

SVEUČILIŠTE U ZAGREBU  
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 479

# **Gusto semantičko prognoziranje regresijom značajki**

Luka Družijanić

Zagreb, srpanj 2022.

## ZAVRŠNI ZADATAK br. 479

Pristupnik: **Luka Družijanić (0036522126)**  
Studij: Elektrotehnika i informacijska tehnologija i Računarstvo  
Modul: Računarstvo  
Mentor: prof. dr. sc. Siniša Šegvić

Zadatak: **Gusto semantičko prognoziranje regresijom značajki**

### Opis zadatka:

Predviđanje semantičke budućnosti u videu neriješen je problem računalnog vida s mnogim zanimljivim primjenama. U posljednje vrijeme najbolji rezultati u tom području postižu se dubokim konvolucijskim modelima. Ovaj rad razmatra rješenje tog problema regresijom budućih konvolucijskih značajki iz konvolucijskih značajki viđenih slika. U okviru rada, potrebno je proučiti konvolucijske arhitekture za semantičko predviđanje cestovnih scena u videu. Oblikovati model za predviđanje budućih značajki. Validirati hiperparametre, prikazati i ocijeniti ostvarene rezultate te provesti usporedbu s rezultatima iz literature. Predložiti pravce budućeg razvoja. Radu priložiti izvorni kod razvijenih postupaka uz potrebna objašnjenja i dokumentaciju. Citirati korištenu literaturu i navesti dobivenu pomoć.

Rok za predaju rada: 10. lipnja 2022.

*Zahvaljujem prof. dr. sc. Siniši Šegviću i mag. ing. Josipu Šariću na prenesenom znanju, savjetima i pomoći.*

# SADRŽAJ

<b>1. Uvod</b>	<b>1</b>
<b>2. Konvolucijski slojevi</b>	<b>2</b>
2.1. Konvolucija . . . . .	2
2.2. Dilatirana konvolucija . . . . .	3
2.3. Deformabilna konvolucija . . . . .	3
<b>3. Modeli za semantičko prognoziranje</b>	<b>6</b>
3.1. Treniranje modela . . . . .	7
<b>4. Eksperimenti</b>	<b>9</b>
4.1. Cityscapes skup podataka . . . . .	9
4.2. Implementacija . . . . .	9
4.3. Srednji omjer presjeka i unije . . . . .	10
<b>5. Rezultati</b>	<b>11</b>
5.1. Odabir broja prošlih slika . . . . .	13
5.2. Kvalitativni rezultati . . . . .	13
5.3. Vizualizacija gradijenata po ulaznim slikama . . . . .	17
<b>6. Zaključak</b>	<b>22</b>
<b>Literatura</b>	<b>23</b>

# 1. Uvod

Predviđanje budućnosti bitan je aspekt umjetne inteligencije, posebice u industriji samovozećih automobila. Razni sustavi bi mogli donijeti bolje odluke kada bi, osim razmatranja prošlosti i sadašnjosti, mogli imati i informacije o budućim događajima. Semantičko prognoziranje, kao jedan način predviđanja budućnosti, bi moglo imati ključnu ulogu u razvoju upravljačkih sustava za autonomna vozila.

Tri su osnovne razine na kojima možemo raditi prognoziranje: slike, semantičke predikcije (S2S) i tenzori značajki (F2F). Prognoziranje na razini slika je težak problem [7], no za potrebe inteligentnih sustava nisu ni potrebne same slike. Pokazalo se boljim preskočiti korak prognoziranja slike, te prognozirati semantičke predikcije [4]. Konačno, možemo raditi i prognoziranje na razini tenzora značajki modela semantičke segmentacije [5] [9]. U ovom radu razmatramo prognoziranje tenzora značajki s raznim tipovima konvolucijskih mreža — običnim, dilatiranim i deformabilnim [2].

Nekoliko modela za semantičku segmentaciju ostvaruje više od 80% mIoU na Cityscapes test skupu podataka, no takvi modeli zahtijevaju previše računalne snage za potrebe prognoziranja. Stoga, kao i u [9], koristimo nešto slabiji, ali brži model opisan u [8]. Ovaj model je pogodan za F2F prognoziranje jer ima značajke malih dimenzija. Koristimo malo oslabljenu inačicu modela, bez lateralnih veza, kako bismo mogli imati samo jedan F2F model koji koristi samo najapstraktnije i najsažetije značajke.

## 2. Konvolucijski slojevi

### 2.1. Konvolucija

Konvolucijski slojevi osnovni su alat u dubokim neuronskim mrežama. 2D konvolucija sastoji se od dva koraka: 1) uzorkovanje ulaznih značajki poljem  $R$ ; 2) sumiranje uzorkovanih vrijednosti pomnoženih težinom  $\mathbf{w}$ .

Primjerice, za konvolucijski sloj s jezgrom  $3 \times 3$  (slika 2.1a), polje  $R$  definiramo kao:

$$R = \{(-1, -1), (-1, 0), (-1, 1), (0, -1), (0, 0), (0, 1), (1, -1), (1, 0), (1, 1)\}$$

Tada za svaku lokaciju  $\mathbf{p}_0$  na izlaznoj mapi značajki  $\mathbf{y}$  računamo vrijednost na sljedeći način [2]:

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in R} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n). \quad (2.1)$$

Matricu težina  $\mathbf{w}$  nazivamo *jezgra*, te se njeni parametri uče tijekom treniranja. Jezgru efektivno "posmičemo" po cijeloj širini i dužini ulazne mape značajki, računamo umnožak odgovarajućih elemenata jezgre i ulaza, te konačni zbroj smještamo u izlaznu mapu značajki  $\mathbf{y}$ . Obično u svakom sloju imamo više jezgri koje posmičemo po ulazu, čime na izlazu dobivamo *tenzor značajki*.

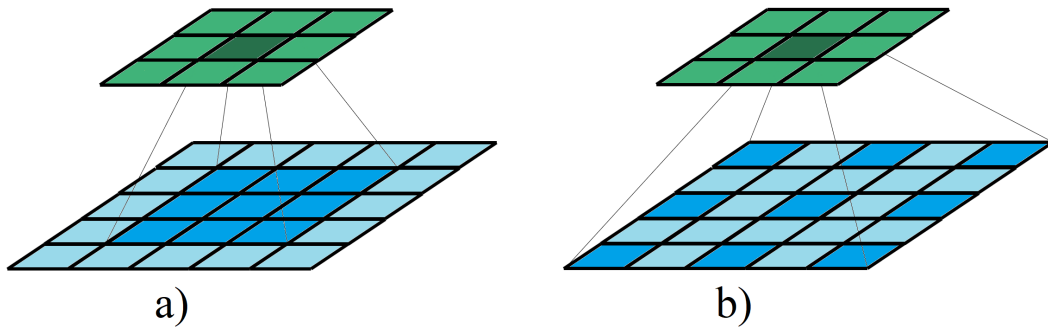
U odnosu na potpuno povezane slojeve, gdje svaka izlazna aktivacija ovisi o svakoj ulaznoj, u konvoluciji svaka aktivacija izlazne mape značajki ovisi samo o nekoliko susjednih aktivacija ulazne mape značajki. Dodatno, budući da se svaka aktivacija jedne izlazne mape značajki računa s istom jezgrom, konačna konvolucija bit će ekvivarijantna s obzirom na pomak. Također, značajno smo smanjili broj potrebnih težina. U većini slučajeva, ovo su nam sve poželjne karakteristike.

## 2.2. Dilatirana konvolucija

Ipak, novi problem nam je puno manje receptivno polje pojedinih izlaza. Obično se receptivno polje proširi dodavanjem slojeva sažimanja, no to nam ne odgovara u prognoziranju jer želimo zadržati rezoluciju mapa značajki.

Umjesto sažimanja, ovdje koristimo dilatirane konvolucije. Sada nam elementi polja  $R$  neće više biti susjedni, već će biti rašireniji. Primjerice, za konvoluciju s jezgrom  $3 \times 3$  i dilatacijom 2 (slika 2.1b):

$$R = \{(-2, -2), (-2, 0), (-2, 2), (0, -2), (0, 0), (0, 2), (2, -2), (2, 0), (2, 2)\}$$



Slika 2.1: Obična (a) i dilatirana (b) konvolucija s  $3 \times 3$  jezgrom

## 2.3. Deformabilna konvolucija

Konačno, poboljšanje konvolucija koje ovdje razmatramo su *deformabilne konvolucije* [2]. Polju  $R$  dodajemo pomake  $\{\Delta \mathbf{p}_n | n = 1, \dots, N\}$ , gdje je  $N = |R|$ . Jednadžba 2.1 tada postaje [2]:

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in R} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n + \Delta \mathbf{p}_n). \quad (2.2)$$

Sada uzorkujemo ulaz nepravilnom mrežom. Pomak  $\Delta \mathbf{p}_n$  će uobičajeno sadržavati decimalne vrijednosti, pa uzorkovanje  $\mathbf{x}(\mathbf{p})$  iz jednadžbe 2.2 implementiramo bilinearnom interpolacijom:

$$\mathbf{x}(\mathbf{p}) = \sum_{\mathbf{q}} G(\mathbf{q}, \mathbf{p}) \cdot \mathbf{x}(\mathbf{q}), \quad (2.3)$$

gdje je  $\mathbf{p}$  neka decimalna lokacija (za jednadžbu 2.2,  $\mathbf{p} = \mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n$ ),  $\mathbf{q}$  pobrojava sve cjelobrojne lokacije u mapi značajki  $\mathbf{x}$ , a  $G$  je dvodimenzijska jezgra bilinearne interpolacije, koju možemo rastaviti na dvije jednodimenzijske jezgre:

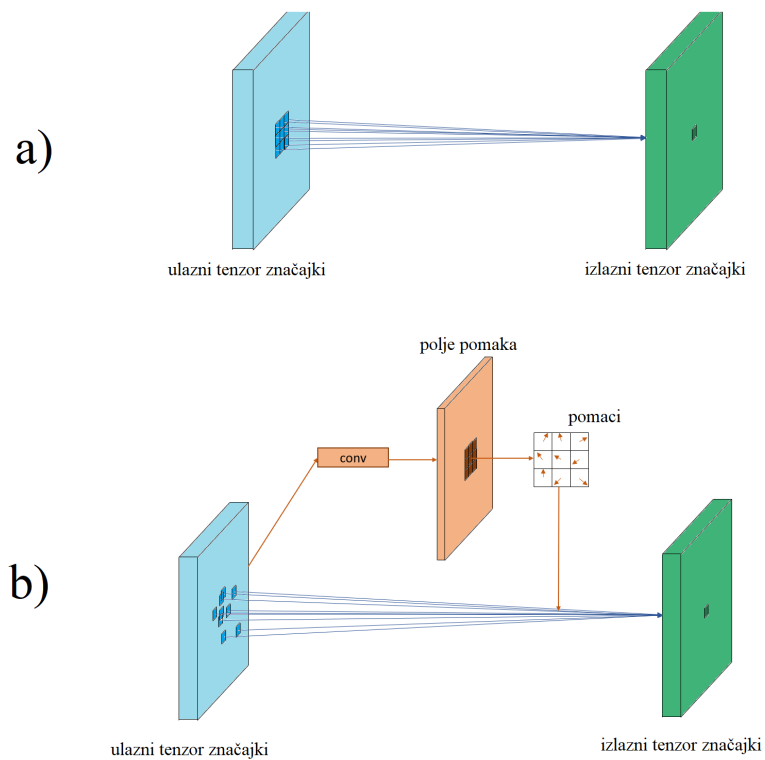
$$G(\mathbf{q}, \mathbf{p}) = g(q_x, p_x) \cdot g(q_y, p_y), \quad (2.4)$$

gdje je  $g(a, b) = \max(0, 1 - |a - b|)$ .  $G(\mathbf{q}, \mathbf{p})$  je rijetka matrica — jednaka je 0 za većinu vrijednosti  $\mathbf{q}$ , pa je jednadžba 2.3 i dalje brza za izračunati.

Pomake  $\Delta\mathbf{p}_n$  ćemo dobiti iz obične konvolucije primijenjene na isti tenzor značajki. Jezgra konvolucije će biti istih dimenzija i dilatacija kao i jezgra odgovarajuće deformabilne konvolucije. Izlazno polje pomaka će biti iste rezolucije kao i ulazna mapa značajki, ali s  $2N$  kanala jer trebamo  $N$  2D pomaka. Tijekom treniranja, parametre konvolucije za pomake inicijalno postavljamo na 0, te se oni uče u isto vrijeme kao i same deformabilne konvolucije.

Deformabilnom konvolucijom dobijemo konvoluciju s prilagodljivim lokacijama uzorkovanja. Umjesto da sami definiramo konvoluciji gdje smije tražiti informacije u ulazu, sada se nadamo da će konvolucija sama naučiti, ovisno u ulazu koji dobije, koje elemente ulaza će razmatrati. Deformabilnu konvoluciju možemo smatrati kao dodatno poopćenje običnih i dilatacijskih konvolucija, koja će sada sama odabrati veličinu i oblik svoje dilatacije ovisno o ulazu, stoga očekujemo bolje rezultate.



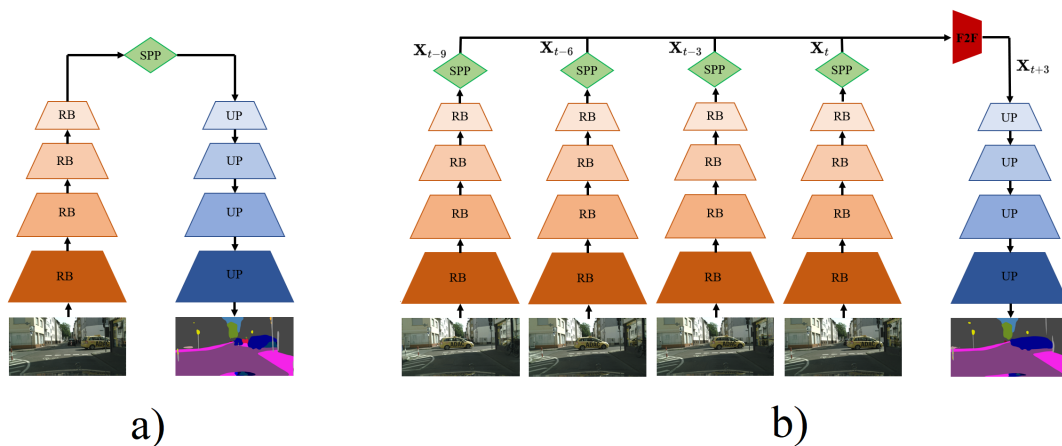


**Slika 2.2:** Obična konvolucija (a) za isti element izlazne mape značajki uvijek uzorkuje iste elemente ulazne mape značajki, dok deformabilna konvolucija (b) prvo iz ulazne mape značajki računa pomake, te onda s tim pomacima uzorkuje ulaznu mapu značajki

### 3. Modeli za semantičko prognoziranje

U ovom radu razmatramo modele za semantičko prognoziranje koji se sastoje od sljedećih dijelova:

1. ekstraktor značajki (ResNet-18),
2. F2F model za prognoziranje značajki s običnim, dilatiranim ili deformabilnim konvolucijama,
3. put naduzorkovanja opisan u [8], ali bez lateralnih veza.



**Slika 3.1:** Arhitektura modela za semantičku segmentaciju (a) i modela za semantičko prognoziranje (b). Oba modela sadrže ResNet-18 ekstraktor značajki (narančasto) i put naduzorkovanja (zeleno, plavo), dok model za prognoziranje sadrži i F2F model za prognoziranje (crveno).

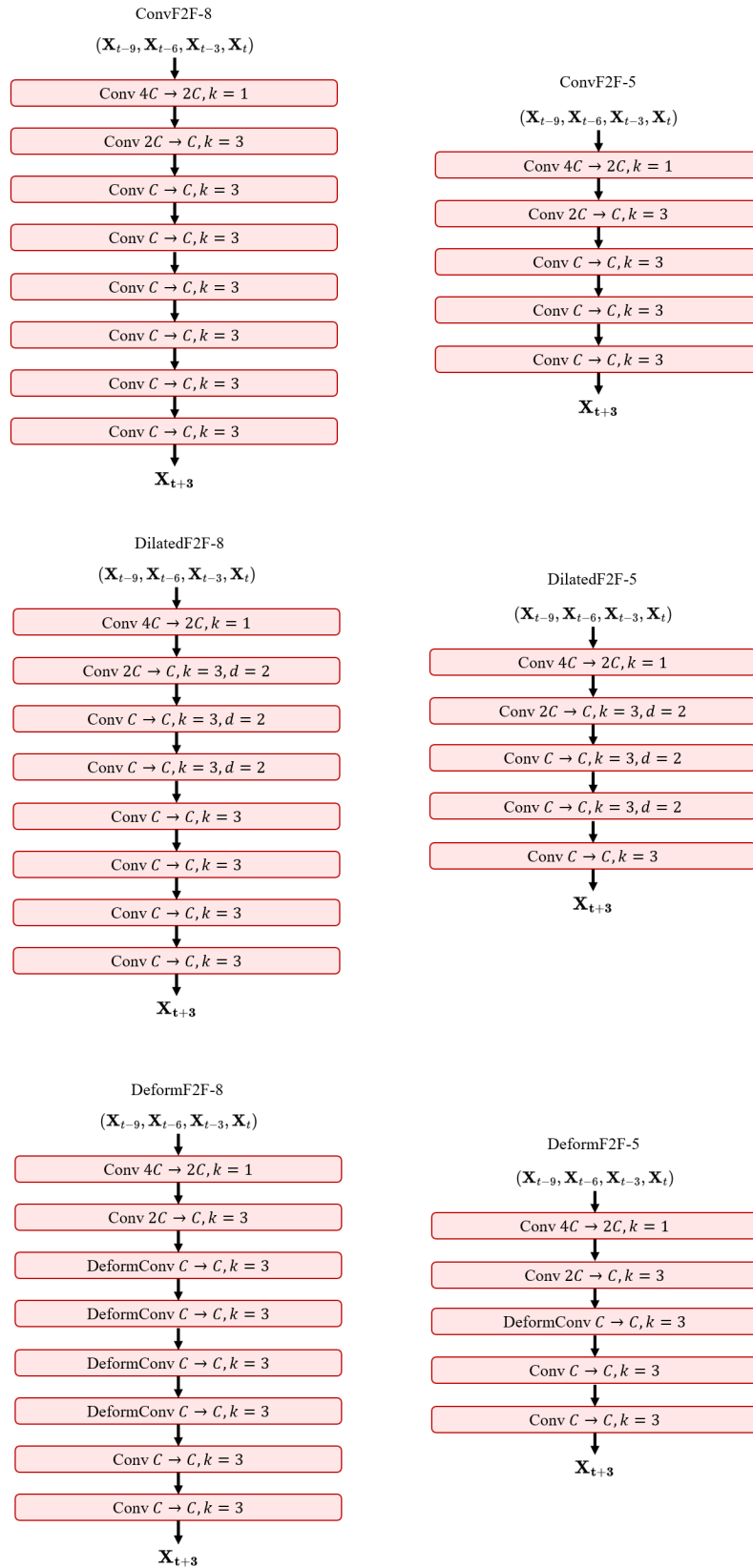
Ekstraktor značajki (narančasti trapezi i zeleni rombi na slici 3.1) prima originalne slike na ulaz, te vraća tenzore značajki (označene s  $X_t$ ,  $X_{t-3}$ ,  $X_{t-6}$ ,  $X_{t-9}$ ). Tenzore značajki iz četiri slike iz prošlosti konkatenujemo u jedan tenzor kojeg predajemo F2F modelu (crveni trapez na slici 3.1), koji prognozira budući tenzor značajki ( $X_{t+3}$  ili  $X_{t+9}$ ). Put naduzorkovanja tada iz budućeg tenzora značajki  $X_{t+3}$  izvodi konačnu semantičku segmentaciju prognozirane scene.

Ekstraktor značajki i put naduzorkovanja će nam za svaki eksperiment biti isti, jedino što mijenjamo je F2F model. Detaljne arhitekture različitih F2F modela koje smo koristili u eksperimentima opisane su na slici 3.2. Svaki crveni pravokutnik predstavlja jednu konvoluciju — obična ili dilatirana konvolucija označena je s *Conv*, a deformabilna [2] s *DeformConv*. Iza toga slijedi broj ulaznih i izlaznih kanala, gdje je sa  $C$  označen broj kanala značajki koje očekuje put naduzorkovanja (u našem slučaju,  $C = 128$ ). Sa  $k$  je označena veličina jezgre konvolucija, a sa  $d$  dilatacija (ukoliko nije napisana, podrazumijeva se da dilatacije nema —  $d = 1$ ). Primjerice, "Conv  $2C \rightarrow C, k = 3, d = 2$ " označava konvoluciju sa  $2C$  ulaznih kanala,  $C$  izlaznih kanala,  $3 \times 3$  jezgrom i dilatacijom 2. Iza svakog sloja osim zadnjeg nalazi se ReLU prijenosna funkcija. Svaki sloj sadrži nadopunu takvu da se zadrži ista rezolucija kroz sve slojeve.

U radu razmatramo tri tipa arhitekture konvolucijskih mreža za F2F — obične konvolucijske mreže (ConvF2F-N na slici 3.2), dilatirane konvolucijske mreže (DilatedF2F-N) i deformabilne konvolucijske mreže (DeformF2F-N), gdje  $N$  označava ukupan broj konvolucijskih slojeva u mreži.

### 3.1. Treniranje modela

Treniranje počinje od javne parametrizacije ekstraktora značajki za ImageNet [3]. Istovremeno treniramo ekstraktor značajki i put naduzorkovanja (model prikazan na slici 3.1a) na označenim slikama [8]. S dobivenim treniranim modelom za svaku sliku u skupu podataka računamo značajke. Primijetimo da sada uopće ne trebamo imati označene slike — našem F2F modelu tijekom učenja predajemo samo skup od 5 tenzora značajki (četiri iz prošlosti te jednu iz budućnosti), te očekujemo da će model naučiti mapirati prošle značajke u jednu buduću.



**Slika 3.2:** Arhitektura raznih F2F modela — iza svakog konvolucijskog sloja osim zadnjeg nalazi se ReLU prijenosna funkcija. Svaki sloj ima nadopunu takvu da se održi jednaka rezolucija značajki.

## 4. Eksperimenti

### 4.1. Cityscapes skup podataka

Eksperimente provodimo na Cityscapes skupu podataka [1]. Cityscapes se sastoji od 2975 slika za treniranje, 500 za validaciju i 1525 za testiranje. Svaka slika sadrži jednu scenu iz gradskog prometa snimljenu tijekom dana u dobrim vremenskim uvjetima. Svaka slika je ručno označena, te je svaki piksel slike svrstan u jedan od 19 razreda. Uz svaku označenu sliku imamo i kratki (1.8 sekundi) video isječak koji se sastoji od još 19 slika prije i 10 slika nakon označene.

Slike su rezolucije  $1024 \times 2048$ , no ovdje eksperimente provodimo na upola manjoj rezoluciju,  $512 \times 1024$ . Svaki tenzor značajki  $\mathbf{X}$  je tada veličine  $128 \times 16 \times 32$ . F2F modelu predajemo skup tenzora značajki scena razmaknutih 0.18s veličine  $512 \times 16 \times 32$ , označeno sa  $(\mathbf{X}_t, \mathbf{X}_{t-3}, \mathbf{X}_{t-6}, \mathbf{X}_{t-9})$ . Model predviđa značajke  $\mathbf{X}_{t+3}$  scene udaljene 0.18s (kratkoročno), ili značajke  $\mathbf{X}_{t+9}$  scene 0.54s (srednjoročno) u budućnosti.



**Slika 4.1:** Primjer jedne slike iz Cityscapes skupa podataka — lijevo je originalna slika, a desno je ista slika sa oznakama razreda

### 4.2. Implementacija

Tenzori značajki Cityscapes slika u pola rezolucije unaprijed su izračunati koristeći već trenirani model iz [8] i spremljeni na disk. F2F model treniramo kroz 160 epoha sa L2 gubitkom i optimizatorom Adam (sa stopom učenja  $5e-4$ ).

### 4.3. Srednji omjer presjeka i unije

Za evaluaciju uspješnosti našeg modela koristimo srednji omjer presjeka i unije (engl. *mean Intersection over Union*, mIoU). Za jedan razred definiramo omjer presjeka i unije (IoU) kao:

$$IoU = \frac{TP}{TP + FP + FN}, \quad (4.1)$$

gdje sa TP označavamo broj ispravno klasificiranih piksela razreda, sa FP broj piksela koji su pogrešno klasificirani kao odabrani razred, te sa FN broj piksela odabranog razreda koji su pogrešno klasificirani kao neki drugi razred.

Dobivena mjera, poznata još i kao Jaccardov indeks, dobar je pokazatelj koliko dobro naš model klasificira odabrani razred. Da bismo izračunali kako naš model klasificira više (ili sve) razreda, računamo omjer presjeka i unije za svaki odabrani razred, te uzimamo njihovu aritmetičku sredinu, čime dobivamo traženi srednji omjer presjeka i unije, kojeg ćemo u ostatku rada označavati kao "mIoU".

Radi efikasnosti izvođenja, sve predikcije se akumuliraju u matricu zabune, gdje oznaci TP sada odgovaraju vrijednosti dijagonale, a FP i FN sada odgovaraju sumama vrijednosti u pojedinim redcima i stupcima.

U Cityscapes skupu podataka imamo i poseban razred s oznakom 255 — *void*. Konceptualno, taj razred označava neklasificirane objekte u sceni koje ne očekujemo da naš model klasificira. Taj razred ignoriramo u izračunu mIoU.

## 5. Rezultati

U tablici 5.1 prikazani su rezultati raznih modela za semantičko prognoziranje na Cityscapes skupu za validaciju. Osim mIoU nad svim razredima, u kontekstu prognoziranja posebno nam je zanimljiv i mIoU za 8 razreda koji predstavljaju kretajuće objekte (engl. *moving objects*, označen kao mIoU-MO).

Prvi dio tablice prikazuje modele označene s "Prorok" i "Kopija zadnje segmentacije". Prorok je model prikazan na slici 3.1a kojem predajemo mapu značajki upravo tražene buduće scene. Budući da su značajke koje predajemo proroku upravo one koje F2F model pokušava predvidjeti, uspješnost proroka je gornja granica uspješnosti F2F modela za semantičko prognoziranje. S druge strane, ako istom modelu predamo zadnju mapu značajki iz prošlosti, dobit ćemo "kopiju zadnje segmentacije", što nam predstavlja donju granicu uspješnosti naših modela. Drugi dio tablice prikazuje rezultate iz literature. Treći dio tablice prikazuje rezultate odabranih konvolucijskih modela. Četvrti dio tablice prikazuje iste te modele, ali trenirane na 3 uzorka iz svakog video isječka.

Loša uspješnost kopije zadnje segmentacije pokazuje koliko je prognoziranje težak zadatak, posebice u slučaju srednjoročnog prognoziranja. U našim eksperimentima, DeformF2F-8 postiže najbolje rezultate, što je i očekivano budući da su deformabilne konvolucije praktički generalizacije dilatiranih i nedilatiranih.

Unatoč korištenju istog proroka, rezultati su lošiji od [9] — razlog tomu je što ovdje eksperimente provodimo na upola manjoj rezoluciji, gdje i prorok postiže slabije rezultate (66.1 mIoU ovdje, 72.5 mIoU u [9]). U [9] je svaka konvolucija DeformF2F modela bila deformabilna, dok se u našim eksperimentima pokazalo boljim (za otprilike 0.3 postotna boda) zadržati obične konvolucije u nekoliko slojeva modela. Primjećujemo i da treniranje na više uzoraka po svakom isječku značajno unaprjeđuje rezultate.

**Tablica 5.1:** Rezultati modela za semantičko prognoziranje na Cityscapes skupu za validaciju

	Kratkoročno		Srednjoročno	
	mIoU	mIoU-MO	mIoU	mIoU-MO
Prorok	66.1	64.2	66.1	64.2
Kopija zadnje segmentacije	49.9	45.6	37.2	28.3
Šarić DeformF2F-5 [9]	63.4	61.5	50.9	46.4
Šarić DeformF2F-8 [9]	64.4	62.2	52.0	48.0
Šarić DeformF2F-8-FT [9]	<b>64.8</b>	<b>62.5</b>	<b>52.4</b>	<b>48.3</b>
Luc Dil10-S2S [4]	59.4	55.3	47.8	40.8
Luc F2F [5]	-	61.2	-	41.2
Luc S2S [5]	-	55.4	-	42.4
ConvF2F-5	56.3	52.5	40.6	32.5
DilatedF2F-5	55.4	51.1	40.8	32.5
DeformF2F-5	57.4	54.3	43.4	36.0
ConvF2F-8	55.8	51.8	40.2	32.1
DilatedF2F-8	55.0	50.7	38.6	29.5
DeformF2F-8	57.8	54.6	<b>44.8</b>	<b>39.7</b>
ConvF2F-5 (3 uzorka)	57.7	54.1	-	-
DilatedF2F-5 (3 uzorka)	56.8	53.1	-	-
DeformF2F-5 (3 uzorka)	58.5	55.9	-	-
ConvF2F-8 (3 uzorka)	57.7	54.0	-	-
DilatedF2F-8 (3 uzorka)	57.6	54.5	-	-
DeformF2F-8 (3 uzorka)	<b>59.7</b>	<b>57.4</b>	-	-

**Tablica 5.2:** Rezultati modela za semantičko prognoziranje na Cityscapes skupu za učenje

	Kratkoročno		Srednjoročno	
	mIoU	mIoU-MO	mIoU	mIoU-MO
Prorok	78.6	79.5	78.6	79.5
Kopija zadnje segmentacije	58.3	55.2	44.0	37.4
DeformF2F-8	73.9	73.7	58.4	56.6
DeformF2F-8 (3 uzorka)	73.8	73.8	-	-

Tablica 5.2 prikazuje uspješnost modela DeformF2F-8 na skupu za učenje. Primitimo da je razlika između uspješnosti proroka i DeformF2F-8 modela na skupu za



učenje vrlo slična kao i njihova razlika na skupu za validaciju, što nam pokazuje da se model nije prenaučio na skupu za učenje te da dobro generalizira.

## 5.1. Odabir broja prošlih slika

Tablice 5.3 i 5.4 prikazuju kako broj ulaznih slika iz prošlosti utječe na uspješnost modela DeformF2F-8. U oba slučaja, daleko najlošiju uspješnost (za otprilike 6 do 8 postotnih bodova) dobijemo sa samo jednom slikom iz prošlosti. Model tada vrlo teško može zaključiti kako se neki objekt kreće — možda može zaključiti smjer kretanja iz orijentacije objekta, ali ne može nikako saznati i brzinu. S druge strane, razlika između 4, 3 ili 2 dane slike iz prošlosti nije tako velika (otprilike 0.1-1.0 postotna boda), ali uglavnom više slika daje bolje rezultate.

**Tablica 5.3:** Kratkoročni rezultati modela DeformF2F-8 u ovisnosti o broju prošlih slika

	broj slika	Kratkoročno	
		mIoU	mIoU-MO
DeformF2F-8 (3 uzorka)	4	59.7	57.4
	3	59.6	57.5
	2	59.3	57.4
	1	53.0	50.6

**Tablica 5.4:** srednjoročni rezultati modela DeformF2F-8 u ovisnosti o broju prošlih slika

	broj slika	Srednjoročno	
		mIoU	mIoU-MO
DeformF2F-8	4	44.8	39.7
	3	44.9	39.4
	2	44.9	38.7
	1	38.2	30.5

## 5.2. Kvalitativni rezultati

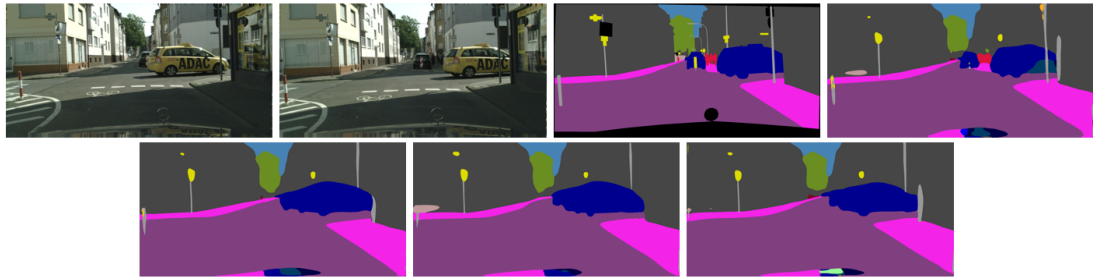
Slike 5.1 i 5.2 prikazuju nekoliko primjera kratkoročnog i srednjoročnog semantičkog prognoziranja. Redom, pojedinačne slike u svakom skupu slika prikazuju:

1. zadnju sliku iz prošlosti,
2. sliku iz budućnosti čiju semantičku segmentaciju model treba prognozirati,
3. istinite, tražene oznake,
4. semantičku segmentaciju proroka,
5. semantičku segmentaciju modela ConvF2F-8
6. semantičku segmentaciju modela DilatedF2F-8
7. semantičku segmentaciju modela DeformF2F-8

Sve slike prikazuju scenu s objektom koji se kreće. U kratkoročnom prognoziranju, uočavamo da sva tri modela uspijevaju pogoditi lokaciju objekta, s manjim razlikama.

U skupu slika 5.1a vidimo žuti taksi koji se kreće te parkirani auto u pozadini. Žuti taksi u slikama iz prošlosti djelomično prekriva auto u pozadini, te tijekom prognoziranja sva tri modela rade istu pogrešku — oba vozila smatraju istim objektom te neispravno popunjavaju prazninu među njima.

U skupu slika 5.1b vidimo bijeli kamion koji se kreće kroz scenu. Modeli ispravno predviđaju lokaciju kamiona. Međutim, modeli za semantičko prognoziranje nasljeđuju probleme svog proroka — i prorok i modeli označuju prednji kraj kamiona kao "auto" (svijetlija nijasna plave).



(a)



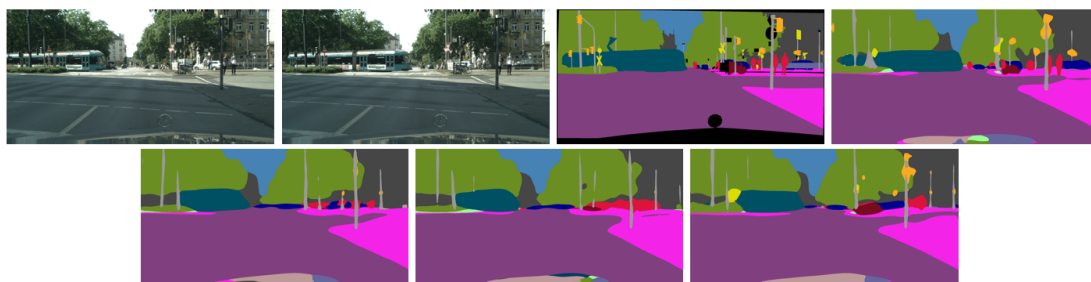
(b)

**Slika 5.1:** Primjer kratkoročnog prognoziranja. Redom, slike prikazuju 1) zadnju sliku iz prošlosti 2) sliku iz budućnosti koju treba prognozirati 3) istinite, tražene oznake 4) segmentaciju proroka 5) prognoziranje ConvF2F 6) prognoziranje DilatedF2F 7) prognoziranje DeformF2F

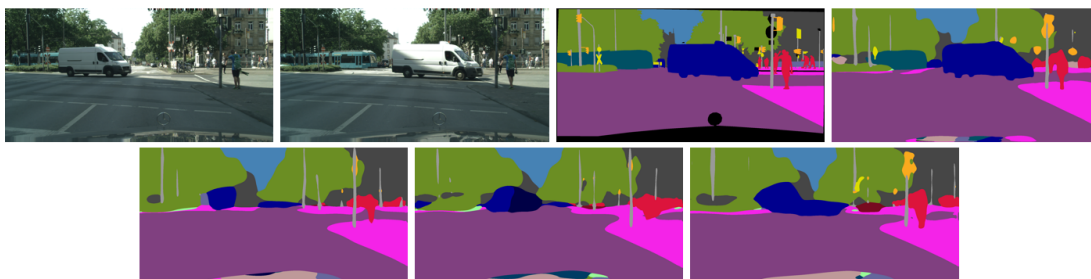
Modeli značajno lošije uspijevaju na srednjoročnom prognoziranju. Na skupu slika 5.2a vidimo tramvaj kojem modeli ne uspijevaju točno predvidjeti lokaciju. Ovdje također vidimo značajnije razlike između pojedinih modela — primjerice, DeformF2F-8 zadržava veću razinu detalja kod manjih objekata u pozadini.

Na skupu slika 5.2b teško predviđaju lokaciju i oblik auta. ConvF2F-8 skoro pa u potpunosti gubi auto u prognoziranju, dok DeformF2F-8 ipak uspijeva donekle zadržati oblik auta.

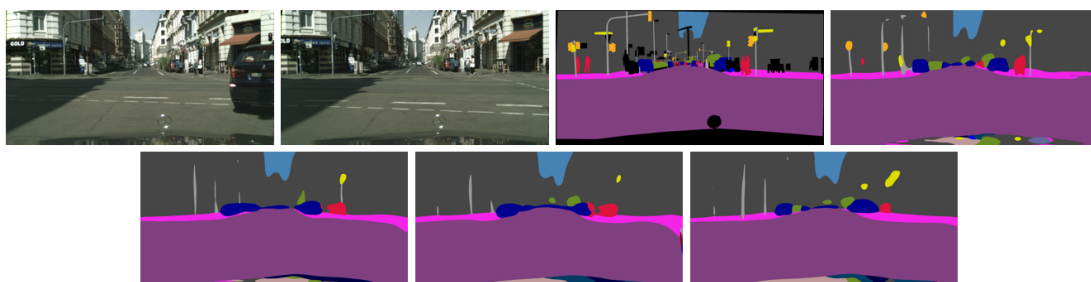
Na slici iz prošlosti u skupu 5.2c vidimo auto koji upravo skreće desno, te ga ne vidimo na budućoj slici. Sva tri modela uspješno sada predviđaju izgled scene iza auta, što nam pokazuje da su F2F modeli dobri u predviđanju neviđenih dijelova.



(a)



(b)



(c)

**Slika 5.2:** Primjer srednjoročnog prognoziranja. Redom, slike prikazuju 1) zadnju sliku iz prošlosti 2) sliku iz budućnosti koju treba prognozirati 3) istinite, tražene oznake 4) segmentaciju proroka 5) prognoziranje ConvF2F 6) prognoziranje DilatedF2F 7) prognoziranje DeformF2F

### 5.3. Vizualizacija gradijenata po ulaznim slikama

Očekujemo da neće svi pikseli slike jednako utjecati na određeni piksel konačne segmentacije modela. U ovom poglavlju pokazat ćemo koji pikseli slika iz prošlosti su najviše utjecali na odabrani piksel konačne segmentacije modela DeformF2F-8.

Odabiremo jedan piksel na slici,  $(i, j)$ , označen sa zelenom bojom na slikama 5.3 i 5.4. Za taj piksel nam naš model za semantičko prognoziranje vraća logite, od kojih računamo  $y_{i,j} = \log(\max(\text{softmax}(\text{logiti})))$

Za svaki piksel slika iz prošlosti,  $x_{n,c,i,j}$ , računamo parcijalnu derivaciju  $\frac{\partial y_{i,j}}{\partial x_{n,c,i,j}}$ . Ovdje  $n$  označava broj slike iz prošlosti,  $c$  broj RGB kanala, a  $i$  i  $j$  lokaciju piksela.

Za sve  $n$ ,  $i$  i  $j$  sada računamo

$$\frac{\partial y_{i,j}}{\partial x_{n,c,i,j}} = \left| \sum_{c=0}^2 x_{n,c,i,j} \right|, \quad (5.1)$$

čime smo konačno dobili apsolutnu vrijednost gradijenta za svaki od piksela. Nadalje, računamo prag —  $k$ -ti najveći gradijent u zadnjoj slici iz prošlosti (u našem slučaju,  $k = 3000$ , čime dobivamo 0.15% označenih piksela). Na svakoj od slika iz prošlosti sada možemo označiti sve piksele čija je apsolutna vrijednost gradijenta veća od praga.

Umjesto jednadžbe 5.1, imalo bi smisla uzeti i normu vektora gradijenta kao konačnu vrijednost gradijenta piksela. Kada bismo izračunali standardnu devijaciju koordinata označenih piksela po  $x$  i  $y$  osi na ulaznim slikama, dobili bismo procjenu efektivnog receptivnog polja (engl. *Effective receptive field*, ERF) [6] modela za odabrani piksel. Dodatno, možemo procijeniti ERF modela kao aritmetičku sredinu ERF-ova modela na središnjim pikselima svih slika u skupu podataka.

Slike 5.3 i 5.4 prikazuju vizualizaciju gradijenata po ulaznim slikama na kratkoročnim i srednjoročnim predviđanjima. Odabrani piksel označen je zelenom bojom, a pikseli s najvećim gradijentom crvenom. Redom, pojedinačne slike u svakom skupu slika prikazuju:

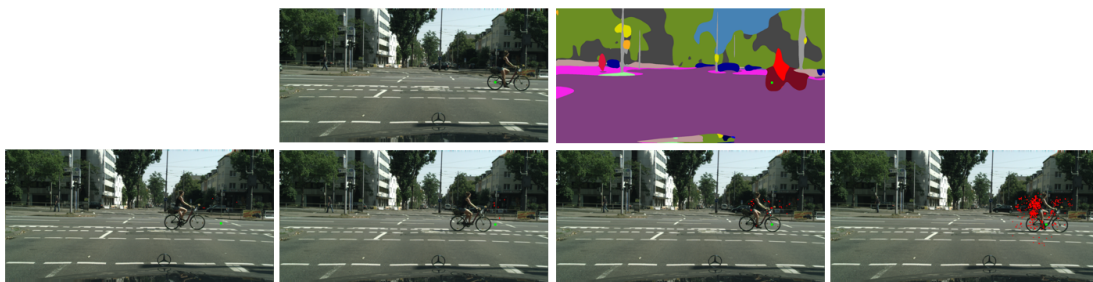
1. sliku iz budućnosti čiju semantičku segmentaciju model treba prognozirati,
2. semantičku segmentaciju modela DeformF2F-8,
3. četiri slike iz prošlosti, redom od najstarije do najnovije

Primjećujemo da je većina označenih piksela u najkasnijoj slici iz prošlosti, te da često odgovaraju upravo granicama različitih objekata.

Primjerice, na slikama 5.4b vidimo auto koji skreće, nestaje iz scene te otkriva pozadinu. Iako smo odabrali piksel koji na prošlim slikama odgovara autu, primjećujemo da je model odlučio promatrati upravo dijelove pozadine koje vidi uz auto, te vrlo malen broj crvenih piksela se nalazi na autu. S druge strane, na slici 5.4c vidimo da će model ipak, kada shvati da je riječ o pozadini ili stacionarnom objektu, pretražiti baš okolinu tog piksela.



(a)



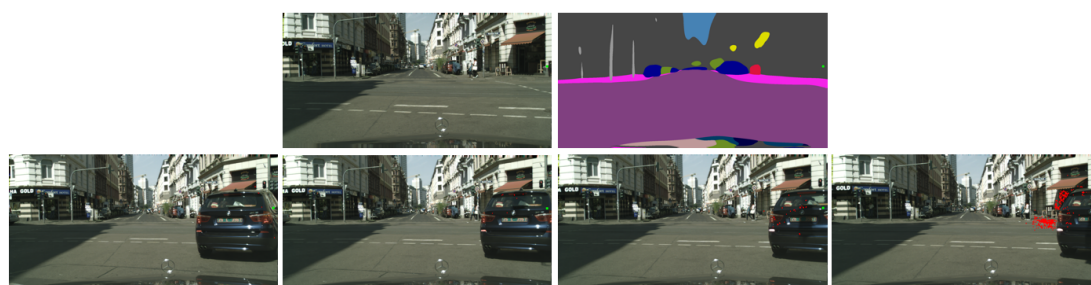
(b)

**Slika 5.3:** Vizualizacija gradijenata odabranog piksela (označen zelenom bojom) u kratkoročnom semantičkom prognoziranju. Slike redom prikazuju zadnju sliku iz budućnosti, semantičku segmentaciju modela DeformF2F-8, te četiri slike iz prošlosti sa označenim najznačajnijim pikselima (označeni crvenom bojom)

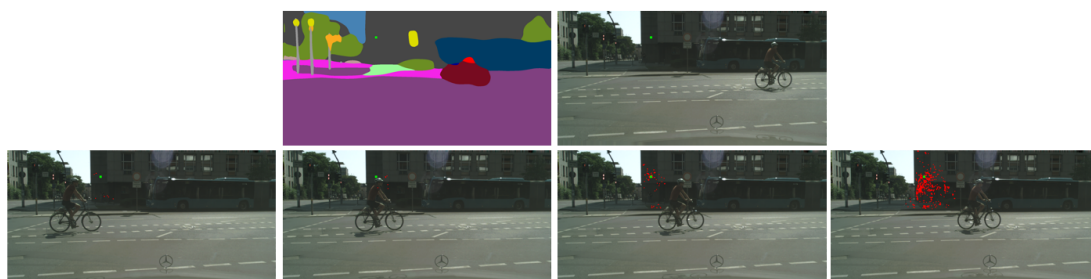
Na slikama 5.5 i 5.6 vidimo vizualizaciju gradijenata po ulaznim slikama za deformabilne (gore) i dilatirane (dolje) konvolucije. Vidimo jasne prednosti deformabilnih konvolucija nad dilatiranim — deformabilne konvolucije su fleksibilnije pri uzorkovanju, prilagođavaju se slici, dok dilatirane često gledaju samo okolinu odabranog piksela. Također, dilatirane konvolucije pridjeljuju veću važnost starijim slikama iz prošlosti, što ih nekad dovodi do krivih zaključaka.



(a)



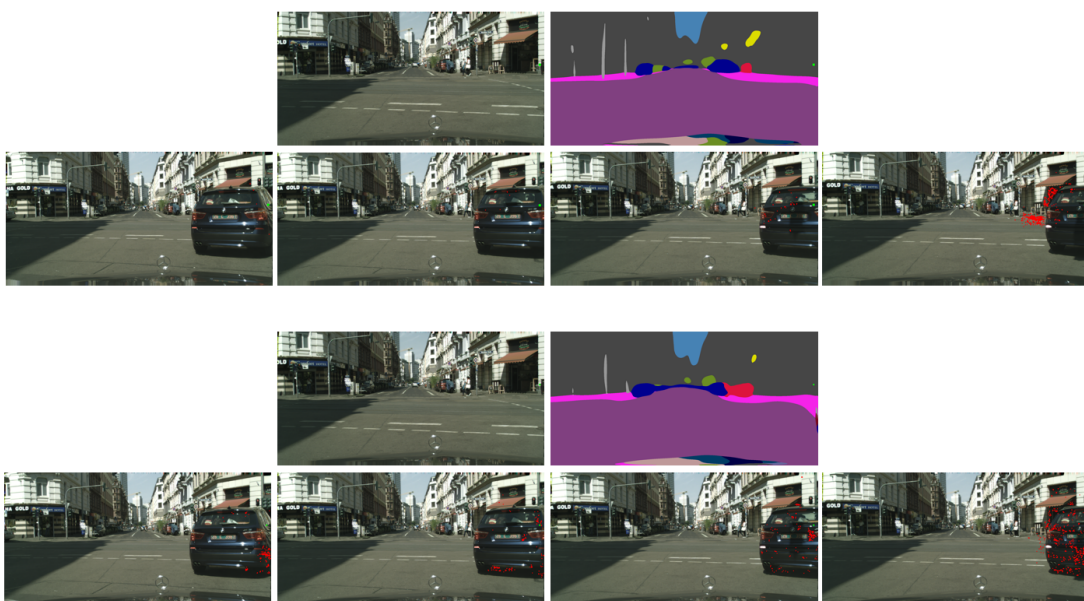
(b)



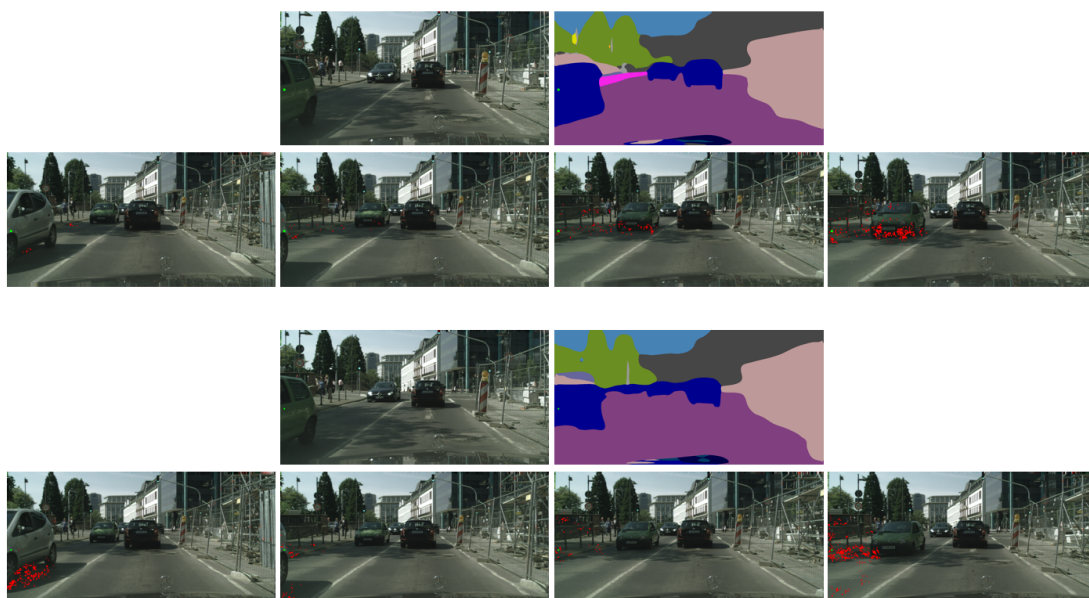
(c)

**Slika 5.4:** Vizualizacija gradijenata odabranog piksela (označen zelenom bojom) u srednjoročnom semantičkom prognoziranju. Slike redom prikazuju zadnju sliku iz budućnosti, semantičku segmentaciju modela DeformF2F-8, te četiri slike iz prošlosti sa označenim najznačajnijim pikselima (označeni crvenom bojom)





**Slika 5.5:** Usporedba gradijenata po ulaznim slikama deformabilnih (gore) i dilatiranih (dolje) konvolucija. Deformabilna konvolucija se prilagođava slici, prepoznaje da treba gledati pozadinu, dok dilatirana konvolucija pogrešno obraća pažnju na auto koji skreće.



**Slika 5.6:** Usporedba gradijenata po ulaznim slikama deformabilnih (gore) i dilatiranih (dolje) konvolucija. Deformabilna konvolucija prepoznaje kretanje auta, te gleda na ispravno mjesto na slici da vidi gdje se auto nalazi u prošlosti. Dilatirana konvolucija ne dokuči do stare lokacije auta, te gleda samo okolinu odabranog piksela.

## 6. Zaključak

U ovom radu razmotrili smo implementacije modela semantičkog prognoziranja značajki (F2F). Kao osnovicu smo koristili model za semantičku segmentaciju bez lateralnih veza, što nam omogućuje da koristimo samo jedan F2F modul na značajkama najmanje rezolucije. Usporedili smo razne tipove konvolucijskih arhitektura za F2F prognoziranje. Eksperimente smo proveli na Cityscapes skupu podataka, te smo ustanovili da deformabilne konvolucije daju značajno bolje rezultate od običnih ili dilatiranih konvolucija. Semantičko prognoziranje težak je zadatak, te ima mjesta za napredak. Uz deformabilne konvolucije, poboljšanje možemo tražiti u optičkom toku, slojevima pažnje ili dodatnom treniranju naduzorkovanja. Preostaje i istražiti ponašanje F2F modela i na zadatku segmentacije instanci.

# LITERATURA

- [1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, i Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. U *Proceedings of the IEEE conference on computer vision and pattern recognition*, stranice 3213–3223, 2016.
- [2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, i Yichen Wei. Deformable convolutional networks. U *Proceedings of the IEEE international conference on computer vision*, stranice 764–773, 2017.
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, i Li Fei-Fei. Imagenet: A large-scale hierarchical image database. U *2009 IEEE Conference on Computer Vision and Pattern Recognition*, stranice 248–255, 2009.
- [4] Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, i Yann LeCun. Predicting deeper into the future of semantic segmentation. U *Proceedings of the IEEE international conference on computer vision*, stranice 648–657, 2017.
- [5] Pauline Luc, Camille Couprie, Yann Lecun, i Jakob Verbeek. Predicting future instance segmentation by forecasting convolutional features. U *Proceedings of the european conference on computer vision (ECCV)*, stranice 584–599, 2018.
- [6] Wenjie Luo, Yujia Li, Raquel Urtasun, i Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.
- [7] Michael Mathieu, Camille Couprie, i Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.
- [8] Marin Orsic, Ivan Kreso, Petra Bevandic, i Sinisa Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. U *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, stranice 12607–12616, 2019.

- [9] Josip Šarić, Marin Oršić, Tonći Antunović, Sacha Vražić, i Siniša Šegvić. Single level feature-to-feature forecasting with deformable convolutions. U *German Conference on Pattern Recognition*, stranice 189–202. Springer, 2019.

## **Gusto semantičko prognoziranje regresijom značajki**

### **Sažetak**

Predviđanje budućnosti bitan je aspekt umjetne inteligencije. Razmatramo semantičko prognoziranje značajki (F2F), bazirano na modelu za semantičku segmentaciju bez lateralnih veza u naduzorkovanju, što omogućuje F2F modelu da radi samo na značajkama najmanje rezolucije. Nadalje, uspoređujemo nekoliko konvolucijskih arhitektura za F2F prognoziranje. Eksperimenti pokazuju da deformabilne konvolucije postižu bolje rezultate od običnih i dilatiranih konvolucija.

**Ključne riječi:** semantičko prognoziranje, semantička segmentacija, značajke, F2F, deformabilna konvolucija, konvolucija

## **Dense semantic forecasting by feature regression**

### **Abstract**

Predicting the future is an important aspect of artificial intelligence. We explore feature-to-feature (F2F) semantic forecasting, based on a semantic segmentation model without lateral connections in the upsampling path, which allows the F2F model to work with features of lowest resolution. Furthermore, we compare several convolutional architecture for F2F forecasting. Experiments show that deformable convolutions achieve better results than regular and dilated convolutions.

**Keywords:** semantic forecasting, semantic segmentation, features, F2F, deformable convolution, convolution