SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI PROJEKT

Procjena dubine objekata na slikama primjenom MultiScale duboke mreže

Marko Tutić
Voditelj: *Izv. prof. dr. sc. Zoran Kalafatić*

Zagreb, siječanj 2023.

SADRŽAJ

1. Uvod

Procjena dubine je proces određivanja udaljenosti objekata u sceni od točke snimanja same scene. Informacija o dubini može se koristiti za 3D vizualizaciju prostora na slici. Dubina je važna komponenta u mnogim područjima poput robotike, proširene stvarnosti i autonomnih vozila.

Dvije najčešće kategorije procjene dubine su procjena dubine primjenom stereoskopskih kamera i monokularna procjena dubine.

Stereoskopska procjena dubine koristi dvije odvojene kamere kako bi procijenila dubinu, slično kao što i čovjek ostvaruje trodimenzionalnu percepciju zbog dva oka. Takva metoda daje gotovo savršene rezultate zbog mogućnosti jednoznačnog određivanja udaljenosti objekta.

Monokularna procjena dubine koristi samo jednu sliku za procjenu dubine i teži je problem od stereoskopske procjene dubine jer jedna slika ne sadrži dovoljno informacija za jednoznačnu procjenu dubine, odnosno više različitih scena može davati jednake rezultate. Za monokularnu procjenu dubine razvijene su razne tehnike procjene bazirane na dubokom učenju. Jedan od takvih modela obrađen je u sklopu projekta.

Korišteni model se od hijerarhijske topologije mreže u kojoj jedan dio mreže procjenjuje dubinu tako da za jednu točku za koju se procjenjuje dubina ima na raspolaganju cijelu sliku, odnosno potpunu informaciju. Drugi dio mreže nastoji preciznije procijeniti dubinu na temelju izvorne slike i procjene koju je dala prva mreža pri čemu prilikom procjene dubine za neku točku nema na raspolaganju cijelu sliku, već samo manji lokalni dio slike.

2. Opis problema

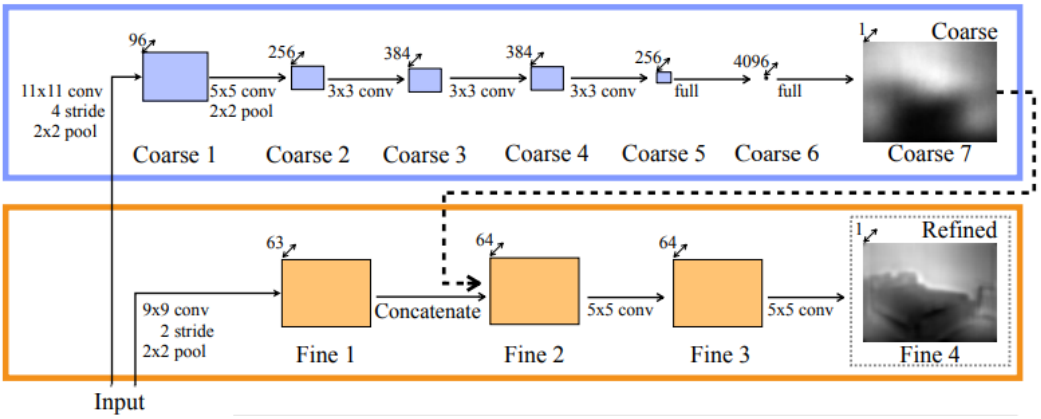
Za monokularnu procjenu dubine isproban je model koji se temelji na dvije komponente: mreža za grubu procjenu dubine na razini cijele slike i mreža za preciznu procjenu dubine [1].

Mreža za grubu procjenu dubine sastoji se od 5 konvolucijskih slojeva za ekstrakciju značajki iz slike i 2 potpuno povezana sloja u kojem prvi sloj sadrži onoliko neurona koliko je piksela u slici, a drugi sloj sadrži 4096 neurona. Iza svakog konvolucijskog sloja dolazi sloj sažimanja primjenom funkcije maksimuma te se iza sloja sažimanja primjenjuje aktivacijska funkcija zglobnice (engl. ReLU).

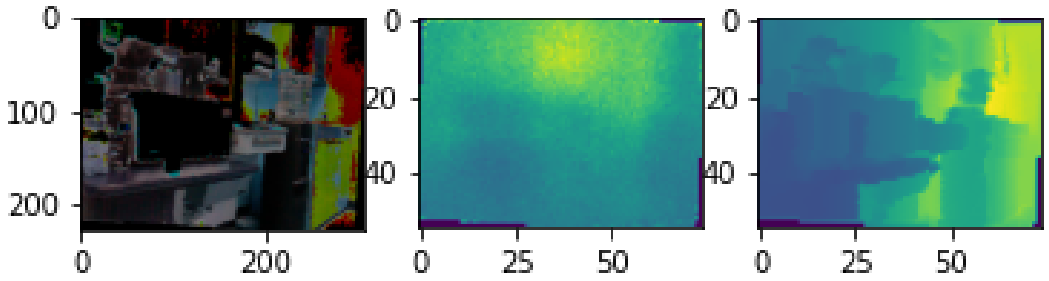
Ideja potpuno povezanih slojeva na kraju mreže je da svaki neuron u takvom sloju ima dostupnu informaciju o cijeloj slici i da na temelju cijele slike procjeni dubinu u toj točki. Konačan izlaz takve mreže je gruba procjena dubine, odnosno mapa dubina na slici. Takva procjena je prilično mutna te je zadatak mreže za finu procjenu dubine. Primjer izlaza takve mreže dan je slikom 2.2. Mapa dubina dobivena takvom mrežom dodaje se na postojeće mape značajki u drugom konvolucijskom sloju.

Druga mreža sadrži 3 konvolucijska sloja te je svrha tog dijela mreže precizno procijeniti dubinu pomoću manjih dijelova slike. Gruba procjena dubine dobivena iz prve mreže služi drugoj mreži kao temeljna informacija za još precizniju procjenu dubine.

Slika 2.1 prikazuje arhitekturu korištenog modela.



Slika 2.1: Arhitektura modela[1]



Slika 2.2: Primjer izlaza modela za grubu procjenu dubine

3. Skup podataka

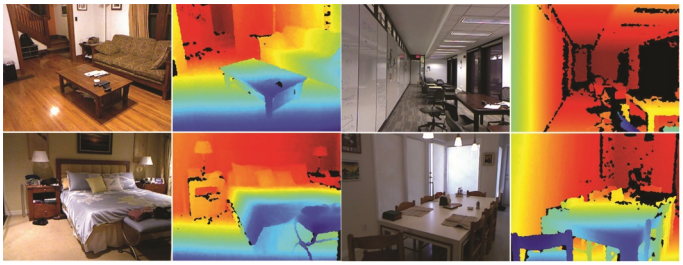
Za učenje modela korišten je skup podataka NYUv2 Depth [2]. Navedeni skup podataka sastoji se od 1449 označenih slika te je uz svaku sliku priložena je i mapa dubina. Sve slike sadrže slike interijera stanova, zgrada i sl koje su snimljene u 464 različitih scena i snimljenih u 3 različita grada. Skup podataka dodatno sadrži 407024 neoznačenih slika no one nisu bile korištene zbog potrebe za oznakama prilikom nadziranog učenja.

Vrijednosti u mapama dubina predstavljaju udaljenost objekta na slici od kamere kojom je slika snimana te je ta udaljenost izražena u metrima. Skup podataka dodatno sadrži mapu oznaka u kojoj se nalaze razredi u koje se pikseli na slici svrstavaju prilikom semantičke segmentacije, no navedene oznake izlaze iz okvira eksperimenta. Slike su uslikane stereoskopskom kamerom Microsoft Kinect te su pomoću nje dobivene mape dubine.

Slike te njihove mape dubina su izvorno dimenzija 640x480 no za potrebe eksperimenta se slikama mijenjaju dimenzije na 304x228. S obzirom na to da model za procjenu dubine vraća mape dubina koje imaju manju veličinu od ulazne slike, mapama dubina se mijenja veličina sa 640x480 na 74x55 kako bi dimenzije izlaza modela i točne oznake odgovarale.

Skup podataka se dijeli na skup za učenje modela, skup za validaciju modela i skup za testiranje. Skup za učenje sadrži 60% slika, skup za validaciju sadrži 20% slika dok skup za testiranje sadrži 20% slika. Podaci su se proširivali tako da su se slika i njena mapa dubina s vjerojatnošću 50% horizontalno zrcalili. Dodatno su se slike nasumično rotirale oko središta u rasponu od -5 ° do 5 °. Ovo proširenje kao rezultat daje 20 puta više slika kao primjere za učenje te se samim time poboljšavaju i svojstva generalizacije.

Slika 3.1 prikazuje primjere slika za koje se procjenjuje dubina i njihovih oznaka.



Slika 3.1: Primjer slika [2]

4. Eksperimenti

Model je bio treniran 3 puta nad istim podacima uz različite funkcije gubitka i načine učenja modela. U prvoj inačici eksperimenta, model je bio treniran tako da se prvo učio dio mreže za grubu procjenu dubine, a zatim dio mreže za finu procjenu. Gubitak koji je bio korišten je srednja kvadratna pogreška i definiran je sljedećom formulom:

$$L(y, y^*) = \frac{1}{N} * \sum_{i=1}^{N} (y_i - y_i^*)^2$$

U drugoj inačici se gubitak modificira tako da je invarijantan na skalu te je sada formula za gubitak sljedeća:

$$L(y, y^*) = \frac{1}{N} * \sum_{i=1}^{N} (\log y_i - \log y_i^* + \frac{1}{N} * \sum_{j=1}^{N} (\log y_j^* - \log y_j))^2$$

Konačno u trećoj inačici se mreža za grubu i finu procjenu dubine uče istovremeno, odnosno tretiraju se kao jedna mreža. U ovoj inačici se koristi srednja kvadratna pogreška kao gubitak.

Model za finu i model za grubu procjenu dubine učili su se 1000 epoha te je veličina jedne mini grupe kod učenja iznosila 32.

Za optimizaciju parametara korišten je stohastički gradijentni spust s veličinom mini grupe 16. Parametri za pojedine slojeve prilagođavali su se različitim intenzitetom. U mreži za procjenu dubine cijele slike korištena je stopa učenja u iznosu od 0.1, dok je za konvolucijske slojeve korištena stopa učenja 0.001. Kod konvolucijskih slojeva za finu procjenu dubine, prvi i treći sloj imali su stopu učenja 0.001, dok je drugi konvolucijski sloj imao stopu učenja 0.01.

Za poboljšavanje svojstava generalizacije, korišten je dropout u prvom potpuno povezanom sloju. Dropout postavlja izlaz slučajno izabranih neurona na 0 i time sprječava daljnju propagaciju gradijenta unatrag. Kako bi se spriječila situacija u kojoj se pojavljuje eksplodirajući gradijent, koristi se gradient clipping. Gradient clipping koristi se tako da je najveća dozvoljena norma vektora gradijenta 1.0.

5. Rezultati

Za mjerenje dobrote nekog modela za monokularnu procjenu korištene su sljedeće metrike: prag, apsolutna relativna razlika, kvadrirana relativna razlika i relativna srednja kvadratna pogreška u linearnom prostoru, log prostoru te u log prostoru invarijantnom na skalu.
Prag je definiran kao postotak piksela koji zadovoljavaju sljedeću nejednakost:

$$max(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}) = \delta < thresh$$

Mjera može poprimiti vrijednosti između 0 i 1. U idealnom slučaju vrijednost δ iznosi 1 za sve piksele te je stoga za valjane rezultate vrijednost $thresh$ potrebno postaviti na vrijednost veću od 1.
Apsolutna relativna razlika dana je sljedećom formulom:

$$\frac{1}{|T|}\sum_{yT}\frac{|y-y^*|}{y^*}$$

Navedena mjera u idealnom slučaju daje vrijednost 0.
Kvadrirana relativna razlika se definira sljedećom formulom:

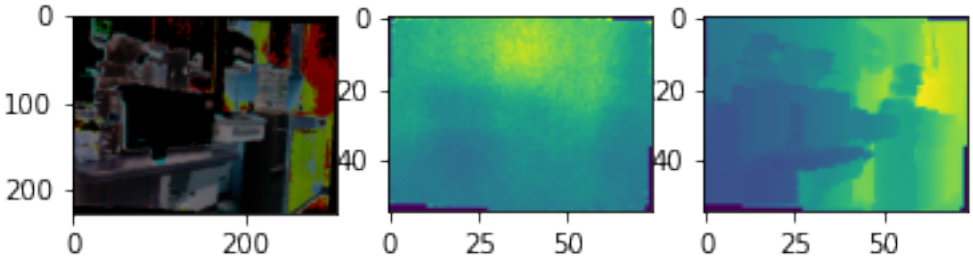
$$\frac{1}{|T|}\sum_{yT}\frac{(y-y^*)^2}{y^*}$$

Formula u idealnom slučaju kao i apsolutna relativna razlika daje 0.
Korijen srednje kvadratne pogreške u linearnom prostoru dana je formulom:

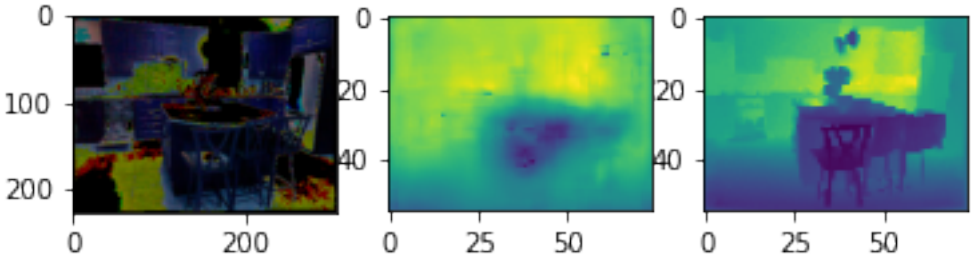
$$\sqrt{\frac{1}{|T|}\sum_{yT}(y-y^*)^2}$$

Korijen srednje kvadratne pogreške u log prostoru dana je formulom:

$$\sqrt{\frac{1}{|T|}\sum_{y\in T}(\log y - \log y^*)^2}$$



Slika 5.1: Primjer izlaza modela za grubu procjenu



Slika 5.2: Primjer izlaza modela

Korijen srednje kvadratne pogreške u log prostoru invarijantnom na skalu dana je formulom:

$$\sqrt{\frac{1}{|T|}\sum_{y\in T}(\log y - \log y^* + \frac{1}{n}\sum_i (\log y_i^* - \log y_i)^2)}$$

Korijen srednje kvadratne pogreške u idealnom slučaju poprima vrijednost 0.

U tablici 5.1 prikazani su rezultati za sva 3 eksperimenta:

Tablica 5.1: Rezultati na skupu za testiranje

	MSE - separate training	Scale invariant MSE - separate training	Scale invariant MSE - joint training
Threshold ($\delta = 1.25$)	0.089	0.456	0.001
Threshold ($\delta = 1.25^2$)	0.233	0.756	0.002
Threshold ($\delta = 1.25^3$	0.458	0.904	0.007
Abs rel diff	0.583	0.628	0.834
Sqr rel diff	0.863	0.581	1.892
RMSE (linear)	1.701	1.088	2.473
RMSE (log)	0.769	0.411	1.735
RMSE (log, scale invariant)	0.360	0.354	0.427

Slika 5.1 prikazuje primjer izlaza komponente za grubu procjenu dubine. Slika 5.2 prikazuje izlaz cijelog modela.

6. Zaključak

Monokularna procjena dubine kompleksan je problem koji zahtjeva veliku količinu podataka i veliku količinu vremena koje je potrebno utrošiti na učenje modela.

Model u kojemu je dio za grubu procjenu i dio za finu procjenu treniran zasebno te je korištena srednja kvadratna pogreška kao funkcija gubitka daje najbolje rezultate. Naučeni model daje grubu procjenu dubine objekata na slici, no model nije dovoljno precizan da bi se koristio u stvarnim sustavima poput robota ili autonomnih vozila gdje se zahtjeva vrlo precizna procjena dubina ili se najčešće koriste sustavi sa stereoskopskim kamerama za procjenu dubine.

Treniranje modela s korištenjem prosječne kvadratne pogreške invarijantne na skaliranje za funkciju gubitka rezultiralo je lošijim rezultatima od treniranja modela sa srednjom kvadratnom pogreškom kao funkcijom gubitka te pokazuje važnost odabira dobre funkcije gubitka.

Treniranje cijelog modela odjednom također je dalo lošije rezultate od modela koji koristi srednju kvadratnu pogrešku kao funkciju gubitka što ukazuje na važnost djela mreže namijenjenog grubu procjenu dubine te njeno pažljivo treniranje.

Niske vrijednosti funkcije gubitka na skupu za treniranje i visoke vrijednosti na testnom skupu ukazuju na to da je model sklon prenaučivanju ako skup podataka ne sadrži dovoljno primjera i nisu odabrane prikladne optimizacijske metode i metode regularizacije.

7. Literatura

[1] David Eigen, Christian Puhrsch, i Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. URL https://arxiv.org/abs/1406.2283.

[2] Pushmeet Kohli Nathan Silberman, Derek Hoiem i Rob Fergus. Indoor segmentation and support inference from rgbd images. U *ECCV*, 2012.