SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Lokalno uparivanje značajki primjenom pažnje

David Kerman
Voditelj: *prof.dr.sc. Siniša Šegvić*

Zagreb, svibanj 2023.

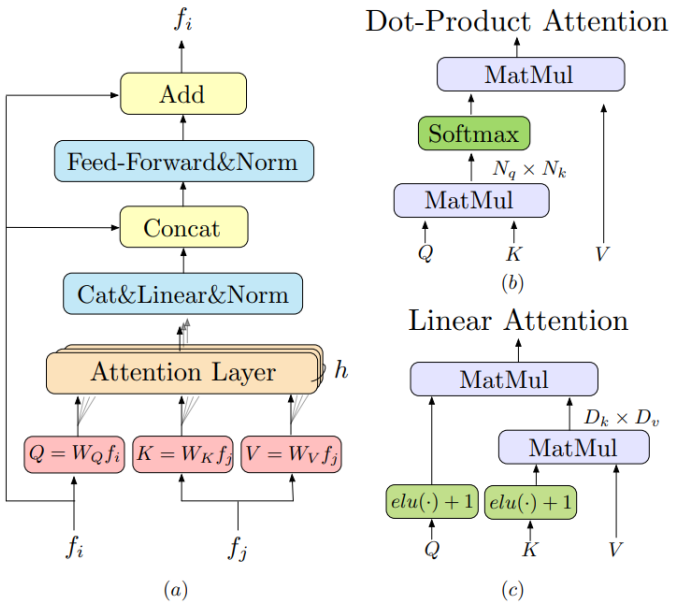
SADRŽAJ

1. Uvod

Računalni vid kao jedna od temeljnih grana umjetne inteligencije omogućava računalima da izluče značajnu informaciju iz slika, videa ili sličnih vizualnih ulaza. Područje stereoskopske rekonstrukcije ima za cilj odrediti trodimenzionalan položaj točaka koje su promatrane u dvije ili više slika, te kao takvo ima vrlo široke primjene uključujući one za određivanje sličnosti slikovnih okana. Novi pristupi u ovom području temelje se na dubokim modelima temeljenim na pažnji, tzv. transformerima. U ovom radu razmatramo metodu LoFTR koja prvo pronalazi gruba podudaranja te ih kasnije profinjuje na izlaznoj rezoluciji. Za razliku od gustih metoda koje koriste volumnu cijenu za traženje korespondencija, LoFTR koristi standardnu (eng.self-attention) i unakrsnu pažnju (eng.cross-attention) kako bi odredio opisnike značajki koji se nalaze u obje slike (stereo slučaj). Globalno receptivno polje, koje sa sobom donose slojevi pažnje, omogućuje ovom pristupu uspostavljanje gustih podudaranja u regijama slabih tekstura.

U uvom radu bit će predstavljeni duboki modeli temeljeni na pažnji i metoda LoFTR koja se oslanja na te modele. Nadalje, bit će provedeni eksperimenti učenja na slikama iz skupa KITTI 2015, čije ćemo rezultate evaluirati računanjem rekonstrukcijske točnosti.

2. Transformeri



Slika 2.1: Slika prikazuje transformer koder i slojeve pažnje u LoFTR-u. Pod **(a)** prikazan je tzv. transformer koder, pri čemu h označava broj glava korištenih u pažnji. Pod **(b)** prikazana je standardan sloj pažnje računske složenosti $O(N^2)$. Pod **(c)** prikazan je sloj linearne pažnje računske složenosti $O(N)$. Izvor: [1]

2.1. Sloj pažnje (engl. attention layer)

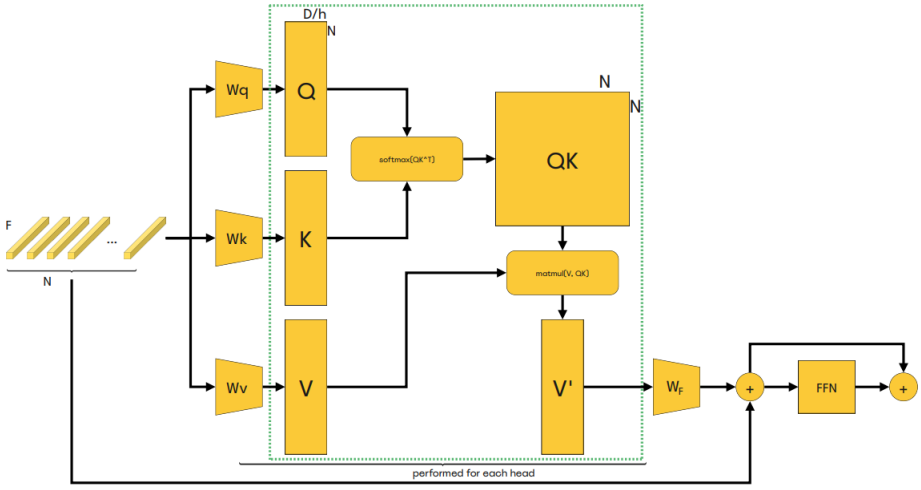
Arhitektura LoFTR sastoji se od sekvencijalno povezanih slojeva (eng. encoder layer). Slika 2.1 pokazuje da se slojevi LoFTR-a sastoje od unakrsne pažnje i potpuno povezanog miješanja značajki. Ulaze unakrsne pažnje nazivamo upitima Q, ključevima K i vrijednostima V. Pažnja usrednjuje informaciju iz vektora V prema sličnosti odgovarajućih elemenata upita i ključeva. Matricu kosinusne sličnosti dobivamo umnoškom upita i ključeva koja nakon toga prolazi kroz softmax. Primjenom softmaksa male vri-

jednosti postat će manje, dok će velike vrijednosti sličnosti postati veće. Računski graf prikazan je na slici 2.1.b.
Formalno, sloj pažnje može se prikazati sljedećom jednadžbom:

$$\text{Attention}(Q, K, V) = softmax(QK^T)V \tag{2.1}$$

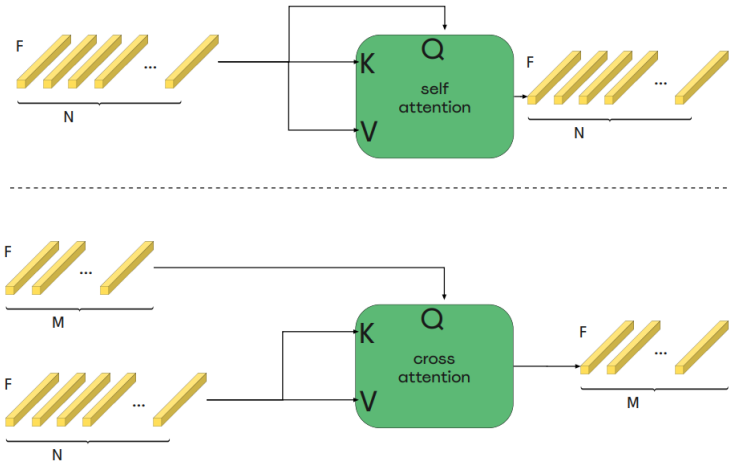
Izlazni vektori su normalizirane mješavine vektora vrijednosti otežana mjerama sličnosti između odgovarajućih upita i ključeva.

U izgradnji transformera koriste se dvije vrste pažnji. Jednoulazna pažnja obrađuje vektore Q, K i V koji su dobiveni iz istog ulaznog tenzora značajki. Unakrsna pažnja obrađuje vektore Q, K i V koji su dobiveni iz različitih ugrađivanja.

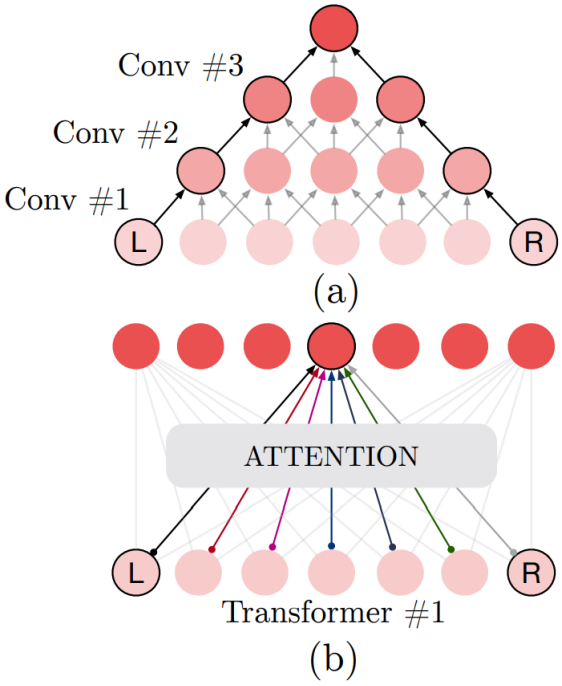


Slika 2.2: Jednoulazna pažnja (eng. self-attention) tvori upite, ključeve i vrijednosti iz istog ulaznog tenzora. Izvor: [4]

Slojevi pažnje imaju globalno receptivno polje. Slika 2.4 uspoređuje receptivna polja konvolucija i pažnje. Pretpostavimo da nam je cilj uspostaviti odnos između piksela L i R kako bismo izlučili njihovu zajedničku reprezentaciju značajki. Kako konvolucije razmatraju samo lokalnu ovisnost, puno konvolucijskih slojeva mora biti nagomilano kako bismo uspjeli ostvariti tu poveznicu. Globalno receptivno polje koje donose transformeri omogućuje da se ta poveznica ostvari jednim slojem pažnje.



Slika 2.3: Usporedba jednoulazne (gore) i unakrsne pažnje (dolje). Izvor: [4]



Slika 2.4: Uspredba receptivnih polja prikaz receptivnog polja konvolucijskih slojeva koje je označeno pod **(a)**, te receptivnog polja transformera prikazano prikazano pod **(b)**. Izvor:[1]

2.1.1. Linearna pažnja

Ukoliko označimo duljinu vektora Q i K s N i dimenzije korištenih značajki s D matrični umnožak između Q i K u transformeru ima kvadratnu složenost s obzirom na broj značajki ulaza. Vrlo je nepraktično izravno primijenjivati takvu operaciju u kontekstu lokalnog uparivanja značajki. Stoga kako bi se ovaj problem ublažio, ko-

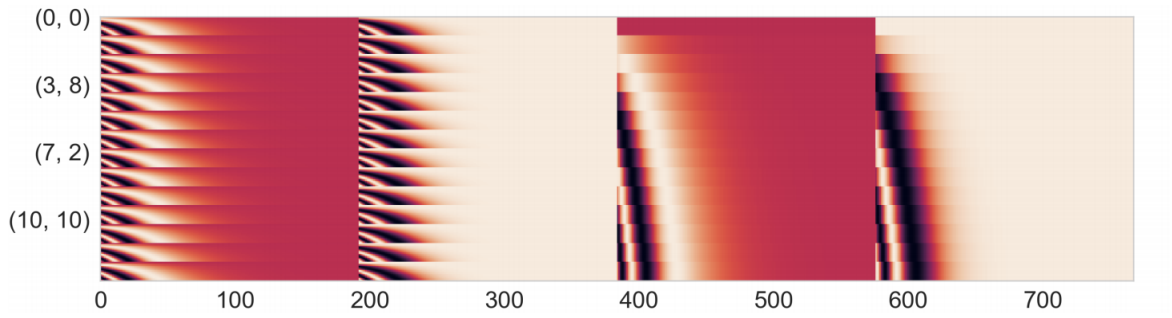
riste se učinkovite varijante slojeva pažnje u transformerima. Tzv. linearna pažnja predlaže da se računska složenost reducira do $O(N)$ tako što se zamijeni eksponencijalna jezgra korištena u originalnom sloju pažnje s alternativnom jezgrenom funkcijom $sim(Q, K) = \Phi(Q) \cdot \Phi(K^T)$, pri čemu $\Phi(\cdot) = elu(\cdot) + 1$.
Ova operacija prikazana je na slici 2.1.c. Koristeći asocijativnost operatora matričnog množenja, množenje između $\Phi(K^T)$ i V može se izvršiti prvo. Kako je $D \ll N$, računska složenost je reducirana na $O(N)$.

2.2. Pozicijsko kodiranje

Kako modeli temeljeni na pažnji nemaju domenski specifične pristranosti potrebno je ugraditi korisne pristranosti na razini ulaza. Jedna od tih pristranosti je pozicijsko kodiranje. Intuitivno, pozicijsko kodiranje pridaje svakom elementu jedinstvenu pozicijsku informaciju u sinusoidnom formatu. U računalnom vidu koristi se 2D formulacija jer se kodiraju pozicije u 2D formatu (x, y). Svaka os zasebno je kodirana prema sljedećim izrazima:

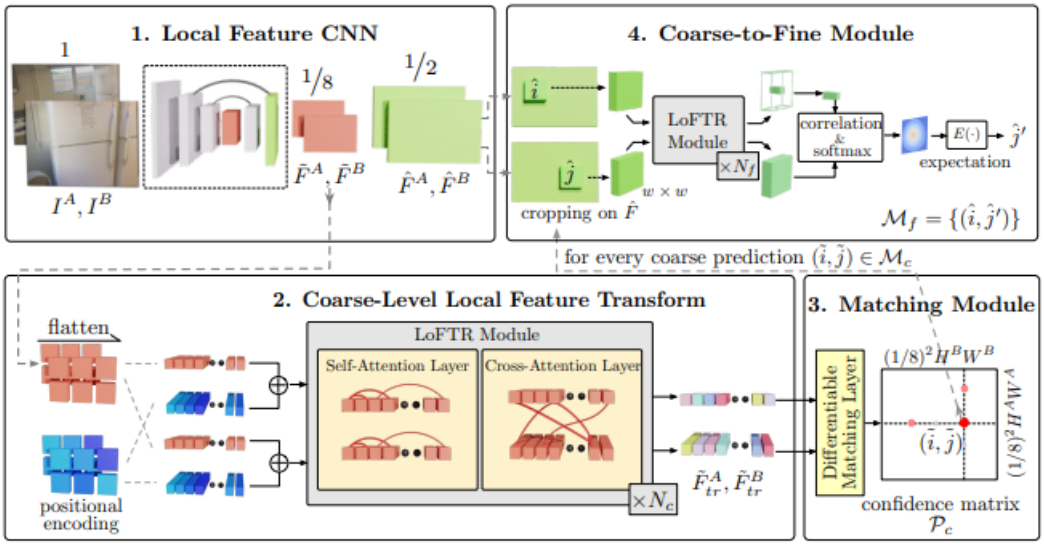
$$
\begin{aligned}
&sin(\omega_k \cdot x) && i = 4k \\
&cos(\omega_k \cdot x) && i = 4k + 1 \\
&sin(\omega_k \cdot y) && i = 4k + 2 \\
&cos(\omega_k \cdot y) && i = 4k + 3
\end{aligned}
$$

U izrazima iznad je $\omega_k = \frac{1}{10000^{\frac{2k}{d}}}$, d broj kanala, a i označava indeks kanala.



Slika 2.5: Ilustracija pozicijskog kodiranja na 2D mreži. Na osi ordinata označene su pozicije koje se kodiraju, dok je na osi apcisa označen broj kanala. Izvor:[4]

3. Metoda LoFTR



Slika 3.1: Metoda LoFTR sastoji se od: 1. Konvolucijski modul (eng. Local Feature CNN) 2. Transformiranje značajki grube razine 3. Modul za podudaranje transformiranih značajki 4. Određivanje podudaranja s podpikselskom razinom Izvor:[1]

Na slici 3.1 prikazan je pregled metode LoFTR. Metoda ima četiri komponente. Prva komponenta je konvolucijski modul (eng. Local Feature CNN) koji izlučuje konvolucijske značajki grube razine $\tilde{F}^A$ i $\tilde{F}^B$, te ih naduzorkuje u značajke $\hat{F}^A$ i $\hat{F}^B$ iz para slike I^A i I^B (Odjeljak 3.1). Nakon toga mape značajki grube razine se izravnavaju, te im se dodaje pozicijsko kodiranje opisano u prijašnjem poglavlju. Takve značajke zatim obrađuje modul za lokalnu transformaciju značajki (LoFTR), koji ima N_c slojeva jednoulazne i unakrsne pažnje (Odjeljak 3.2). Za podudaranje transformiranih značajki metoda LoFTR primijenjuje diferencijalni sloj podudaranja, što dovodi do matrice pouzdanosti P_c. Podudaranja u P_c-u odabiru se prema pragovima pouzdanosti i kriterijima međusobnog najbližeg susjeda, što daje grubu predikciju podudaranja M_c (Odjeljak 3.3). Za svaku odabranu predikciju grube razine $(\tilde{i}, \tilde{j}) \in M_c$, iz mape značajki fine razine izrezuje se lokalni prozor veličine $w \times w$. Gruba podudaranja će se

unutar ovog lokalnog prozora poboljšati na podpikselsku razinu te će tvoriti konačnu predikciju podudaranja M_f.

3.1. Izlučivanje lokalnih značajki

Konvolucijski modul izlučuje guste lokalne značajke iz obje ulaznih slika A i B. Konvolucijske neuronske mreže posjeduju induktivnu pristranost u obliku translacijske ekvivarijantnosti i lokalnosti, koje su dobro prilagođene za izlučivanje lokalnih značajki. [1]

Arhitektura LoFTR postiže učinkovito učenje i zaključivanje obradom na smanjenoj rezoluciji i naknadnim naduzorkovanjem prema arhitekturi FPN [5]. LoFTR ima samo dva koraka naduzorkovanja pa je rezolucija značajki $\tilde{F}^A$ i $\tilde{F}^B$ 8 puta manja od rezolucije ulaznih slika, što kraće označavamo s $\frac{R}{8}$. Te značajke bilinearno se naduzorkuju na rezoluciju $\frac{R}{2}$.

3.2. LoFTR modul

Nakon izlučivanja lokalnih značajki, $\tilde{F}^A$ i $\tilde{F}^B$ prolaze kroz LoFTR modul kako bismo naglasi pozicijsku i kontekstnu ovisnost lokalnih značajki. Intuitivno, LoFTR modul transformira konvolucijske značajke $\tilde{F}^A$ i $\tilde{F}^B$ u metrička ugrađivanja koja su pogodna za podudaranje. Takve dobivene značajke predstavljene su sa $\tilde{F^A_{tr}}$ i $\tilde{F^B_{tr}}$. [1]

3.3. Modul za grubo podudaranje

LoFTR predviđa dvije izvedbe diferencijabilnog podudaranja: sloj optimalnog transporta (OT), te dualni softmax operator. Matrica sličnosti (eng.score matrix) $\mathcal{S}$ između transformiranih značajki računa se prema jednadžbi $S(i,j) = \frac{1}{\tau}\left\langle \tilde{F^A_{tr}}(i), \tilde{F^B_{tr}}(j) \right\rangle$. U okviru optimalnog transporta, $-\mathcal{S}$ se može iskoristiti kao matrica cijene parcijalnog problema dodjele (eng. partial assignment problem). Alternativno, može se upotrijebiti dualni softmax kako bi izračunali vjerojatnost mekog podudaranja zajedničkog najbližeg susjeda. Formalno, prilikom korištenja dualnog softmaksa, vjerojatnost podudaranja $\mathcal{P}_{\text{c}}$ dobiva se kao:

$$\mathcal{P}_{\text{c}}(i,j) = \text{softmax}(S(i,\cdot))_j \cdot \text{softmax}(S(\cdot,j))_i \tag{3.1}$$

Koristeći matricu pouzdanosti $\mathcal{P}_{\text{c}}$, odabiremo podudaranja s pouzdanošću većom od

praga θ_c i nakon toga koristimo kriterij zajedničkog najbližeg susjeda (MNN) koji filtrira moguće stršeće vrijednosti za gruba podudaranja. Predikcije podudaranja grube razine mogu se opisati sljedećim izrazom [1]:

$$\mathcal{M}_{\mathrm{c}} = \left\{(\tilde{i}, \tilde{j}) | \forall (\tilde{i}, \tilde{j}) \in \mathrm{MNN}(\mathcal{P}_{\mathrm{c}}), \mathcal{P}_{\mathrm{c}}(\tilde{i}, \tilde{j}) \geq \theta_c \right\} \tag{3.2}$$

3.4. Modul za fino podudaranje

Ovaj modul naduzorkuje gruba metrička ugrađivanja $\tilde{F_{tr}^A}$ i $\tilde{F_{tr}^B}$ na ulaznu rezoluciju. Za taj proces koristi se pristup baziran na korelaciji. Za svako ulazno ugrađivanje $(\tilde{i}, \tilde{j})$ najprije s lociraju te pozicije $(\hat{i}, \hat{j})$ u mapama značajki fine razine $\hat{F}^A$ i $\hat{F}^B$. Nakon lokalizacije izrezuju se dva seta lokalnih prozora veličine $w \times w$. Nadalje, manji LoFTR modul s N_f slojeva pažnje transformira izrezane značajke, rezultirajući dvjema transformiranim lokalnim mapama značajki $\hat{F_{tr}^A}(i)$ i $\hat{F_{tr}^B}(j)$ centriranih u $\hat{i}$ i $\hat{j}$ respektivno. Nakon toga se korelira središnja lokacija od $\hat{F_{tr}^A}(i)$ sa svim lokacijama u $\hat{F_{tr}^B}(j)$ rezultirajući toplinskom mapom koja reprezentira vjerojatnost podudaranja svakog piksela u susjedstvu od $\hat{j}$ s $\hat{i}$. Računajući očekivanje preko vjerojatnosne distribucije dobivamo konačne pozicije $\hat{j}'$ s podpikselskom točnosti na slici I^B. Skupljanjem svih podudaranja $\left\{(\hat{i}, \hat{j}')\right\}$ dobivamo konačna podudaranja fine razine $\mathcal{M}_{\mathrm{f}}$. [1]

3.5. Gubitak

Konačan gubitak sastoji se od gubitka grube i fine razine: $\mathcal{L} = \mathcal{L}_{\mathrm{c}} + \mathcal{L}_{\mathrm{f}}$. Funkcija gubitka za grubu razinu $\mathcal{L}_{\mathrm{c}}$ je negativna log-izglednost preko matrice $\mathcal{P}_{\mathrm{c}}$ koju smo dobili slojem optimalnog transporta ili dualnog softmaks operatora. Za izračun vrijednosti stvarne (eng. ground-truth) matrice pouzdanosti tijekom učenja koriste se dubinske mape. Stvarna podudaranja grube razine $\mathcal{M}_{\mathrm{c}}^{\mathrm{gt}}$ definirana su kao zajednički najbliži susjedi dvaju setova mreža (eng. grid) na $\frac{1}{8}$ rezolucije. Udaljenost tih dvaju mreža izračunana je preko reprojekcijske udaljenosti njihovih centralnih lokacija. Reprojekcijsku udaljenost definira udaljenost između projicirane točke i korespondentne točke druge mreže.

Funkcija gubitka za finu razinu $\mathcal{L}_{\mathrm{f}}$ zapravo je $\mathcal{L}_2$ gubitak. Za svaku točku $\hat{i}$ mjeri se njezina nesigurnost računanjem varijance korespondirajuće toplinske mape. Cilj je optimizirati profinjene pozicije koje imaju nisku nesigurnost, što rezultira težinskom funkcijom gubitka. [1]

3.6. Implementacijski detalji

Težine koje smo koristili za inicijalizaciju modela dobivene su učenjem modela na MegaDepth skupu. Konvolucijska neuronska mreža za izlučivanje lokalnih značajki koristi ResNet-18 kao okosnicu. Veličina prozora w postavljena je na 5. Transformirane mape značajki $\hat{F}_{tr}^{A}$ i $\hat{F}_{tr}^{B}$ naduzorkovane su i konkatenirane s $\hat{F}^{A}$ i $\hat{F}^{B}$ prije prolaska kroz LoFTR modul fine razine.

4. Motivacija

Referencirajući se na prijašnji rad na ovom problemu, metoda koja koristi kosinusnu sličnost u visokodimenionalnom prostoru [2] kako bi uspostavila korespondencije daje dobre rezultate na mjestima postojanih tekstura dok se loše ponaša na područjima slika bez teksture, te područjima s releksivnim površinama. Na slici 4.1. moguće je primijetiti navedene probleme na staklima automobila, dijelovima šumskog raslinja te područjima slike koja prikazuju cestu. Navedeni problemi naveli su nas na razmatranje

Slika 4.1: Prikaz rekonstrukcijske točnosti modela za ugrađivanje okana u visokodimenzionalan prostor. Lijeve slike prikazuju slike iz lijeve referentne kamere. U sredini se nalaze procijenjene mape dispariteta. Na mapi dispariteta što je svjetliji piksel to je vei disparitet, a tamniji što je manji. Zdesna se nalaze rekonstrukcijske točnosti, pri čemu su crnom bojom označeni pikseli uz nepoznat disparitet, plavom bojom pogrešno rekonstruirani pikseli, te zelenom bojom točno rekonstruirani pikseli. Izvor: [3]

metode LoFTR čiji bi dobri rezultati u područjima slike bez teksture mogli doprinijeti poboljšanju rezultata na problemu stereoskopske rekonstrukcije. Potrebno je napomenuti da metoda LoFTR inicijalno ima za cilj uspostaviti podudaranja nad nerektificiranim slikama. Stoga će biti zanimljivo vidjeti kako se LoFTR ponaša na problemu stereoskopske rekonstrukcije iz rektificiranih slika.

5. Eksperimentalni rezultati

Eksperimente učenja provodili smo na podatkovnom skupu KITTI 2015, te su težine bile inicijalizirane predtreniranim modelom na MegaDepthu. Provedena su dva eksperimenta učenja. U prvom smo stvarnu (eng. ground-truth) matricu pouzdanosti (eng. confidence matrix) računali koristeći identičan postupak iz članka [1]. Stvarna podudaranja grube razine $\mathcal{M}_c^{gt}$ bila su definirana kao zajednički najbliži susjedi dviju mreža (eng. grid) na $\frac{1}{8}$ rezolucije. Udaljenost tih dviju mreža izračunata je preko reprojekcijske udaljenosti njihovih centralnih lokacija (objašnjeno u prijašnjem poglavlju). Za taj izračun bile su nam potrebne transformacijske matrice između lijeve i desne kamere, te matrica intrinsičnih parametara. Intrinsične matrice bile su dobivene QR-dekompozicijom iz projekcijske matirce. U drugom eksperimentu stvarna podudaranja grube razine računali smo iz stvarnih (eng. ground-truth) disparitetnih mapa.

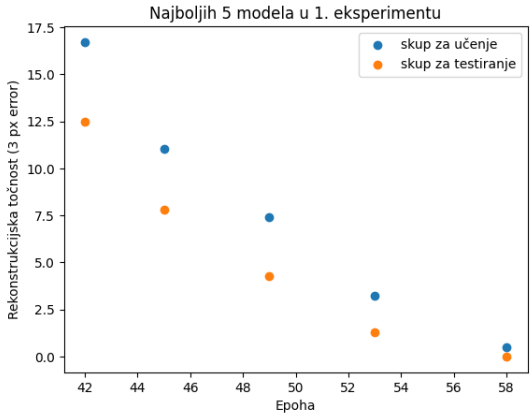
Za evaluacijsku metriku koristili smo rekonstrukcijsku točnost samo na pikselima koje nam je dao model na izlazu, tzv. pogrešku 3 piksela (3px error). Zbog ograničenosti resursa pohrane na platformi na kojoj se učio model, postupak učenja za oba eksperimenta vratio nam je pet najboljih modela prema validacijskoj AUC metrici. U svakoj epohi bi se računao gubitak na skupu za učenje te specifična metrika AUC s pragom (5, 10 i 15) na validacijskom skupu.

U tablici ispod prikazane su rekonstrukcijske točnosti eksperimenata. Model koji je učen na MegaDepthu imao je rekonstrukcijsku točnost na našem skupa za učenje 19.99%, dok je na skupu za testiranje imao 14.15%. Taj model ćemo smatrati referentim za postupak učenja.

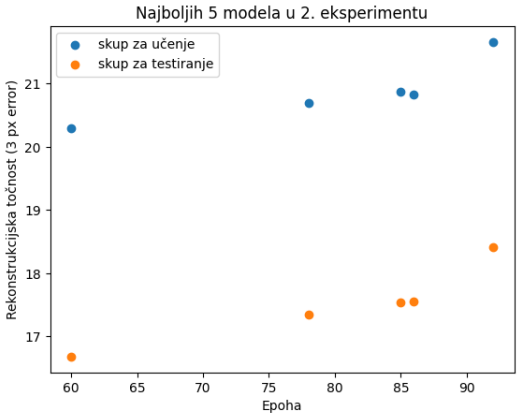
model	**skup za učenje**	**skup za testiranje**
outdoor MegaDepth	19.99%	14.15%
1. eksperiment najbolji	16.70%	12.50%
2. eksperiment najbolji	21.66%	18.42%

Na slici 5.1 prikazane su rekonstrukcijske točnosti za pet najboljih modela iz prvog eksperimenta. Vidljivo je da model degradira, odnosno njegova rekonstrukcijska točnost pada. Jedan od mogućih razloga su krivo izračunati intrinsični parametri, koji su dobiveni QR-dekompozicijom iz projekcijske matrice nakon rektifikacije. Najbolji model prema AUC metrici na validacijskom skupu ima rekonstrukcijsku točnost na skupu za učenje 16.70%, dok na skupu za testiranje ima 12.50%.

Na slici 5.2. prikazane su rekonstrukcijske točnosti za pet najboljih modela iz drugog eksperiment. U drugom eksperimentu vidljiv je pozitivan trend rasta rekonstrukcijske točnosti, što sugerira uspjeh učenja. Potrebno je napomenuti da je model bio treniran 100 epoha.

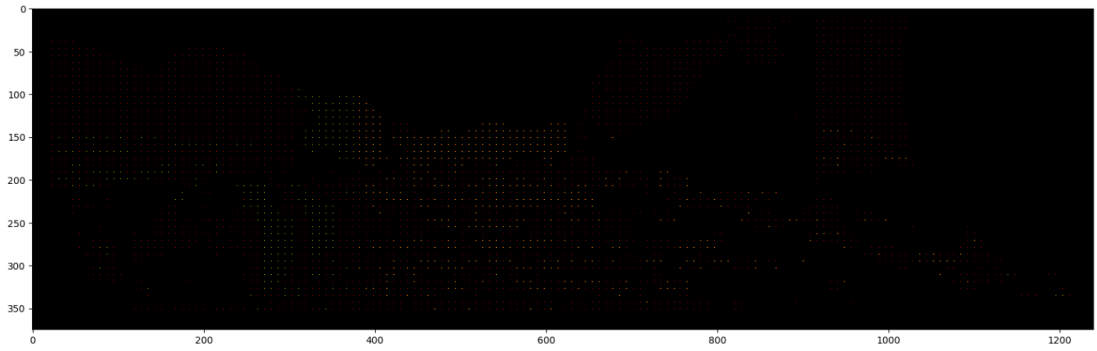


Slika 5.1: Graf koji prikazuje rekonstrukcijsku točnost pet najboljih modela iz prvog eksperimenta. Plavom bojom označene su rekonstrukcijske točnosti na skupu za učenje, a narančastom na skupu za testiranje.

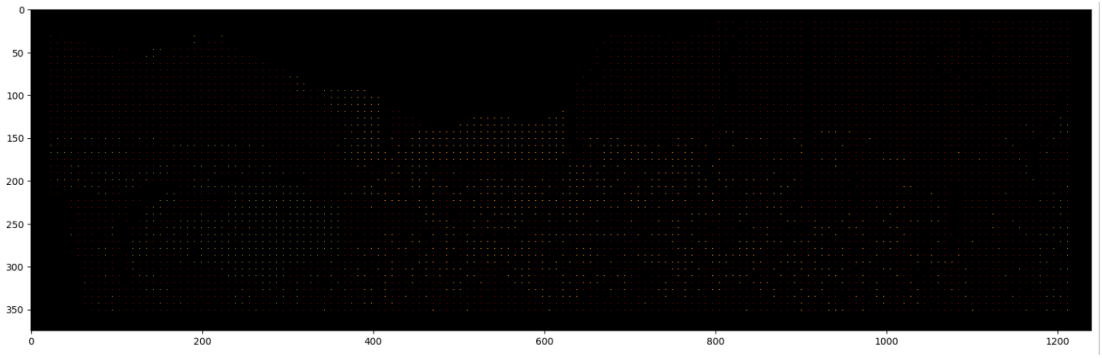


Slika 5.2: Graf koji prikazuje rekonstrukcijsku točnost pet najboljih modela iz drugog eksperimenta. Plavom bojom označene su rekonstrukcijske točnosti na skupu za učenje, a narančastom na skupu za testiranje.

Na slikama 5.3 i 5.4 prikazane su rekonstrukcijske točnosti na 100. okviru iz skupa KITTI modela treniranog na Megadepthu i najboljeg modela iz drugog eksperimenta. Viljivo je da je model učenjem na skupu KITTI poboljšao rekonstrukcijsku točnost i istovremeno povećao broj uparivanja.



Slika 5.3: Prikaz rekonstrukcijske točnosti modela treniranog na MegaDepthu na 100. okviru iz podatkovnog skupa KITTI. Crvenom bojom označen je krivo rekonstruiran piksel, žutom točno rekonstruiran, a crnom piksel za kojeg nema predikcije. Za detalje potrebno je uvećati.



Slika 5.4: Prikaz rekonstrukcijske točnosti najboljeg modela iz drugog eksperimenta na 100. okviru iz podatkovnog skupa KITTI. Crvenom bojom označen je krivo rekonstruiran piksel, žutom točno rekonstruiran, a crnom piksel za kojeg nema predikcije. Za detalje potrebno je uvećati.

Također, pokušali smo isprobati različite pragove za rekonstrukcijsku točnost da vidimo koliko naš model griješi. Vidljiv je linearan trend povećanja rekonstrukcijske točnosti uz porast praga. Prosječna razlika između rekonstrukcijske točnosti mjerene s 3 piksela i one sa 7 piksela je 4 postotna boda.

6. Zaključak i budući rad

Referencirajući se na prijašnji rad i pripadne rezultate na problemu stereoskopske rekonstrukcije, htjeli smo isprobati metodu LoFTR koja je temeljena na slojevima pažnje. U tu svrhu najprije su opisani fundamenti transformera, odnosno modela temeljenih na pažnji. Nadalje, detaljno je opisana metoda LoFTR sa svim pripadnim modulima, gubitkom i implementacijskim detaljima.

Provedena su dva eksperimenta, prvi koji je stvarne (eng. ground truth) matrice pouzdanosti (indirektno gruba uprivanja) izračunao koristeći intrinsične parametre koji su dobiveni QR-dekompozicijom iz projekcijske matrice. U tom eksperimentu evidentno je da je model degradirao tijekom epoha učenja, što sugerira neuspjeh učenja. U drugom eksperimenta matrica pouzdanosti dobivena je direktno koristeći disparitetne mape. Ovakva postavka omogućila je napredak modela kroz epohe učenja. Najbolji model iz drugog eksperimenta je prema izračunu rekonstrukcijske točnosti bio bolji od modela prednaučenog na MegaDepthu.

Rezultati ovog rada sugeriraju da globalno receptivno polje transformera ima potencijal na problemu stereskopske rekonstrukcije. Za poboljšanje trenutačnih rezultata i budući rad predlaže se modifikacija LoFTR-a kojom bismo injektirali informaciju o rektificiranosti slike. Potrebno je izmijeniti modul za gruba uparivanja u kojem bi gruba uparivanja trebalo tražiti samo po epipolarnoj liniji.

7. Sažetak

LoFTR je novi pristup za lokalno podudaranje značajki u slikama koji se oslanja na transformere umjesto tradicionalnih detektora značajki poput SIFT-a ili SURF-a. Slojevi pažnje relativno su novi pristup u izgradnji dubokih modela za računalni vid. U ovom radu predlaže se postupak treniranja LoFTR modela na skupu podataka KITTI stereo 2015. Skup podataka KITTI stereo 2015 sastoji se od 200 stereo slika s pripadajućim mapa dispariteta, koje se koriste za treniranje i testiranje algoritama za određivanje dispariteta u scenama u stvarnom vremenu. Rezultati eksperimenata vrednovani su izračunom rekonstrukcijske točnosti na odvojenom testnom skupu kako bi procijenili generalizacijsku performansu naučenog modela.

8. Literatura

[1] Sun et al. Loftr: Detector-free local feature matching with transformers. 2021. URL https://arxiv.org/pdf/2104.00680.pdf.

[2] Yann LeCun Jure Žbontar. Stereo matching by training a convolutional neural network to compare image patches. 2016.

[3] David Kerman. Korespondencijska ugradivanja za stereoskopsku rekonstrukciju. 2022. URL http://www.zemris.fer.hr/~ssegvic/project/pubs/kerman22bs.pdf.

[4] Marin Oršić. Attention based architectures in computer vision. 2023. URL http://www.zemris.fer.hr/~ssegvic/vision/cv_transformers.pdf.

[5] Ross Girshick Kaiming He Bharath Hariharan Serge Belongie Tsung-Yi Lin, Piotr Dollár. Feature pyramid networks for object detection. 2017.