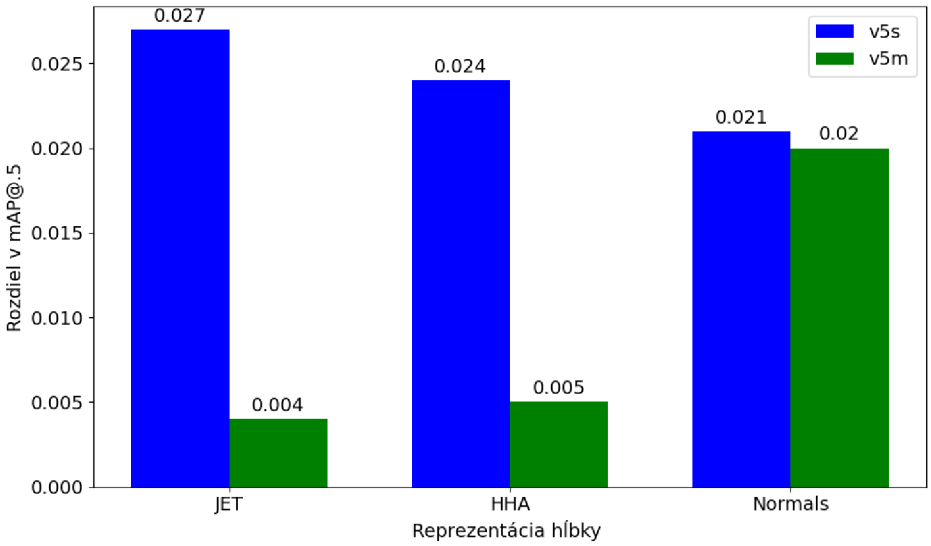
Class	Jet	HHA	Normals	RGB
All	0.394	0.395	**0.410**	0.390
Bathtub	0.521	0.412	**0.536**	0.316
Bed	0.486	**0.525**	0.508	0.540
Bookshelf	0.456	**0.473**	0.432	0.473
Box	0.0828	**0.101**	0.0916	0.089
Chair	0.522	0.442	**0.540**	0.475
Counter	**0.433**	0.381	0.405	0.414
Desk	**0.158**	0.102	0.109	0.0938
Door	0.281	**0.322**	0.296	0.367
Dresser	0.0322	**0.0643**	0.042	0.136
Garbage Bin	**0.516**	0.462	0.462	0.641
Lamp	0.512	**0.554**	0.490	0.462
Monitor	0.528	0.502	**0.608**	0.479
Night Stand	0.561	0.637	**0.736**	0.672
Pillow	0.347	**0.403**	0.374	0.303
Sink	0.339	**0.413**	0.385	0.354
Sofa	0.478	0.520	**0.538**	0.434
Table	0.210	0.203	**0.212**	0.155
Television	**0.501**	0.388	0.475	0.460
Toilet	0.521	**0.607**	0.550	0.556

Tabuľka 4.4: Metriky AP@.5 upraveného YOLOv5m s využitím skorého zlúčenia pri rozlíšení 416×416 v rôznych hĺbkových reprezentáciách.

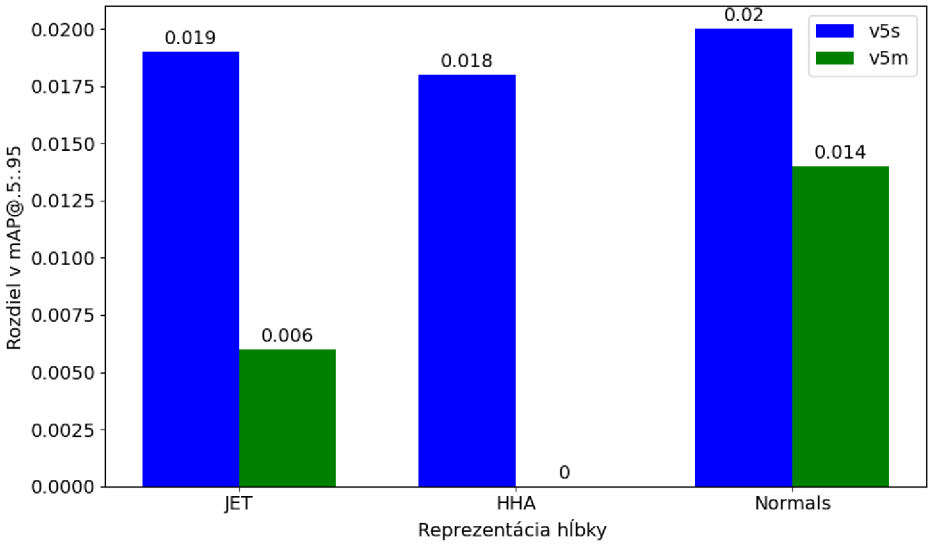
Model	Jet	HHA	Normals	RGB
YOLOv5s	0.189	0.188	**0.190**	0.170
YOLOv5m	0.215	0.209	**0.223**	0.209

Tabuľka 4.5: Metrika mAP@.5:.95 upravených modelov využívajúcich skoré zlúčenie pri rozlíšení 416×416 v rôznych hĺbkových reprezentáciách.

Na výsledkoch z vyššie uvedených tabuliek môžme pozorovať, že pridanie hĺbky ako štvrtého kanálu má pozitívny vplyv na presnosť detekcie. Pri modeli YOLOv5s bol oproti základnému RGB modelu zaznamenaný viditeľný nárast mAP@.5 aj mAP@.5:.95 pri všetkých hĺbkových reprezentáciách. Pri modeli YOLOv5m bol nárast v mAP výrazný hlavne pri reprezentácií hĺbky povrchovými normálami. Pri ostatných reprezentáciách bol nárast miernejší. Toto zlepšenie lepšie vykresľujú grafy 4.2 a 4.3, ktoré zobrazujú zlepšenie metriky mAP pri hĺbkových modeloch využívajúcich skoré zlúčenie oproti základnému modelu bez použitia hĺbkových dát naprieč rôznymi reprezentáciami hĺbky.



Obr. 4.2: Rozdiel v mAP@.5 medzi modelom využívajúcim metódu skorého zlúčenia a základným RGB modelom pri rozlíšení 416×416.



Obr. 4.3: Rozdiel v mAP@.5:.95 medzi modelom využívajúcim metódu skorého zlúčenia a základným RGB modelom pri rozlíšení 416×416.

Modely YOLOv5s a YOLOv5m boli následne natrénované znovu s využitím vyššieho rozlíšenia vstupných obrázkov, konkrétne 640×640. Výsledné metriky AP@.5 pre rôzne hĺbkové reprezentácie a triedy objektov je možné vidieť v tabuľkách 4.6 a 4.7. Opäť bola porovnaná aj celková metrika mAP@.5:.95, ktorej hodnoty sú uvedené v tabuľke 4.8.

Class	Jet	HHA	Normals	RGB
All	0.386	0.385	**0.397**	0.365
Bathtub	**0.382**	0.306	0.360	0.243
Bed	0.494	0.506	**0.513**	0.449
Bookshelf	**0.447**	0.373	0.421	0.372
Box	0.0833	0.114	**0.118**	0.0928
Chair	0.456	**0.489**	0.485	0.423
Counter	0.367	0.360	**0.392**	0.366
Desk	0.124	0.110	**0.128**	0.136
Door	0.291	**0.293**	0.289	0.351
Dresser	0.175	0.147	**0.179**	0.0655
Garbage Bin	**0.502**	0.489	0.484	0.489
Lamp	0.510	**0.552**	0.505	0.486
Monitor	**0.533**	0.501	0.527	0.567
Night Stand	0.649	0.661	**0.726**	0.690
Pillow	0.323	0.367	**0.389**	0.326
Sink	0.346	**0.383**	0.342	0.218
Sofa	0.416	0.474	**0.485**	0.400
Table	0.169	**0.192**	0.161	0.158
Television	0.477	**0.512**	0.447	0.576
Toilet	**0.589**	0.492	0.583	0.523

Tabuľka 4.6: Metriky AP@.5 upraveného YOLOv5s s využitím skorého zlúčenia pri rozlíšení 640×640 v rôznych hĺbkových reprezentáciách.

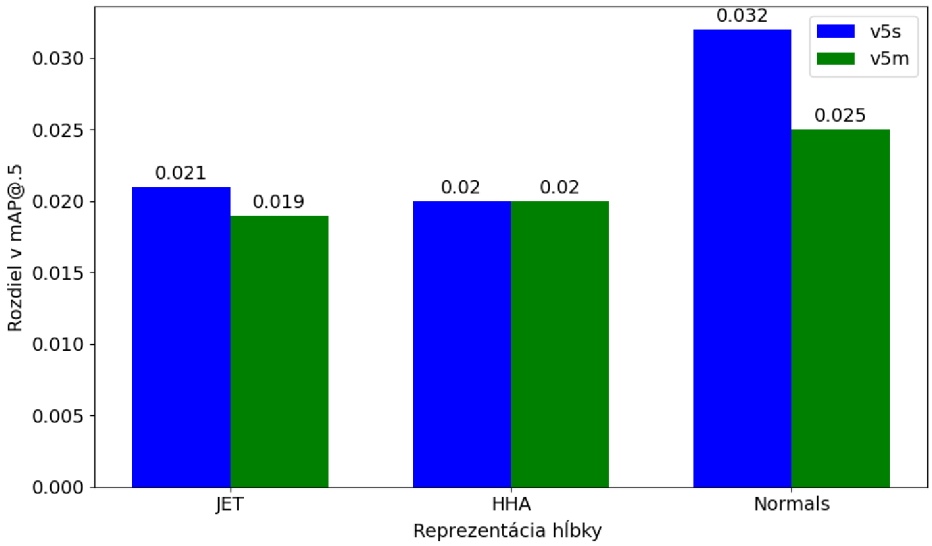
Class	Jet	HHA	Normals	RGB
All	0.403	0.404	**0.409**	0.384
Bathtub	**0.397**	0.372	0.384	0.381
Bed	0.522	**0.532**	0.507	0.512
Bookshelf	**0.504**	0.489	0.416	0.440
Box	**0.134**	0.0941	0.0636	0.114
Chair	**0.534**	0.498	0.508	0.413
Counter	0.322	**0.380**	0.343	0.315
Desk	**0.165**	0.164	0.152	0.116
Door	0.297	0.326	**0.360**	0.318
Dresser	0.158	0.121	**0.164**	0.100
Garbage Bin	0.465	0.483	**0.627**	0.547
Lamp	0.579	**0.593**	0.559	0.552
Monitor	0.487	**0.529**	0.490	0.510
Night Stand	0.673	**0.747**	0.679	0.710
Pillow	0.372	0.388	**0.440**	0.288
Sink	0.386	0.324	**0.391**	0.343
Sofa	0.456	0.511	**0.538**	0.435
Table	**0.228**	0.213	0.190	0.139
Television	0.360	0.377	**0.401**	0.483
Toilet	**0.616**	0.543	0.562	0.573

Tabuľka 4.7: Metriky AP@.5 upraveného YOLOv5m s využitím skorého zlúčenia pri rozlíšení 640 × 640 v rôznych hĺbkových reprezentáciách.

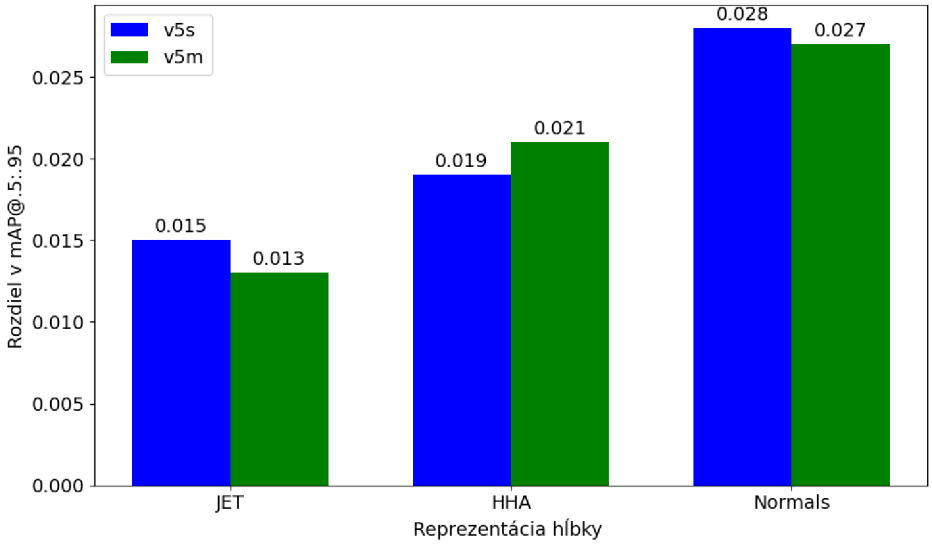
Model	Jet	HHA	Normals	RGB
YOLOv5s	0.193	0.197	**0.206**	0.178
YOLOv5m	0.215	0.223	**0.229**	0.202

Tabuľka 4.8: Metrika mAP@.5:.95 upravených modelov využívajúcich skoré zlúčenie pri rozlíšení 640 × 640 v rôznych hĺbkových reprezentáciách.

Výsledky z vyššie uvedených tabuliek 4.6, 4.7 a 4.8 opäť ukazujú, že integrácia hĺbkových dát pomocou štvrtého kanálu pozitívne ovplyvňuje detekčné schopnosti modelov YOLOv5s a YOLOv5m aj pri rozlíšení 640 × 640. Zlepšenie metrík mAP obidvoch veľkostí modelu YOLO bolo v porovnaní s nižším rozlíšením vstupu viac konzistentné naprieč všetkými hĺbkovými reprezentáciami, čo je možné lepšie pozorovať na obrázku 4.4 pre mAP@.5 a obrázku 4.3 pre mAP@.5:.95.



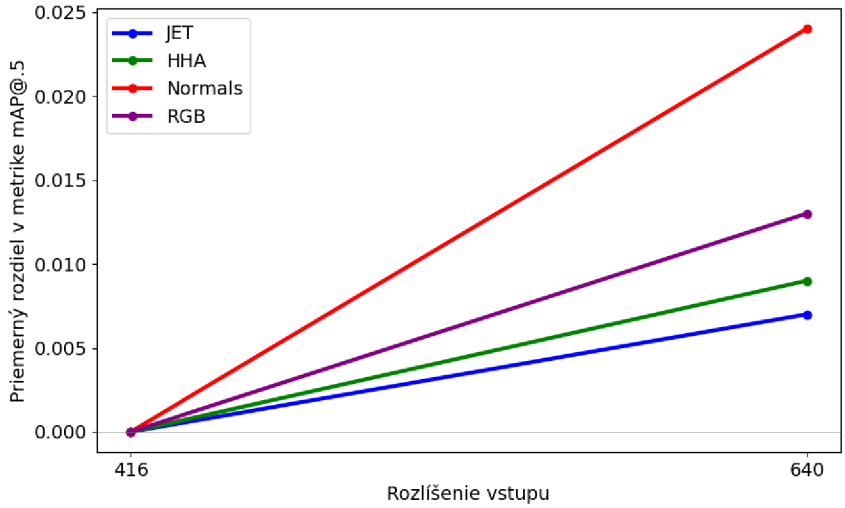
Obr. 4.4: Rozdiel v mAP@.5 medzi modelom využívajúcim metódu skorého zlúčenia a základným RGB modelom pri rozlíšení 640×640.



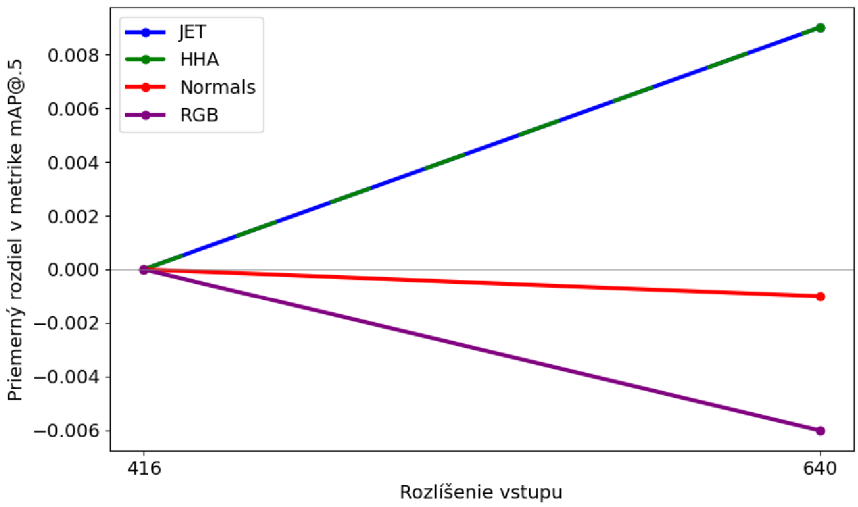
Obr. 4.5: Rozdiel v mAP@.5:.95 medzi modelom využívajúcim metódu skorého zlúčenia a základným RGB modelom pri rozlíšení 640×640.

Pri hlbšej analýze vplyvu zväčšenia rozlíšenia vstupných obrazov na presnosť detekcie objektov je možné pozorovať zaujímavé trendy. Pri modeli YOLOv5s viedlo zvýšenie rozlíšenia z 416 na 640 k pozitívnemu ovplyvneniu presnosti u všetkých testovaných reprezentácií vrátane základného RGB, čo je znázornené na obrázku 4.6. Toto zlepšenie naznačuje, že väčšie rozlíšenie poskytuje menšiemu modelu viac informácií, čo umožňuje lepšie využitie hĺbkových dát pri detekcii objektov. Pri väčšom modeli YOLOv5m však zväčšenie rozlíšenia z 416 na 640 malo rozdielne účinky, čo môžme vidieť na obrázku 4.7. Zatiaľ čo metódy JET a HHA vykázali mierne zlepšenie presnosti, u metód Normals a RGB došlo k poklesu

presnosti. Toto môže naznačovať, že väčší z modelov môže byť viac citlivý na šum a zmenu charakteristiky vizuálnych a hĺbkových prvkov spôsobenú zväčšením originálneho rozlíšenia.



Obr. 4.6: Vplyv zmeny rozlíšenia vstupných obrazov na metriku mAP@.5 menšieho modelu YOLOv5s využívajúceho metódu skorého zlúčenia.



Obr. 4.7: Vplyv zmeny rozlíšenia vstupných obrazov na metriku mAP@.5 väčšieho modelu YOLOv5m využívajúceho metódu skorého zlúčenia.

Metóda neskorého zlúčenia

V tejto časti som skúmal efektivitu metódy neskorej fúzie, kde sa spracovanie hĺbkových a RGB dát vykonáva nezávisle pomocou dvoch oddelených chrbticových sietí ako je bližšie popísané v časti 3.2.

Experimenty boli podobne ako pri metóde skorého zlúčenia vykonané pomocou modelov YOLOv5s a YOLOv5m s cieľom zistiť, ako veľkosť a komplexnosť modelu ovplyvňuje jeho schopnosť spracovať a integrovať príznaky z dvoch rôznych dátových domén. V experimentoch boli využité rovnaké hĺbkové interpretácie ako v prípade skorej fúzie, t. j. farebná paleta Jet, kódovanie HHA a reprezentácia pomocou povrchových normál. Tieto modely boli znovu testované pri dvoch rôznych rozlíšeniach: 416×416 a 640×640, čo umožnilo preskúmať, ako rozlíšenie vstupných obrazov ovplyvňuje účinnosť fúzie informácií z dvoch rozdielnych domén.

Výsledky experimentov, ktoré porovnávajú efektivitu neskorej fúzie v modeloch YOLOv5s a YOLOv5m pri rozlíšení 416×416, sú prezentované v tabuľkách 4.9 a 4.10, kde sú uvedené hodnoty metriky AP@.5 pre rôzne hĺbkové reprezentácie a triedy objektov. V tabuľke 4.11 je možné nájsť celkovú metriku mAP@.5:.95 pre obidve veľkosti modelu naprieč rôznymi hĺbkovými reprezentáciami.

Class	**Jet**	**HHA**	**Normals**	**RGB**
All	**0.392**	0.386	0.384	0.352
Bathtub	**0.616**	0.449	0.471	0.350
Bed	0.480	0.512	**0.532**	0.510
Bookshelf	**0.499**	0.419	0.386	0.403
Box	**0.0892**	0.0419	0.0839	0.0634
Chair	**0.480**	0.472	0.471	0.416
Counter	0.379	**0.416**	0.407	0.358
Desk	**0.185**	0.110	0.132	0.0757
Door	0.323	**0.335**	0.298	0.290
Dresser	0.0973	0.167	**0.190**	0.0564
Garbage Bin	**0.452**	0.450	0.429	0.376
Lamp	0.494	0.483	**0.526**	0.451
Monitor	0.466	0.463	**0.544**	0.520
Night Stand	0.629	**0.647**	0.538	0.608
Pillow	0.335	0.375	**0.388**	0.247
Sink	0.324	**0.352**	0.297	0.274
Sofa	0.431	0.449	**0.466**	0.451
Table	0.171	**0.176**	0.167	0.147
Television	0.482	0.495	**0.500**	0.521
Toilet	0.514	**0.519**	0.474	0.575

Tabuľka 4.9: Metriky AP@.5 upraveného YOLOv5s s využitím metódy neskorého zlúčenia pri rozlíšení 416×416 v rôznych hĺbkových reprezentáciách.

Class	Jet	HHA	Normals	RGB
All	**0.414**	0.4	0.394	0.390
bathtub	**0.517**	0.495	0.414	0.316
bed	0.496	**0.508**	0.495	0.540
bookshelf	**0.462**	0.446	0.451	0.473
box	**0.0892**	0.0769	0.0727	0.089
chair	0.489	0.486	**0.509**	0.475
counter	0.443	0.392	**0.461**	0.414
desk	0.166	0.143	**0.173**	0.0938
door	0.342	**0.348**	0.299	0.367
dresser	**0.198**	0.14	0.16	0.136
garbage bin	**0.475**	0.415	0.385	0.641
lamp	0.518	0.476	**0.554**	0.462
monitor	0.489	**0.582**	0.577	0.479
night stand	**0.696**	0.567	0.588	0.672
pillow	**0.401**	0.339	0.389	0.303
sink	0.376	**0.41**	0.281	0.354
sofa	0.478	**0.527**	0.479	0.434
table	**0.214**	0.198	0.167	0.155
television	0.45	**0.513**	0.489	0.460
toilet	**0.563**	0.545	0.550	0.556

Tabuľka 4.10: Metriky AP@.5 upraveného YOLOv5m s využitím metódy neskorého zlúčenia pri rozlíšení 416 × 416 v rôznych hĺbkových reprezentáciách.

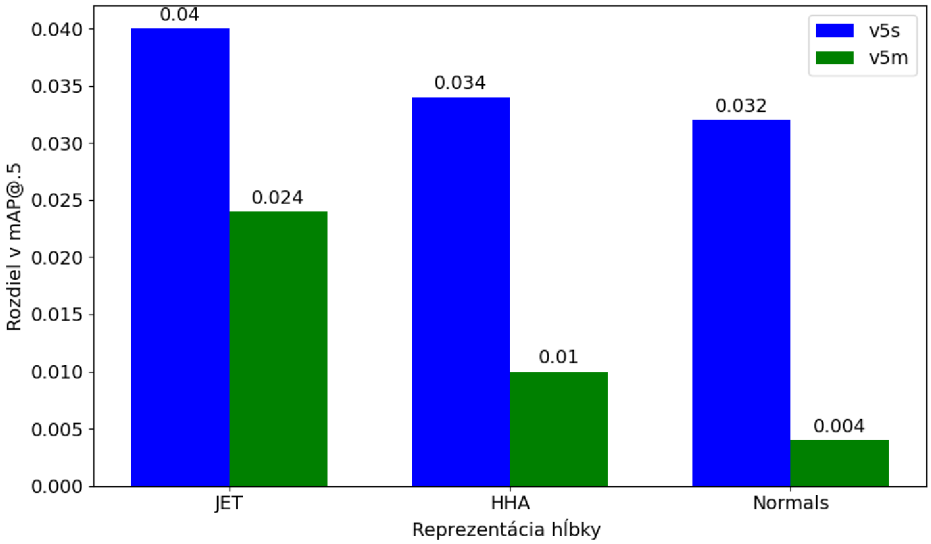
Model	Jet	HHA	Normals	RGB
YOLOv5s	**0.198**	0.194	0.186	0.170
YOLOv5m	**0.231**	0.216	0.210	0.209

Tabuľka 4.11: Metrika mAP@.5:.95 upravených modelov s využitím metódy neskorého zlúčenia pri rozlíšení 416 × 416 v rôznych hĺbkových reprezentáciách.

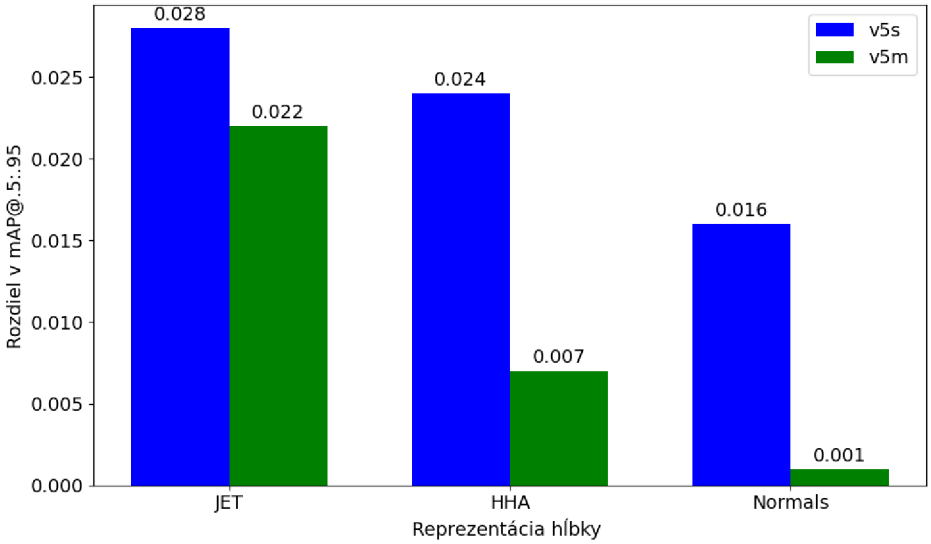
Dosiahnuté výsledky naznačujú, že integrácia hĺbkových dát do detekčných modelov pomocou metódy neskorého zlúčenia prispieva k zlepšeniu presnosti detekcie v porovnaní s použitím čisto RGB modelov. Zaujímavým pozorovaním je, že menší model YOLOv5s vykazoval výraznejšie zlepšenie než jeho väčší ekvivalent YOLOv5m. Tento rozdiel môže byť spojený s pomerom veľkosti modelu k pridaným hĺbkovým príznakom prostredníctvom paralelnej chrbticovej siete.

Menšie modely, ako je YOLOv5s, majú menej parametrov, čo znamená, že pridanie hĺbkových príznakov cez paralelnú chrbticovú sieť môže mať výraznejší vplyv na ich presnosť. Na druhej strane, väčší model YOLOv5m už v základe obsahuje viac parametrov, ktoré mu umožňujú efektívnejšie spracovávať komplexné príznaky. Pridanie hĺbkových príznakov preto nevedie k tak výraznému zlepšeniu, pretože tento model už má dostatočne bohatý parametrický základ pre spracovanie dostupných dát.

Spomínané zlepšenie metrík mAP pri modeloch, ktoré využívajú na integráciu hĺbkových dát metódu neskorej fúzie oproti základným RGB modelom je znázornené na obrázkoch 4.8 a 4.9.



Obr. 4.8: Rozdiel v mAP@.5 medzi modelom neskorej fúzie a základným RGB modelom pri rozlíšení 416 × 416.



Obr. 4.9: Rozdiel v mAP@.5:.95 medzi modelom neskorej fúzie a základným RGB modelom pri rozlíšení 416 × 416.

Modely vo veľkosti YOLOv5s a YOLOv5m boli opätovne trénované pri použití zvýšeného rozlíšenia vstupných obrázkov, presne 640 × 640 pixelov. Výsledné metriky AP@.5 pre rozličné hĺbkové reprezentácie a kategórie objektov je možné nájsť v tabuľkách 4.12

a 4.13. Rovnako bolo pre obidve veľkosti modelov vykonané porovnanie celkovej metriky mAP@.5:.95, ktorej hodnoty sú zobrazené v tabuľke 4.14.

Class	**Jet**	**HHA**	**Normals**	**RGB**
All	0.391	0.371	**0.400**	0.365
Bathtub	**0.611**	0.371	0.521	0.243
Bed	0.507	0.512	**0.522**	0.449
Bookshelf	0.362	0.409	**0.418**	0.372
Box	0.0608	0.0723	**0.0936**	0.0928
Chair	0.441	**0.465**	0.456	0.423
Counter	0.313	**0.343**	0.325	0.366
Desk	**0.160**	0.141	0.120	0.136
Door	0.316	**0.358**	0.321	0.351
Dresser	0.115	0.140	**0.155**	0.0655
Garbage Bin	0.382	0.366	**0.519**	0.489
Lamp	0.534	0.480	**0.539**	0.486
Monitor	0.495	**0.533**	0.523	0.567
Night Stand	0.614	**0.690**	0.575	0.690
Pillow	**0.455**	0.383	0.405	0.326
Sink	0.352	0.300	**0.417**	0.218
Sofa	0.423	0.393	**0.484**	0.400
Table	0.129	0.148	**0.160**	0.158
Television	**0.594**	0.461	0.526	0.576
Toilet	**0.562**	0.489	0.515	0.523

Tabuľka 4.12: Metriky AP@.5 upraveného YOLOv5s s využitím metódy neskorého zlúčenia pri rozlíšení 640×640 v rôznych hĺbkových reprezentáciách.

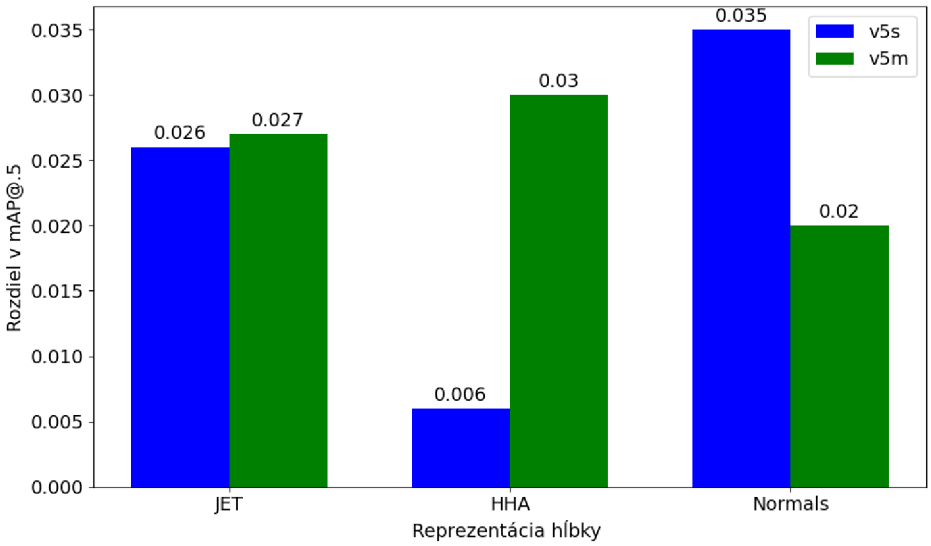
Class	Jet	HHA	Normals	RGB
All	0.411	**0.414**	0.404	0.384
Bathtub	0.434	**0.441**	0.400	0.381
Bed	0.493	0.456	**0.519**	0.512
Bookshelf	0.393	0.426	**0.462**	0.440
Box	**0.125**	0.0851	0.0959	0.114
Chair	**0.517**	0.487	0.498	0.413
Counter	0.361	**0.364**	0.337	0.315
Desk	0.192	**0.217**	0.117	0.116
Door	**0.367**	0.358	0.358	0.318
Dresser	0.107	**0.182**	0.179	0.100
Garbage Bin	0.393	0.439	**0.510**	0.547
Lamp	0.532	**0.550**	0.547	0.552
Monitor	**0.616**	0.522	0.566	0.510
Night Stand	0.682	0.681	**0.695**	0.710
Pillow	**0.426**	0.391	0.349	0.288
Sink	0.381	0.390	**0.406**	0.343
Sofa	0.442	**0.504**	0.489	0.435
Table	0.171	0.167	**0.195**	0.139
Television	0.589	**0.621**	0.403	0.483
Toilet	**0.593**	0.588	0.541	0.573

Tabuľka 4.13: Metriky AP@.5 upraveného YOLOv5m s využitím metódy neskorého zlúčenia pri rozlíšení 640 × 640 v rôznych hĺbkových reprezentáciách.

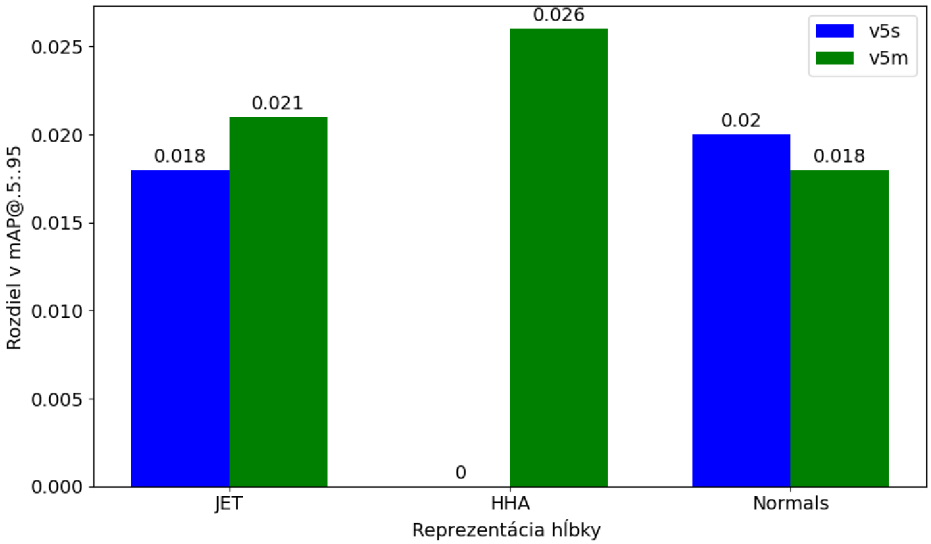
Model	Jet	HHA	Normals	RGB
YOLOv5s	**0.196**	0.178	0.198	0.178
YOLOv5m	**0.223**	0.228	0.220	0.202

Tabuľka 4.14: Metrika mAP@.5:.95 upravených modelov s využitím neskorej fúzie pri rozlíšení 640 × 640 v rôznych hĺbkových reprezentáciách.

Výsledky uvedené v tabuľkách opäť potvrdzujú, že použitie metódy neskorého zlúčenia na intregráciu hĺbkových dát do tréningového procesu má pozitívny vplyv na detekčné schopnosti modelov YOLOv5s a YOLOv5m aj pri zväčšení rozlíšenia vstupných obrazov. Toto zlepšenie je konzistentné vo väčšine reprezentácií a modelov, pričom jedinú výnimku tvorí reprezentácia HHA pri menšom z modelov YOLOv5s, kde bolo zlepšenie pomerne zanedbateľné. Dosiahnuté zlepšenia mAP sú znázornené na obrázkoch 4.10 a 4.11.



Obr. 4.10: Rozdiel v mAP@.5 medzi modelom neskorej fúzie a základným RGB modelom pri rozlíšení 640×640.

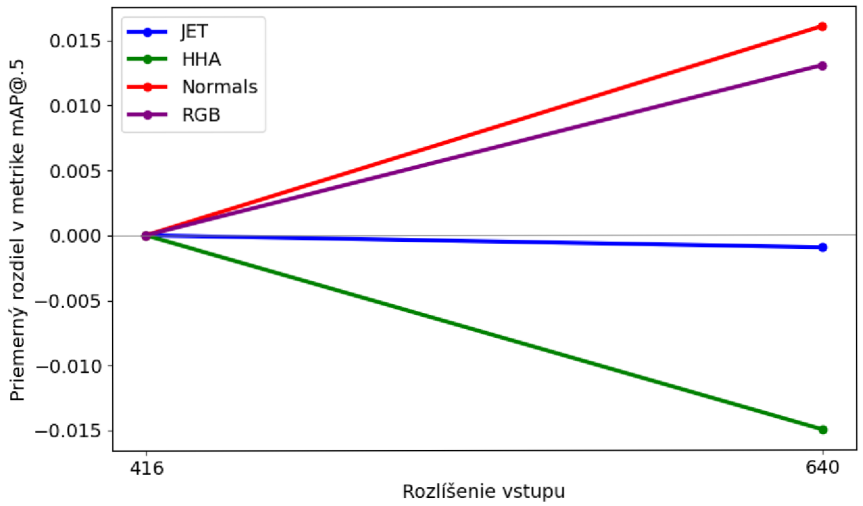


Obr. 4.11: Rozdiel v mAP@.5:.95 medzi modelom neskorej fúzie a základným RGB modelom pri rozlíšení 640×640.

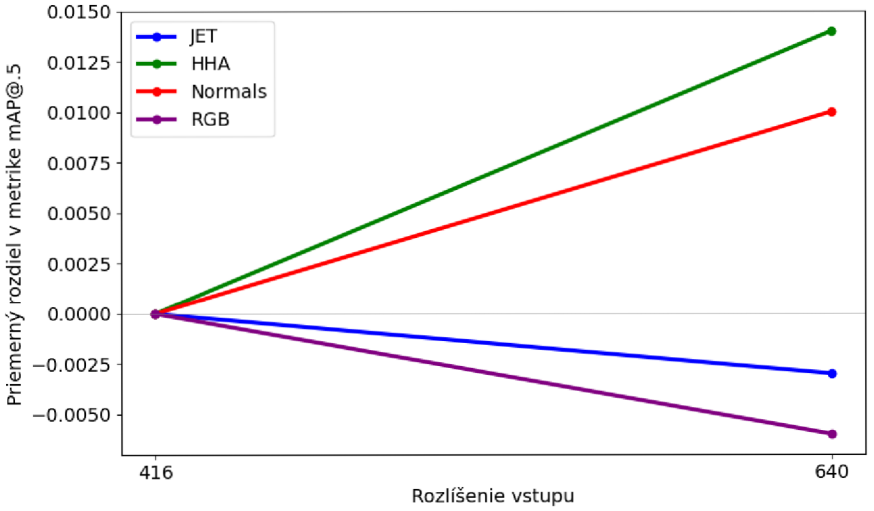
Pri analýze vplyvu zmeny rozlíšenia vstupných obrazov na presnosť detekcie objektov s použitím metódy neskorého zlúčenia v modeloch YOLOv5s a YOLOv5m možno pozorovať rozdielne trendy. Pri menšom modeli YOLOv5s zlepšenie presnosti zaznamenala metóda Normals, zatiaľ čo metóda HHA vykázala pokles. Tieto výsledky sú znázornené na obrázku 4.12.

Pre väčší model YOLOv5m, ako ukazuje obrázok 4.13, metóda HHA naopak vykázala výrazné zlepšenie, ale pokles zaznamenal základný RGB model. Výsledky naznačujú, že

pri neskorom zlúčení môže byť presnosť detekcie viac ovplyvnená charakteristikami metód spracovania obrazu a ich citlivosťou na šum a detailnosť v obrazových dátach.



Obr. 4.12: Vplyv zmeny rozlíšenia vstupných obrazov na metriku mAP@.5 menšieho modelu YOLOv5s využívajúceho metódu neskorého zlúčenia.



Obr. 4.13: Vplyv zmeny rozlíšenia vstupných obrazov na metriku mAP@.5 väčšieho modelu YOLOv5m využívajúceho metódu neskorého zlúčenia.

Porovnanie výsledkov metód skorého a neskorého zlúčenia

Táto časť sa venuje porovnaniu výsledkov skorej a neskorej fúzie z predchádzajúcich experimentov. Cieľom je analyzovať a vyhodnotiť, ako rôzne prístupy fúzie ovplyvňujú presnosť detekcie objektov v kontexte rôznych reprezentácií hĺbkových dát. Zameriavame sa na met-

riku mAP@.5, ktorá bola vyhodnotená pre rôzne kombinácie rozlíšení, modelov a metód fúzie. Výpočet priemerných hodnôt mAP@.5 a ich rozdielov oproti základnej hodnote RGB umožňuje hlbšie pochopenie vplyvu skorej a neskorej fúzie na presnosť detekcie.

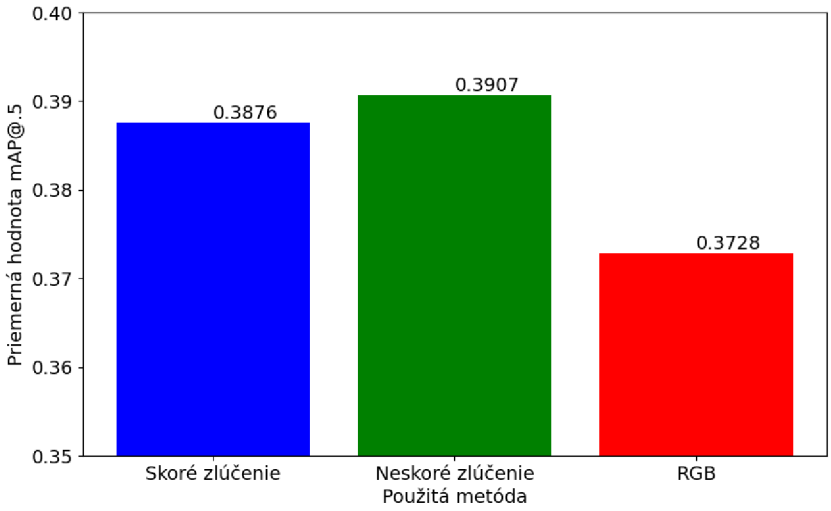
Výsledky mAP@.5 za rôzne scenáre použitia metód fúzie, modelov a rozlíšení sú zhrnuté v tabuľke 4.15. V tejto tabuľke je možné vidieť, že neskoré zlúčenie vo väčšine prípadov dosahuje lepšie výsledky než skoré zlúčenie. Ďalšími dôležitými údajmi sú priemerné hodnoty mAP@.5 pre jednotlivé reprezentácie hĺbky a ich rozdiely oproti metóde RGB, ktoré sú prezentované v tabuľke 4.16. Pri vizualizácii týchto rozdielov pomáha obrázok 4.14, ktorý zobrazuje priemerné hodnoty mAP@.5 medzi skorým a neskorým zlúčením a základným RGB modelom. Ďalšie detaily o vplyve rôznych hĺbkových reprezentácií na výsledky detekcie sú zobrazené na obrázku 4.15 a obrázku 4.16, ktoré poskytujú porovnanie priemerných rozdielov mAP@.5 podľa rôznych hĺbkových reprezentácií a metód zlúčenia.

Vstup	**Model**	**Zlúčenie**	**Jet**	**HHA**	**Normals**	**RGB**
416	v5s	Skoré zlúčenie	0.379	0.376	0.373	0.352
		Neskoré zlúčenie	0.392	0.386	0.384	0.352
	v5m	Skoré zlúčenie	0.394	0.395	0.410	0.390
		Neskoré zlúčenie	0.414	0.400	0.394	0.390
640	v5s	Skoré zlúčenie	0.386	0.385	0.397	0.365
		Neskoré zlúčenie	0.391	0.371	0.400	0.365
	v5m	Skoré zlúčenie	0.403	0.404	0.409	0.384
		Neskoré zlúčenie	0.411	0.414	0.404	0.384

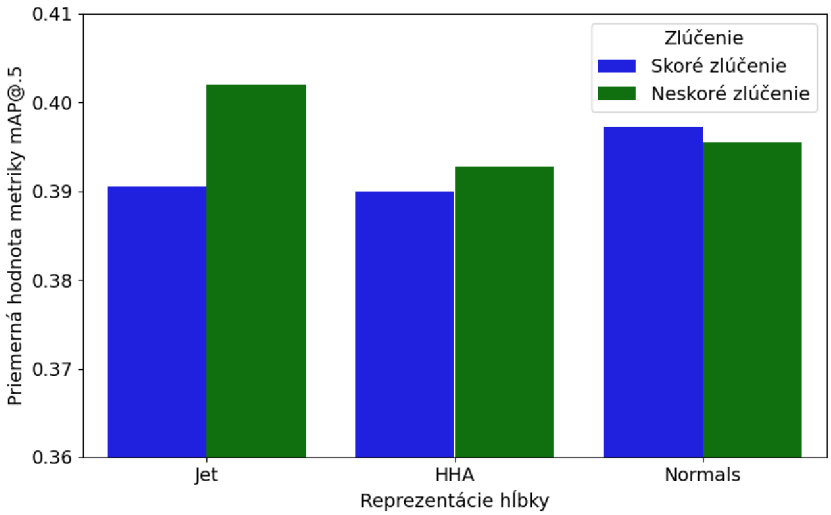
Tabuľka 4.15: Zoskupené zobrazenie metrík mAP@.5 pri použití rôznych metód fúzie podľa veľkosti vstupu a modelu s rozlíšením rôznych reprezentácií hĺbky.

Metóda	**Reprezentácia hĺbky**	**Priemerná výkonnosť**	
		mAP@.5	**Rozdiel oproti RGB**
Skoré zlúčenie	Jet	0.3905	0.01775
	HHA	0.3900	0.01725
	Normals	**0.3973**	**0.02450**
Neskoré zlúčenie	**Jet**	**0.4020**	**0.02925**
	HHA	0.3928	0.02000
	Normals	0.3955	0.02275

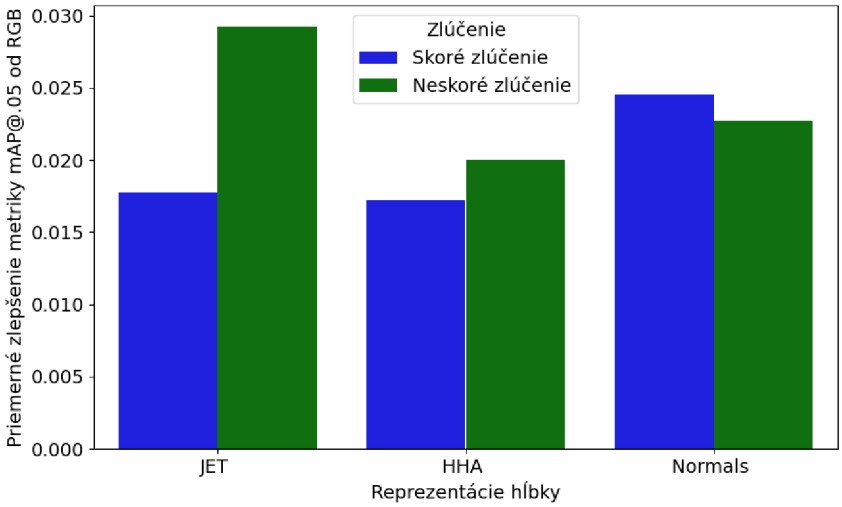
Tabuľka 4.16: Priemerná hodnota mAP@.5 a priemerný rozdiel od základnej hodnoty RGB podľa metódy zlúčenia.



Obr. 4.14: Priemerné hodnoty mAP@.5 rôznych veľkostí modelu a vstupného rozlíšenia podľa metódy zlúčenia.



Obr. 4.15: Priemerný rozdiel mAP@.5 podľa metódy zlúčenia v rôznych reprezentáciách hĺbky.

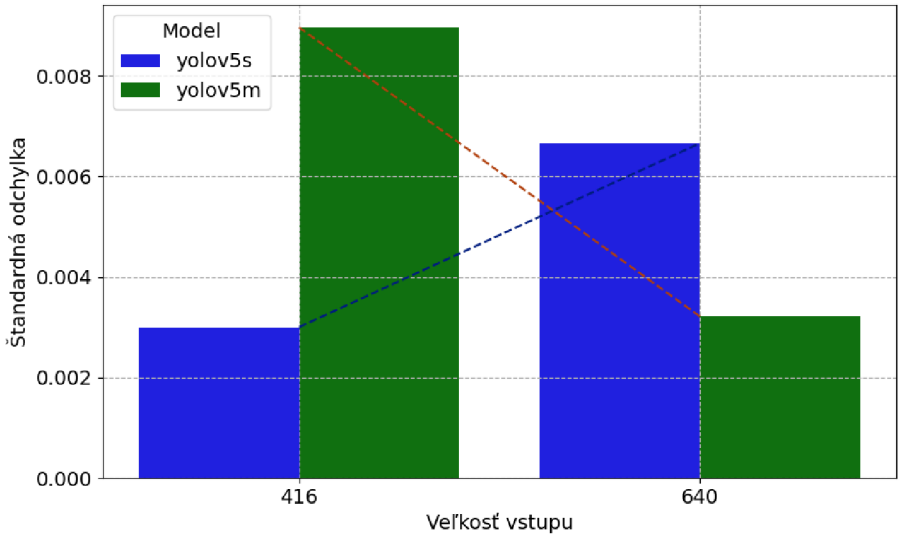


Obr. 4.16: Priemerné zlepšenie metriky mAP@.5 od základnej hodnoty RGB podľa metódy zlúčenia.

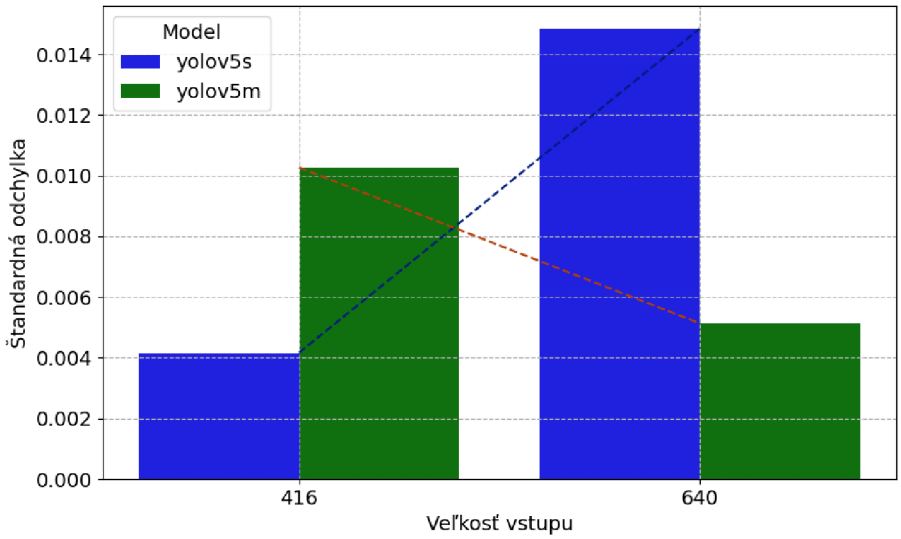
Z výsledkov analýzy jasne vyplýva, že integrácia hĺbkových dát pozitívne ovplyvňuje presnosť detekčných modelov. Najvyššia priemerná presnosť dosiahnutá v našich experimentoch bola zaznamenaná pri použití neskorej fúzie, konkrétne metódou JET, kde mAP@.5 dosiahla hodnotu 0.402. Prístup neskorej fúzie sa ukazuje ako efektívnejší v zlepšovaní výkonnosti modelov, avšak je dôležité pripomenúť, že je tiež náročnejší na výpočtový výkon.

Posledné porovnanie sa zameriava na analýzu citlivosti modelov a metód zlúčenia na rôzne reprezentácie hĺbky. Trendy vykreslené grafmi 4.17 a 4.18 ukazujú, že pri menšej verzií modelu sa s rastúcim rozlíšením vstupu zvyšuje štandardná odchýlka medzi rôznymi reprezentáciami hĺbky, čo naznačuje zvýšenú citlivosť na rozdiely v dátach. Naopak, u väčšieho modelu YOLOv5m tieto rozdiely s rastúcim rozlíšením vstupu klesajú, čo možno pripísať väčšiemu počtu parametrov a tým pádom lepšej schopnosti modelu generalizovať.

Pri porovnaní hodnôt štandardnej odchylky skorej (obr. 4.17) a neskorej fúzie (obr. 4.18) na rôzne reprezentácie hĺbky je zrejmé, že rozdiely medzi reprezentáciami hĺbky sú výraznejšie pri neskorej fúzii. Túto väčšiu citlivosť metódy možno vysvetliť tým, že zachováva pôvodné charakteristiky jednotlivých kanálov, zatiaľ čo pri skorej fúzii sa všetky reprezentácie hĺbky najskôr prevedú do odtieňov šedi, čo môže viesť k menšej premenlivosti výsledkov.



Obr. 4.17: Štandardná odchylka reprezentácií hĺbky pre modely využívajúce metódu skorého zlúčenia.



Obr. 4.18: Štandardná odchylka reprezentácií hĺbky pre modely využívajúce metódu neskorého zlúčenia.

4.2 Experiment s upravenými modelmi YOLOv5 na dátovej sade Washington RGB-D

Predchádzajúce experimenty naznačili, že integrácia hĺbkových dát do detektorov objektov sád môže zlepšiť presnosť detekcie. Cieľom tohto experimentu bolo preto overiť, či integrácia

hĺbkových dát môže zlepšiť presnosť detekcie aj na inej ako pôvodnej dátovej sade. Tento experiment preto skúma výkon upravených modelov, ktoré využívajú s metódy skorého a neskorého zlúčenia v porovnaní so základnými modelmi, ktoré hĺbkové dáta nevyužívajú. Pre tento účel bol zvolený model YOLOv5s a reprezentácie hĺbky Jet a HHA, nakoľko reprezentácia pomocou povrchových normál nebola pri tejto dátovej sade dostupná. Výkonnosť modelov sa pri tomto experimente porovnáva s využitím metriky mAP@.5:.95. Vyšší prah metriky mAP v tomto prípade lepšie znázorňuje rozdiel jednotlivých prístupov, nakoľko pri nižšom prahu mAP@.5 sa výkonnosť modelov blíži maximálnym hodnotám.

Detaily o výkonnosti metód skorého a neskorého zlúčenia sú zhrnuté v tabuľkách, kde sú predstavené celkové metriky mAP@.5:.95 ako aj metriky AP@.5:.95 pre konkrétne triedy objektov. Výsledky pre metódu skorého zlúčenia sú zhrnuté v tabuľke 4.17 a výsledky pre metódu neskorého zlúčenia sú zas obsiahnuté v tabuľke 4.18.

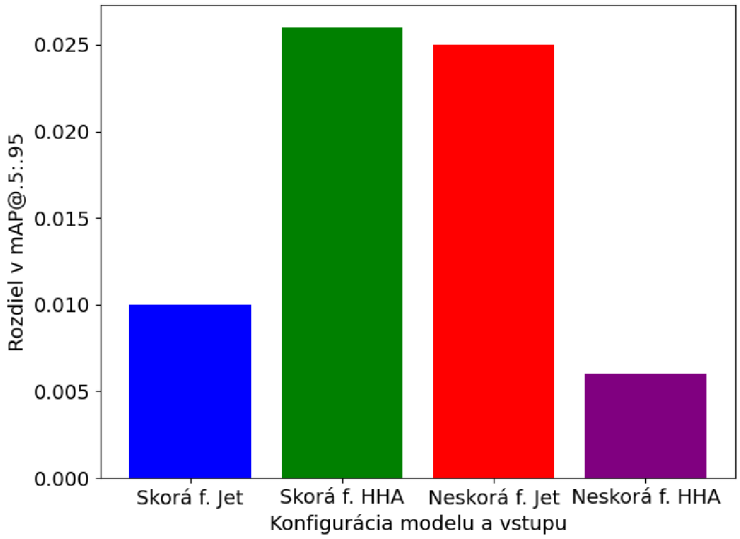
Class	**Jet**	**HHA**	**RGB**
All	0.763	**0.779**	0.753
Bowl	0.765	**0.786**	0.750
Cap	0.800	**0.822**	0.802
Cereal Box	0.828	**0.851**	0.817
Coffee Mug	0.746	**0.767**	0.731
Flashlight	**0.732**	0.728	0.715
Soda Can	0.705	**0.718**	0.702

Tabuľka 4.17: Metriky mAP@.5:.95 modelu YOLOv5s pre metódu skorého zlúčenia.

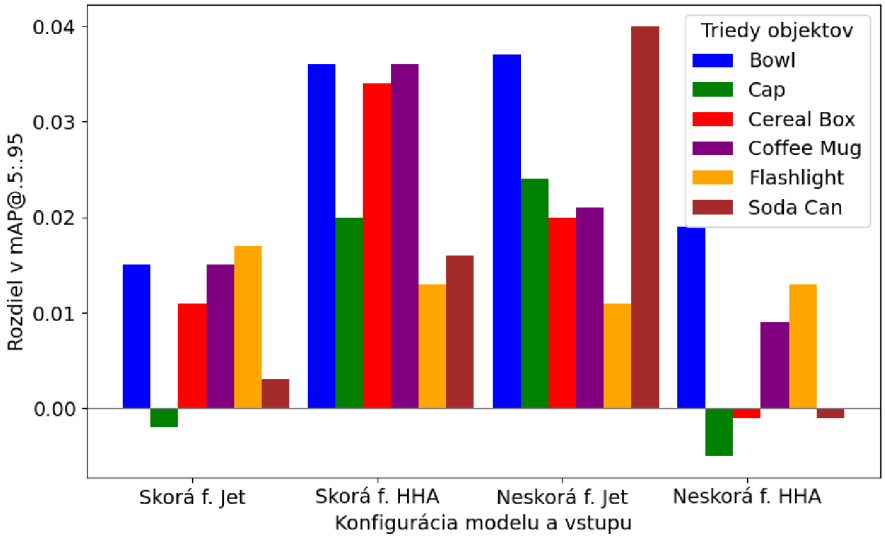
Class	**Jet**	**HHA**	**RGB**
All	**0.778**	0.759	0.753
Bowl	**0.787**	0.769	0.750
Cap	**0.826**	0.797	0.802
Cereal Box	**0.837**	0.816	0.817
Coffee Mug	**0.752**	0.740	0.731
Flashlight	0.726	**0.728**	0.715
Soda Can	**0.742**	0.701	0.702

Tabuľka 4.18: Metriky mAP@.5:.95 modelu YOLOv5s pre metódu neskorého zlúčenia.

Z analýzy tabuliek 4.17 a 4.18 vyplýva, že modely s integráciou hĺbkových dát, či už s metódou skorého alebo neskorého zlúčenia, vykazujú lepšiu výkonnosť v porovnaní s modelmi, ktoré sa spoliehajú len na RGB dáta. Toto tvrdenie potvrdzuje fakt, že priemerná metrika mAP@.95 sa aj pri odlišnej dátovej sade zvýšila pri každej variácii metódy a hĺbkových dát. Zaujímavým pozorovaním je aj fakt, že rovnako ako pri predchádzajúcom experimente aj v tomto prípade pri metóde neskorého zlúčenia s rozlíšením 640×640 mala reprezentácia HHA najhoršie výsledky. Rozdiely vo výsledkoch sú zreteľnejšie v grafoch nižšie. Obrázok 4.19 ukazuje celkové zlepšenie metriky mAP@.5:.95 pre modely využívajúce hĺbkové dáta oproti štandardnému RGB modelu. Obrázok 4.20 zas zobrazuje rozdiely metrík AP@.5:.95 naprieč triedami obsiahnutými v dátovej sade.



Obr. 4.19: Zlepšenie metriky mAP@.5:.95 modelov YOLOv5s využívajúcich hĺbkové dáta v rôznych variáciach oproti základnému RGB modelu.



Obr. 4.20: Rozdiel metrík AP@.5:.95 modelov YOLOv5s využívajúcich hĺbkové dáta v rôznych variáciach oproti základnému RGB modelu naprieč triedami.

4.3 Experiment s pridaním hĺbkových dát do modelu YOLOv8

V rámci predošlých experimentov bolo zistené, že integrácia hĺbkových dát pomocou metód skorého a neskorého zlúčenia zlepšila presnosť detekcie modelov YOLOv5 na dátových

sadách NYUv2 a Washington RGBD. Na základe týchto pozitívnych výsledkov bol realizovaný posledný experiment, ktorý sa zameriava na overenie efektívnosti využitia hĺbkových dát na alternatívnom modeli - upravenom YOLOv8, ktorý využíva štyri kanály a teda predstavuje metódu skorého zlúčenia, ktorá bola bližšie opísaná v časti 3.2. V experimente boli použité modely YOLOv8n, YOLOv8s a YOLOv8m natrénované na dátovej sade NYU, pričom boli hodnotené výkonnostné metriky mAP@.5 a mAP@.5:.95. Detaily a špecifické výsledky tohto experimentu sú zhrnuté v tabuľkách 4.19 a 4.20.

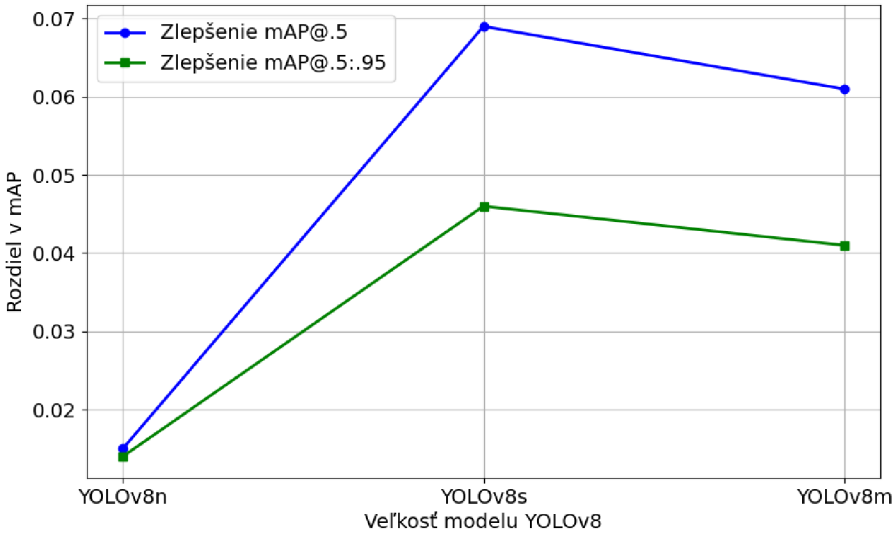
Veľkosť modelu	**RGBD**	**RGB**
YOLOv8n	0.352	0.337
YOLOv8s	0.405	0.336
YOLOv8m	0.423	0.362

Tabuľka 4.19: Metriky mAP@.5 pre model využívajúci hĺbkové dáta a základný RGB model naprieč rôznymi veľkosťami modelu YOLOv8.

Veľkosť modelu	**RGBD**	**RGB**
YOLOv8n	0.211	0.197
YOLOv8s	0.249	0.203
YOLOv8m	0.275	0.234

Tabuľka 4.20: Metriky mAP@.5:.95 pre model využívajúci hĺbkové dáta a základný RGB model naprieč rôznymi veľkosťami modelu YOLOv8.

Výsledky experimentu jasne demonštrujú, že integrácia hĺbkových dát do modelov YOLOv8 priniesla výrazné zlepšenie obidvoch výkonnostných metrík. Najvýraznejšie zlepšenie metrík bolo zaznamenané pri modeloch YOLOv8s a YOLOv8m a je vizuálne zobrazené na obrázku 4.21 nižšie.



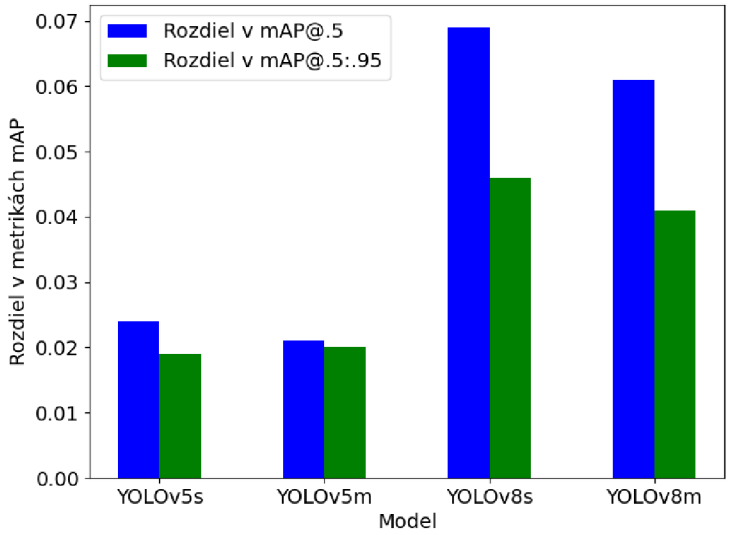
Obr. 4.21: Zlepšenie metrík mAP pri využití hĺbkových dát oproti základným RGB modelom YOLOv8.

Zlepšenia zaznamenané v modeloch YOLOv8 boli následne porovnané so zlepšeniami, ktoré boli dosiahnuté pri modeli YOLOv5 využívajúcom skorú fúziu. Toto porovnanie je možné nájsť v tabuľke 4.21.

Model	**Zlepšenie mAP@.5**	**Zlepšenie mAP@.5:.95**
YOLOv5s	0.024	0.019
YOLOv5m	0.021	0.020
YOLOv8s	0.069	0.046
YOLOv8m	0.061	0.041

Tabuľka 4.21: Rozdiely v metrikách mAP@.5 a mAP@.5:.95 medzi YOLO modelmi využívajúcimi hĺbkové dáta a základnými RGB modelmi.

Toto porovnanie ukázalo, že pri modeloch YOLOv8 dochádza k zlepšeniu presnosti vďaka dostupným hĺbkovým dátam oveľa viac ako pri modeloch YOLOv5. Zlepšenie presnosti, ktoré hĺbkové dáta poskytujú, je nielen konzistentné, ale aj výrazné, čo naznačuje potenciál hĺbkovej integrácie pre budúce generácie objektových detektorov. Tieto zistenia sú zobrazené na obrázku 4.22, ktorý porovnáva zlepšenia metrík mAP pri využití hĺbkových dát v oboch modelových radách.



Obr. 4.22: Zlepšenie metrík mAP pri využití hĺbkových dát oproti základným RGB modelom YOLOv5 a YOLOv8.

Tento experiment podporuje predchádzajúce zistenia o význame hĺbkových dát pre zvýšenie presnosti detekčných modelov a zdôrazňuje potenciál metódy skorej fúzie pre vylepšenie výkonu štandardných detektorov objektov.

Kapitola 5

Záver

Táto práca poskytuje komplexný pohľad na detekciu objektov v obraze s využitím hĺbkových informácií. V úvode bola zmienená dôležitosť integrácie hĺbkových dát do procesov detekcie a rozpoznávania objektov. Hĺbkové dáta poskytujú rozsiahle dodatočné informácie o vzdialenostiach objektov od senzora a ich polohách v priestore, čo by malo viesť k lepšej efektivite detektorov. Práca sa následne zameriava na viaceré koncepty spracovania obrazu s dôrazom na hĺbkové dáta, ich reprezentáciu formou point cloudových dát či hĺbkových máp a spôsoby ich získavania pomocou špecializovaných senzorov. V tomto kontexte bol poskytnutý prehľad o doterajšom výskume rôznych metód vhodných na spracovanie point cloudových dát a RGB-D obrazov. Predstavené boli rôzne prístupy zlúčenia farebného obrazu a hĺbky, rôzne metódy reprezentácie hĺbky v hĺbkových mapách a tiež niektoré výzvy spojené so spracovaním hĺbkových dát. Teoretickú časť tejto práce uzatvára prehľad o jednokrokových detektoroch objektov, s dôrazom na detektory série YOLO vo verziách YOLOv5 a YOLOv8. Na základe získaných poznatkov sa následne práca zaoberá návrhom riešenia. V tejto časti sú opísané konkrétne metódy, implementácie algoritmov, metriky a úpravy modelov ktoré boli využité v experimentálnej časti. Ako zdroje hĺbkových dát potrebných pre výskum boli predstavené dve dátové sady, NYU Depth Dataset V2 a Washington RGB-D. Práca následne prechádza od teoretickej časti na časť experimentálnu, ktorej cieľom bolo overiť efektivitu využitia hĺbkových dát pri detekcií objektov s použitím rôznych metód.

Prvý experiment bol zameraný porovnanie dvoch prístupov k integrácií hĺbkových dát do modelov YOLOv5. Dva rozličné prístupy predstavovali metódy skorej a neskorej fúzie. Analýza výsledkov preukázala, že neskorá fúzia, pri ktorej sa hĺbkové a RGB dáta spracúvajú nezávisle a kombinujú až v neskorších vrstvách modelu, sa ukázala ako efektívnejšia. Toto zistenie môže byť dôsledkom toho, že tento prístup zdvojnásobuje počet parametrov modelu, čo umožňuje modelu efektívnejšie kombinovať informácie z rôznych druhov dát. Zvýšenie počtu parametrov modelu však výrazne zvyšuje výpočetné nároky na tréning aj inferenciu. Z analýzy týchto zistení vyplýva, že metóda neskorého zlúčenia môže nájsť praktické uplatnenie pri riešeniach, pri ktorých sa kladie najväčší dôraz na presnosť detekcie a pri ktorých nie je dôležité spracovávať dáta v reálnom čase s využitím obmedzených výpočetných zdrojov.

Na druhej strane, skorá fúzia, kde sú hĺbkové dáta integrované priamo k RGB kanálom ešte pred spracovaním modelom, ponúkla taktiež uspokojivé výsledky, hoci o niečo nižšie než pri neskorej fúzii. Tento prístup vyniká najmä svojou nižšou výpočtovou náročnosťou,

čo ho robí ideálnym pre praktické aplikácie s obmedzenými zdrojmi alebo s požiadavkou spracovávať dáta v reálnom čase.

Experiment tiež poukázal na fakt, že rozdiely v efektívnosti medzi rôznymi hĺbkovými reprezentáciami sa znižujú s rastúcou veľkosťou modelov YOLOv5. Pri menších modeloch, ako YOLOv5s, boli rozdiely medzi jednotlivými reprezentáciami hĺbky výraznejšie, zatiaľ čo pri väčšom modeli YOLOv5m boli tieto rozdiely menej výrazné. Toto zistenie môže svedčiť o tom, že väčšie modely majú lepšiu schopnosť generalizácie a môžu efektívnejšie využívať dostupné informácie z hĺbkových dát bez ohľadu na ich špecifickú formu.

Cieľom druhého bolo zistiť, či pozitívne výsledky získané na dátovej sade NYU Depth v2 možno replikovať v iných podmienkach a kontextoch. Experiment potvrdil, že pridanie hĺbkových dát do modelov YOLOv5 zlepšuje presnosť detekcie v porovnaní so štandardnými RGB modelmi. Zaujímavosťou pri tomto experimente bol opakujúci sa trend: reprezentácia HHA s využitím metódy neskorej fúzie pri modeli YOLOv5s a rozlíšení 640×640 dosiahla podobne ako v prípade prvej dátovej sady najhorší výsledok. Tento opakovaný jav naznačil, že menšia varianta modelu má problém pri spracovaní dát, ktoré táto reprezentácia dát poskytuje.

Záverečný experiment sa zaoberal overením účinnosti integrácie hĺbkových dát do moderných detekčných modelov, konkrétne na architektúre YOLOv8. Cieľom bolo zistiť, či je možné replikovať dosiahnuté výsledky aj na iných modeloch, a tým overiť univerzálnosť predchádzajúcich zistení. Experimenty s upravenými modelmi YOLOv8, ktoré využívali hĺbkové dáta ukázali výrazné zlepšenie presnosti oproti základným modelom, ktoré tieto dáta nevyužívali. Najvýraznejšie zlepšenie nastalo pri veľkostiach S a M. Modernejšie modely YOLOv8 tiež vykázali výrazne vyššiu efektivitu pri využívaní hĺbkových dát v porovnaní s rovnako veľkými modelmi YOLOv5. Toto zistenie naznačuje, že moderné architektúry modelovej rady YOLO sú pri integrácií hĺbkových dát efektívnejšie ako ich predchodcovia.

Integrácia hĺbkových dát do procesov detekcie objektov zvýšila presnosť detekcie v každom z realizovaných experimentov. Výsledky rôznych metód sa síce líšia ale zlepšenie, ktoré prináša pridanie hĺbkových dát je konzistentné naprieč rôznymi modelmi a metódami. Na základe tohto môžme usúdiť, že hĺbkové dáta môžu byť pre oblasť počítačového spracovania obrazu veľkým prínosom. Preto by som chcel spomenúť niekoľko nápadov, ktorými sa môžu zaoberať budúce výskumy v tejto oblasti.

Súčasné modely, ako YOLO a iné varianty, sú primárne navrhnuté pre spracovanie dvojrozmerných obrazov. Výskum by sa mohol zamerať na vývoj nových architektúr alebo adaptácie existujúcich, ktoré by od základu integrovali hĺbkové dáta. Tieto modely by mohli využívať hĺbkovú informáciu nielen ako doplnok, ale ako základný stavebný prvok, čo by mohlo umožniť hlbšie pochopenie scény a lepšiu lokalizáciu objektov v priestore. Ďalej by bolo vhodné hlbšie preskúmať metódy zlúčenia príznakov z hĺbkových a RGB dát. Súčasné prístupy ako skorá a neskorá fúzia sú efektívne, no existuje potenciál pre vývoj adaptívnych fúznych metód. Tieto by mohli dynamicky meniť spôsob, akým sa dáta kombinujú, na základe kontextu alebo špecifických charakteristík scény. Napríklad, systém by mohol používať hĺbkové dáta intenzívnejšie v situáciách, kde RGB dáta neposkytujú dostatok informácií (napríklad pri splývaní objektu s pozadím), a naopak zameriavať sa viac na RGB dáta, keď hĺbkové dáta nie sú dostatočne presné kvôli odrazenému svetlu alebo iným faktorom. Prínos by mohlo priniesť aj preskúmanie možného využitia metódy preneseného učenia pre modely pracujúce s hĺbkou.

V závere možno povedať, že integrácia hĺbkových dát do detekčných systémov predstavuje zásadný pokrok v oblasti počítačového videnia. Táto práca poskytla cenné základy,

ktoré ukazujú výrazný potenciál pre ďalší rozvoj a praktické aplikácie. Čoraz lepšie porozumenie a využívanie hĺbkových dát môže významne prispieť k presnosti a efektivite moderných detekčných systémov. S budúcim vývojom v tejto oblasti môžeme očakávať inovácie, ktoré transformujú nielen akademický výskum, ale aj širokú škálu aplikácií v priemysle a každodennom živote. Táto oblasť nepochybne stojí na prahu významných objavov a zlepšení, ktoré otvoria nové možnosti v technológiách strojového videnia.

Literatúra

[1] Aakerberg, A., Nasrollahi, K., Rasmussen, C. B. a Moeslund, T. B. Depth value pre-processing for accurate transfer learning based RGB-D object recognition. In: SCITEPRESS Digital Library. *International Joint Conference on Computational Intelligence*. 2017, s. 121–128.

[2] Bochkovskiy, A., Wang, C. a Liao, H. M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *CoRR*. 2020, abs/2004.10934. Dostupné z: https://arxiv.org/abs/2004.10934.

[3] Cao, Y., Shen, C. a Shen, H. T. Exploiting depth from single monocular images for object detection and semantic segmentation. *IEEE Transactions on Image Processing*. IEEE. 2016, zv. 26, č. 2, s. 836–846.

[4] Cetinkaya, B., Kalkan, S. a Akbas, E. *Does depth estimation help object detection?* 2022.

[5] Chen, H. a Li, Y. Progressively Complementarity-Aware Fusion Network for RGB-D Salient Object Detection. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, s. 3051–3060. DOI: 10.1109/CVPR.2018.00322.

[6] Chen, X. *Depth2HHA-python: Converting depth maps to HHA encodings* [https://github.com/charlesCXK/Depth2HHA-python]. 2018.

[7] Chen, X., Ma, H., Wan, J., Li, B. a Xia, T. Multi-view 3d object detection network for autonomous driving. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2017, s. 1907–1915.

[8] Diaz, A. *5 Aspects of Fully Convolutional Networks for Semantic Segmentation* [online]. [cit. 2023-01-17]. Dostupné z: https://www.marscrowd.com/blog/text/fully-convolutional-networks-for-semantic-segmentation/.

[9] Ding, X., Li, B. a Wang, J. Geometric property-based convolutional neural network for indoor object detection. *International Journal of Advanced Robotic Systems*. 2021, zv. 18, č. 1, s. 1729881421993323. DOI: 10.1177/1729881421993323. Dostupné z: https://doi.org/10.1177/1729881421993323.

[10] Dobo, C. a Benyó, B. Intelligent Assisting Tools for Endodontic Treatment. In:. Marec 2019. DOI: 10.5772/intechopen.84900.

[11] Du, H., Henry, P., Ren, X., Cheng, M., Goldman, D. et al. RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. In: *Proceedings of the 13th international conference on Ubiquitous computing, Beijing, China*. 2011.

[12] Dwivedi, P. *Semantic Segmentation — Popular Architectures* [online]. [cit. 2023-03-16]. Dostupné z: https://towardsdatascience.com/semantic-segmentation-popular-architectures-dff0a75f39d0.

[13] Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M. A. a Burgard, W. Multimodal Deep Learning for Robust RGB-D Object Recognition. *CoRR*. 2015, abs/1507.06821. Dostupné z: http://arxiv.org/abs/1507.06821.

[14] Farhadi, A. a Redmon, J. Yolov3: An incremental improvement. In: Springer Berlin/Heidelberg, Germany. *Computer vision and pattern recognition*. 2018, sv. 1804, s. 1–6.

[15] Gonzalez, R. C. a Woods, R. E. *Digital image processing (3rd Edition)*. Upper Saddle River, N.J.: Pearson Education, 2007. ISBN 978-0131687288.

[16] Grilli, E., Menna, F. a Remondino, F. A review of point clouds segmentation and classification algorithms. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Copernicus GmbH. 2017, zv. 42, s. 339–344.

[17] Gupta, S., Arbeláez, P. a Malik, J. Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 2013, s. 564–571. DOI: 10.1109/CVPR.2013.79.

[18] Gupta, S., Girshick, R., Arbeláez, P. a Malik, J. *Learning Rich Features from RGB-D Images for Object Detection and Segmentation*. 2014.

[19] Han, J., Zhang, D., Cheng, G., Liu, N. a Xu, D. Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection: A Survey. *IEEE Signal Processing Magazine*. 2018, zv. 35, č. 1, s. 84–100. DOI: 10.1109/MSP.2017.2749125.

[20] He, K., Zhang, X., Ren, S. a Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*. IEEE. 2015, zv. 37, č. 9, s. 1904–1916.

[21] Hou, Q., Cheng, M., Hu, X., Borji, A., Tu, Z. et al. Deeply supervised salient object detection with short connections. *CoRR*. 2016, abs/1611.04849. Dostupné z: http://arxiv.org/abs/1611.04849.

[22] Hou, S., Wang, Z. a Wu, F. Object detection via deeply exploiting depth information. *Neurocomputing*. Elsevier. 2018, zv. 286, s. 58–66.

[23] Huang, L., Chen, C., Yun, J., Sun, Y., Tian, J. et al. Multi-Scale Feature Fusion Convolutional Neural Network for Indoor Small Target Detection. *Frontiers in Neurorobotics*. 2022, zv. 16. DOI: 10.3389/fnbot.2022.881021. ISSN 1662-5218. Dostupné z: https://www.frontiersin.org/articles/10.3389/fnbot.2022.881021.

[24] Hui, J. *MAP (mean Average Precision) for Object Detection* [online]. 2018 [cit. 2023-01-19]. Dostupné z: https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173.

[25] Hulstaert, L. *A Beginner's Guide to Object Detection* [online]. [cit. 2023-01-17]. Dostupné z: https://www.datacamp.com/tutorial/object-detection-guide.

[26] Ian Goodfellow, A. C. *Deep Learning.* MIT Press, 2016. ISBN 978-0-262-03561-3.

[27] Jin, L., Gao, S., Li, Z. a Tang, J. Hand-Crafted Features or Machine Learnt Features? Together They Improve RGB-D Object Recognition. *Proceedings - 2014 IEEE International Symposium on Multimedia, ISM 2014.* Február 2015, s. 311–319. DOI: 10.1109/ISM.2014.56.

[28] Jocher, G. *YOLOv5 by Ultralytics.* 2020. Dostupné z: https://github.com/ultralytics/yolov5.

[29] Jocher, G., Chaurasia, A. a Qiu, J. *YOLOv8 by Ultralytics.* Január 2023. Dostupné z: https://github.com/ultralytics/ultralytics.

[30] Ju, R.-Y. a Cai, W. *Fracture Detection in Pediatric Wrist Trauma X-ray Images Using YOLOv8 Algorithm.* 2023.

[31] Katsamenis, I., Karolou, E., Davradou, A., Protopapadakis, E., Doulamis, A. et al. *TraCon: A novel dataset for real-time traffic cones detection using deep learning.* Máj 2022. DOI: 10.48550/arXiv.2205.11830.

[32] Kazhdan, M., Bolitho, M. a Hoppe, H. Poisson surface reconstruction. In: *Proceedings of the fourth Eurographics symposium on Geometry processing.* 2006, sv. 7, č. 4.

[33] Ladickỳ, L., Zeisl, B. a Pollefeys, M. Discriminatively trained dense surface normal estimation. In: Springer. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13.* 2014, s. 468–484.

[34] Lai, K., Bo, L. a Fox, D. Unsupervised feature learning for 3d scene labeling. In: IEEE. *2014 IEEE International Conference on Robotics and Automation (ICRA).* 2014, s. 3050–3057.

[35] Lai, K., Bo, L., Ren, X. a Fox, D. A large-scale hierarchical multi-view rgb-d object dataset. In: IEEE. *2011 IEEE international conference on robotics and automation.* 2011, s. 1817–1824.

[36] Levin, A., Lischinski, D. a Weiss, Y. Colorization using optimization. In: *ACM SIGGRAPH 2004 Papers.* 2004, s. 689–694.

[37] Liang, M., Yang, B., Wang, S. a Urtasun, R. Deep continuous fusion for multi-sensor 3d object detection. In: *Proceedings of the European conference on computer vision (ECCV).* 2018, s. 641–656.

[38] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. et al. Feature pyramid networks for object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2017, s. 2117–2125.

[39] Liu, S., Zhang, M., Kadam, P. a Kuo, C.-C. *3D Point Cloud Analysis: Traditional, Deep Learning, and Explainable Machine Learning Methods.* Január 2021. ISBN 978-3-030-89179-4.

[40] Liu, S., Qi, L., Qin, H., Shi, J. a Jia, J. Path aggregation network for instance segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, s. 8759–8768.

[41] Mcmanamon, P. *LiDAR Technologies and Systems*. Júl 2019. ISBN 9781510625396.

[42] Menze, M. a Geiger, A. Object scene flow for autonomous vehicles. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, s. 3061–3070. DOI: 10.1109/CVPR.2015.7298925.

[43] Mesika, A., Ben Shabat, Y. a Tal, A. CloudWalker: Random walks for 3D point cloud shape analysis. *Computers and Graphics*. Elsevier BV. August 2022, zv. 106, s. 110–118. DOI: 10.1016/j.cag.2022.06.001. ISSN 0097-8493. Dostupné z: http://dx.doi.org/10.1016/j.cag.2022.06.001.

[44] Nygård, I. a Vittersø, S. *Improved Sheep Detection - Modifying YOLOv5 to accurately detect grazing sheep in UAV imagery*. NTNU, 2022.

[45] Ophoff, T., Van Beeck, K. a Goedemé, T. Exploring RGB+Depth Fusion for Real-Time Object Detection. *Sensors*. 2019, zv. 19, č. 4. DOI: 10.3390/s19040866. ISSN 1424-8220. Dostupné z: https://www.mdpi.com/1424-8220/19/4/866.

[46] Pan, X., Xia, Z., Song, S., Li, L. E. a Huang, G. 3d object detection with pointformer. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, s. 7463–7472.

[47] Pomerleau, F., Colas, F. a Siegwart, R. 2015.

[48] Qi, C. R., Zhou, Y., Najibi, M., Sun, P., Vo, K. et al. Offboard 3d object detection from point cloud sequences. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, s. 6134–6144.

[49] Rafique, A. A., Jalal, A. a Kim, K. Statistical multi-objects segmentation for indoor/outdoor scene detection and classification via depth images. In: IEEE. *2020 17th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*. 2020, s. 271–276.

[50] Rahman, M. M., Tan, Y., Xue, J. a Lu, K. RGB-D object recognition with multimodal deep convolutional neural networks. In: IEEE. *2017 IEEE International Conference on Multimedia and Expo (ICME)*. 2017, s. 991–996.

[51] Reading, C., Harakeh, A., Chae, J. a Waslander, S. L. Categorical depth distribution network for monocular 3d object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, s. 8555–8564.

[52] Redmon, J., Divvala, S. K., Girshick, R. B. a Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *CoRR*. 2015, abs/1506.02640. Dostupné z: http://arxiv.org/abs/1506.02640.

[53] Richard, S. *Computer Vision: Algorithms and Applications*. Springer London, Limited, 2010. ISBN 9781848829343.

[54] Shi, Y., Guo, Y., Mi, Z. a Li, X. Stereo CenterNet-based 3D object detection for autonomous driving. *Neurocomputing*. Elsevier. 2022, zv. 471, s. 219–229.

[55] Sikudova, E., Cernekova, Z., Benesova, W., Berger Haladova, Z. a Kucerova, J. *Počítačové videnie: Detekcia a rozpoznavanie objektov. - 1. vydanie.* Január 2014. 268 s. ISBN 978-80-87925-06-5.

[56] Silberman, N., Hoiem, D., Kohli, P. a Fergus, R. Indoor segmentation and support inference from rgbd images. In: Springer. *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*. 2012, s. 746–760.

[57] Song, S., Lichtenberg, S. P. a Xiao, J. SUN RGB-D: A RGB-D scene understanding benchmark suite. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, s. 567–576. DOI: 10.1109/CVPR.2015.7298655.

[58] Song, S. a Xiao, J. *Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images*. 2016.

[59] Tian, H., Chen, Y., Dai, J., Zhang, Z. a Zhu, X. Unsupervised object detection with lidar clues. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, s. 5962–5972.

[60] Wang, A., Cai, J., Lu, J. a Cham, T.-J. MMSS: Multi-modal Sharable and Specific Feature Learning for RGB-D Object Recognition. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, s. 1125–1133. DOI: 10.1109/ICCV.2015.134.

[61] Wang, A., Cai, J., Lu, J. a Cham, T.-J. Modality and component aware feature fusion for RGB-D scene classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, s. 5995–6004.

[62] Wang, C.-Y., Liao, H.-Y. M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W. et al. CSPNet: A new backbone that can enhance learning capability of CNN. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, s. 390–391.

[63] Wang, Y. a Ye, J. An overview of 3d object detection. *ArXiv preprint arXiv:2010.15614*. 2020.

[64] Ye, E. S. a Malik, J. Object detection in rgb-d indoor scenes. *Univ. California Berkeley, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2013–3*. 2013.

[65] Zhang, Y., Yin, M., Wang, H. a Hua, C. Cross-level multi-modal features learning with transformer for rgb-d object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*. IEEE. 2023.

[66] Zhao, Z.-Q., Zheng, P., Xu, S.-T. a Wu, X. Object Detection With Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*. 2019, zv. 30, č. 11, s. 3212–3232.

Príloha A

Obsah priloženej SD karty

Priložená SD karta obsahuje:

- datasets/ - priečinok s dátovými sadami
- models/ - priečinok s natrénovanými modelmi
- src/ - priečinok s pomocnými scriptami
- RGBD_YOLOv8/ - priečinok s modelom YOLOv8 schopným spracovávať štvorkanálový vstup
- RGBD_YOLOv5/ - priečinok s využitými modelmi YOLOv5, ktoré využívajú metódu skorého a neskorého zlúčenia
- tex/ - priečinok so zdrojovými súbormi textovej časti práce
- bp.pdf - textová časť práce