Monokularna Estimacija Dubine

Lukić, Matea

SVEUČILIŠTE U SPLITU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET U SPLITU

DIPLOMSKI RAD

Monokularna Estimacija Dubine

Matea Lukić
Mentor: *Dr.sc. Saša Mladenović*

Split, rujan 2024.

Temeljna dokumentacijska kartica

Diplomski rad
Sveučilište u Splitu
Prirodoslovno-matematički fakultet
Odjel za matematiku
Ruđera Boškovića 33, 21000 Split, Hrvatska

Monokularna estimacija dubine

Matea Lukić

SAŽETAK
Monokularna estimacija dubine omogućuje dobivanje dubinske mape iz jedne RGB slike, procjenjujući dubinu svakog piksela, koja predstavlja udaljenost subjekta od kamere. Ovaj diplomski rad temelji se na projektu optimizacije transporta kutija, gdje je bilo potrebno automatizirati određivanje oblika zaštitnog materijala putem strojnog učenja i računalnog vida, pri čemu je estimacija dubine bila ključni međukorak.

Rad opisuje analizu podataka za procjenu dubine te probleme u radu s takvim skupom podataka, predlažući normalizaciju kroz detekciju ključnih točaka. Koriste se alati poput PyCharm-a, TensorFlow-a, i OpenCV-a.

Monokularna estimacija dubine temeljena na dubokom učenju odabrana je kao najprikladniji pristup. Razvijen je model neuronske mreže koji procjenjuje dubinsku mapu uz korištenje funkcija gubitka poput SSIM-a i L1 loss-a. Evaluacija modela pokazuje točnost od 88%, uz mogućnost poboljšanja kroz preneseno učenje i finu prilagodbu parametara.

Zaključuje se da je procjena dubine ključna za razvoj softvera za optimizaciju transporta kutija.

Ključne riječi:

- Monokularna estimacija dubine
- Struktura iz pokreta

- Podudaranje stereo vizije
- Duboko učenje
- Konvolucijska neuronska mreža
- GAN
- Dubinska mapa
- Temeljna istina
- Predikcija i rekonstrukcija
- Time of Flight (ToF) kamera
- Funkcija gubitka
- Konvolucijski sloj
- Konvolucija
- Filter (kernel)
- Normalizacija podataka
- Distorzija (zakrivljenost)
- Skala boja
- Detekcija ključnih točaka
- SSIM metoda za mjerenje sličnosti dvaju slika
- L1 funkcija gubitka
- Depth Smoothness funkcija gubitka (Funkcija gubitka glatkoće dubine)
- Konvergencija
- Lokalni/Globalni minimum
- Konveksnost
- Mean, varijanca, kovarijanca, standardna devijacija
- Svjetlina
- Kontrast
- Struktura
- Hessian matrica
- Gradijentni spust
- Optimizator (Adam)

– Evaluacija modela

Rad je pohranjen u knjižnici Prirodoslovno-matematičkog fakulteta, Sveučilišta u Splitu.
Rad sadrži: 56 stranica, 21 grafičkih prikaza, 3 tablice i 9 literaturnih navoda.
Izvornik je na hrvatskom jeziku.

Mentor: Dr. sc. Saša Mladenović, redoviti profesor

Ocjenjivači:
Dr. sc. Saša Mladenović, redoviti profesor
Dr. sc. Milica Klaričić-Bakula, redoviti profesor
Dr. sc. Jurica Perić, redoviti profesor

Rad prihvaćen: rujan 2024.

Basic Documentation Card

Master Thesis
University of Split
Faculty of Science
Department of Mathematics
Ruđera Boškovića 33, 21000 Split, Croatia

Monocular Depth Estimation

Matea Lukić

ABSTRACT

Here is the translation:

Monocular depth estimation enables the creation of a depth map from a single RGB image by estimating the depth of each pixel, which represents the distance of the subject from the camera. This thesis is based on a project aimed at optimizing the transport of boxes, where it was necessary to automate the determination of the protective material's shape using machine learning and computer vision, with depth estimation being a crucial intermediate step.

The thesis describes the analysis of data used for depth estimation and the challenges encountered when working with such datasets, proposing normalization through keypoint detection. Tools such as PyCharm, TensorFlow, and OpenCV are utilized.

Monocular depth estimation based on deep learning was chosen as the most suitable approach. A neural network model was developed to estimate the depth map using loss functions like SSIM and L1 loss. The model evaluation showed an accuracy of 88%, with potential for improvement through transfer learning and fine-tuning of parameters.

It is concluded that depth estimation is essential for developing software to optimize box transport.

Key words:

- Monocular depth estimation
- Structure from motion
- Stereo vision matching
- Deep learning
- Convolutional neural network
- GAN
- Depth map
- Ground truth
- Prediction and reconstruction
- Time of Flight (ToF) camera
- Loss function
- Convolutional layer
- Convolution

- Filter (kernel)
- Data normalization
- Distortion (curvature)
- Color scale
- Keypoint detection
- SSIM method for measuring similarity between two images
- L1 loss function
- Depth smoothness loss function
- Convergence
- Local/Global minimum
- Convexity
- Mean, variance, covariance, standard deviation
- Brightness
- Contrast
- Structure
- Hessian matrix
- Gradient descent
- Optimizer (Adam)
- Model evaluation

Thesis deposited in the library of Faculty of Science, University of Split.
Thesis consists of: 56 pages, 21 figures, 3 tables, and 9 references.
Original language: Croatian.

Mentor: Saša Mladenović, Ph.D., Professor

Reviewers:
Saša Mladenović, Ph.D., Professor

Milica Klaričić- Bakula, Ph.D., Professor
Jurica Perić, Ph.D., Professor

Thesis accepted: September, 2024.

SADRŽAJ

1. Uvod

Estimacija dubine služi kako bi se konstruirala 3D scena iz jednostavnih 2D slika. Dubina je u ovom slučaju predočena bojama, točnije skalom boja. Najjednostavnija takva skala koju možemo zamisliti je ona koja ide od potpuno bijele boje, preko sive do crne. Svakoj nijansi boje možemo pridružiti određenu udaljenost od objektiva kamere te na taj način dobivamo ideju o trodimenzinalnoj reprezentaciji dvodimenzionalne slike. Na primjer, na slici ispod je helikopter, a u pozadini su planine. Kako je helikopter bliže objektivu kamere nego planine, na dubinskoj mapi je prikazan svjetlije sivom bojom, dok su planine nešto tamnije, a nebo ide prema crnoj boji. Dakle, vrijednost 255 na skali predstavlja bijelu boju i najmanju udaljenost od objektiva, odnosno 0 metara. Vrijednost 000 predstavlja crnu boju i najveću udaljenost koju definiramo s obzirom na promatranu scenu.

Slika 1.1: Originalna 2D slika i odgovarajuća dubinska mapa
izvor slike : paperswithcode.com

Cilj monokularne estimacije dubine je procijeniti vrijednost dubine svakog piksela na jednoj (monokularnoj) RGB slici. Odnosno, za ulaz koji je obična RGB slika, ciljani izlaz je odgovarajuća dubinska mapa.

Aplikacija dobivenih dubinskih mapa je veoma široka. Često se koriste pri 3D rekonstrukciji scene, a susrećemo se s njihovom uporabom i kod autonomne vožnje te proširene stvarnosti. Motivacija za ovaj diplomski rad dolazi upravo iz interesa prema ovom području koji se stvorio radeći na projektu čiji je zadatak bio blizak 3D rekonstrukciji scene.

2. Motivacija

Projekt u kojem je važan korak bio monokularna estimacija dubine je zahtjevao informaciju o obliku i volumenu praznog prostora u kutijima s osjetljivim sadržajem kako bi se znao oblik i veličina sigurnosnog materijala (stiropora i kartona) kojeg bi stavljali u kutiji u svrhu fiksiranja sadržaja i sigurnog transporta. Dakle, finalni produkt je trebao biti montiran iznad pokretne trake s kutijama te u realnom vremenu uzimati sliku pojedine kutije i vraćati informaciju o obliku i veličini sigurnosnog materijala za promatranu kutiju. Za rješavanje navedenog problema, odlučeno je implementirati algoritam koji bi procjenjivao dubinske mape svake pojedine kutije. Na temelju tih dubinskih mapa bi se, pomoću 3D point vizualizacije, rekonstruirala scena te izračunao oblik i volumen traženog sigurnosnog materijala. Nakon provedenog istraživanja područja na temu estimacije dubine, odluka je pala na pristup koji se temelji na dubokom učenju. Dakle, algoritam koji procjenjuje dubinu je model neuronske mreže o kojem će više pisati kasnije.

3. Potručje istraživanja i odabir metode

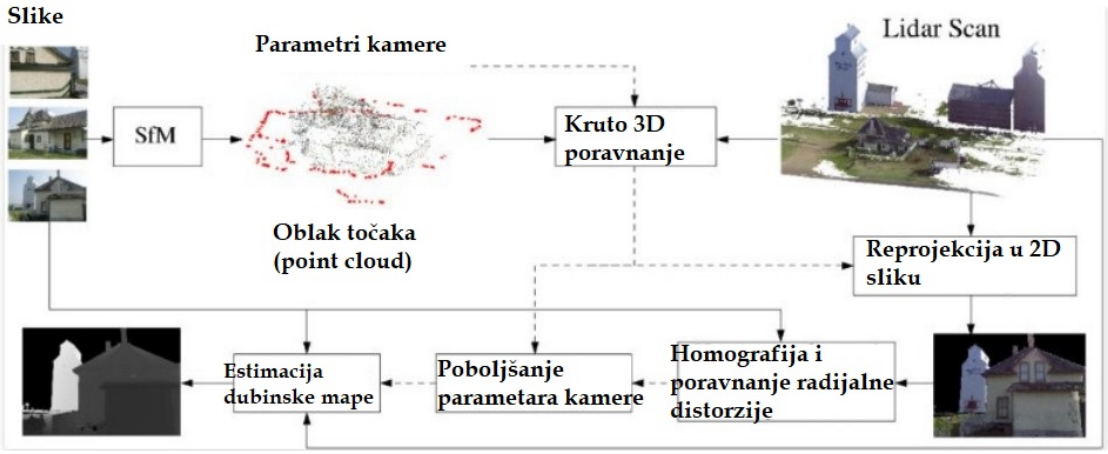
3.1. Tradicionalne metode estimacije dubine

Tradicionalne metode procjene dubine, poput strukture iz pokreta (structure from motion) i podudaranja stereo vizije (stereo vision matching), izgrađene su na podudarnosti značajki višestrukih gledišta. Brzim razvojem dubokih neuronskih mreža, monokularna procjena dubine temeljena na dubokom učenju je u posljednje vrijeme široko proučavana i postignuta je obećavajuća točnost rezultata. Također, dense depth maps se procjenjuju iz pojedinačnih slika dubokim neuronskim mrežama na "end-to-end" način. Kako bi se poboljšala točnost estimacije dubine naknadno se predlažu različite vrste mrežnih okvira, funkcije gubitaka i strategije obuke.

3.1.1. Structure from motion (SFM) depth estimation - Estimacija dubine iz strukture pokreta

Procjena dubine pomoću SFM-a proizlazi iz ideje da smo u mogućnosti percipirati i strukturalno razumjeti 3D okolinu krećući se oko nje. Kada se promatrač pomiče, objekti oko njega pomiču se različitim količinama ovisno o njihovoj udaljenosti od promatrača. To je poznato kao paralaksa gibanja, a iz te se dubinske informacije mogu generirati točan 3D prikaz svijeta oko nas. U računalnom vidu to se postiže tako što kamera u pokretu snima scenu i mjeri preklapanje između promjene prikaza u svakom vremenskom koraku. Mreža dubine koristi se za razumijevanje paralakse kretanja. Dok se mreža poza koristi za predviđanje promjene u promatranju između okvira.

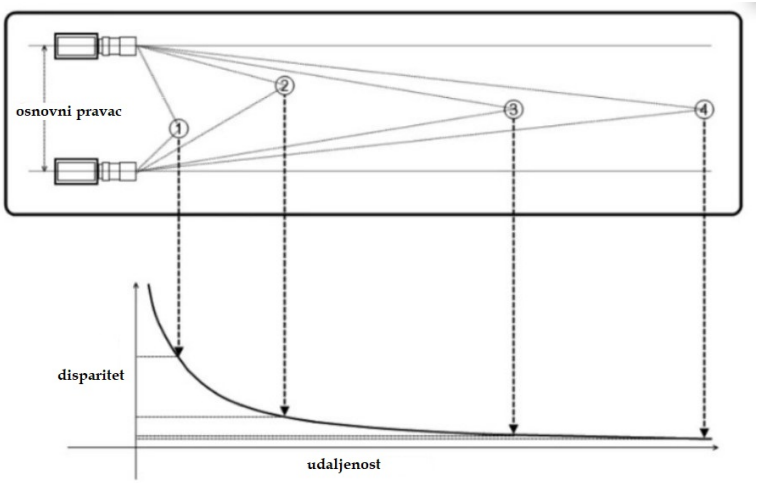
Ovdje je dan primjer jednog okvira za procjenu dense mapa dubine.



Slika 3.1: Primjer SFM estimacije dubine
izvor slike : labsites.rochester.edu

3.2. Estimacija dubine pomoću podudaranja stereo vizije (stereo vision matching)

Stereovizija je tehnika koja se koristi za procjenu dubine točkastog objekta $'P'$ iz kamere pomoću dvije kamere. Osnova stereo vida slična je 3D percepciji u ljudskom vidu i temelji se na triangulaciji zraka iz više točaka gledišta. Stereovizija je široko područje, a ovdje je prikazan graf koji je ključ za razumijevanje procesa estimacije dubine. Dubina je obrnuto proporcionalna disparitetu. Što je disparitet veći, objekt je bliži osnovnoj liniji (baseline) kamere. Što je disparitet manji, objekt je udaljeniji od osnovne linije.



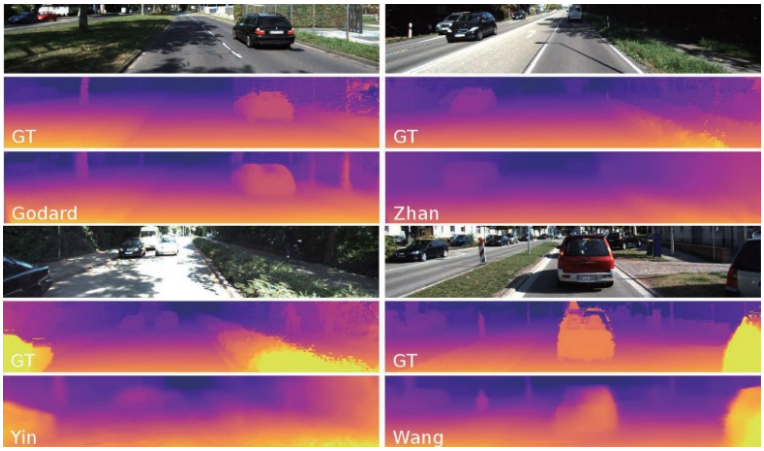
Slika 3.2: Estimacije dubine pomoću stereovizije
izvor slike : medium.com

Disparitet je proporcionalan osnovnoj vrijednosti. Ovo je lako vizualizirati. Ako imamo malu osnovnu udaljenost između dvije kamere, tada će razlika/odstupanje između dvije slike biti mala. Kako budemo povećavali osnovnu vrijednost, razlika će se povećavati.

3.3. Metode estimacije dubine temeljene na dubokom učenju

Ovo poglavlje daje pregled trenutnih metoda procjene dubine temeljene na dubokom učenju. Razne neuronske mreže pokazale su svoju učinkovitost u rješavanju monokularne procjene dubine. To su, na primjer, konvolucijske neuronske mreže (CNN), povratne neuronske mreže (RNN), varijacijski autoenkoderi (VAE) i generativne adversarial mreže (GAN).

Glavni cilj ovog pregleda je pružiti intuitivno razumijevanje glavnih algoritama koji su dali značajan doprinos monokularnoj procjeni dubine. Pregledajmo neke povezane radove u monokularnoj procjeni dubine iz aspekta metoda učenja, uključujući funkciju gubitka i dizajn okvira mreže. Neki primjeri procjene monokularne dubine na temelju dubokog učenja prikazani su na slici dolje.



Slika 3.3: Primjer procjene monokularne dubine. GT se odnosi na ground truth dubinske karte. Dubinske karte predviđaju se iz duboke neuronske mreže, Zhan i dr., Yin i Shi , i Wang et al.. Kao što je prikazano na ovim vizualima, 3D strukture objekata, poput drveća, ulica i automobila, mogu se učinkovito percipirati iz pojedinačnih slika dubokim mrežama dubine izvor slike : Monocular depth estimation based on deep learning: An overview ZHAO ChaoQiang, SUN QiYu, ZHANG ChongZhen, TANG Yang* QIAN Feng

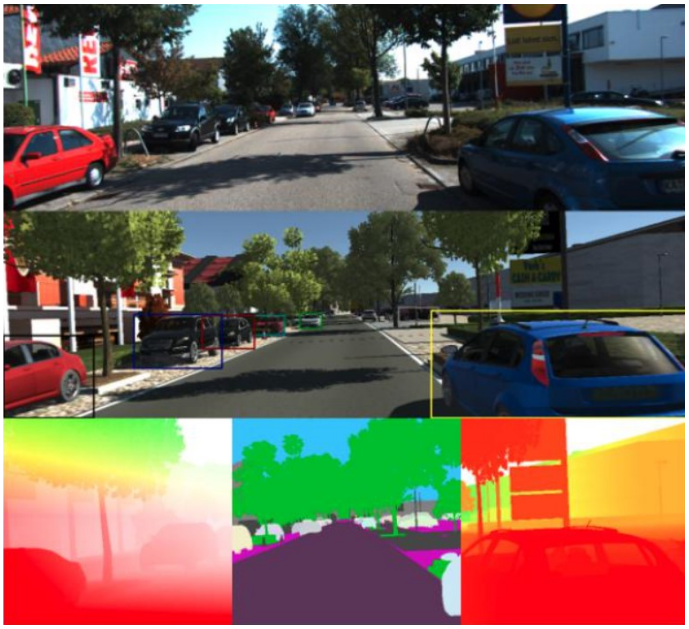
Za početak, nabrojat ćemo nekoliko široko korištenih skupova podataka i indikatora evaluacije u procjeni dubine temeljene na dubokom učenju. Nadalje, pregledavamo neke predstavnike postojećih metode prema različitim načinima obuke: supervizirani, nesupervizirani i polusupervizirani.

3.3.1. Skupovi podataka i pokazatelji evaluacije procjena

Skupovi podataka

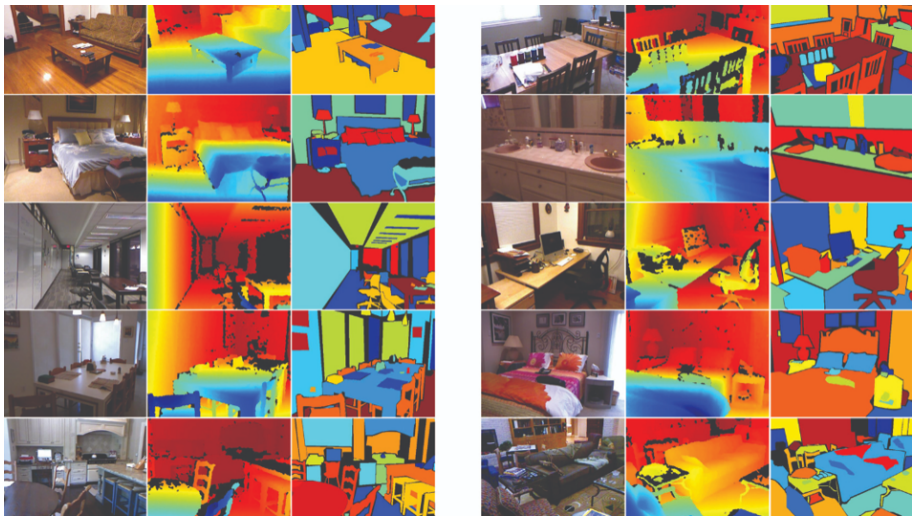
KITTI je najveći i najčešće korišten skup podataka za podzadatke računalnog vida. To je također najčešća referentna vrijednost i primarni skup podataka za obuku u nenadziranoj i polunadziranoj monokularnoj procjeni dubine. Slike iz kategorija “grad”, “stambeni objekti” i “cesta” su prikupljeni u skupu podataka KITTI, te je 56 scena u skupu podataka KITTI podijeljeno u dva dijela, 28 za obuku i drugih 28 za testiranje. Svaka se scena sastoji od parova stereo slika. Podaci su prikupljeni LIDAR-om[1].

[1]Lidar (akronim od engl. Light Detection and Ranging: svjetlosno zamjećivanje i klasifikacija) je optički mjerni instrument koji odašilje laserske zrake koje se odbijaju od vrlo sitnih čestica raspršenih u Zemljinoj atmosferi (aerosola, oblačnih kapljica i drugo) i potom registriraju u optičkom prijamniku (obično teleskopu). Drugi naziv za lidar je optički radar (eng. light radar) i laserski radar.



Slika 3.4: Skup podataka KITTI

NYU Depth skup podataka usredotočen je na okruženja na vratima, postoje 464 scene u zatvorenom prostoru. Za razliku od skupa podataka KITTI, koji prikuplja ground truth slike s LIDAR-om, skup podataka dubine NYU-a monokularne video sekvence scena i ground truth dubine su dobivene pomoću RGB-D kamere. To je uobičajeno mjerilo i primarni skup podataka za obuku u nadziranoj monokularnoj procjeni dubine. Ove scene u zatvorenom su podijeljene u 249 jedan za obuku i 215 jedan za testiranje.



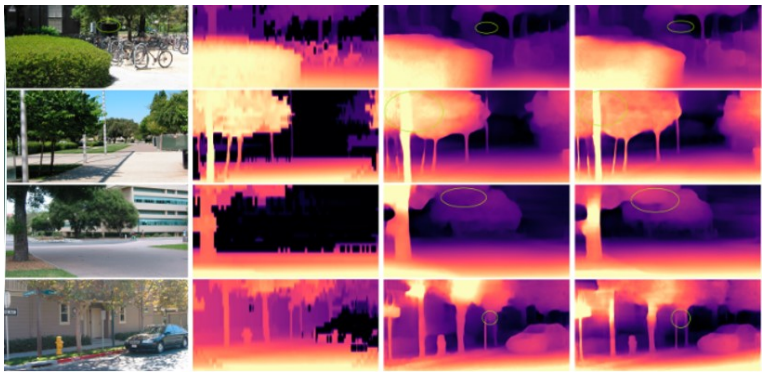
Slika 3.5: Skup podataka NYU Depth

Cityscapes Skup podataka uglavnom se fokusira na zadatke semantičke segmen-

tacije. Ima 5000 lijepo anotiranih slika i 20 000 ugrubo anotiranih slika. Ovaj skup podataka sastoji se od set stereo video sekvenci, koje su prikupljane u 50 gradova nekoliko mjeseci. Budući da ovaj skup podataka ne sadrži ground truth informaciju o dubini, primjenjuje se samo na proces obuke nekoliko nenadziranih metoda procjene dubine. Podaci o obuci sastoje se od 22973 prikupljenih parova stereo slika.

Slika 3.6: Skup podataka Cityscapes

Make3D skup podataka sastoji se samo od monokularnih RGB slika i dubinskih slika te nema stereo slike, što se razlikuje od gore navedenih skupova podataka. Budući da nema monokularnih sekvenci ili parova stereo slika u ovom skupu podataka, polunadzirano i nenadzirano učenje metode ga ne koriste kao set za obuku, dok ga nadzirane metode obično uzimaju za obuku.



Slika 3.7: Skup podataka Make3D

Mjere evaluacije

Kako bi se ocijenila i usporedila izvedba raznih mreže za procjenu dubine, općeprihvaćena evaluacija metoda je predložena s pet pokazatelja evaluacije: RMSE[2], RMSE log, Abs Rel, Sq Rel i Accuracy. Ovi pokazatelji su formulirani kao:

$$\begin{aligned}
&\text{- RMSE } = \sqrt{\frac{1}{|N|}\sum_{i\in N}\left\|d_i - d_i^*\right\|^2} \\
&\text{- RMSE log} = \sqrt{\frac{1}{|N|}\sum_{i\in N}\left\|\log\left(d_i\right) - \log\left(d_i^*\right)\right\|^2} \\
&\text{- Abs Rel } = \frac{1}{|N|}\sum_{i\in N}\frac{\left|d_i - d_i^*\right|}{d_i^*} \\
&\text{- Sq Rel } = \frac{1}{|N|}\sum_{i\in N}\frac{\left\|d_i - d_i^*\right\|^2}{d_i^*} \\
&\text{- Accuracies } = \%od d_i s.t. \max\left(\frac{d_i}{d_i}, \frac{d_i^*}{d_i}\right) = \delta < thr
\end{aligned}$$

,gdje je d_i predviđena vrijednost dubine piksela i, a d_i^* označava ground truth dubine. N označava ukupan broj piksela sa stvarnim vrijednostima dubine, a thr označava treshold.

3.3.2. Metode monokularne estimacije dubine temeljene na dubokom učenju

Budući da ljudi mogu koristiti apriorne informacije o svijetu, oni mogu percipirati dubinske informacije iz jedne slike. Inspirirani time, istraživački radovi postižu procjenu dubine jedne slike kombinirajući neke prethodne informacije, poput odnosa između nekih geometrijskih struktura (nebo, tlo, zgrada). S obećavajućim rezultatima u obradi slike, CNN-ovi su također pokazali snažnu sposobnost točne procjene mapa dubine iz pojedinačne (monokularne) slike. Provedena istraživanja daju ideju o tome koje vrste dubinskih mreža bi se trebale koristiti za monokularnu procjenu dubine na temelju četiri objavljene metode (MonoDepth, SfMLearner, Semodepth i LKVOLearner). Duboke neuronske mreže mogu se smatrati crnom kutijom. Duboka mreža će naučiti neke strukturne informacije za zaključivanje uz pomoć nadziranih signala. Međutim, jedan od najvećih izazova dubokog učenja je nedostatak skupova podataka s ground truth mapama dubine, jer ih je skupo nabaviti. Stoga je u ovom odjeljku dat pregled metoda monokularne procjene dubine s aspekta korištenja ground truth slika. Promatrat ćemo : nadzirane metode, nenadzirane metode i polu-nadzirane metode. Iako se proces obuke nenadziranih i polunadziranih metoda oslanja na monokularne videozapise ili parove stereo slika, obučene dubinske mreže predviđaju mape dubine iz pojedinačnih slika

[2]root mean square error

tijekom testiranja. Sažet ćemo postojeće metode prema podacima korištenih za treniranje, nadziranih signala i arhitekture u tablici (TABLICA 1). Također, prikupljeni su kvantitativni rezultati nenadziranih i polunadziranih algoritama na skupu podataka KITTI u drugoj tablici (TABLICA 2).

Metode	Godine	Skup podataka	Način nadziranja (Supervised manner (Sup))			Glavni doprinosi
			Sup	Semi-sup	Unsup	
Eigen et al.	2014	RGB + Depth	√			CNNs
Li et al.	2015	RGB + Depth	√			Hierarchical CRFs
Liu et al.	2015	RGB + Depth	√			Continuous CRF
Wang et al.	2015	RGB + Depth	√			Multi-task, hierarchical CRFs
Shelhamer et al.	2015	RGB + Depth	√			Fully CNNs
Eigen et al.	2015	RGB + Depth	√			Multi-task
Szegedy et al.	2015	RGB + Depth	√			Inception Module
Mousavian et al.	2016	RGB + Depth	√			Multi-task
Roy et al.	2016	RGB + Depth	√			RFs
Mayer et al.	2016	RGB + Disparity	√			Multi-task
Laina et al.	2016	RGB + Depth	√			Residual learning
Jung et al.	2017	RGB + Depth	√			Adversarial learning
Kendall et al.	2017	Stereo images + Disparity	√			Disparity loss
Zhang et al.	2018	RGB + Depth	√			Task-attentional, BerHu loss
Xu et al.	2018	RGB + Depth	√			Continuous CRF, structured attention
Lore et al.	2018	RGB + Depth	√			Conditional GAN
Fu et al.	2018	RGB + Depth	√			Ordinal regression
Facil et al.	2019	RGB + Depth	√			Transferability
Wofk et al.	2019	RGB + Depth	√			Lightweight network
Garg et al.	2016	Stereo images		√		Stereo framework
Chen et al.	2016	RGB + Relative depth annotations		√		The wild scene
Godard et al.	2017	Stereo images		√		Left-right consistency loss
Kuznietsov et al.	2017	Stereo images + LiDAR		√		Direct image alignment loss
Poggi et al.	2018	Stereo images		√		Trinocular assumption
Ramirez et al.	2018	Stereo images + Semantic Label		√		Multi-task
Aleotti et al.	2018	Stereo images		√		GAN
Pilzer et al.	2018	Stereo images		√		Cycled generative network
Luo et al.	2018	Stereo images		√		Stereo matching
He et al.	2018	Stereo images +LIDAR		√		Weak-supervised framework
Pilzer et al.	2019	Stereo images		√		Knowledge distillation
Tosi et al.	2019	Stereo images		√		Stereo matching
Chen et al.	2019	Stereo images		√		Multi-task
Fei et al.	2019	Stereo images + IMU + Semantic Label		√		Multi-task, physical information
Feng et al.	2019	Stereo images		√		Stacked-GAN
Wang et al.	2018	Mono. sequences		√		Direct VO
Zhan et al.	2018	Stereo sequences		√		Deep feature reconstruction
Li et al.	2018	Stereo sequences		√		Absolute scale recovery
Wang et al.	2019	Stereo sequences		√		Multi-task
Zhao et al.	2019	Stereo images + Synthesized GT		√		Domain adaptation, cycle GAN
Wu et al.	2019	Mono. sequences + LIDAR		√		Attention mechanism, GAN
Zhou et al.	2017	Mono. sequences			√	Monocular framework, mask network
Vijayanarasimhan et al.	2017	Mono. sequences			√	Multi-task
Yang et al.	2017	Mono. sequences			√	Multi-task
Mahjourian et al.	2018	Mono. sequences			√	ICP loss
Yin and Shi	2018	Mono. sequences			√	Multi-task
Zou et al.	2018	Mono. sequences			√	Multi-task
Kumar et al.	2018	Mono. sequences			√	GAN
Sun et al.	2019	Mono. sequences			√	Cycle-consistent loss
Wang et al.	2019	Mono. sequences			√	Geometry mask
Bian et al.	2019	Mono. sequences			√	Scale-consistency
Casser et al.	2019	Mono. sequences			√	Multi-task
Ranjan et al.	2019	Mono. sequences			√	Multi-task
Chen et al.	2019	Mono. sequences			√	Multi-task
Gordon et al.	2019	Mono. sequences			√	Multi-task
Li et al.	2019	Mono. sequences			√	GAN, LSTM, mask
Almalioglu et al.	2019	Mono. sequences			√	GAN, LSTM

TABLICA 1 Sažetak procjene monokularne dubine temeljene na dubokom učenju. "Mono." odnosi se na "Monokular", a "multi-tasks" znači da osim poze i procjena dubine, postoje i drugi zadaci koji se zajednički obučavaju u okviru, kao što su semantička segmentacija, segmentacija pokreta, optički protok, kretanje objekata, normalna površina itd...

Metoda	Godina	Obrazac za treniranje	Cap (m)	Što manje, to bolje				Accuracy: Što veće, to bolje		
				Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25^1$	$\delta < 1.25^2$	$\delta < 1.25^3$
Garg et al. L12 Aug8 $\times$ cap 50 m	2016	Semi-sup	50	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Godard et al.	2017	Semi-sup	80	0.148	1.344	5.927	0.247	0.862	0.960	0.964
Kuznietsov et al.	2017	Semi-sup	80	0.113	0.741	4.621	0.189	0.803	0.922	0.986
Poggi et al.	2018	Semi-sup	80	0.126	0.961	5.205	0.220	0.835	0.941	0.974
Ramirez et al. [68]	2018	Semi-sup	80	0.143	2.161	6.526	0.222	0.850	0.939	0.972
Aleotti et al.	2018	Semi-sup	80	0.119	1.239	5.998	0.212	0.846	0.940	0.976
Pilzer et al.	2018	Semi-sup	80	0.152	1.388	6.016	0.247	0.789	0.918	0.965
Luo et al.	2018	Semi-sup	80	0.094	0.626	4.252	0.177	0.891	0.965	0.984
He et al.	2018	Semi-sup	80	0.110	1.085	5.628	0.199	0.855	0.949	0.981
Pilzer et al.	2019	Semi-sup	80	0.098	0.831	4.656	0.202	0.882	0.948	0.973
Tosi et al.	2019	Semi-sup	80	0.111	0.867	4.714	0.199	0.864	0.954	0.979
Chen et al.	2019	Semi-sup	80	0.118	0.905	5.096	0.211	0.839	0.945	0.977
Feng et al.	2019	Semi-sup	80	0.065	0.673	4.003	0.136	0.944	0.979	0.991
Zhou et al.	2017	Unsup	80	0.208	1.768	6.865	0.283	0.678	0.885	0.957
Yang et al.	2017	Unsup	80	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian et al.	2018	Unsup	80	0.163	1.240	6.221	0.250	0.762	0.916	0.968
Yin and Shi	2018	Unsup	80	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Zou et al.	2018	Unsup	80	0.150	1.124	5.507	0.223	0.806	0.933	0.973
Wang et al.	2019	Unsup	80	0.158	1.277	5.858	0.233	0.785	0.929	0.973
Bian et al.	2019	Unsup	80	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Casser et al.	2019	Unsup	80	0.109	0.825	4.750	0.187	0.874	0.958	0.983
Ranjan et al.	2019	Unsup	80	0.140	1.070	5.326	0.217	0.826	0.941	0.975
Chen et al.	2019	Unsup	80	0.100	0.811	4.806	0.189	0.875	0.958	0.982
Li et al.	2019	Unsup	80	0.150	1.127	5.564	0.229	0.823	0.936	0.974
Almalioglu et al.	2019	Unsup	80	0.150	1.141	5.448	0.216	0.808	0.939	0.975

TABLICA 2 Rezultati monokularne dubine polu-nadziranih i nenadziranih metoda na skupu podataka KITTI. "Cap" označava gornju granicu predviđenih dubina, a "sup" se odnosi na "nadzirano".

3.3.3. Nadzirana monokularna estimacija dubine

Osnovni model za nadzirane metode Nadzorni signal nadziranih metoda temelji se na ground truth mapi dubine, tako da se monokularna procjena dubine može smatrati regresijskim problemom. Duboke neuronske mreže su dizajnirane za predviđanje mapa dubine iz pojedinačnih slika. Razlike između predviđene i stvarne dubinske karte se koriste za nadgledanje obuke mreža. Funkcija gubitka, $\mathscr{L}2$, je dana na ovaj način :

$$\mathscr{L}_2(d,d^*) = \frac{1}{N}\sum_i^N \|d-d^*\|_2^2.$$

Dubinske mreže uče informacije o dubini scene približavanjem ground truth informacijama. Metode se temelje na različitim arhitekturama i funkcijama gubitaka. Koliko znamo, problem monokularne procjene dubine se da riješiti koristeći CNN. Pred-

ložena arhitektura koja je sastavljena od dvokomponentnih blokova (globalna mreža grubog mjerila i lokalna mreža finog mjerila) dizajnirana je za predviđanje dubinske karte iz jedne slike na "end-to-end" način. Tijekom procesa obuke, koristi se ground truth dubine $d*$ kao nadzirani signal, a dubinska mreža predviđa logaritam dubina kao $log(d)$. Funkcija gubitka treninga postavljena je na sljedeći način:

$$\mathscr{L}(d,d^*) = \frac{1}{N}\sum_i^N y_i^2 - \frac{\lambda}{N^2}\left(\sum_i^N y_i\right)^2,$$

gdje je $y_i^2 = log(d)log(d)$. λ se odnosi na faktor ravnoteže i postavljen je na 0,5. Mreža grubog mjerila se najprije uvježbava i zatim se fina mreža osposobljava za pročišćavanje rezultata, odnosno - fiksiranje parametara grube mreže. Eksperimenti pokazuju da je fina mreža učinkovita za pročišćavanje mapa dubine procijenjenih mrežom grubog mjerila. Predložen je opći okvir za bavljenje zadacima kao što su procjena mapa dubine, normalna procjena površine i predviđanje semantičke oznake iz jedne slike. Za procjenu dubine, predlaže se dodatna funkcija gubitka za promicanje lokalne strukturne dosljednosti:

$$\mathscr{L}_s = \frac{1}{N}\sum_i^n \left[(\nabla_x D_i)^2 + (\nabla_y D_i)^2\right],$$

gdje je $D_i = log(d_i) - log(d_i^*)$, a ∇ je vektorski diferencijalni operator. Ova funkcija izračunava gradijente razlike između estimirane dubine i ground truth dubine u vodoravnom i okomitom smjeru. Kako taj optički tok uspješno rješen putem nadziranog učenja koristeći CNN, optički protok mreže je proširen na disparitet i procjenu toka scene. Predloženi okvir za monokularnu estimaciju dubine optimizira faktorizaciju za oporavak ulazne slike. Inspirirani izvanrednom izvedbom ResNeta, stručnjaci su uveli rezidualno učenje kako bi se naučio odnos mapiranja između karata dubine i pojedinačnih slika. Dakle, njihove mreže su dublje od prethodnih radova u procjeni dubine s većom točnosti. Osim toga, potpuno povezani slojevi u ResNetu su zamijenjeni blokovima za povećanje uzorkovanja radi poboljšanja razlučivosti estimirane karte dubine. Tijekom procesa treniranja, korišten je obrnuti Huber (Berhu) kao nadzirani signal dubinske mreže te postiže bolji rezultat od $\mathscr{L}2$ gubitka. Berhu gubitak je:

$$\mathscr{L}_{\text{Berhu}}(d,d^*) = \begin{cases} |d-d^*|, & \text{if } |d-d^*| \le c, \\ \frac{(d-d^*)^2+c^2}{2c}, & \text{if } |d-d^*| > c, \end{cases}$$

,gdje je c prag (treshold) koji je postavljen na $\frac{1}{5}max_i(|\text{d - d*}|)$. Ako je $|x| < c$, Berhu gubitak jednak je $\mathscr{L}1$ normi. Berhu gubitak jednak je $\mathscr{L}2$ kada je $|x|$ izvan navedenog

raspona. Zbog dubljih potpunih konvolucijskih mreža i poboljšane funkcije gubitka, ova metoda postiže bolji rezultat od prethodnih radova s manje parametara i podataka o obuci. Ovdje je fokus na primjeni procjene dubine u otkrivanju prepreka. Umjesto predviđanja dubine iz jedne slike, predloženi potpuni CNN okvir koristi i monokularnu sliku i odgovarajući optički protok za procjenu karte dubine.

Kasnije su se znanstvenici uhvatili u koštac s izazovom opažanja procjene dubine pojedinačne slike istražujući novi algoritam. Drugačiji je po tome što ne koristi ground truth dubine kao nadzirane signale, već su njihove mreže uvježbane na anotacijama relativne dubine. Budući da monokularni pogled sadrži malo geometrijskih detalja, osmišljen je okvir dubokog učenja za učenje strukture iz parova stereo slika. Osim toga, dizajnirana mreža predviđa kartu dispariteta umjesto karte dubine, a disparitet temeljne istine koristi se za nadzor:

$$\mathscr{L}(I, I^*) = \frac{1}{N}\sum_i^N \|Dis_i - Dis_i^*\|_1$$

gdje Di_{s_i} označava estimirani disparitet piksela i, a Di_{s*_i} je odgovarajuća ground truth (temeljna istina). S obzirom na sporu konvergenciju i lokalna optimalna rješenja uzrokovana minimiziranjem srednje kvadratne pogreške u log-prostoru tijekom učenja, procjena monokularne dubine smatra se problemom regresije. Kako nesigurnost predviđenih vrijednosti dubine raste zajedno s vrijednostima ground truth dubine, bolje je dopustiti veće pogreške u procjeni pri predviđanju većih vrijednosti dubine, koje se ne mogu dobro riješiti uniformom strategijom diskretizacije (UD). Stoga, spacing-increasing strategija diskretizacije (SID) predložena je za diskretizaciju dubine i optimizaciju procesa treninga. U svrhu poboljšanja transportabilnosti dubinske mreže na različitim kamerama, uveden je model kamere u mrežu za procjenu dubine. Na taj način je poboljšana mogućnost generalizacije mreža. Iako se navedenim metodama postiže izvanredna točnost, velik broj parametara kod ovih metoda ima ograničenu primjenu mreže u praksi, posebno na ugrađenim sustavima. Stoga je donešeno riješenje - dizajniran je mrežni okvir za automatsko kodiranje. U međuvremenu se primjenjuje obrezivanje mreže kako bi se smanjila računalna složenost i poboljšalo izvođenje u stvarnom vremenu.

Metode temeljene na uvjetnim slučajnim poljima (conditional random fields)

Umjesto korištenja dodatne mreže za pročišćavanje rezultata, postoji metoda usavršavanja temeljena na kondicionalnim slučajnim poljima (CRFs), koja se također naširoko koristi za semantičku segmentaciju. Zbog kontinuirane karakteristike dubine između piksela, CRF može poboljšati procjenu dubine uzimajući u obzir dubinu for-

miranja susjednih piksela, tako da je CRF model široko primijenjen u procjeni dubine. Duboki okvir CNN-a dizajniran je za regresiju karte dubine od višerazinskih patchova (zakrpa) slike na razini super-piksela. Zatim, karta dubine se pročišćava od razine super-piksela do razine piksela preko hijerarhijskog CRF-a, a energetska funkcija je:

$$\mathbf{E}(\mathbf{d}) = \sum_{i \in S} \phi_i(d_i) + \sum_{(i,j) \in \varepsilon_i} \phi_{ij}\left(d_i, d_j\right) + \sum_{c \in P} \phi_c(d_c)$$

gdje S označava skup super-piksela, $\in s$ se odnosi na skup parova superpiksela koji dijele zajedničku granicu. P označava skup patcheva (zakrpa) na razini piksela. $E(d)$ se sastoji od tri dijela:

- 1) podatkovni izraz za izračunavanje kvadratne udaljenosti između vrijednosti dubine d i mrežne regresirane dubine $\overline{d}$;
- 2) izraz glatkoće za nametanje relevantnosti između susjednih super-piksela
- (3) autoregresijski model za opisivanje lokalne relevantne strukture u karti dubine

Iste godine, sličan okvir koji istražuje duboku CNN s kontinuiranim CRF-om, nazvan dubokim konvolucijskim neuralnim poljima, predložen je za rješavanje problema monokularne procjene dubine. Metodu udruživanja super piksela su predložili za ubrzanje konvolucijske mreže. To pomaže u dizajniranju dublje mreže kako bi se poboljšala točnost procjene dubine. Kasnije je predstavljen okvir za zajedničku procjenu mape dubine na razini piksela i semantičke oznake jedne slike. Zbog strukturalne konzistencije između dubinske karte i semantičkih oznaka, interakcija između dubinske i semantičke informacije se koristi za poboljšanje izvedbe procjene dubine. Zadaci dubine i semantičkog predviđanja zajednički se treniraju nadziranim signalom:

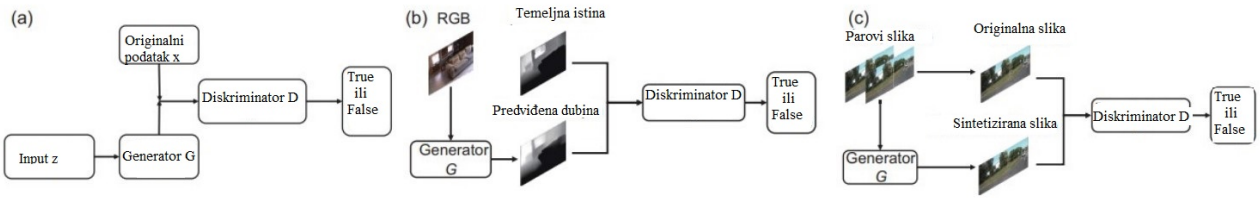
$$\mathscr{L}(I, I^*) = \frac{1}{N}\sum_i^N \left(\log(d_i) - \log\left(d_i^*\right)\right)^2 - \lambda \frac{1}{N}\sum_i^N \log\left(P\left(l_i^*\right)\right)$$
$$P\left(l_i^*\right) = \exp\left(z_{i,I_i}\right) / \sum \exp\left(z_{i,l_1}\right)$$

,gdje $l*$ stoji za ground truth (temeljnu istinu) semantičkih oznaka, dok se l_i odnosi na predviđene oznake (labele). z_i, l_i^* označava izlaz semantičkog čvora. Da bi se dodatno preciziralo procijenjenu dubinu, uvedena je dvoslojna hijerarhija CRF za ažuriranje pojedinosti o dubini izvlačenjem čestih predložaka za svaku semantičku kategoriju, što dovodi do činjenice da njihove metode ne mogu raditi dobro kako se broj klasa povećava. Zbog toga je kreiran spojeni okvir za simultanu procjenu karata dubine i semantičke oznake iz jedne slike, te ova dva zadatka dijele prikaz značajki na visokoj razini slika izdvojenih iz CNN-a. Koristi se potpuno povezani CRF koji je spojen s dubokim CNN-om za poboljšanje interakcije između karata dubine i semantičke oznake.

Stoga se njihova metoda obučava na način od kraja do kraja (end-to-end) i 10 puta brže od prethodno spomenute metode. Nakon toga, predložen je modul usmjeren na zadatak koji bi obuhvatio interakciju i poboljšao performanse mreža, što se razlikuje od prethodnih modula. Slično, integriran je kontinuirani CRF model u dubinski CNN okvir za obuku od kraja do kraja. Osim toga, strukturirani attention model u kombinaciji s CRF modelom predložen za jačanje prijenosa informacija između odgovarajućih značajki. Random Forest (RF) model također je uveden u zadatke monokularne procjene dubine i učinkovito provodi točnost dubine procjena.

Metode temeljene na adversarial (kontradiktornom) učenju

Zbog izvanredne izvedbe u generiranju podataka, adversarial učenje postalo je popularan smjer istraživanja posljednjih godina. Okviri adversarial učenja u procjeni dubine prikazani su na slici ispod ovog teksta.



Slika 3.8: (a) Okvir sirovog (raw) GAN-a. Generator sirovog GAN-a ima mogućnost generiranja podataka iz vektora z, a diskriminator je dizajniran za razlikovanje pravih od lažnih podataka. Podaci koje generira generator imaju istu distribuciju podataka kao i pravi podaci. (b) Okvir nadziranih metoda temeljen na GAN-u. U nadziranim metodama temeljenim na GAN-u, dubinske karte predviđene generatorom (dubinska mreža) i stvarne karte dubine šalju se diskriminatoru tijekom obuke. c) Okvir nenadziranih i polunadziranih metoda temeljen na GAN-u. U bazi GAN-a nenadzirane i polunadzirane metode, zbog nedostatka pravih dense mapa dubine, RGB slike (sintetizirane algoritmom za rekonstrukciju pogleda u generatoru) i stvarne slike šalju se diskriminatoru. Generator uzima parove slika, poput isječaka slike u nenadziranoj metodi ili parove stereo slika u polu-nadziranim metodama, za procjenu karata dubine iz pojedinačnih slika i sintetiziranih RGB slika
izvor slike : Monocular depth estimation based on deep learning: An overview ZHAO ChaoQiang, SUN QiYu, ZHANG ChongZhen, TANG Yang* QIAN Feng

Različite vrste adversarial okvira učenja, poput složenog GAN-a, conditional GAN-a i cikličkog GAN-a, uvode se u zadatke procjene dubine i imaju pozitivan utjecaj na procjenu dubine. Adversarial učenje se uvodi u zadatke monokularne procjene dubine. Generator se sastoji od GlobalNet-a i Refinement Net-a, a te su mreže dizajnirane za

procijenu globalne i lokalne 3D strukture iz jedne slike. Zatim se koristi diskriminator za razlikovanje predviđene mape dubine od pravih. Ovaj je oblik uobičajen te se koristi u nadziranim metodama. Konfrontacija između generatora G i diskriminatora D olakšava uvježbavanje okvira temeljenog na problemu min-max:

$$\min_G \max_D \mathbb{B}_{x\sim P_s,}[\log D(x)] + \mathbb{E}_{\hat{x}\sim P_G}[\log(1 - D(\hat{x}))]$$

,gdje je x ground truth mapa dubine , a $\hat{x}$ se odnosi na kartu dubine predviđenu generatorom. Slično, uvjetni (conditional) GAN je također korišten za monokularnu procjenu dubine. Razlika od prethodno spomenutog je da je sekundarni GAN uveden kako bi se na temelju njega dobila preciznija karta dubine slika i gruba procijenjena karta dubine. Budući da se nadzire ground truth, nadzirane metode mogu učinkovito naučiti funkcije za mapiranje 3D strukture i informacije o njihovom mjerilu iz pojedinačnih slika. Međutim, ove nadzirane metode ograničene su označenim (labelanim) skupovima za treniranje, koje je teško i skupo nabaviti.

3.3.4. Nenadzirana monokularna procjena dubine

Umjesto korištenja temeljne istine, koju je skupo nabaviti, uzimaju se u obzir geometrijska ograničenja između okvira kao nadzorni signal tijekom procesa obuke nenadzirane metode.

Osnovni model za nenadzirane metode Nenadzirane metode uvježbane su monokularnim sekvencama slika, a geometrijska ograničenja izgrađena su na projekciji između susjednih okvira:

$$p_{n-1} \sim \mathbf{K} T_{n\to n-1} D_n(p_n) \mathbf{K}^{-1} p_n$$

,gdje p_n označava piksel na slici I_n, a p_{n-1} se odnosi na odgovarajući piksel od p_n na slici I_{n-1}. K je intrinzička matrica kamere, koja je poznata. $D_n(p_n)$ označava dubinsku vrijednost piksela p_n, a $T_{n\to n-1}$ predstavlja prostornu transformaciju između I_n i I_{n-1}. Dakle, ako su $D_n(p_n)$ i $T_{n\to n-1}$ poznati, podudarnosti između piksela na različitim slikama (I_n i I_{n-1}) uspostavljaju se pomoću funkcije projekcije.

Inspirirani ovim ograničenjem, znanstvenici su projektirali duboku mrežu za predviđanje mape dubine $\hat{D}_n$ iz jedne slike I_n, i mreže poza za regresiju transformacije $\hat{T}_{n\to n-1}$ između okvira (In i I_{n-1}). Na temelju izlaza mreža, izgrađene su korespondencije piksela između I_n i I_{n-1}:

$$p_{n-1} \sim \mathbf{K} \hat{T}_{n\to n-1} \hat{D}_n(p_n) \mathbf{K}^{-1} p_n.$$

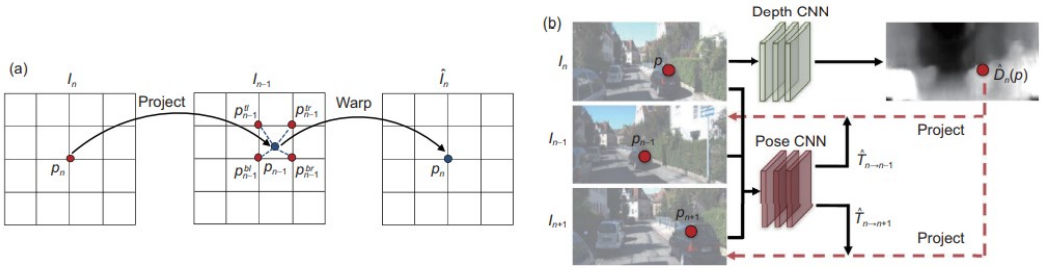
Zatim, fotometrijska pogreška između odgovarajućih piksela izračunava se kao geometrijsko ograničenje. Koristi se sinteza pogleda kao metrika, i rekonstrukcijski gubitak (reconstruction loss) je formuliran kao :

$$\mathscr{L}_{vS} = \frac{1}{N}\sum_{p}^{N}\left|I_n(p) - \hat{I}_n(p)\right|$$

,gdje p označava koordinate piksela. $\hat{I}_n(p)$ označava rekonstruirani okvir. Sličnost strukture temeljena na indeksu strukturne sličnosti (SSIM) također je uvedena u $\mathscr{L}_{vs}$ za kvantificiranje razlike između rekonstruiranih i ciljanih slika:

$$\mathscr{L}_{vs} = \alpha\frac{1-\mathrm{SSIM}\left(I_n - \hat{I}_n\right)}{2} + (1-\alpha)\left|I_n - \hat{I}_n\right|$$

,gdje je α faktor ravnoteže. Osim toga, dokazano je da je učinkovitije izračunati minimalnu vrijednost greške rekonstrukcije od srednje vrijednosti. Algoritam rekonstrukcije pogleda (view reconstruction algorithm) primjenjuje se za rekonstrukciju okvira $\hat{I}_n(p)$ iz I_{n-1} na temelju funkcije projekcije, kao što je prikazano na slici ispod.



Slika 3.9: (a) **Ilustracija procesa savijanja slike**. Proces savijanja slike za rekonstrukciju pogleda u nenadziranim metodama. (b) **Ilustracija monokularne dubine bez nadzora** - Opći okvir monokularnih metoda bez nadzora. Tijekom treninga, dubina D_n i poza $\hat{T}_{n\to n-1}$ predviđena dubinskom mrežom i mrežom položaja koriste se za uspostavljanje projekcijskog odnosa između I_n i I_{n-1}, a zatim se $\hat{I}_n$ rekonstruira postupkom krivljenja slike na temelju projekcije. Razlike između stvarnih I_n i rekonstruiranih $\hat{I}_n$ slika izračunate su kako bi se nadziralo obučavanje mreža.
izvor slike : Monocular depth estimation based on deep learning: An overview ZHAO ChaoQiang, SUN QiYu, ZHANG ChongZhen, TANG Yang* QIAN Feng

Gubitak glatkoće dubine s obzirom na rub (edge-aware depth smoothness loss), usvojen je za poticanje lokalne glatke karte dubine:

$$\mathscr{L}_{\text{smooth}} = \frac{1}{N}\sum_{p}^{N}|\nabla D(p)|\cdot\left(\mathrm{e}^{-|\nabla I(p)|}\right)^{\mathrm{T}},$$

,gdje se T odnosi na operaciju transponiranja. Iako je mreža dubine spojena s mrežom poza tijekom treninga, kao što je prikazano na slici iznad, mreže se mogu koristiti individualno tijekom testiranja.

Metode temeljene na maski objašnjivosti (explainability mask)

Algoritam rekonstrukcije prikaza temeljen na funkciji projekcije oslanja se na pretpostavka statičkog scenarija, tj. pozicija dinamičkih objekata na susjednim okvirima ne zadovoljava funkciju projekcije, što utječe na fotometrijsku grešku i proces treniranja. Stoga se maske naširoko koriste za smanjenje utjecaja dinamičkih objekata na gubitak rekonstrukcije pogleda (view reconstruction loss) L_{vs}. Mreža maski je dizajnirana za smanjenje učinaka dinamičkih objekata na rekonstrukciju pogleda:

$$\mathscr{L}_{vs}^{M} = \frac{1}{N}\sum_{p}^{N} M \left| I_n(p) - \hat{I}_n(p) \right|,$$

,gdje se M odnosi na masku objašnjivosti koju predviđa mreža maski. Budući da ne postoji izravni nadzor za M, trening s gornjim gubitkom L_{vs}^{M} rezultirao bi trivijalnim rješenjem mreže koje predviđa da je M jednak nuli, što je savršeno minimiziran gubitak. Prema tome, regularizacijski član $L_{reg}(M)$ se koristi za poticanje predviđanja različitih od nule minimiziranjem cross-entropy gubitka s konstantnom oznakom 1 na svakoj lokaciji piksela. Osim toga, dizajnirana je mreža maske objekta za procjenu dinamičkih objekata. Razlika je u tome da je gibanje objekta regresirano zajedno s pozom kamere i koristi se za izračunavanje optičkog protoka. Uveden je termin normale površine i normale dubine za nenadzirani okvir kako bi se poboljšala ograničenja na procjenu dubine. Pretvorba između dubine i normale rješava se projektiranjem sloja dubine na normalu i sloja normale na dubinu u mreži dubine. Kao rezultat toga, mreža dubine postiže veću točnost. Provedena su istraživanja o geometrijskim ograničenjima između mapa dubine uzastopnih okvira. Predložen je ICP gubitak (ICP loss) kao uvjet za provedbu dosljednosti procijenjenih karata dubine. Iako se procjena maske temeljena na dubokoj neuronskoj mreži naširoko koristi u prethodno spomenutim metodama te učinkovito smanjuje učinke dinamičkih objekata na pogreške rekonstrukcije, takva procjena ne samo da povećava količinu računanja, već i komplicira obuku mreže. Stoga će dizajnirane maske temeljene na geometriji zamijeniti maske temeljene na dubokom učenju i imati bolji učinak na procjenu dubine. Predložen je termin gubitka u skladu s ciklusom kako bi se u potpunosti iskoristio niz informacija. Zatim je razmotrena praznina regije na rekonstruiranim slikama uzrokovana promjenama prikaza i okluzija piksela generiranih tijekom projekcije. Oni analiziraju proces rekonstrukcije pogleda i utjecaj neusklađenosti piksela na trening. Dakle, imamo dvije maske -

maska projicirane slike i maska ciljane slike, odnosno overleap maska (maska koja se preklapa) i prazna maska, predlažu se za rješavanje razmatranih problema. Osim toga, detaljnija maska je dizajnirana da filtrira trag neusklađenih piksela. Eksperimenti dokazuju učinkovitost predloženih maski.

Postoji maska koja se temelji na ograničenju dosljednosti geometrije. Dizajnirana maska je otkrivena na temelju dosljednosti dubinskih karti susjednih slika. Osim toga, izraz gubitka konzistentnosti ljestvice značajno rješava problem nedosljednosti mjerila između različitih karata dubine.

Metode temeljene na tradicionalnoj vizualnoj odometriji

Umjesto korištenja poze procijenjene mrežom poza, poza regresirana iz tradicionalne izravne vizualne odometrije koristi se kao podrška procjeni dubine. Izravna vizualna odomerija uzima kartu dubine koju generira mreža dubine i isječak od tri okvira za procjenu poza između okvira minimiziranjem fotometrijske pogreške. Zatim, izračunate poze šalju se natrag u okvir obuke. Zbog toga što dubinsku mrežu nadziru točnije poze, točnost procjene dubine značajno je poboljšana.

Metode temeljene na multi-task okviru

Nedavni pristupi uvode dodatne mreže za multi-task u osnovni okvir, poput optičkog toka, gibanja objekta i intrinsične matrice kamere. Stoga, geometrijski odnos između različitih zadataka koristi se kao dodatni nadzorni signal, koji jača uvježbavanje cijelog okvira. Predložen je okvir zajedničkog učenja za dubinu, ego-motion i optički tijek zadataka. Predloženi nenadzirani okvir sastoji se od dva dijela:

- 1) rekonstrukcija krute (rigid) strukture za rekonstrukciju krute scene
- 2) lokalizator nekrutog (non-rigid) kretanja za dinamičku obrada predmeta

ResFlowNet dizajniran je u drugom dijelu za učenje zaostalog nekrutog protoka. Stoga, točnost sva tri zadatka je poboljšana odvajanjem krute i nekrute (rigid and nonrigid) scene i eliminiranje outliera kroz predloženi gubitak adaptivne geometrijske konzistencije (adaptive geometric consistency loss). Budući da je strujno polje krutih područja generirano procjenom dubine i položaja, pogreške uzrokovane dubinom ili položajem procjena se propagiraju na predviđanje protoka. Stoga je dizajnirana dodatna mreža za procjenu optičkog protoka. Osim toga, predložili su gubitak dosljednosti između zadataka (cross-task consistency loss) kako bi se ograničila dosljednost između procijenjenih protoka (iz mreže) i generiranih protoka (iz dubine i procjena poze). Kasnije su znanstvenici dodatno proširili okvir za više zadataka, a segmentacija kretanja je zajednički trenirana s drugim zadacima (dubina, poza, protok) u nenadziranom načinu. Više zadataka čini proces treninga kompliciranijim, pa uvode kompetitivnu suradnju za

koordinaciju procesa treninga i postizanja izvanrednih performansi. Nešto kasnije, dolazi do razmatranja kretanja dinamičnih objekata u scenama. Uvodi se mreža gibanja objekata za predviđanje gibanja pojedinačnih objekata, a ta mreža uzima segmentirane slike kao ulaz. Budući da se gore navedene metode temelje na preduvjetima poznatih intrinzičnih parametara kamere, to ograničava primjenu mreže na nepoznate kamere. Stoga se proširuje mrežu poza za procjenu intrinzičnog parametra kamere i dodatno se smanjuju preduvjeti tijekom treninga.

Metode temeljene na advesarial (kontradiktornom) učenju

Okvir kontradiktornog učenja također se uvodi u nenadziranu monokularnu procjenu dubine. Budući da nema prave dubinske karte u nenadziranoj obuci, nije moguće koristiti kontradiktorno (adversarial) učenje. Stoga, umjesto da pomoću diskriminatora razlikujemo stvarne i predviđene karte dubine, slike sintetizirane algoritmom rekonstrukcije pogleda (view reconstruction algorithm) i stvarne slike smatraju se ulazom za diskriminator. Generator se sastoji od mreža poze i mreža dubine, a izlazna mreža koristi se za sintetiziranje slika rekonstrukcijom pogleda (view reconstruction). Zatim, diskriminator je dizajniran za razlikovanje stvarnih i predviđenih karata dubine. Budući da vremenske informacije pomažu da se poboljša performans mreže, LSTM modul upoznaje se s mrežom poza i mrežom dubine za kontakt s kontekstualnim informacijama. Nadalje, dizajnirana je dodatna mreža za uklanjanje nedostataka algoritma za rekonstrukciju prikaza (view reconstruction algorithm). Kako bi se dobilo više 3D značajki, informacije između okvira ekstrahiranih LSTM-om i pojedinačnih slika se usvojaju zajedno za procjenu dubine. U usporedbi s nadziranim i polunadziranim metodama, nenadzirane metode uče dubinske informacije iz geometrijskih ograničenja umjesto temeljne istine (ground truth). Stoga, proces treninga oslanja se na monokularne sekvence snimljene kamerom, a učenje bez nadzora korisno je za praktičnu primjenu metoda bez nadzora. Učenjem iz monokularnih sekvenci, koje ne sadrže podatke o apsolutnoj skali, nenadzirane metode pate od dvosmislenosti ljestvice, nedosljednosti ljestvice i drugih problema.

3.3.5. Polunadzirane (Semi-supervised) metode za monokularnu estimaciju dubine

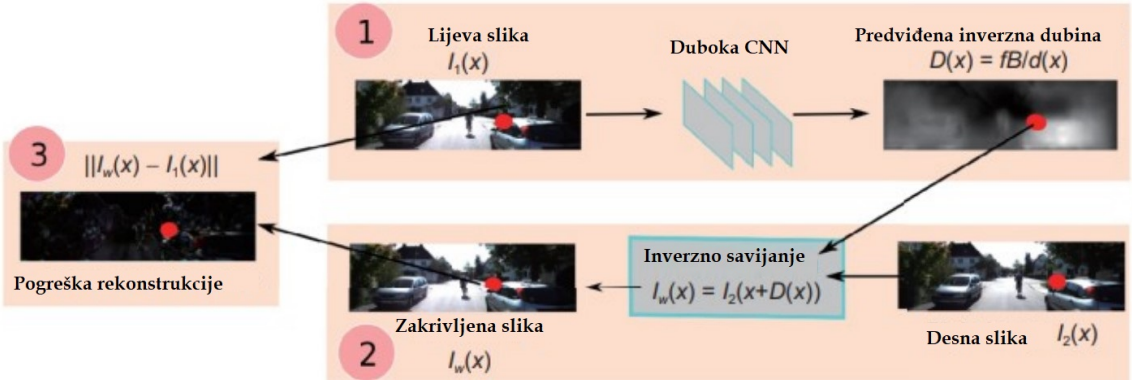
Budući da tijekom treninga nema potrebe za temeljnom istinom, izvedba nenadziranih metoda još je daleko od izvedbe nadziranih metoda. Osim toga, i nenadzirane metode pate od raznih problema, poput dvosmislenosti razmjera i nedosljednosti razmjera. Stoga su polunadzirane metode predložene da se dobije veća točnost procjene uz sma-

njenje ovisnosti o skupim temeljnim istinama (ground truths). Osim toga, informacije o ljestvici mogu se saznati od polunadziranih signala.

Obuka na parovima stereo slika je slična slučaju monokularnih videozapisa. Glavna je razlika u tome hoće li transformacija između dva okvira (lijevo-desno slike ili prednje-stražnje slike) biti poznata. Stoga neke studije uzimaju u obzir okvir temeljen na parovima stereo slika kao nenadgledane metode, dok ih drugi tretiraju kao polusupervizirane metode. U ovom pregledu smatramo ih polunadziranim metodama te su poze između lijevo-desnih slika nadzirani signali tijekom treninga.

Osnovni model za polunadzirane metode

Modeli polunadziranih metoda su uvježbani na procjeni parova stereo slika karte dispariteta (inverzne karte dubine) između lijeve i desne slike. Zatim, mapa dispariteta D_{is} izračunata iz predviđene inverzne dubine koristi se za sintezu lijeve slike iz desne slike inverznim savijanjem, kao što je prikazano na slici.



Slika 3.10: Opći okvir polu-nadzirane monokularne procjene dubine na temelju parova stereo slika. Mreža dubine uzima lijevu sliku za predviđanje njezine inverzne karte dubine na razini piksela (ili mape dispariteta), a predviđena inverzna karta dubine koristi se za rekonstrukciju lijeve slika od desne slike algoritmom obrnutog savijanja. Pogreška rekonstrukcije izračunata je za nadzor procesa treninga.

izvor slike : Monocular depth estimation based on deep learning: An overview ZHAO ChaoQiang, SUN QiYu, ZHANG ChongZhen, TANG Yang* QIAN Feng

Slično nenadziranim metodama, koriste se razlike između sintetiziranih slika I_w i stvarnih slika I_l kao nadzirani signal i za ograničavanje procesa obuke:

$$\begin{aligned}\mathscr{L}_{\text{recons}} &= \sum_{p} \|I_l(p) - I_w(p)\|^2 \\ &= \sum_{,p} \|I_l(p) - I_r(p + \text{Dis}(p))\|^2,\end{aligned}$$

,gdje je I_r odgovarajuće desna slika. Karta dubine d može se prenijeti iz karte predviđenog dispariteta kroz: $d = fB/D$, gdje je f lokalna duljina kamere, a B se odnosi na udaljenost između lijeve i desne kamere. Na temelju gornjeg okvira, korišten je izraz gubitka glatkoće (smoothness loss) kako bi poboljšao kontinuitet mapa dispariteta. Kasnije je poboljšan i gornji mrežni rad i funkcija gubitka. Karta desnog dispariteta Dis^r se predviđa zajedno s lijevom kartom dispariteta Dis^l i koristi se za rekonstrukciju desne slike iz lijeve slike. Osim toga, oni predstavljaju gubitak lijevo-desne disperzije konzistencije (left-right disparity consistency loss) kako bi se ograničila dosljednost između lijevog i desnog dispariteta:

$$\mathscr{L}_{lr} = \frac{1}{N}\sum_p \left|\mathrm{Dis}'(p) - \mathrm{Dis}^r\left(p + \mathrm{Dis}'(p)\right)\right|.$$

Osim toga, SSIM je uveden kako bi se ojačala sličnost strukture između sintetiziranih slika i stvarnih slika, a funkcija gubitka slična je jednadžbi. Kao rezultat toga, eksperimenti pokazuju njihovu učinkovitost poboljšanja, a performanse nadmašuju prethodne radove. Uzimajući u obzir da gornji okvir trpi od okluzija i lijevog ruba slike, na temelju okvira s trinokularnim pretpostavkama predložen je okvir za zadatke predviđanja dubine i semantike zajedno. Dizajniran je dodatni tok dekodera za procjenu semantičkih oznaka (labels) i obučen je na nadzirani način. Nadalje, termin diskontinuiteta među domenama (cross-domain discontinuity term) temeljen na predviđenoj semantičkoj slici se primjenjuje kako bi se poboljšala glatkoća predviđene karte dubine, koja pokazuje bolju izvedbu od prethodnih uvjeta gubitka glatkoće. Slično, postoji rad gdje se, također, koristi semantička segmentacija za poboljšanje procjene monokularne dubine. Procjena dubine i semantička segmentacija koriste dijeljeni mrežni okvir te se izmjenjuju prema uvjetu. Novi pojam lijevo-desne semantičke dosljednosti (left-right semantic consistency) predložen je za izvođenje procjene dubine s obzirom na regiju te poboljšava točnost i robusnost oba zadatka.

Metode temeljene na stereo usklađivanju Jedna od najpoznatijih stereo matching metoda u polunadziranom učenju se sastoji od mreže sinteze pogleda temeljene na Deep3D (view synthesis network based on Deep3D). Služi za estimaciju desne slike preko (od) lijeve slike, koja se razlikuje od gore navedenih radova. Štoviše, mreža stereo usklađivanja dizajnirana je za uzimanje neobrađene lijeve i sintetizirane desne slike za regresiju karte dispariteta. Tijekom treninga, mreža sinteze pogleda je nadzirana sirovim desnim slikama kako bi se poboljšala kvaliteta gradnje. Predviđene mape dispariteta (disparity maps) se koriste za rekonstrukciju lijevih slika iz procijenjenih desnih slika. Slično, imamo pristup u kojem se također iskorištava strategija stereo poduda-

ranja kako bi poboljšala izvedba i robusnost monokularne procjene dubine. Značajke s različitih stajališta sintetizirane su izvođenjem stereo usklađivanja, čime se postižu izvanredne performanse. Njihov mrežni okvir sastoji se od tri dijela, a to su:

- multi-scale ekstraktor značajki za ekstrakciju značajki na visokoj razini (multi-scale feature extractor for high-level feature extraction)
- mreža dispariteta za predviđanje mape dispariteta
- i mreža preciziranja za preciziranje dispariteta

U usporedbi s modelom koji je prvi naveden kao popularna metoda stereo usklađivanja, mreže predložene u ovoj metodi se zajednički obučavaju dok se prethodno navedene treniraju samostalno. Stoga je pojednostavljivanja složenost obuke kod mreža u ovoj metodi naspram mreža u prethodno navedenoj stereo matching metodi.

Metode temeljene na adversarial učenju i destilaciji znanja (knowledge distillation)

Kombiniranje naprednih mrežnih okvira, npr. adversarial learninga i destilacije znanja, postaje popularno i može značajno poboljšati izvođenje. Okvir destilacije znanja sastoji se od dvije neuronske mreže, mreže učitelja i mreže učenika. Mreža učitelja je složenija od mreže učenika. Svrha destilacije znanja je prenijeti znanje koje je naučila mreža učitelja u mrežu učenika, tako da se funkcije koje je naučio veliki model sažimaju u manje i brže modele. Destilacija znanja je korištena za prijenos informacija iz mreže za usavršavanje u učeničku mrežu. S obzirom na učinkovitost treninga sa sintetičkim slikama, usvojen je okvir cycle (cikličkog) GAN-a za transformaciju između sintetičke i stvarne domene za proširenje skupa podataka. Predložena je i mreža koje se naziva "geometry-aware symmetric domain adaptation network" (GASDA) kako bi se bolje iskoristili sintetički podaci. Ta mreža uči iz oznaka temeljne istine (ground truth labels) u sintetičkoj domeni te uči i epipolarnu geometriju stvarne domene, čime se postižu zavidni rezultati. Arhitektura generatora je poboljšana korištenjem modula prostorne korespondencije za podudaranje značajki i mehanizma pažnje za ponovno ponderiranje značajki.

Metode temeljene na rijetkoj temeljnoj istini (sparse ground truth)

Za jačanje nadziranih signala, sparse ground truth je široko uključena u okvir obuke. Reprezentativni rad za ovu metodu je usvojio ground truth dubinu koju je prikupio LIDAR za polunadzirano učenje. Osim toga, i lijeve i desne dubinske karte (D_l, D_r) procjenjuju CNN-ovi. Nadzorni signal temeljen na LIDAR podacima (G_l,G_r) formulira se na sljedeći način:

$$\mathscr{L}_{\text{recons}} = \sum_{p\in\Omega_Z} \|D_l(p) - G_l(p)\|_i$$
$$= \sum_{p\in\Omega_2} \|D_i(p) - G_i(p)\|_i$$

,gdje se $\omega_{Z,l}$ odnosi na skup piksela s dostupnim temeljnim istinama, a $\|*\|_\delta$ označava Berhuovu normu. Slično tome, drugi popularan pristup u ovom području je uveo gubitak između predviđene dubinske karte i LIDAR podataka kao dodatni signal. Štoviše, usvajaju se i fizičke informacije polu-nadzirane metode. Zatim se, u idućem značajnom radu, upotrebljava globalna orijentacija izračunata iz inercijskih mjerenja kao priori informacija za ograničavanje normalnih vektora na površine objekata. Općenito, normalni vektori na površine objekata su paralelni ili okomiti na smjer gravitacije, što se može lako izračunati iz karata procijenjene dubine. Stoga se značajno poboljšava točnost procjene dubine.

Polunadzirane metode postižu veću točnost od nenadziranih metoda zbog polunadziranih signala, te se informacije o mjerilu mogu naučiti iz tih signala. Međutim, točnost polunadziranih metoda uvelike se oslanja na temeljne istine, kao što su poza i LIDAR podaci, no svakako ih je lakše nabaviti nego skupe guste karte dubine.

4. Diplomski rad

4.1. Alati i postavke

4.1.1. Jezik, sučelje, biblioteke

Za uređivanje skripti, PyCharm se pokazao kao zahvalno razvojno okruženje. Vodič kojim je inspiriran algoritam modela neuronske mreže za estimaciju dubije je implementiran u "Google Colab" bilježnici. Takvo sučelje je pogodno za rad s podacima koji su dostupni online te kada nije potrebno spremati model nakon svake epohe. Kako je ovaj projekt služio za estimaciju dubine na konkretnom skupu podataka pohranjenom na privatnom tvrdom disku, prirodno je bilo odlučiti se za razvijanje softwarea lokalno na osobnom računalu. Također, Pycharm se vrlo jednostavno povezuje na GitHub, što omogućava veoma dobro praćenje, dokumentiranje, recikliranje i kontroliranje razvoja softwarea. Najbolje što takvo sučelje za git repozitorij omogućava jeste razvijanje i pohrana projekta u oblaku.

Programski jezik u kojemu je najzgodnije izgraditi rješenje za projekt ovakve prirode jest, naravno, Python. U Pythonu su implementirane najpoznatije biblioteke za strojno učenje i računalni vid koje su, dakako, i ovdje korištene. Radi se o bibliotekama kao što su : tensorflow, keras, matplotlib, seaborn, numpy i openCV.

Za vizualizaciju modela neuronske mreže je korištena aplikacija Netron.

Za treniranje je korišten Docker[1] Container tensorflow:2.6.0-gpu.

[1]Docker container image je lagani, samostalni, izvršni paket softvera koji uključuje sve što je potrebno za pokretanje aplikacije: kod, vrijeme izvođenja, sistemske alate, sistemske biblioteke i postavke Docker je platforma dizajnirana da pomogne programerima u izgradnji, dijeljenju i pokretanju modernih aplikacija. Docker riješava zamorno prirešivanje postavki, tako da se developer može usredotočiti na kod.

4.1.2. Softverske ovisnost

Zbog korištenja Docker Containera, bilo je potrebno imati mašinu s Linux operacijskim sustavom. Za svrhu treniranja je korišten server koji je imao instaliran Linus OS, dok je ostatak razvoja projekta bio na računalu s operacijskim sustavom Windows.

4.1.3. Hardverske ovisnosti

Server na kojemu su se pokretali treninzi je imao jedan GPU, odnosno jedinicu za grafičku obradu. To je znatno ubrzavalo treninge. Za pohranu podataka je poslužio SSD od 2TB.

4.2. Priprema podataka

4.2.1. Razumijevanja podataka

Za početak je trebalo skupiti podatke na kojima bi se spomenuta neuronska mreža trenirala. Skup podataka se sastojao od oko 20000 tripleta (trojki) slika. Svaku trojku su činile jedna monokularna slika slikana jednostavnom RGB kamerom te RBG slika sa svojom odgovarajućom dubinskom mapom koje su produkt ToF senzora.

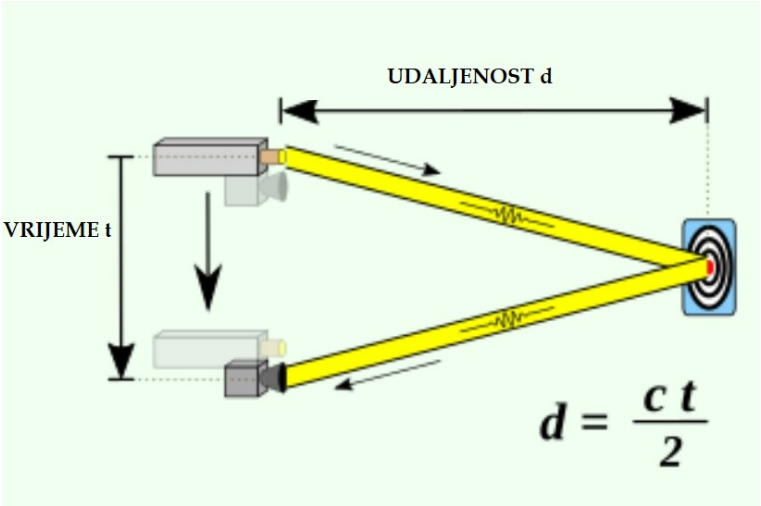
Vrlo je bitno da svaka trojka ima zajedničku identifikacijsku oznaku kako bi se u obradi podataka moglo što lakše baratati elementima tih uređenih parova slika. ID je bio sadržan u imenu datoteke. Također, korisno je kreirati JSON datoteku koja sadrži riječnik u kojemu su podaci za svaki pojedini par slike i dubinske mape.

Slika 4.1: Primjer tripleta RGB slike obične kamere, RGB slike i odgovarajuće dubinske mape ToF senzora

ToF kamera

Obične RGB slike su slikane jednostavnom RGB kamerom, dok su dubinske mape dobivene ToF (Time of Flight) kamerom. ToF senzor baca svjetlosni signal kojeg daje

laser ili LED te mjeri vrijeme povratnog putovanja tog signala. Na taj način se dobiva raspon udaljenosti od kamere.



Slika 4.2: Vrijeme leta svjetlosnog signala koji se odbija od metu
izvor slike : wikipedia.org

Na slici se može vidjeti kako funkcionira senzor te kako se mjeri udaljenost subjekta od objektiva. Formula je, dakle : $d = \frac{c*t}{2}$, pri čemu je d udaljenost, a t vrijeme povratnog putovanja svjetlosnog signala.

4.2.2. Transformacija i normalizacija podataka

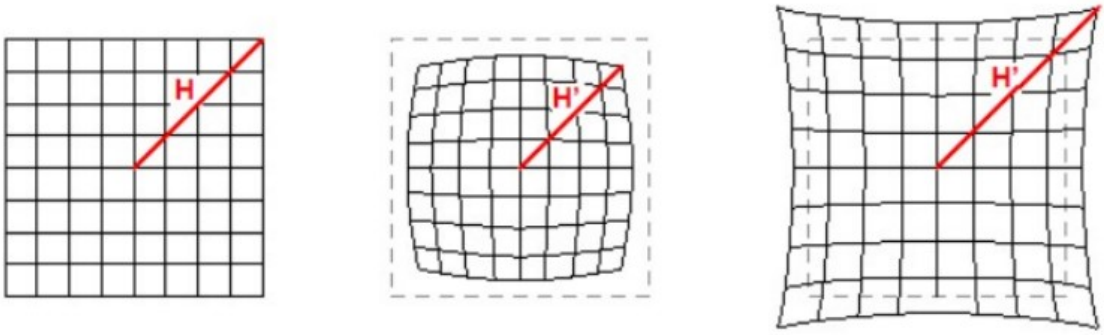
Cilj je bio transformirati i preoblikovati RGB sliku iz obične kamere i dubinsku mapu ToF kamere, kako bismo dobili parove od kojih bi se sastojao skup podataka za trening i validaciju modela estimacije dubine.

Relevantna informacija na slikama je bio sadržaj kutije i sama kutija. Pozadina, koju je činio pod i pokretna traka, je bila višak informacija. Zbog toga je trebalo provesti normalizaciju u vidu rezanja viška informacija oko kutije. Taj zadatak je obavljen tako što je, za početak, dio običnih RGB slika iz skupa podataka anotiran na način da su označeni gornji vrhovi kutija. Potom je tim anotiranim skupom slika nahranjen i istreniran model neuronske mreže za detekciju ključnih točaka. Pomoću tog modela je anotiran ostatak podataka te su sve informacije slikama, pripadnim identifikacijskim oznakama i pripadnim ključnim točkama pohranjene kao riječnik u JSON datoteku. Kada su koordinate ključnih točaka bile pohranjene, preostalo je, uz pomoć openCV paketa, implementirati algoritam koji bi spajao ključne točke tako da crta dužine koje bi obrubile kutiju. Posljednji korak u ovoj fazi je bio odrezati višak informacija i os-

taviti samo ono što je od interesa, a to je kutija sa svojim sadržajem. Ovaj postupak je analogno ponovljen na RGB slikama ToF kamere, te su predikcije ključnih točaka korištene za rezanje slika odgovarajućih dubinskih mapa ToF kamere.

Slika 4.3: Detektirane ključne točke na jednom tripletu slika iz skupa podataka

Normalizacija podataka je, također, provedena na nivou "neizobličenja" (undistortion). Slike dobivene iz RGB kamere su bile izobličene (distorted[2] image) te ih je se moralo ispraviti kako bi se mogle precizno upariti sa odgovarajućim slikama ToF kamere. Prvi korak je bilo mjerenje distorzije, a zatim je bilo potrebno koristiti transformacije za ispravljanje slika. Na slici je demonstriran problem izobličenja (distorzije) :



Slika 4.4: Neizobličena mreža, negativna distorzija i pozitivna distorzija
izvor slike : www.image-engineering.de

Izobličenje se mjeri sljedećom formulom:

$$D = \frac{\Delta H}{H} \cdot 100 = \frac{H^* - H}{H} \cdot 100$$

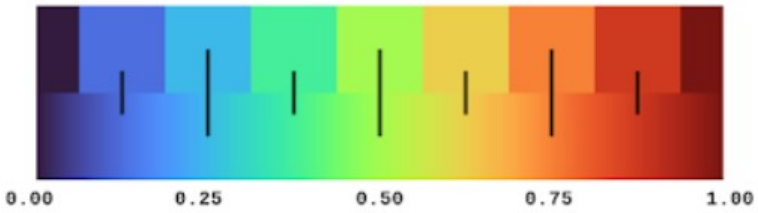
[2]Izobličenje slike je kada se čini da su ravne linije slike neprirodno deformirane ili zakrivljene, stvarajući različite vrste izobličenja, uključujući bačvaste, jastučaste i valne oblike. Izobličenje je često rezultat geometrije leće i može značajno narušiti kvalitetu slike

,gdje je :

- H* - udaljenost točke od središta slike
- H - nominalna udaljenost točke od središta slike na temelju proširene pravilne mreže

Algoritam za neizobličenje preslikava koordinate izlazne neizobličene slike na ulaznu sliku kamere pomoću koeficijenata izobličenja. Kao ulazne podatke za algoritam poništavanja, potrebno je bilo specificirati intrinzičnu matricu i koeficijente izobličenja koji opisuju izobličenje slike koje treba ispraviti. Intrinzična matrica sastoji se od žarišne duljine, optičkog središta (također poznatog kao glavna točka) i koeficijenta zakrivljenosti. Koeficijenti izobličenja matematički modeliraju radijalna i tangencijalna izobličenja[3]. Idući korak je izračunavanje ulaznih parametara iz navedenih dimenzija izlazne slike i objekta koji opisuje unutarnju matricu kamere, koeficijente izobličenja i žarišne duljine kamere u x- i y-smjerovima. Implementirana funkcija uklanja radijalna i tangencijalna izobličenja leće na ulaznoj slici pomoću izračunatih parametara.

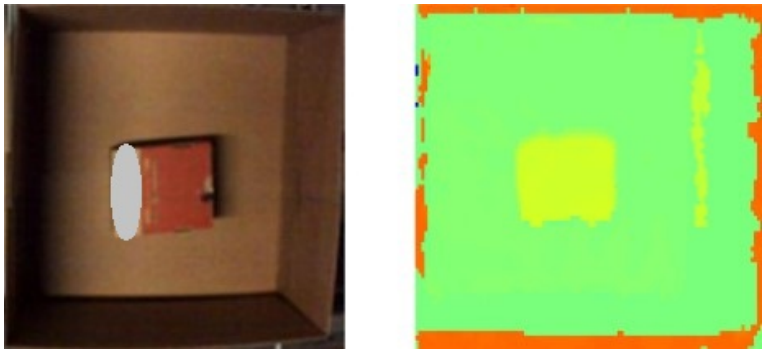
Još jedan važan razlog za normalizaciju podataka na opisan način jest to da ToF kamere često budu podešena na način da se spektar ponavlja nakon određene udaljenosti. Za skup podataka koji su korišteni za ovaj projekt, informacija o udaljenosti nakon koje se spektar ponavlja nije bila dostupna. Zbog toga je normalizacija dobro došla jer kada je izlučen samo dio koji prikazuje kutiju, jasno se vidi da su boje (udaljenosti od objektiva) iz jednog spektra koji ide od bijele do crne boje. Mapa boja je, inače, gore spomenuta skala boja koja reprezentira udaljenost. U projektima koji su predmet interesa ovog diplomskog rada, koristi se JET colormap.



Slika 4.5: JET colormap
izvor slike : ai.googleblog.com

Kada su podaci izrezani te je promijenjena mapa boja, konačno su dobiveni parovi slika i dubinskih mapa koji su spremni za ulaz u generator podataka.

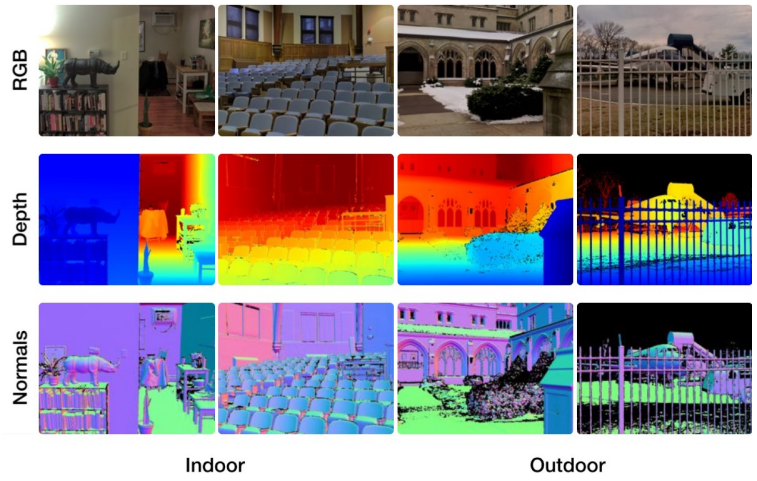
[3]Radijalno izobličenje nastaje kada se svjetlosne zrake savijaju više u blizini rubova leće nego u njenom optičkom središtu. Tangencijalno izobličenje nastaje kada leća i ravnina slike nisu paralelne.



Slika 4.6: Par RGB slike i odgovarajuće dubinske mape u JET mapi boja

4.2.3. Generator podataka

Inspiracija za implementaciju neuronske mreže dolazi iz primjera monokularne estimacije dubine rubrike "vision" na stranici keras.io. U tom tutorijalu se koristi skup podataka DIODE: A Dense Indoor and Outdoor Depth Set.



Slika 4.7: Skup podataka DIODE
izvor slike : diode-dataset.org

Kako bi se algoritam iz vodiča mogao upotrijebiti na vlastitom skupu podataka, potrebno je bilo uraditi određene modifikacije. Najprije je trebalo prilagoditi klasu za generiranje batcheva parova slika koji bi zajedno činili skup za treniranje, odnosno validaciju. Generator je jednostavno uzimao parove slika te im mijenjao veličinu i svojstva prema zadanim parametrima.

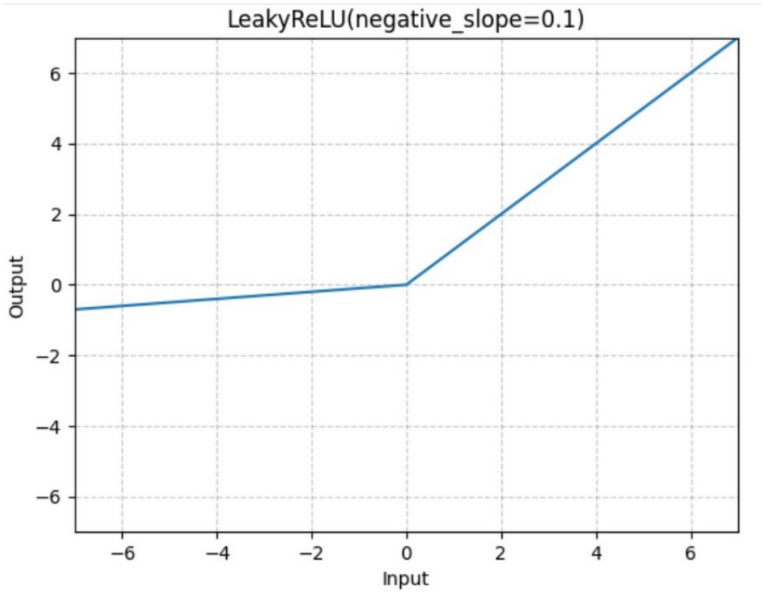
4.3. Model

Kada je generator spreman za konverziju skupa podataka u format pogodan za ulaz u neuronsku mrežu, modeliranje te mreže može početi. Model koji je korišten u ovom diplomskom radu se sastoji od, redom, četiri downscale bloka, jednog bottleneck bloka, četiri upscale bloka te posljednjeg - konvolucijskog sloja. Downscale blokovi su dobili ime po tome što se kod njih konvolucijskim slojevima kao argument šalje broj filtera koji je kod prvog downscale bloka najmanji definirani - na primjer : 16, a kod posljednjeg downscale bloka je broj filtera najveći - na primjer : 256. Kod upscale blokova je situacija obrnuta. Blokovi su građeni od konvolucijskih 2D slojeva, aktivacijskih LeakyReLU slojeva te slojeva za normalizaciju batchova. LeakyReLU aktivacijska funkcija inače je definirana na način:

$$LeakyReLU(x) = max(0,x) + negativeSlope * min(0,x)$$

ili

$$LeakyReLU = \begin{cases} x, & \text{if } x \geqslant 0 \\ negativeSlope * x, & \text{inače} \end{cases}$$



Slika 4.8: Primjer grafa LeakyReLU funkcije
izvor slike : pytorch.org

4.4. Funkcije gubitka

Sva se magija događa tijekom treninga – tijekom optimizacije funkcije gubitka. Funkcija gubitka je kombinacija (zbroj) tri loss funkcije. Te funkcije su SSIM loss, L1 loss i depth smoothness loss (loss glatkoće dubine). Svaka funkcija gubitka se množi svojom težinom. Te težine su hiperparametri koji se mogu mijenjati jer je moguće da će s promijenjenim vrijednostima trening biti stabilniji, a rezultati točniji.

4.4.1. SSIM metoda i funkcija gubitka

Mjera indeksa strukturne sličnosti (structural similarity index measure; SSIM) je metoda za predviđanje percipirane kvalitete digitalne televizije i kinematografske slike, kao i drugih vrsta digitalnih slika i videa. SSIM se koristi za mjerenje sličnosti između dvije slike. To je puna referentna metrika; drugim riječima, mjerenje ili predviđanje kvalitete slike temelji se na početnoj nekomprimiranoj slici ili slici bez izobličenja kao referentnoj.

SSIM je model temeljen na percepciji koji razmatra degradaciju slike kao percipiranu promjenu u strukturnim informacijama, dok također uključuje važne perceptualne pojave, uključujući izraze maskiranja svjetline i maskiranja kontrasta. Razlika u odnosu na druge tehnike kao što su MSE ili PSNR je u tome što ovi pristupi procjenjuju apsolutne pogreške. Strukturna informacija je ideja da pikseli imaju snažnu međuovisnost, posebno kada su prostorno blizu. Ove ovisnosti nose važne informacije o strukturi objekata u vizualnoj sceni. Maskiranje osvjetljenja je fenomen pri kojem su izobličenja slike (u ovom kontekstu) manje vidljiva u svijetlim područjima, dok je maskiranje kontrasta fenomen pri kojem izobličenja postaju manje vidljiva tamo gdje postoji značajna aktivnost ili "tekstura" na slici.

U modelu koji je korišten za estimaciju dubine u slučaju koji je predmet ovog rada, SSIM loss je definiran kao srednja vrijednost razlike $1 - SSIM(target, predikcija)$. SSIM funkcija je definirana u tensorflow.image biblioteci. Target je "ground truth" tenzor koji reprezentira originalnu dubinsku mapu. Predikcija je tenzorska reprezentacija estimacije koju model vraća kao izlaz.

Algoritam

SSIM indeks izračunava se na različitim prozorima slike. Mjera između dva prozora x i y zajedničke dimenzije N je :

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

gdje je :

- μ_x mean uzorka pixela od x
- μ_y mean uzorka pixela od y
- σ_x^2 varijanca x :
- σ_y^2 varijanca y_i
- σ_{xy} kovarijanca x i y.
- $c_1 = (k_1 L)^2, c_2 = (k_2 L)^2$ dvije varijable za stabilizaciju dijeljenja sa slabim nazivnikom;
- L dinamički raspon vrijednosti piksela (obično je to $2^{\#\text{bits per pixel}} - 1$);
- $k_1 = 0.01$ i $k_2 = 0.03$ prema zadanim postavkama.

Komponente formule

SSIM formula temelji se na tri usporedna mjerenja između uzoraka x i y:

- svjetlina (l)
- kontrast (c)
- struktura (s).

Pojedinačne funkcije usporedbe su :

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}$$
$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}$$
$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}$$

pri čemu je $c_3 = \frac{c_2}{2}$.

SSIM je tada ponderirana kombinacija tih usporednih mjera:

$$\mathrm{SSIM}(x,y) = l(x,y)^\alpha \cdot c(x,y)^\beta \cdot s(x,y)^\gamma$$

Postavljanjem vrijednosti α, β, γ na 1 , formula se može svesti na gore prikazani oblik.

Matematička svojstva

SSIM zadovoljava identitet nerazlučivosti i svojstva simetrije, ali ne i nejednakost trokuta ili nenegativnost, te stoga nije funkcija udaljenosti. Međutim, pod određenim uvjetima, SSIM se može pretvoriti u normaliziranu korijensku mjeru MSE, koja je funkcija udaljenosti. Kvadrat takve funkcije nije konveksan, ali je lokalno konveksan i kvazikonveksan, što SSIM čini mogućim ciljem za optimizaciju.

Aplikacija formule

Kako bi se procijenila kvaliteta slike, ova se formula obično primjenjuje samo na svjetlinu, iako se također može primijeniti na vrijednosti boja (npr. RGB) ili kromatske (npr. YCbCr) vrijednosti. Rezultirajući SSIM indeks je decimalna vrijednost između -1 i 1, gdje 1 označava savršenu sličnost, 0 označava da nema sličnosti, a -1 označava savršenu anti-korelaciju. Za sliku se obično izračunava pomoću kliznog Gaussovog prozora veličine 11x11 ili blok prozora veličine 8x8. Prozor se može pomicati piksel po piksel na slici kako bi se stvorila mapa kvalitete SSIM slike. U slučaju procjene kvalitete videa, autori predlažu korištenje samo podskupine mogućih prozora kako bi se smanjila složenost izračuna.

4.4.2. L1 funkcija gubitka

L1 funkcija gubitka koristi se za minimiziranje pogreške koja je zbroj svih apsolutnih razlika između prave vrijednosti i predviđene vrijednosti.

$$\text{L1LossFunction} = \sum_{i=1}^{n} \left| y_{\text{true}} - y_{\text{predicted}} \right|$$

U modelu je definirana je kao reducirana srednja vrijednost (reduce mean) od apsolutne razlike targeta i predikcije. Reducirana srednja vrijednost, inače, izračunava srednju vrijednost elemenata kroz dimenzije tenzora.

4.4.3. Depth smoothness loss

Treća funkcija gubitka je vrlo zanimljiva. Zove se depth smoothness loss ili funkcija gubitka glatkoće dubine. Ta funkcija je zapravo reducirana srednja vrijednost (reduce mean) zbroja apsolutnih glatkoća varijabli x i y. Glatkoća varijable x je gradijednt slike dx pomnožen s težinom varijable x. Slika čiji se gradijent uzima je predikcija modela neuronske mreže. Težina varijable x je eksponencijalna funkcija kojoj je baza eulerov

broj e, a eksponent je reducirani mean apsolutne vrijednosti komponente dx uređenog para gradijenata slike (dy,dx) iz uređenog para gradijenata slike (dy,dx). Slika čiji gradijent uzimamo je originalna slika dubinske mape. Analogno se definira glatkoća varijable y. Dakle, depth smoothness loss f funkcija je definirana na sljedeći način :

$$DepthSmoothnessLoss = reduceMEAN\left(|\text{smoothness}_x|\right) + reduceMEAN\left(\left|\text{smoothness}_y\right|\right)$$

pri čemu je :

- $\text{smoothness}_x = dx_{pred} * \text{weights}_x$ - glatkoća varijable x
- $\text{smoothness}_y = dy_{pred} * \text{weights}_y$ - glatkoća varijable y
- $\text{weights}_x = e^{reduceMEAN(|\text{dx}_{true}|)}$ - težine varijable x
- $\text{weights}_y = e^{reduceMEAN(|\text{dy}_{true}|)}$ - težine varijable y
- $dy_{true}, dx_{true} = \nabla_{image}\left(\text{target}\right)$ - Gradijenti originalne dubinske mape
- $dy_{pred}, dx_{pred} = \nabla_{image}\left(\text{pred}\right)$ - Gradijenti estimirane dubinske mape

Gradijent slike

U ovoj funkciji su jasne sve varijable osim gradijenata slika. Gradijent slike je usmjerena promjena intenziteta ili boje na slici. Gradijent slike jedan je od temeljnih gradivnih blokova u obradi slike. Na primjer, detektor rubova Canny koristi gradijent slike za detekciju rubova. U grafičkom softveru za uređivanje digitalnih slika, pojam gradijent ili gradijent boje također se koristi za postupno miješanje boja koje se može smatrati ravnomjernom gradacijom od niskih do visokih vrijednosti, kao što se koristi od bijele do crne. Drugi naziv za to je progresija boja.

Matematički gledano, gradijent funkcije s dvije varijable (ovdje funkcija intenziteta slike) u svakoj točki slike je 2D vektor s komponentama danim derivacijama u vodoravnom i okomitom smjeru. Na svakoj točki slike, vektor gradijenta pokazuje u smjeru najvećeg mogućeg povećanja intenziteta, a duljina vektora gradijenta odgovara brzini promjene u tom smjeru.

Budući da je funkcija intenziteta digitalne slike poznata samo u diskretnim točkama, derivacije ove funkcije ne mogu se definirati osim ako pretpostavimo da postoji temeljna kontinuirana funkcija intenziteta koja je uzorkovana na točkama slike. Uz neke dodatne pretpostavke, derivacija kontinuirane funkcije intenziteta može se izračunati kao funkcija na uzorkovanoj funkciji intenziteta, tj. digitalnoj slici. Aproksimacije ovih izvedenih funkcija mogu se definirati s različitim stupnjevima točnosti. Najčešći

način aproksimacije gradijenta slike je konvolucija slike s jezgrom, kao što je Sobelov operator ili Prewittov operator.

Gradijenti slika često se koriste u kartama i drugim vizualnim prikazima podataka kako bi se prenijele dodatne informacije. GIS alati koriste progresiju boja za označavanje nadmorske visine i gustoće naseljenosti.

Gradijenti slike mogu se koristiti za izdvajanje informacija iz slika. Gradijentne slike stvaraju se iz izvorne slike (općenito konvolviranjem s filtrom) u tu svrhu. Svaki piksel gradijentne slike mjeri promjenu intenziteta te iste točke na izvornoj slici, u određenom smjeru. Da bi se dobio cijeli raspon smjera, izračunavaju se slike gradijenta u smjerovima x i y.

Jedna od najčešćih upotreba je otkrivanje rubova. Nakon što su slike gradijenta izračunate, pikseli s velikim vrijednostima gradijenta postaju mogući rubni pikseli. Pikseli s najvećim vrijednostima gradijenta u smjeru gradijenta postaju rubni pikseli, a rubovi se mogu pratiti u smjeru okomitom na smjer gradijenta. Jedan primjer algoritma za otkrivanje rubova koji koristi gradijente je detektor rubova Canny.

Gradijenti slike u kompjuterskom vidu se mogu koristiti za robusno podudaranje značajki i tekstura. Različito osvjetljenje ili svojstva kamere mogu uzrokovati da dvije slike iste scene imaju drastično različite vrijednosti piksela. To može uzrokovati neuspjeh algoritama za podudaranje vrlo sličnih ili identičnih značajki. Jedan od načina da se to riješi je izračunavanje tekstura ili potpisa značajki na temelju gradijentnih slika izračunatih iz izvornih slika. Ti su gradijenti manje osjetljivi na promjene osvjetljenja i kamere, tako da su pogreške pri podudaranju smanjene.

Gradijent slike je vektor njezinih parcijala:

$$\nabla f = \begin{bmatrix} g_x \\ g_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

pri čemu je :

- $\frac{\partial f}{\partial x}$ derivacija u odnosu na x (gradijent u smjeru x)
- $\frac{\partial f}{\partial y}$ derivacija u odnosu na y (gradijent u smjeru y)

Derivacija slike može se aproksimirati konačnim razlikama. Ako se koristi središnja razlika, za izračun $\frac{\partial f}{\partial y}$ možemo primijeniti 1-dimenzionalni filter na sliku $\mathbf{A}$ konvolucijom:

$$\frac{\partial f}{\partial y} = \begin{bmatrix} -1 \\ +1 \end{bmatrix} * \mathbf{A}$$

gdje $*$ označava operaciju 1-dimenzionalne konvolucije. Ovaj filtar 2×1 pomaknut

će sliku za pola piksela. Da bismo to izbjegli, sljedeći 3×1 filter

$$\begin{bmatrix} -1 \\ 0 \\ +1 \end{bmatrix}$$

može se koristiti. Smjer gradijenta može se izračunati formulom:

$$\theta = \tan^{-1}\left[\frac{g_y}{g_x}\right]$$

a veličina je dana sa:

$$\sqrt{g_y^2 + g_x^2}$$

4.4.4. Definirana funkcija gubitka

Konačna funkcija gubitka je zbroj triju navedenih funkcija gubitaka koje su ponderirane odgovarajućim težinama. Za svaku funkciju gubitka je definiran skalar koji ima ulogu težine. Te težine su hiperparametri koje možemo mijenjati i evaluacijom modela pratiti kako se predikcije koje daje model poboljšavaju, odnosno pogoršavju s obzirom na inicijalizirane težine. Dakle, konačna funkcija gubitka izgleda ovako :

$$\text{loss} = \text{weight}_{\text{ssim}} * \text{ssimLoss} + \text{weight}_{\text{l1}} * \text{l1Loss} + \text{weight}_{\text{edgeLoss}} * \text{depthSmoothnessLoss}$$

4.5. Trening i evaluacija

Model je treniran 10000 epoha. Korišten je optimizator Adam. Adam je optimizacijski algoritam koji se može koristiti umjesto klasične stohastičke procedure spuštanja gradijenta za ažuriranje iteracije težine mreže na temelju podataka o obuci. Autori opisuju Adama kao kombinaciju prednosti dva druga proširenja stohastičkog gradijentnog spuštanja.

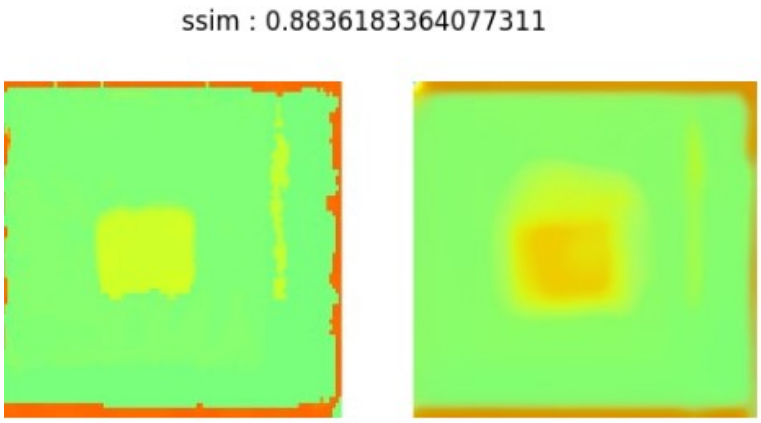
- Adaptivni gradijentni algoritam (AdaGrad) koji održava stopu učenja po parametru koja poboljšava izvedbu na problemima s rijetkim gradijentima (npr. prirodni jezik i problemi s računalnim vidom). Propagacija srednje kvadratne vrijednosti
- (RMSProp) koja također održava stope učenja po parametru koje se prilagođavaju na temelju prosjeka nedavnih veličina gradijenata za težinu (npr. koliko brzo se mijenja). To znači da algoritam dobro radi na mrežnim i nestacionarnim problemima (npr. buka).

Adam shvaća prednosti i AdaGrada i RMSProp-a. Umjesto prilagodbe stopa učenja parametara na temelju prosječnog prvog trenutka (srednja vrijednost) kao u RMSProp, Adam također koristi prosjek drugih trenutaka gradijenata (necentrirana varijanca).

Za kompajliranje je korištena funkcija gubitka Sparse Categorical Crossentropy koja izračunava loss krosentropije između originalnih podataka i predviđanja.

Pri treningu je spreman istrenirani model nakon svake epohe. Za evaluaciju je implementiran algoritam koji učitava sve spremljene modele i radi prekdikcije na skupu podataka za evaluaciju. Glavna mjera za evaluaciju je korišten SSIM. Za svaki model je dobivena predikcija i izračunat SSIM za svaku pojedinu sliku skupa podataka za evaluciju. Svi SSIM-ovi su zbrojeni i "pobjednički" model je bio onaj koji imaju maksimalnu vrijednost sume SSIM-ova.

Na slici možemo vidjeti kako je izgledala estimacija dubine jednog od spremljenih modela pri treniranju. Iznad slike stoji SSIM vrijednost koja je u ovom slučaju veća od 0.88. To se može interpretirati kao da je sličnost između originalne dubinske mape i estimacije oko 88%.



Slika 4.9: Originalna dubinska mapa (lijevo) i estimirana dubinska mapa (desno)

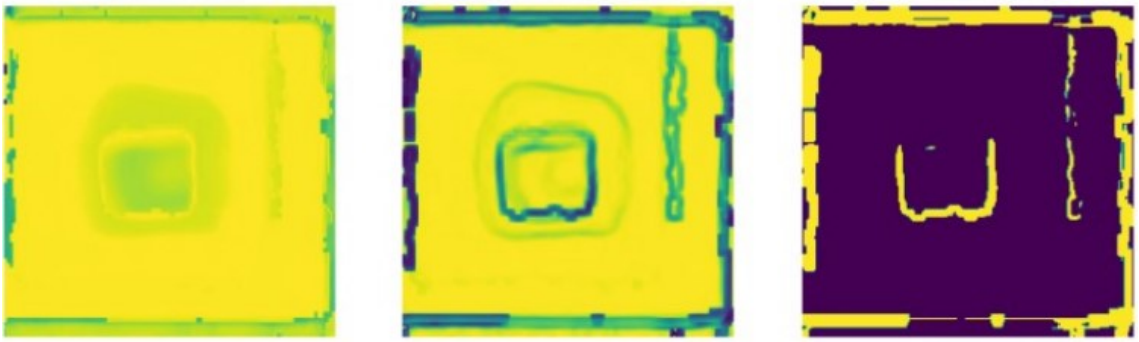
Druga mjera koja je korištena za evaluaciju je razlika. Na slici dolje, možemo vidjeti vizuale koji prikazuju razliku, SSIM razliku i SSIM treshold između dvije slike na Slici 4.8. Prikazi spomenutih razlika se mogu shvatiti kao mape pogrške (error maps). Razlika je definirana kao :

$$D = 255 - \|img1 - img2\|$$

Dakle, od 255 se oduzima apsolutna razlika originalne dubinske mape i estimirane mape. U RGB reprezentaciji, (255,255,255) je bijela boja, dok na crno-bijeloj (greyscale) skali, 255 predstavlja bijelu boju, a 0 crnu. Dakle, kako bi se dobila razlika koju

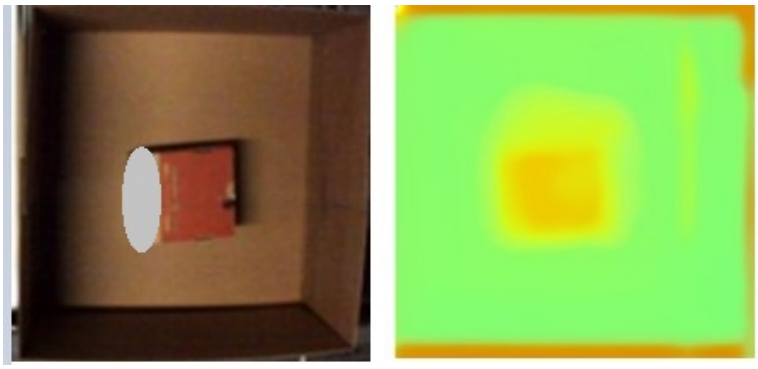
se može lako i vizualizirati, od vrijednosti bijele boje se oduzima apsolutna razlika boja odgovarajućih piksela dvaju slika. Kada ne bi bilo razlike na slikama, kao izlaz bismo dobili bijelu pozadinu. Svjetliji tonovi na vizualu razlike su poželjni, a tamniji ukazuju na značajnija odudaranja između promatranih slika. Rezultat ove mjere je, također, zadovaljavajuć - to se vidi i na slici razlike koja je svjetlije (žute) pozadine sa zelenkastim obrisima u sredini koji nisu drastično u kontrastu sa pozadinom.

Na srednjoj slici koja prikazuje SSIM razliku, može se vidjeti razlika srednjih vrijednosti indeksa strukturne sličnosti. Na takvom grafičkom prikazu se lakše uočavaju razlike u rubovima. Kod estimacije se, golim okom, teže zamijećuju pogreške jer se rubovi izglađuju te se gubi oštrina. Zbog toga je SSIM razlika zahvalna mjera, kako za vizualizaciju, tako i za analizu samog modela.



Slika 4.10: Razlika, SSIM razlika, SSIM treshold na originalnoj i na estimiranoj mapi dubine

Dakle, na kraju za uzlaz kao na lijevoj slici, dobivamo izlaz kao na desnoj slici (Slika 4.10.) :



Slika 4.11: Ulazna slika i estimirana dubinska mapa

4.6. Moguća poboljšanja

Vizualna inspekcija rezultata i estimacija preko gore navedenih mjera pokazuju zadovoljavajuće rezultate na ovoj razini. Međutim, postoji dosta prostora za poboljšanje i podizanje indeksa strukturne sličnosti na 100%. Model se može poboljšati zamjenom dijela mreže s unaprijed obučenim mrežama kao što su DenseNet ili ResNet. Takav pristup se naziva Transfer Learning (učenje prenošenjem). Funkcije gubitaka igraju važnu ulogu u rješavanju ovog problema. Podešavanje funkcija gubitaka može dovesti do značajnog poboljšanja. To se odnosi na igranje s različitim kombinacijama vrijednosti težina za funkcije gubitka, definicija konačne funckije gubitka te definicija pojedinačnih loss funkcija od kojih se konačna loss funkcija sastoji.

5. Zaključak

Monokularna estimacija dubine se pokazala kao adekvatno rješenje za problem mjerenja volumena praznog prostora u kutijama. Donekle rješava i sestrinski problem određivanja oblika zaštitnog materijala prije konstrukcije. Za taj dio problema bi bilo dobro poboljšati performans modela modela za estimaciju dubine. Moram komentirati da skup podataka dobivenih iz ToF kamere nije najbolji mogući. Čini mi se da takva kamera daje precizne informacije kada se radi o dubini slika na kojima je scena s većim udaljenostima - na primjer, slika prostorije ili scena sa helikopterom na planini koja je prikazana na početku ovog rada. Smatram da za problem kutija, čija se visina mjeri u centimetrima, treba senzor koji daje preciznije informacije o dubini, odnosno senzor koji može uočiti razlike u visinama predmeta koji se razlikuju za centimetar. Što se tiče samog modela, mišljenja sam da prototip dosta dobro radi s obzirom na podatke na kojima je treniran. Što se tiče pristupa - monokularna estimacija dubine koja se temelji na dubokom učenju, sigurno je dosta dobro rješenje u vidu automatizacije. Razlog odabira ove metode jest to što je end-to-end način razvijanja dobrodošao. Također, istraživanja su pokazala da je ovaj pristup zahvalniji od tradicionalnih što se tiče točnosti estimacije dubine te mogućnosti modificiranja i poboljšavanja modela.

6. Literatura

[1] Victor Basu. Monocular depth estimation. https://keras.io/examples/vision/depth_estimation/, Kolovoz 2021.

[2] Amlaan Bhoi. Monocular depth estimation: A survey. https://arxiv.org/abs/1901.09402, Siječanj 2019.

[3] Michael Firman Gabriel Brostow Clement Godard, Oisin Mac Aodha. Digging into self-supervised monocular depth estimation. https://openaccess.thecvf.com/content_ICCV_2019/papers/Godard_Digging_Into_Self-Supervised_Monocular_Depth_Estimation_ICCV_2019_paper.pdf.

[4] Vasileios Belagiannis Federico Tombari Nassir Navab Iro Laina, Christian Rupprecht. Deeper depth prediction with fully convolutional residual networks. https://arxiv.org/abs/1606.00373.

[5] David Jacobs. Image gradients. https://www.cs.umd.edu/~djacobs/CMSC426/ImageGradients.pdf, Jesen 2005.

[6] Reza Mahjourian Anelia Angelova Vincent Casser, Soeren Pirk. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. https://arxiv.org/pdf/1811.06152v1.

[7] Reza Mahjourian Anelia Angelova Vincent Casser, Soeren Pirk. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. https://arxiv.org/pdf/1606.00373v2, 2019.

[8] Sheikh Simoncell Wang, Zhou; Bovik. Image quality assessment: from error visibility to structural similarity, Siječanj 2004.

[9] Chunxiao Fana Hui Yub Yue Minga, Xuyang Menga. Deep learning for monocular depth estimation: A review. https://pure.port.ac.

uk/ws/portalfiles/portal/26286067/Deep_Learning_for_Monocular_Depth_Estimation_A_Review_pp.pdf.

7. Sažetak

Monokularna estimacija dubine je tehnika za dobivanje dubinske mape jedne (monokularne) RGB slike. Na dubinskoj mapi su procijenjene vrijednosti dubine svakog piksela na slici. Dubina je zapravo udaljenost subjekta od objektiva kamere, a na slici je izražena skalom boja. Za dobivanje dubinskih mapa se koriste ToF senzori.

Ovaj diplomski rad je temeljen na projektu koji je za cilj imao optimizirati i osigurati transport kutija. Za ispunjavanje tog zadatka, potrebno je bilo odrediti oblik i veličinu zaštitnog materijala (stiropor) kojim bi se punili paketi. Ideja je bila to riješiti i automatizirati pomoću strojnog učenja i računalnog vida. Estimacija dubine je bila preduvjet za konstrukciju sigurnosnog materijala i zbog toga je važan međukorak.

Ovaj diplomski rad opisuje strukturu, analizu i obradu podataka koji se koriste pri procjeni dubine. Također, opisuje probleme s kojima se developer može susresti pri radu sa takvim skupom podataka te daje prijedloge za rješavanje tih problema. U ovom radu je predložena normalizacija u vidu ekstrakcije bitnog sadržaja sa slike preko detekcije ključnih točaka. Predloženi su i adekvatni alati, tehnologije i postavke. Tu se, kao ključni pojmovi, mogu izdvojiti PyCharm, Python, git, tensorflow, keras, openCV, docker kontejneri, Linux i GPU.

Područje istraživanja na temu Estimacije dubine je bogato različitim pristupima. Neki od tih pristupa su : estimacija dubine pomoću strukture iz pokreta (structure from motion -SFM) te estimacija dubine pomoću podudaranja stereo vizije (stereo vision matching). Za ovaj problem se pristup koji nudi monokularna estimacija dubine temeljena na dubokom učenju činio najprikladniji. Stoga je opisan jedan model neuronske mreže koji daje procjenjuje dubinsku mapu ukoliko ga nahranimo ulaznim podacima koji su originalne slike čije se dubinske mape želi procijeniti. U tom modelu, najvažnije je razumijeti funkcije gubitka, njihovu matematičku pozadinu i koncept. Funkcija gubitka je kombinacija ponderiranih individualnih loss funkcija. Riječ je o funkcijama : SSIM loss, L1 loss te depth smothness loss.

Model je evaluiran indeksom strukturne sličnosti te razlikom dvaju slika u vidu razlike vrijednosti piksela. Pokazalo se da je sličnost između originalne dubinske mape

te one koju vraća neuronska mreža, u prosjeku, oko 88%. Takvim rezultatom smo zadovoljni iako su poboljšanja performansa uvijek dobrodošla. Kao mogućnosti za poboljšanje su navedene opcije modificiranja modela u vidu prenesenog učenja i promijene parametara u funkciji gubitka.

Može se zaključiti da je estimacija dubine krucijalan korak pri konstrukciji softwarea za gore navedenu problematiku te da je ovaj prototip veoma dobar "base line" za daljnji razvoj robota koji bi optimizirao transport paketa s osjetljivim sadržajem.