Združeno prognoziranje značajki i njihova pomaka za predviđanje semantičke budućnosti u videu

Šarić, Josip

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Josip Šarić

ZDRUŽENO PROGNOZIRANJE ZNAČAJKI I NJIHOVA POMAKA ZA PREDVIĐANJE SEMANTIČKE BUDUĆNOSTI U VIDEU

DOKTORSKI RAD

Zagreb, 2022.

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Josip Šarić

ZDRUŽENO PROGNOZIRANJE ZNAČAJKI I NJIHOVA POMAKA ZA PREDVIĐANJE SEMANTIČKE BUDUĆNOSTI U VIDEU

DOKTORSKI RAD

Mentor: prof. dr. sc. Siniša Šegvić

Zagreb, 2022.

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Josip Šarić

JOINT FORECASTING OF FEATURES AND FEATURE MOTION FOR SEMANTIC FUTURE PREDICTION IN VIDEO

DOCTORAL THESIS

Supervisor: Professor Siniša Šegvić, PhD

Zagreb, 2022

Doktorski rad izrađen je na Sveučilištu u Zagrebu Fakultetu elektrotehnike i računarstva, na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave.

Mentor: prof. dr. sc. Siniša Šegvić

Doktorski rad ima: 86 stranica

Doktorski rad br.: _____

O mentoru

Siniša Šegvić rođen je 1971. u Splitu. Osnovnu školu i gimnaziju završio je u Zadru osim osmog razreda osnovne škole kojeg je pohađao u Milanu. Diplomirao je elektrotehniku na zagrebačkom ETF-u (1996), a tamo je i magistrirao (2000.) i doktorirao (2004.) te se zaposlio kao docent od 2006. godine.

Bio je postdoktorski istraživač na institutu IRISA u Rennesu (2006.-2007.) te na TU Graz (2007.-2008.). Vodio je tri istraživačka projekta Hrvatske zaklade za znanost (MultiCLOD, MASTIF, ADEPT) te više industrijskih istraživačkih projekata koje su financirale tvrtke Rimac automobili, RoMB, MicroBlink te Promet i prostor. Sudjelovao je u istraživačkom centru izvrsnosti DataCross, na nekoliko ERDF projekata (SafeTram, MAS, A-UNIT) kao i na jednom projektu iz programa FP7 (ACROSS).

Njegovi istraživački i profesionalni interesi uključuju računalni vid, strojno učenje, razumijevanje scena, i gustu predikciju dubokim konvolucijskim modelima. Objavio je 5 radova na vrhunskim konferencijama računalnog vida i umjetne inteligencije (3xCVPR, ECCV, NeurIPS) te 10 radova u časopisima koje indeksira SCI. Recenzent je na vrhunskim konferencijama računalnog vida i umjetne inteligencije (CVPR, ECCV, ICCV, AAAI) kao i u znanstvenim časopisima u područjima računalnog vida, inteligentnih transportnih sustava i robotike. Sudjelovao je u industrijskom razvoju kao tehnički konzultant. Mentorira više doktoranada koje financiraju europski projekti, nacionalni projekti i privatne tvrtke. Njegova istraživačka grupa postigla je zapažene rezultate na nekoliko natjecanja u računalnom vidu (WildDash, Robust vision challenge, Cityscapes, Fishyscapes i SegmentMeIfYouCan).

Siniša Šegvić odlično govori engleski i talijanski, i ima osnovne komunikacijske vještine na francuskom. Oženjen je i ima troje djece. Član je IEEE.

About the Supervisor

Siniša Šegvić was born in 1971 in Split, Croatia. He completed elementary school and high school in Zadar, Croatia, with one year abroad in Milano, Italy. He received the BS degree in electrical engineering (9 semesters) in 1996 as well as the MS and PhD degrees in 2000 and 2004. He has been employed at UniZg-FER as an assistant professor since 2006.

He was a postdoc researcher at IRISA, Rennes and at TU Graz. He led three research projects of the Croatian Science Foundation (MultiCLOD, MASTIF, ADEPT) as well as several industrial research projects funded by Rimac automobili, RoMB technologies, MicroBlink etc. He has participated in the research center of excellence DataCross, several ERDF projects (SafeTram, MAS, A-UNIT) as well as on one FP7 project (ACROSS).

His research and professional interests include computer vision, visual recognition, scene

understanding, and dense prediction with deep convolutional models. He has published several papers at top conferences (CVPR, ECCV, NeurIPS) and scientific journals. He has been a reviewer at top conferences (CVPR, ECCV, ICCV, AAAI, ICLR) as well as in scientific journals in the fields of computer vision, intelligent transportation systems and robotics. He participated in the industrial development as a technical consultant. He advises several PhD students funded by EU projects, national projects and private companies. His research group has achieved notable results while participating at computer vision challenges such as WildDash, Robust vision challenge, Cityscapes and Fishyscapes.

Siniša Šegvić speaks english and italian very well, and has basic communication skills in french. He is married and has three children. He is a member of IEEE.

Zahvala

Hvala prof. dr. sc. Siniši Šegviću na vjeri, podršci i vođenju kroz doktorski studij. Podučio me zaista svačemu, a posebno sam zahvalan na lekcijama iz računalnog vida, programiranja, Linuxa, Latexa, pisanja i najvažnije - života!

Hvala svim trenutnim i bivšim članovima naše istraživačke grupe. Hvala Ivanu K., Petri, Marinu O., Ivanu G., Marinu K., Mateju, Ivi, Jeleni i Nenadu. Činili ste ugodnu radnu okolinu koja inspirira i potiče na izvrsnost, a u isto vrijeme bili pažljivi i brižni prijatelji!

Hvala tvrtki Rimac Technology na financiranju i podršci tijekom ovoga istraživanja. Posebno hvala Sachi i Tonćiju na savjetima, pomoći i idejama!

Hvala pokojnom ocu Zdenku i majci Željki. Oni su svoje živote posvetili nama djeci i našoj dobrobiti. Neka ova disertacija bude vječni dokaz njihove neprocjenjive žrtve. Hvala mojoj braći: Branku, Tomislavu i Stipi na podršci i razumijevanju. Hvala Leonu i njegovoj obitelji na svakom obliku podrške tijekom cijelog mog školovanja. Hvala obitelji Jonjić na prihvaćanju i potpori. Hvala i ostaloj obitelji koja me pratila svojim mislima i molitvama.

Hvala svim mojim prijateljima. Posebno hvala Kuprešacima sa Save! Tona briga ostala je na onim klupama, a još toliko ih je izbrisala naša unikatna šala i zdrava iglena!

Najveće hvala mojoj supruzi Katarini koja me svakodnevno tješi, bodri i gura naprijed! Bez nje ovo ne bi bilo moguće.

Sažetak

Gusto semantičko prognoziranje anticipira događaje u videu predviđanjem semantike na razini piksela u budućem neopaženom slikovnom okviru. Ova disertacija posebnu pažnju posvećuje metodama gustog prognoziranja na razini apstraktnih značajki dubokog modela. Predlažemo novu metodu prognoziranja značajki koja je primjenljiva na različite zadatke i arhitekture za gusto raspoznavanje u jednoj slici. Predložena metoda sadrži dva modula. Modul F2M (eng. *features to motion*) prognozira gusto polje pomaka kojim se značajke iz prošlosti deformiraju u budućnost. Modul F2F (eng. *features to features*) izravno regresira buduće značajke i zbog toga ima mogućnost zamišljanja u novootkrivenim dijelovima scene. Kombinirani model F2MF može detektirati dijelove slike u kojima će se pojaviti novootkriveni dijelovi scene. U njima model favorizira izravno prognoziranje modulom F2F te prognoziranje defomiranjem u dijelovima scene gdje se može uspostaviti korespondencija. Naša metoda F2MF prognozira najsažetiju i najapstraktniju moguću reprezentaciju modela za gusto raspoznavanje u jednoj slici. Koristimo deformabilne konvolucije i prostorno-vremenske korelacijske koeficijente izračunate između značajki vremenski susjednih slikovnih okvira. Provodimo eksperimente za prognoziranje semantičke segmentacije, segmentacije instanci i panoptičke segmentacije. Rezultati naše metode postižu stanje tehnike u točnosti prognoziranja na podatkovnom skupu Cityscapes.

Ključne riječi: računalni vid, duboko učenje, gusto semantičko prognoziranje, predviđanje budućnosti

Summary

Joint Forecasting of Features and Feature Motion for Semantic Future Prediction in Video

Modern robots operate in complex and dynamic environments which often involve humans. Visual perception is a key component of such autonomous robot systems. It enables a robot to understand the world and its surrounding environment. The goal of perception is to extract a numerical representation from raw sensor data that is useful for planning or decision-making. That representation should not only describe the past and the current state of the environment, but also the future. Anticipation of the future events is especially important because the decisions we make now are fully manifested only in the future. In some applications, such as autonomous driving, forecasting can be of vital importance. It enables making timely decisions some of which could prevent accidents and save lives.

This thesis investigates models for dense semantic forecasting in video. Such models take multiple images on input and provide dense semantic predictions for an unobserved future frame. The predictions usually correspond to either semantic segmentation, instance segmentation, or panoptic segmentation. There are several motivations for research of dense semantic forecasting. For example, it has been shown that it may decrease control errors of an autonomous vehicle. Dense semantic forecasting may also enable better representation learning from unsupervised video. Future semantics can be used as guidance for RGB forecasting, which also can be learned in an unsupervised manner. Forecasting is also important for latency cancellation in real-time systems. In a vanilla setup, the decision-making process is based on obsolete semantic data because of latency of the camera and the semantic processing. Short-term forecasting can neglect that effect and provide timely semantics.

In order to clarify, let us discuss the relations between dense recognition and forecasting. In recognition, the predictions are aligned in time with the input. In forecasting, the predictions are aligned with an unobserved future frame. Hence, models for dense semantic forecasting are required to perform well both in recognition and motion prediction. Accordingly, many dense semantic forecasting systems build on some single-frame recognition model and combine it with some kind of forecasting model.

The forecasting model maps past representations into the future ones. We can categorize the dense semantic forecasting approaches with respect to the type of the representation involved. Naive approach perform the forecasting in the image space. It observes the past RGB images and forecasts the future image. Consequently, we denote it as I2I (image-to-image). The future semantics is obtained by applying the desired recognition model to the forecasted image. However, this approach promotes error propagation from RGB forecasting to the future semantics. Furthermore, most of the decision-making systems are mainly concerned with the

semantics of a future scene, rather than its appearance. Therefore, from the application point of view, dealing with a hard problem such as RGB forecasting seems suboptimal. The second group of approaches perform the forecasting in the space of semantic predictions. They are therefore called S2S (semantics-to-semantics). These approaches first apply the recognition model to all of the past observed frames in order to recover the corresponding semantic predictions. Then the forecasting model maps the past predictions into the future. Experiments from the literature suggest that this approach achieves better performance than the I2I approach. However, operating on semantic predictions poses specific challenges to the design of the forecasting model. First, it requires to establish temporal correspondence across video. However, this is not easily achieved by looking at the semantics since it abstracts out all texture. Furthermore, S2S hinders efficient processing due to high resolution of semantic predictions. The third group of approaches forecast the future optical flow by observing the flow between the past frames. They are therefore called M2M (motion-to-motion). The forecasted flow warps semantic predictions from the most recent frame into the future. These approaches are based on an assumption that the future can be entirely explained by transforming from the past. This is not completely correct since future images often contain emergent novel scenery that has not been observed in the past. Additionally, this approach completely ignores the semantics when forecasting the future motion. This is suboptimal since the future motion is highly correlated with local semantics (a pixel on a car moves differently than a pixel on a person). The fourth group of approaches forecast the hidden feature representation of the recognition model. They are therefore called F2F (feature-to-feature). Accordingly, the recognition model is split into two parts: the feature extraction module and the semantic formation module. The inference pipeline is as follows. First, the feature extraction module computes feature representation for each of the past frames. Second, the forecasting model maps these past features into the corresponding future representation. Finally, the semantic formation module computes the future semantic predictions from the forecasted features. Note that the forecasting model operates on a latent representation of a deep model and is not familiar with the type of the semantic predictions. Consequently, we say that feature forecasting is task agnostic: it can be easily embedded into different recognition architectures that produce different semantic predictions. Furthermore, F2F forecasting promises efficient learning and inference. If the forecasted representation is spatially condense, the forecasting is likely to require less computational resources comparing to the other approaches. Finally, F2F forecasting is also favorable for establishing the correspondence through time. Previous work shows that latent convolutional features may preserve some information about the scene appearance which is valuable for establishing correspondence.

This thesis consolidates the results of the previous research on dense semantic forecasting and extends the previous work on F2F forecasting. We propose a novel method for feature-to-feature forecasting called F2MF (feature-to-motion-and-feature). Our method brings multiple

novelties with respect to the previous work. First, we propose a single-level feature forecasting as opposed to computationally expensive forecasting of a feature pyramid. We achieve that by designing a special single-frame dense recognition models with reduced number of encoder-decoder skip connections. Our experiments reveal that single-level forecasting can be efficient and accurate. Second, we demonstrate effectiveness of deformable convolutions and spatio-temporal correlation for feature forecasting. Third, we propose a novel feature-to-motion (F2M) module which predicts a deformation field which warps past feature tensors into the combined future representation. Finally, we combine F2M forecasting with classic feature-to-feature (F2F) forecasting and demonstrate their complementary nature. In the following paragraphs we describe each part of the proposed dense semantic forecasting system in more details.

Our system builds on the desired single-frame dense recognition model extends it with the proposed feature forecasting model. As in the literature, we split the single-frame recognition model in two parts: the feature extraction module and the semantic formation module. We experiment with different recognition models specialized for semantic segmentation, instance segmentation and panoptic segmentation. Our forecasting approach requires that the feature extraction module provide a single-level representation that is concise and semantically rich. To achieve that, we use single-frame recognition models with certain adaptations which as described next.

For semantic segmentation, we use the SwiftNet baseline model without skip connections. The model consists of an ImageNet pretrained backbone, spatial pyramid pooling (SPP) and the upsampling path. The backbone and the SPP form the feature extraction module, while the upsampling path and the final classification layer form the semantic formation module. We experiment with two architectures which differ in terms of capacity. The more powerful instance has a DenseNet-121 backbone, 512 channels in SPP and 256 channels in the upsampling path. The weaker instance has a ResNet-18 backbone, and 128 channels in the SPP and the upsampling path.

For instance segmentation, we use Mask R-CNN C4. This model requires no adaptation, because the detection module is applied to a single feature tensor which corresponds to the output of the third residual block. Consequently, the feature extraction module consists of the first three residual blocks of the ResNet-50 backbone. The semantic formation module includes the region proposal network, the last residual block and the detection and segmentation modules.

For panoptic segmentation, we use Panoptic Deeplab with a single skip connection. The feature extraction module incorporates the first three residual blocks of the ResNet-50 backbone. The semantic formation module includes the last residual block, and two separate decoders for class-agnostic instance segmentation and semantic segmentation.

At this point we discuss the details of the proposed feature-to-motion-and-feature (F2MF) forecasting model. The model takes $T = 4$ past feature tensors extracted by the the front part

of the single-frame model. The goal of the F2MF is to predict the feature representation from the unobserved future image. The inference starts by concatenating the input features across the semantic dimension. The concatenated features are mixed with the fusion module that corresponds to a single convolutional unit. Note that a single convolutional unit consists of a batch normalization, convolutional layer and ReLU activation. In parallel with the fusion, the features are processed by the correlation module.

The correlation module first separately embeds features from each time-instant into a vector space with enhanced metric properties. Then, we compute spatio-temporal correlation coefficients for the three pairs of temporally adjacent feature tensors. The coefficients of a single input pair are arranged in a third-order tensor. This tensor contains d^2 correlation coefficients in each image location $\mathbf{p}$. These coefficients represent similarity between the feature vector at location $\mathbf{p}$ from the first input and feature vectors from the second input inside the $d \times d$ neighborhood centered at $\mathbf{p}$. The final output of the correlation module is a concatenation of the three correlation tensors.

The correlation features are then concatenated with the outputs of the fusion module. The compound representation is further processed by a sequence of six units with deformable convolutions. The resulting representation is shared between the F2F and the F2M modules. The same shared representation is also used to predict the blending weights used for combining the predictions from the two modules. The weights are computed by a single convolutional unit with five feature maps on output. The first four maps quantify the contributions of the four F2M forecasts which originate from warping each of the past input representations. The last map quantifies the contribution of the F2F forecast.

The F2M module follows the assumption that the future can be explained by a geometric transformation of the past. In practice, that transformation corresponds to backward warping according to the predicted feature flow. Thus, the F2M module predicts the feature flow from unobserved future features towards each of the four past feature tensors. The flow is inferred from the shared representation through a single convolutional unit with $4 \times 2 = 8$ feature maps. Four coming from the four input feature tensors, and two from displacement across the x and y axis. The warping proceeds separately for each past time-instant according to the corresponding forecasted flow. Warping towards multiple past instants enables the model to choose the most appropriate moment for explaining a certain part of the scene. This is especially useful in some (dis)occlusion patterns. Nevertheless, the basic assumption about explaining the future solely by transforming from the past is only partially true. The future is unpredictable and novel parts of the scene often emerge. Moreover, deforming from the past is hard in such regions, also at disadvantage whenever it is difficult to establish the correspondence. Because of that, we combine F2M and F2F forecasting, as we describe next.

The F2F module forecasts the future features directly from the shared representation. It con-

sists of a single convolutional unit where the number of output channels has to be equal to the number of channels of a single input feature tensor. This module is not restricted to geometrical reconstruction from the past. Because of that, this approach has a chance to develop ability to imagine emergent parts of the scene. There are some differences between our F2F module and similar modules from the literature. First, our module always targets a single-level representation of the unobserved future image. Second, instead of regular or dilated convolutions, we use deformable convolutions which have a better potential for modeling geometric transformations. Third, our F2F module operates on top of a representation that is enriched with spatio-temporal correlation features.

Finally, our combined forecasting module combines the direct feature forecasting (F2F) and the forecasting by deforming from the past (F2M). Hence, we denote our approach as F2MF forecasting. We hypothesize that F2F and F2M might be complementary. Consequently, we form the final forecasted predictions as a weighted sum according to dense blending weights. This encourages specialization of the F2F module for the novel parts of the scene, and allows the F2M module to concentrate on parts of the scene which do not contain emergent scenery. However, attaching the loss only to the compound forecast may weaken the learning signal for individual F2F and F2M modules. Consequently, we add two auxiliary losses which take into account only forecasts by the individual modules. Each of the three loss terms is expressed as the L2 distance between the forecasted features and extracted features obtained from the future frame. Note that this type of training does not require human-made annotations. Therefore, the forecasting model can be trained on unlabeled video. However, we do need the labels for supervised learning of the single-frame recognition model.

We provide a comprehensive experimental evaluation of our method. Most of our experiments are based on the Cityscapes dataset. Cityscapes is suitable for forecasting experiments sine each labeled image is accompanied by the corresponding unlabeled video clip. We evaluate the short-term (3 timesteps, 0.18s) and the mid-term (9 timesteps, 0.54s) forecasting performance by comparing the future semantic predictions with ground truth. We use mean intersection over union (mIoU) for semantic segmentation, average precision (AP) for instance segmentation, and panoptic quality (PQ) for panoptic segmentation.

Our method achieves competitive performance across all three tasks in both short-term and mid-term forecast. We achieve state-of-the-art performance on short-term instance segmentation forecasting. To illustrate the level of the forecasting accuracy, we point out that in short-term forecasting we reach around 69 mIoU points, and around 58 mIoU points in mid-term forecasting. Qualitative experiments support our hypothesis about the complementary nature of F2F and F2M forecasting. Specifically, we visualize per-pixel blending weights which reveal whether the corresponding part of the scene was forecasted either by the F2M or the F2F module. We noticed that in novel parts of the scene the model tends to trust the F2F module more,

while in the static parts the model prefers the F2M module. The same hypothesis is justified by another validation experiment. We show that in certain groups of pixels the F2F module achieves better accuracy, while in others it is the opposite (F2M prevails). This suggests that it makes sense to combine these two approaches. Other validation experiments show that the compound model outperforms the individual modules. We demonstrate a significant boost in performance when using the deformable convolutions, and also when including the correlation features. We investigate the optimal number of observed input frames, and show that it is indeed four frames. We also evaluate the performance of our approach in long-term forecasting (up to 1.44s into the future) by applying our model autoregressively. We measure the forecasting performance under the presence of domain shift. In particular, we train our system on Cityscapes and evaluate it on on CamVid. These experiments revealed graceful performance degradation. Finally, computational analysis reveals that our method is more efficient than all approaches from the literature, as well as that it may even be applicable in real-time.

To conclude, we have shown that single-level feature forecasting can achieve competitive dense-semantic-forecasting performance, while being more efficient than all previous approaches. We have empirically demonstrated the complementary nature of direct feature forecasting and the forecasting by deforming the past representations. We have shown that deformable convolutions and the spatio-temporal correlation features can significantly improve the forecasting performance. The proposed method opens several directions for future work. For example, it can be extended towards multi-modal future prediction, which is especially important for the long-term forecasting.

Keywords: computer vision, deep learning, dense semantic forecasting, future prediction

Sadržaj

Poglavlje 1

Uvod

U moderno vrijeme roboti djeluju u kompleksnim i dinamičnim okruženjima koja na razne načine uključuju i ljude. Vizualna percepcija ključna je komponenta takvih autonomnih robotskih sustava. Ona robotu omogućuje razumijevanje svijeta oko sebe. Robot koji to ne može, ograničen je na obavljanje malog broja jednostavnih zadataka u strogo kontroliranim uvjetima (npr. robotska ruka u tvornici automobila). S druge strane, napredni roboti mogu obavljati različite zadatke i u do tada nepoznatim okolinama. Na primjer, roboti danas mogu: dostavljati hranu [*], prevoziti putnike [†], ili pomoći ljudima u pronalasku željenog odredišta [1].

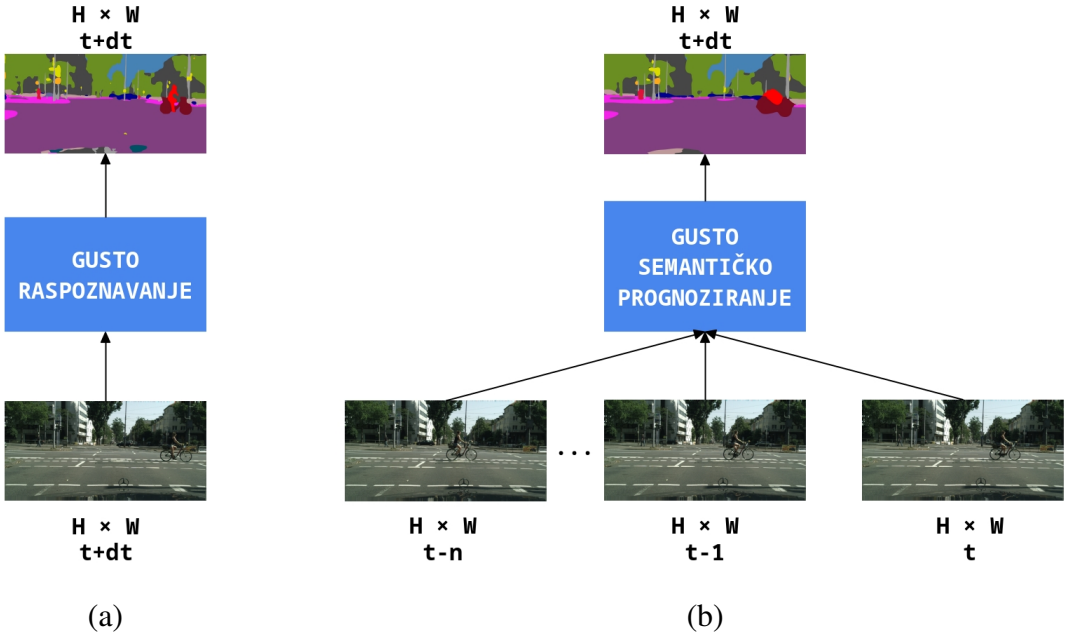
U autonomnom sustavu percepcija je zadužena za obradu podataka prikupljenih senzorima [2]. Cilj percepcije je sirove podatke pretvoriti u numeričku reprezentaciju koja je korisna za planiranje ili donošenje odluke o sljedećoj akciji. Jedan od najdostupnijih i najraširenijih senzora u autonomnim sustavima je kamera koja prikuplja digitalne slike [2]. Iako je većini ljudi lako donijeti složene zaključke iz digitalne slike, ona u svome sirovom obliku računalu predstavlja samo niz nepovezanih brojeva. Zato je potrebno sliku prikupljenu kamerom prikladno obraditi i učiniti računalu korisnom. Obradom digitalnih slika bavi se interdisciplinarno područje poznato kao računalni vid. Cilj računalnog vida je iz slike izlučiti korisne informacije. Modeli računalnog vida specijalizirani za različite zadatke čine jezgru sustava za percepciju. Primjerice, prije nego što donese sljedeću odluku o skretanju ili ubrzavanju autonomni automobil vjerojatno će: prepoznati ostale sudionike u prometu (detekcija objekata), odrediti prohodni dio ceste (semantička segmentacija), procijeniti udaljenost do automobila ispred (procjena dubine) i obaviti druge potrebne analize. Osim analizi onoga što se već dogodilo, inteligentan sustav će svoju odluku prilagoditi razmatranju onoga što bi se tek moglo dogoditi. Da bi to bilo moguće, sustav za percepciju mora imati sposobnost anticipacije budućnosti. Prognoziranje budućnosti je važno jer odluke koje donosimo sada u potpunosti se manifestiraju tek u budućnosti. Autonomni sustav koji prognozira budućnost upravljačke odluke može donositi ranije. Pose-

[*] https://www.youtube.com/watch?v=P_zRwq9c8LY

[†] https://www.youtube.com/watch?v=__EoOvVkEMo

bice u vožnji, pravovremene odluke mogu značiti razliku između života i smrti. Anticipacija budućnosti važan je dio i sustava za navigaciju [3]. Klasični pristupi prognoziraju dinamiku scene eksplicitnim modeliranjem gibanja kamere i objekata [3]. Ipak, takvi pristupi skloni su greškama zbog nedovoljne preciznosti oblikovanih modela gibanja. Stoga su novija istraživanja usmjerena na pokušaje implicitnog praćenja dinamike scene dubokim učenjem u videu [4].

Ova disertacija bavi se modelima za prognoziranje gustih semantičkih predikcija u budućem slikovnom okviru na temelju promatranih okvira iz prošlosti. Za razliku od klasičnih pristupa [3], nećemo se oslanjati na eksplicitno modeliranje geometrijskih odnosa u sceni. Suprotno, naša želja je istražiti pristupe koji mogu učiti s kraja na kraj, izravno iz podataka, kako bismo došli korak bliže robotima koji će se moći snalaziti u prostoru kao što to čine ljudi. Cilj je, dakle, budućnost opisati na semantičkoj razini. Primjerice, u obliku semantičke segmentacije [5], segmentacije instanci [6] ili panoptičke segmentacije [7] budućeg, neopaženog slikovnog okvira. Slika 1.1 ilustrira razliku između gustog raspoznavanja u jednoj slici i gustog semantičkog prognoziranja. Prilikom raspoznavanja, ulaz modela čini jedna slika, a izlaz guste predikcije koje su prostorno poravnate s odgovarajućim pikselima u ulaznoj slici. Prilikom prognoziranja, ulaz modela čini niz slika iz prošlosti, a izlaz guste predikcije koje su prostorno poravnate s odgovarajućim pikselima u budućoj (neopaženoj) slici. Stoga model za prognoziranje nije opterećen samo raspoznavanjem ulaznih piksela, nego i praćenjem njihova gibanja i predviđanjem budućih lokacija. Prognoziranje je očigledno značajno kompleksniji problem. Štoviše, modeli za raspoznavanje u jednoj slici često su integralni dio sustava za prognoziranje.



Slika 1.1: Usporedba gustog semantičkog raspoznavanja u jednoj slici (a) i gusto semantičkog prognoziranja (b).

Iako je raspoznavanje samo po sebi već dovoljno kompleksno, prognoziranje nosi dodatan

skup izazova. Prilikom raspoznavanja često je potrebno zaključiti kako neki skup piksela čini jednu cjelinu. Primjerice, jedan primjerak objekta prilikom segmentacije instanci. Prognoziranje dodatno zahtjeva uspostavu korespondencije u vremenskoj dimenziji, što je nužno kako bi se ispravno odredio položaj objekta u budućnosti. Pored toga, prilikom prognoziranja model se susreće s velikim pomacima objekata u prostoru slike te mogućim pojavama novootkrivenih dijelova scene. Slika 1.2 ilustrira ove dvije pojave. Prikazane su dvije scene iz podatkovnog skupa Cityscapes [8], te za svaku scenu po dvije slike nastale u vremenskom razmaku jednakom približno polovici sekunde. Stupac (a) prikazuje scenu u kojoj automobil koji nosi kameru skreće, te zbog toga dominira rotacijsko gibanje kamere. Kao posljedica toga i dodatnog gibanja osobe na slici, nastaje pomak veći od 400 piksela između lokacije osobe u dva zabilježena trenutka. Veličina toga pomaka ilustrirana je usmjerenom zelenom linijom u budućoj slici (dolje), dok je desni rub osobe u obje slike zabilježen isprekidanom crvenom linijom. Veliki pomaci su problematični za konvolucijske modele zbog toga što se vizualni dokaz za postojanje nekog objekta u prošlosti nalazi daleko od mjesta same predikcije u budućnosti. Stupac (b) prikazuje scenu u kojoj automobil ispred kamere skreće udesno i napušta scenu. Kao posljedica toga u budućoj slici nastaje novootkrivena regija koja je označena crvenim isprekidanim pravokutnikom. Model za prognoziranje obično promatra nekoliko slikovnih okvira iz prošlosti, te je u ovoj situaciji izgledno da označeni dio scene nije vidio u niti jednom od njih. U ovakvim situacijama model mora zamišljati, te na osnovu konteksta prognozirati što bi se moglo nalaziti u novootkrivenom

(a) (b)

Slika 1.2: Ilustracija izazova prilikom gustog semantičkog prognoziranja na dvije scene iz skupa Cityscapes. Vremenski razmak između gornje i donje slike je 0.5 sekundi. Stupac (a) prikazuje mogućnost pojave velikih pomaka objekata u prostoru slike. Stupac (b) prikazuje mogućnost pojave novootkrivenih dijelova scene zbog pomaka objekata.

dijelu scene. Idealno bi model za prognoziranje prepoznao takvu situaciju i na svome izlazu to detektirao kao regiju visoke nepouzdanosti. Modeli za raspoznavanje u jednoj slici nisu suočeni s ovakvim zahtjevima te obično obavljaju izravno mapiranje iz piksela u semantičku predikciju, dok sustavi za prognoziranje zbog ovoga moraju imati i neka generativna svojstva. Štoviše, modeli za dugoročno prognoziranje budućnosti svakako bi u obzir trebali uzeti stohastičnost budućnosti [9] i razmatrati više mogućih ishoda.

Motivacija za istraživanje gustog semantičkog prognoziranja dolazi iz više pravaca. Precizno prognoziranje budućnosti na semantičkoj razini može značajno unaprijediti kvalitetu upravljanja različitim autonomnim sustavima. Prognoziranje omogućava rano opažanje mogućih opasnosti i pruža više vremena za ispravnu reakciju na njih. Simulacijski eksperimenti učenja agenta za autonomnu vožnju pokazuju da agenti koji imaju mogućnost anticipacije budućnosti postižu značajno manju grešku upravljanja od sustava temeljenih samo na analizi trenutnog okvira [10]. Isti eksperimenti otkrivaju i pozitivnu korelaciju između točnosti prognoziranja i kvalitete upravljanja. Takvi rezultati jasno motiviraju razvoj još točnijih metoda semantičkog prognoziranja, jer imaju izravan utjecaj na sigurnost i kvalitetu autonomnih automobila.

Pored praktične primjene u autonomnoj vožnji, motivaciju za dublje istraživanje ovog problema nalazimo i u potencijalu nenadziranog učenja u videu. Video predstavlja neiscrpan izvor podataka za učenje dubokih modela. Slijed slikovnih okvira može biti snažan signal za učenje jer odražava puno zakonitosti o svijetu u kojem živimo. Primjerice, iz videa je jasno da se pikseli automobila gibaju različito od piksela ljudi. Nadamo se da bi model promatrajući video bez ikakvih oznaka, iz dinamike i interakcije između objekata mogao naučiti reprezentaciju koja bi bila prigodna za njihovo raspoznavanje. Istraživanje nenadziranog učenja semantičkog prognoziranja je jedan korak ka tome cilju. Semantičko prognoziranje može biti i dio većeg sustava za prognoziranje RGB okvira [11], koji se pak može učiti nenadzirano s kraja na kraj. U tom slučaju su generativni modeli za prognoziranje RGB okvira uvjetovani budućim gustim semantičkim predikcijama. Uvjet konzistentnosti između semantike i generirane slike modelu služi kao dodatan vodič prilikom kompozicije scene.

Prognoziranje je važno i zbog postojanja latencije u sustavima za rad u stvarnom vremenu [12]. Naime, od trenutka u kojem svjetlosni snop obasja senzor kamere do završetka izračuna semantičkih predikcija može proći nemala količina vremena. To kašnjenje uzrokovano je latencijom same kamere, prijenosom podataka kroz memoriju te samim procesom semantičkog zaključivanja. Zbog toga upravljački podsustav cijelo vrijeme radi sa zapravo zastarjelim podacima. To može biti posebno naglašeno kod autonomnih sustava koji se gibaju velikim brzinama. Metode kratkoročnog semantičkog prognoziranja mogu značajno umanjiti taj problem na način da korigiraju izlazne predikcije tako da odgovaraju izgledu scene u bliskom budućem trenutku [12]. Na taj način bi upravljački podsustav odluke o svojim akcijama donosio na temelju pravovremenih i realnih semantičkih predikcija. U ovim situacijama kašnjenja su dovoljno malena da

budućnost nije višemodalna, nego izgledna. To otvara mogućnost za izravnu primjenu metoda predstavljenih u ovome radu.

Ova disertacija konsolidira rezultate prethodno objavljenog istraživanja semantičkog prognoziranja temeljenog na prognoziranju iz značajki u značajke [13, 14, 15]. Rezultat istraživanja je nova metoda za prognoziranje značajki nazvana F2MF (eng. *features to motion and features*). Metoda ima nekoliko važnih doprinosa u odnosu na prethodne radove [6, 16]. Umjesto računalno skupog prognoziranja piramide značajki pokazali smo da jednorazinsko prognoziranje najsažetije reprezentacije semantičkih modela može biti efikasno i točno. Kako bi to bilo moguće posebno su dizajnirani semantički modeli za predikciju u jednom okviru sa smanjenim brojem preskočnih veza. Predloženom modelu za prognoziranje značajki deformabilne konvolucije olakšavaju imitiranje različitih uzoraka gibanja unutar jednog sloja. Prilagođeni korelacijski modul pospješuje uspostavu korespondencije među značajkama kroz vrijeme. On računa koeficijente sličnosti značajke u trenutku t sa značajkama iz lokalnog susjedstva iz prethodnog vremenskog trenutka $t - dt$. Ti koeficijenti obogaćuju internu reprezentaciju prognostičkog modela informacijama o položaju i pomaku značajke kroz vrijeme, što značajno ograničava njenu poziciju u budućnosti. Predstavili smo i novi modul za prognoziranje značajki F2M (eng. *features to motion*). Značajke iz različitih trenutaka u prošlosti deformiraju se u budućnost predviđenim vektorima pomaka i fuzioniraju predviđenim težinama u jedinstvenu ciljanu prognozu. To modulu F2M omogućuje objašnjavanje dijelova buduće scene izravnim preslikavanjem značajki iz najpogodnijeg prošlog trenutka. Međutim, takvo prognoziranje ne može objasniti pojavu novih dijelova scene. S druge strane, izravno prognoziranje značajki modulom F2F (eng. *features to features*) nije ograničen o rekonstrukcijom iz prošlosti i ima mogućnost zamišljanja. Stoga, predložena metoda F2MF kombinira predikcije ova dva modula prema gustom polju težina koje regresira isti model. Na taj je način olakšano učenje cijelog modela jer se razdvaja objašnjavanje varijance nastale pomakom od varijance nastale pojavom novootkrivenih dijelova scene. Metoda za prognoziranje značajki F2MF je lako primjenjiva na različite zadatke guste predikcije i to je demonstrirano eksperimentima koji istražuju prognoziranje semantičke segmentacije, segmentacije instanci i panoptičke segmentacije. Znanstveni doprinosi disertacije sažeto su opisani u nastavku:

1. Prognoziranje semanti čke budućnosti prirodne scene deformiranjem konvolucijske reprezentacije s obzirom na prognozirano polje pomaka.
2. Kombinirani prognosti čki pristup koji razdvaja utjecaje kretanja od utjecaja prethodno neviđenih dijelova scene.
3. U činkovita implementacija predloženog pristupa zasnovana na deformabilnim konvolucijama, korelacijskim značajkama te prognoziranju najapstraknije i najsažetije reprezentacije modela za predikciju u jednoj slici.

Ostatak ove disertacije strukturiran je na sljedeći način.

U poglavlju 2 opisane su odabrane tehnike računalnog vida korištene prilikom izrade ove disertacije. Poseban dio posvećen je klasifikacijskim konvolucijskim arhitekturama koje čine jezgru modela za gusto raspoznavanje u jednoj slici. Naglasak je na arhitekturama temeljenima na preskočnim vezama i gustoj povezanosti slojeva. Drugi dio poglavlja obrađuje rekonstrukcijske tehnike računalnog vida koje su potrebne za rješavanje problema prognoziranja. Posebno su razmatrane osnove optičkog toka, te korelacijski sloj kao neizostavni modul dubokih arhitektura za optički tok. Naposljetku, detaljno su opisane deformabilne konvolucije koje su korištene u predloženom modulu za prognoziranje značajki.

U poglavlju 3 nalazi se pregled literature podijeljen na dvije cjeline: metode za gustu semantičku predikciju u jednoj slici, te duboki modeli za obradu videa. Prvi dio obrađuje semantičke modele i tehnike bliske modelima korištenima u predloženom sustavu za gusto semantičko prognoziranje. Drugi se dio usredotočuje na metode za obradu videa koje koriste slične koncepte kao naš predloženi sustav za prognoziranje. Posebna pažnja posvećena je RGB prognoziranju izgleda budućih okvira te gustom semantičkom prognoziranju. Detaljno su opisane razlike između predložene metode za gusto prognoziranje i najsličnijih metoda iz literature.

U poglavlju 4 detaljno su opisani korišteni modeli za gusto raspoznavanje u jednoj slici. Prilagođeni modeli SwiftNet, Mask R-CNN, te Panoptic Deeplab dio su predloženih sustava za prognoziranje semantičke segmentacije, segmentacije instanci te panoptičke segmentacije.

Poglavlje 5 detaljno opisuje predloženi prognostički model utemeljen na regresiranju pomaka i izravnom regresiranju značajki (F2MF). Opisane su sve ključne komponente predložene metode. Prvi dio opisuje korelacijski modul i važnost korelacijskih značajki za prognoziranje budućnosti. Nakon toga obrađuje se prognoziranje regresijom pomaka značajki modulom F2M. Razmatraju se razlike između inačica toga modula koje se oslanjaju na unaprijednu odnosno unatražnu deformaciju značajki. Zatim je opisano izravno prognoziranje značajki modulom F2F. Konačno, opisuje se način udruživanja prethodno spomenutih modula u predloženu metodu F2MF.

Poglavlje 6 opisuje eksperimentalno vrednovanje predložene metode. Prvo su opisani eksperimenti prognoziranja semantičke segmentacije, segmentacije instanci i panoptičke segmentacije na podatkovnom skupu Cityscapes. Zatim se opisuju opširni validacijski eksperimenti koji vrednuju doprinose pojedinih komponenti točnosti prognoziranja. Nakon toga razmatra se mogućnost dugoročnog prognoziranja autoregresijskom primjenom predložene metode. Zatim se razmatra sposobnost generalizacije predložene metode u prisutnosti pomaka domene. Konačno se analizira računska složenost predloženog sustava prognoziranja i srodnih metoda iz literature.

Na kraju rada, u poglavlju 7 nalazi se zaključak.

Poglavlje 2

Odabrane tehnike računalnog vida

U ovome poglavlju opisane su odabrane tehnike računalnog vida. Razumijevanje tih tehnika omogućit će lakše razumijevanje predloženog sustava za semantičko prognoziranje. U posljednje vrijeme glavni alat za rješavanje mnogih problema računalnog vida je duboko učenje (eng. *deep learning*). Duboki modeli postali su stanje tehnike u mnogim zadacima računalnog vida: klasifikaciji slika [17], detekciji objekata [18], semantičkoj segmentaciji [19], segmentaciji instanci [20], panoptičkoj segmentaciji [21], procjeni dubine [22], procjeni optičkog toka [23] i dr. Za ovu disertaciju najvažniji elementi dubokog učenja su klasifikacijske konvolucijske arhitekture i deformabilne konvolucije. Pored toga, upoznat ćemo se s pojmovima optičkog toka i korelacijskog sloja te izobličenjem slike s obzirom na gusto deformacijsko polje.

2.1 Klasifikacijske konvolucijske arhitekture

Klasifikacija slika jedan je od osnovnih zadataka računalnog vida. Cilj je odrediti semantičku kategoriju kojoj pripada zadana slika. Primjena dubokog učenja u računalnom vidu započela je upravo kroz ovaj zadatak. Povijesnim trenutkom smatra se dominantna pobjeda dubokog modela AlexNet [17] na natjecanju u klasifikaciji slika na podatkovnom skupu ImageNet-1k [24] 2012. godine. ImageNet-1k sadrži oko milijun slika raspoređenih u 1000 semantičkih razreda. Natjecanje je u to vrijeme bilo posebno zahtjevno zbog značajno slabije računalne moći tadašnjih grafičkih procesora. Model AlexNet sadrži 60 milijuna parametara raspoređenih u 5 konvolucijskih i 3 potpuno povezana sloja. Pobjeda AlexNeta demonstrirala je mogućnost treniranja dubljih modela sa velikim brojem parametara, što je u to vrijeme predstavljalo najveći izazov. Poruka koju su mnogi istraživači izvukli iz ovoga bila je: "dublje je bolje" (eng. *deeper is better*). Istraživanje dubokih modela i oblikovanje novih arhitektura pratili su upravo ovu ideju. Na istom natjecanju dvije godine kasnije pojavila se nova arhitektura pod imenom VGG [25]. Različite varijante predložene arhitekture sadrže od 133 do 144 milijuna parametara raspoređenih u 11 do 19 slojeva. Prilikom treniranja dubljih varijanti ključnom se pokazala ini-

cijalizacija dijela slojeva dubljeg modela parametrima predtreniranog plitkog modela. Sljedeće godine je pobjednički model ResNet [26] sa ukupno 152 sloja nadmašio ljudsku performansu u raspoznavanju slika na ImageNetu. Varijanta istog modela sa 1000 slojeva pokazala je mogućnost treniranja iznimno dubokih modela, ali i pojavu zasićenja točnosti s obzirom na dubinu modela. Validacija klasifikacijskih modela na ImageNetu aktualna je i danas. Točnost modela se kontinuirano podiže novim otkrićima i arhitekturama [27, 28]. Ipak, neke starije arhitekture poput ResNeta ili DenseNeta [29] popularne su i danas zbog široke primjene u drugim zadacima računalnog vida.

Istraživanje klasifikacijskih arhitektura i njihovo testiranje na ImageNetu važno je i za napredak performanse u ostalim zadacima računalnog vida. Klasifikacijski modeli predtrenirani na ImageNetu ključni su dio modernih arhitektura za gustu predikciju [18, 20, 21, 30]. Oni se izravno primjenjuju na ulaznu sliku i služe kao ekstraktori značajki. Takve dijelove modela za gustu predikciju zovemo i okosnica (eng. *backbone*). Ugradnja klasifikacijskih modela u arhitekture za gustu predikciju omogućava prijenos stečenog znanja predtreniranjem na jeftinijim, ali brojnijim oznakama na razini cijele slike. Gradijente gubitka guste predikcije moguće je propagirati skroz do okosnice te na taj način fino podesiti već predtrenirane parametre za ciljni zadatak. Mudro je parametre okosnice inicijalizirati predtreniranim ImageNet težinama. Takva inicijalizacija značajno podiže točnost i olakšava treniranje modela na ciljanom zadatku. Primjerice, predtreniranje na ImageNetu poboljšava segmentacijsku točnost modela SwiftNet na podatkovnom skupu Cityscapes [8] za 5 postotnih bodova metrike mIoU u odnosu na učenje sa slučajnom inicijalizacijom. Više je hipoteza za takvo opažanje. Podatkovni skupovi sa gustim oznakama sadrže značajno manje slika od ImageNeta zbog velikog vremenskog, a onda i novčanog troška gustog označavanja. Model predtreniran na ImageNetu bolje generalizira zbog bogatijeg signala za učenje (kolokvijalno, "vidio" je više slika). Također, predtrenirani model unaprijed može prepoznavati brojne vizualne koncepte od kojih su mnogi prisutni i u ciljanom skupu podataka. Treniranje modela za gustu predikciju sa predtreniranom okosnicom ima priliku blago prilagoditi težine koje prepoznaju korisne vizualne koncepte i potpuno prenamijeniti preostale. To značajno smanjuje potreban broj iteracija treniranja modela na ciljnom zadatku i tako ubrzava eksperimentiranje i razvoj.

Sljedeća potpoglavlja pobliže opisuju klasifikacijske arhitekture ResNet i DenseNet. One su korištene kao okosnice u modelima za gusto raspoznavanje u jednoj slici na kojima se temelje predloženi sustavi za gusto semantičko prognoziranje.

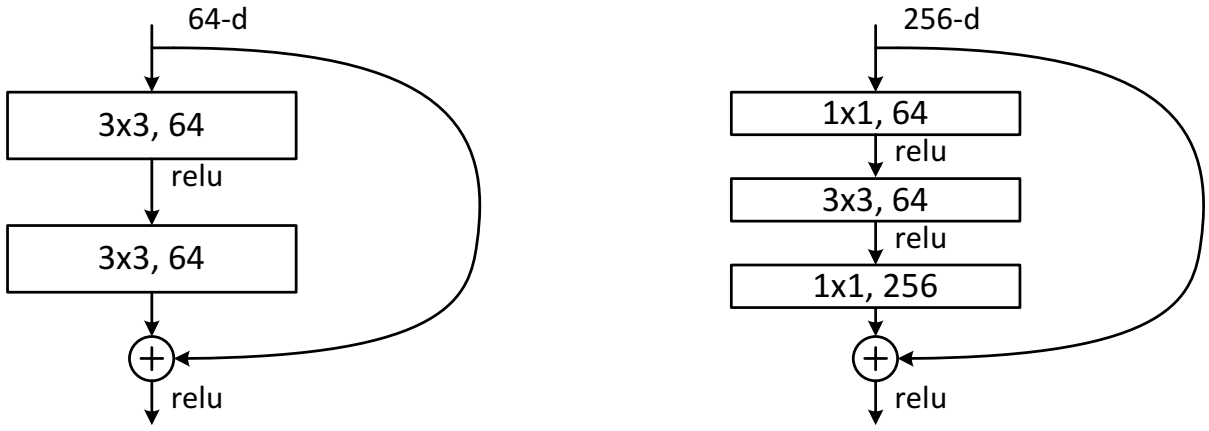
2.1.1 Modeli s rezidualnim preskočnim vezama

Arhitektura ResNet [26] predstavlja obitelj dubokih modela temeljenu na rezidualnim vezama. Preskočne veze na izlaz parametrizirane transformacije F dodaju ulaz x. Na taj način transformacija F čije parametre učimo ne mora izravno učiti zamišljeno željeno mapiranje, nego

samo rezidual između željenog mapiranja i ulaza. Primjerice, ako je željeno mapiranje jednako identitetu, dovoljno je da se parametrizirana transformacija pritegne na nulu.

Pokazalo se da dodavanje rezidualnih veza u model smanjuje pogrešku na skupu za treniranje, ali također pospješuje generalizaciju na skupu za validaciju. Rezidualne veze posebno olakšavaju učenje ranijih slojeva jer gradijent funkcije gubitka do njih sada dolazi skoro izravno. Ipak, važno je napomenuti da ResNet svoju prednost u odnosu na prethodne klasifikacijske arhitekture nije ostvario samo na temelju rezidualnih veza. Važnu ulogu u optimizaciji igraju i normalizirajući slojevi. ResNet svaki konvolucijski sloj uparuje sa po jednim slojem normalizacije nad grupom [31] što optimizacijsku krivulju čini glađom [32]. ResNet je demonstrirao još jedan važan iskorak u dizajnu klasifikacijskih arhitektura. Višestruki potpuno povezani slojevi na izlazu modela zamijenjeni su sa globalnim sažimanjem prosjekom i samo jednim potpuno povezanim slojem. Ovo je značajno smanjilo broj parametara modela i prebacilo težište kapaciteta na brojnije konvolucijske slojeve.

Rezidualne veze dio su osnovnih gradivnih elemenata arhitekture ResNet. Na slici 2.1 shematski su prikazane dvije vrste rezidualnih konvolucijskih jedinica: osnovna rezidualna jedinica (lijevo) i rezidualna jedinica sa uskim grlom (desno). Kod osnovne rezidualne jedinice parametriziranu transformaciju čine dva konvolucijska sloja sa jezgrom 3×3. Kod rezidualne jedinice sa uskim grlom transformacija sadrži tri konvolucijska sloja: prvi sloj sa jezgrom 1×1 koji smanjuje broj kanala, drugi sloj sa jezgrom 3×3 (usko grlo), te posljednji sloj sa jezgrom 1×1 koji restaurira izvorni broj kanala. Ovakva složenija konfiguracija s uskim grlom nastala je zbog potrebe smanjenja računalnog troška osnovne jedinice. Osnovna rezidualna jedinica koristi se kod plićih instanci modela: ResNet-18 i ResNet-34. Rezidualna jedinica sa uskim grlom koristi se kod dubljih instanci modela zbog svoje učinkovitosti: ResNet-50, ResNet-101 te ResNet-152. Kod svih spomenutih modela osnovne jedinice se grupiraju u četiri bloka. Jedan blok čine svi slojevi koji provode obradu na istoj prostornoj rezoluciji. Rezolucije izlaznih tenzora iz četiri



Slika 2.1: Primjer rezidualnih konvolucijskih jedinica - gradivnih elemenata arhitekture ResNet. Lijevo: osnovna rezidualna jedinica (eng. *basic block*). Desno: rezidualna jedinica sa uskim grlom (eng. *bottleneck*). Slika preuzeta iz [26].

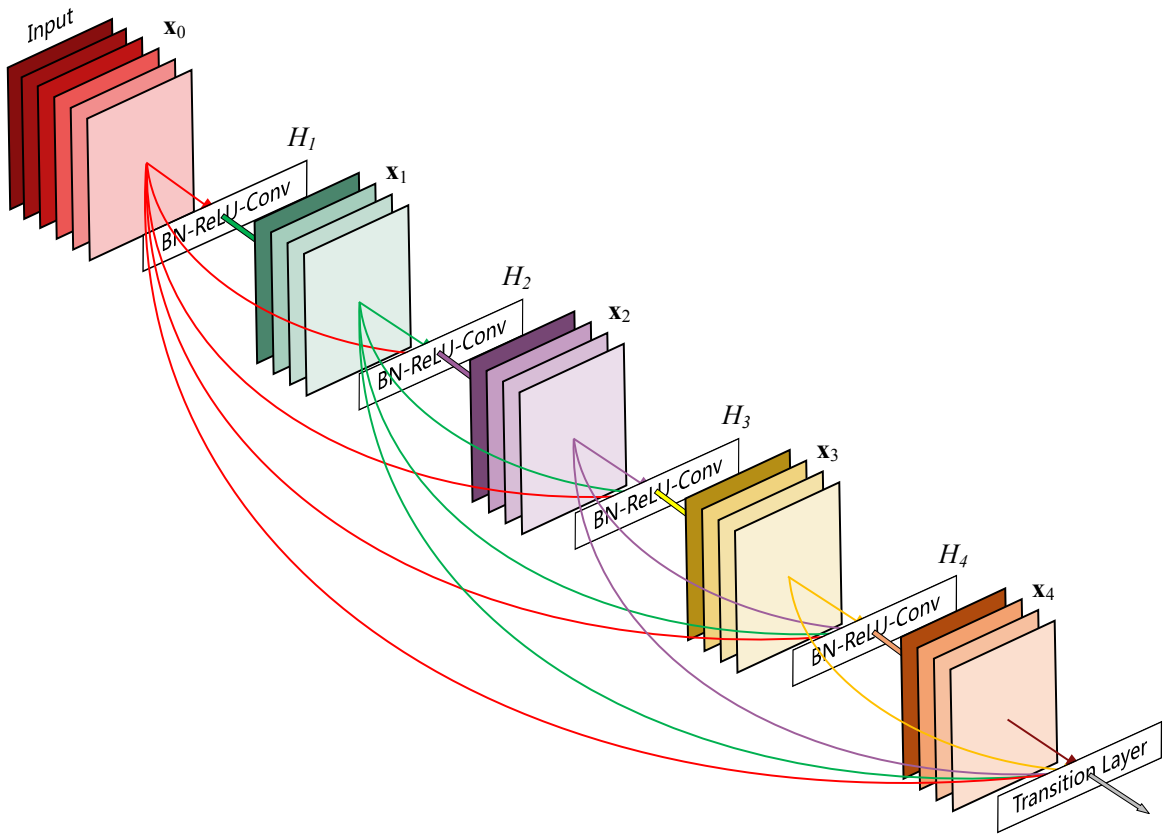
bloka su redom $4\times$, $8\times$, $16\times$ i $32\times$ poduzorkovane u odnosu na ulaznu slike. Dakle, povećanjem dubine bloka smanjuje se prostorna rezolucija značajki, ali povećava njihova semantička dimenzija, odnosno broj kanala. U modelima za gustu predikciju sa okosnicom ResNet uobičajeno je da dekoder na ulazu prima izlaze posljednjeg rezidualnog bloka. Pored toga, dekoderi sa ljestvičastim naduzorkovanjem [33] koriste izlaze iz preostalih blokova za precizniju prostornu lokalizaciju značajki. Modeli sa ResNetom kao okosnicom postizali su stanje tehnike u različitim zadacima: detekciji objekata [18], segmentaciji instanci [20], efikasnoj semantičkoj segmentaciji [30], a popularni su i danas.

2.1.2 Gusto povezani konvolucijski moduli

Arhitektura DenseNet [29] temelji se na gusto povezanim konvolucijskim blokovima. Svaka konvolucijska jedinica unutar gusto povezanog bloka na svome ulazu prima izlaze svih prethodnih jedinica. Jedna konvolucijska jedinica sastoji se od dva konvolucijska sloja: prvog sa jezgrom 1×1 te drugog sa jezgrom 3×3. Izlazi oba sloja se normaliziraju po grupi i aktiviraju zglobnicom. Recentna istraživanja guste povezanosti adresiraju problem prvog sloja u konvolucijskoj jedinici koji ima malu računsku gustoću (omjer množenja i veličine radnog skupa) [34]. Gusto povezani blok određen je brojem konvolucijskih jedinica i faktorom rasta k. Svaka jedinica u bloku doprinosi dijeljenoj reprezentaciji s k mapa značajki. Konačan izlaz bloka računa tranzicijski sloj (eng. *transition layer*) koji konkatenira izlaze svih konvolucijskih jedinica, smanjuje broj kanala primjenom 1×1 konvolucije i provodi prostorno sažimanje. Slika 2.2 ilustrira gusto povezani blok sa četiri konvolucijske jedinice i jednim tranzicijskim slojem. Faktor rasta jednak je $k = 4$.

Postoji više instanci arhitekture DenseNet za klasifikaciju na ImageNetu: DenseNet-121, DenseNet-161, DenseNet-169, DenseNet-201 i DenseNet-264. Slično kao ResNet, svaka od njih ima četiri gusto povezana bloka koji produciraju $4\times$, $8\times$, $16\times$ i $32\times$ poduzorkovane značajke. Instance se razlikuju po broju konvolucijskih jedinica u posljednja dva bloka. U svim blokovima jednu konvolucijsku jedinicu čine dva konvolucijska sloja s pripadajućim normalizacijskim slojem i aktivacijskom funkcijom. Prvi sloj sa jezgrom 1×1 služi kao usko grlo i smanjuje broj ulaznih kanala za drugi sloj sa jezgrom 3×3. Faktor rasta u svim blokovima postavljen je na $k = 32$, osim kod inačice DenseNet-161 kod koje iznosi $k = 48$.

Prednost gusto povezanih blokova nad jednostavnijim rezidualnim blokovima je u tome što favoriziraju rješenja izražena značajkama različite složenosti [35]. Ovakva induktivna pristranost posebno je prikladna za klasifikacijske probleme u kojima su neki vizualni koncepti lakši od drugih (npr. prometni znak vs. autobus). Osim toga DenseNet je i memorijski efikasniji. Mogućnost uštede leži u činjenici da u gusto povezanim blokovima više slojeva dijeli iste ulaze. Naivna implementacija ovakvoga bloka u nekom od okvira za automatsku diferencijaciju pohranit će ulaze svakoga sloja zbog potrebe unatražnog prolaza. Tom prilikom nastat će više



Slika 2.2: Primjer gusto povezanog konvolucijskog bloka - osnovnog gradivnog elementa arhitekture DenseNet. Slika preuzeta iz [29].

kopija istih značajki. Primjerice, identična kopija značajki prvog sloja pojavila bi se u svim konkateniranim tenzorima koji čine ulaze sljedećih slojeva. To je lako izbjeći na način da se te značajke u memoriji pohrane samo jedanput, a prilikom unatražnog prolaza potrebne jednostavne transformacije (npr. normalizacija, konkatenacija ili 1×1 konvolucija) tih značajki izračunaju ponovno. Ponovni izračun usporava proces treniranja za oko 20%, što je cijena koju vrijedi platiti za višestruku uštedu memorije. Memorijski efikasne implementacije DenseNeta [36] dostupne su u mnogim radnim okvirima za računalni vid. Okosnica DenseNet korištena je za memorijski efikasnu semantičku segmentaciju [35] slika visoke rezolucije.

2.2 Rekonstrukcijske tehnike računalnog vida

Računalni vid često se dijeli na raspoznavanje i rekonstrukciju. Raspoznavanje cilja kategoričke predikcije koje obično otkrivaju neko značenje o sceni. Najpoznatiji zadaci raspoznavanja su klasifikacija slika, detekcija objekata, semantička segmentacija i drugi. Rekonstrukcija cilja neke kontinuirane predikcije koje imaju geometrijsku interpretaciju. Pod rekonstrukciju ubrajamo procjenu optičkog toka (gibanje piksela), procjenu dubine (udaljenost piksela od kamere), procjenu gibanja kamere i slično. Nije ideja da se metode raspoznavanja i rekonstrukcije koriste

odvojeno, nego da među njima postoji pozitivna interakcija. Gusto semantičko prognoziranje to svakako zahtjeva. Za ispravnu predikciju budućnosti potrebno je raspoznati što se nalazi u sceni, ali rekonstruirati kako su se objekti ili kamera gibali u prošlosti te procijeniti kako će se scena razviti u budućnosti. Važna rekonstrukcijska tehnika za gusto raspoznavanje svakako je procjena optičkog toka, koja se na poseban način koristi i u sustavu prognoziranja predloženom u ovoj disertaciji.

2.2.1 Optički tok

Optički tok spada u domenu zadataka koji se bave procjenom gibanja u sceni. Cilj optičkog toka je procijeniti gusto 2D-polje vektora pomaka između dva slikovna okvira. Primjenu algoritama za procjenu optičkog toka nalazimo u u različitim postupcima za razumijevanje videa (klasifikacija, prognoziranje, pretraživanje). Srodne metode rekonstrukcije gibanja važne su u vizualnoj odometriji [37], praćenju objekata [38] i dr. Isprva se čini kako bi se problem semantičkog prognoziranja mogao jednostavno riješiti preobrazbom prethodnih predikcija s obzirom na procijenjeni optički tok. Međutim, optički tok rekonstruira gibanje piksela između dva promatrana okvira. Za željenu preobrazbu bilo bi potrebno imati optički tok između trenutnog i budućeg neopažanog okvira, što je za rekonstrukciju nemoguće. Potrebno je dakle prognozirati budući optički tok. Iako izravna primjena za prognoziranje nije moguća, neke ideje iz optičkog toka poput korelacijskog sloja i deformacije tokom moguće je primijeniti i u modelima za prognoziranje značajki.

2.2.2 Korelacijski sloj

U posljednje vrijeme najpreciznji modeli za procjenu optičkog toka zasnovani su na dubokom učenju [23, 39, 40]. Duboki modeli omogućuju treniranje korespondencije s kraja na kraj i imaju mogućnost zamišljanja optičkog toka u zaklonjenim i otkrivenim dijelovima scene. Tipična duboka arhitektura za procjenu optičkog toka uključuje konvolucijsku okosnicu za ekstrakciju značajki i korelacijski sloj.

Korelacijski sloj računa sličnosti među isječcima ulaznih tenzora značajki. Na ulazu prima dva tenzora značajki F_1 i F_2 dimenzija $C \times H \times W$, gdje je C broj kanala, H visina, a W širina. Izlazni tenzor O je dimenzija $H' \times W' \times H \times W$. Jedan element tenzora O_{ijkl} predstavlja mjeru sličnosti između značajki F_1 na lokaciji k,l sa značajkama iz F_2 na lokaciji i,j. Mjera sličnosti zapravo odgovara korelacijskom koeficijentu između isječaka značajki centriranih u odgovarajućim lokacijama. Dimenzije H' i W' ne moraju nužno odgovarati visini i širini ulaznog tenzora, pa tako ni i,j ne mora nužno odgovarati lokaciji isječka u tenzoru F_2. Primjerice, usporedba isječaka može se provesti s korakom (eng. *stride*) različitim od jedan. Globalna usporedba isječaka (svaki sa svakim) je računalno skupa, iako je neke metode provode [40]. Zbog toga

se usporedba može ograničiti na lokalno susjedstvo oko lokacije k, l. U tom slučaju H' i W' odgovaraju veličini lokalnog susjedstva. Dodatno pojednostavljenje moguće je napraviti postavljanjem veličine isječka na 1×1 kada izravno uspoređujemo vektore značajki s C elemenata preuzete iz različitih lokacija ulaznih tenzora.

Korelacijski koeficijent određen je kao skalarni produkt između izravnatih vektora dvaju isječaka značajki. Ako su vektori normalizirani, onda skalarni produkt odgovara mjeri kosinusne sličnosti. Izračun korelacijskih koeficijenata može se promatrati kao poseban slučaj konvolucije: filter odnosno konvolucijska jezgra je isječak značajki iz F_1 kojeg pomičemo po različitim lokacijama ulaznog tenzora F_2. To nas podsjeća na činjenicu da korelacijski sloj zapravo nema parametre koji se uče. Korelacijski sloj dio je predloženog korelacijskog modula za prognoziranje značajki koji je pobliže opisan u poglavlju 5.

2.2.3 Unaprijedna i unatražna deformacija

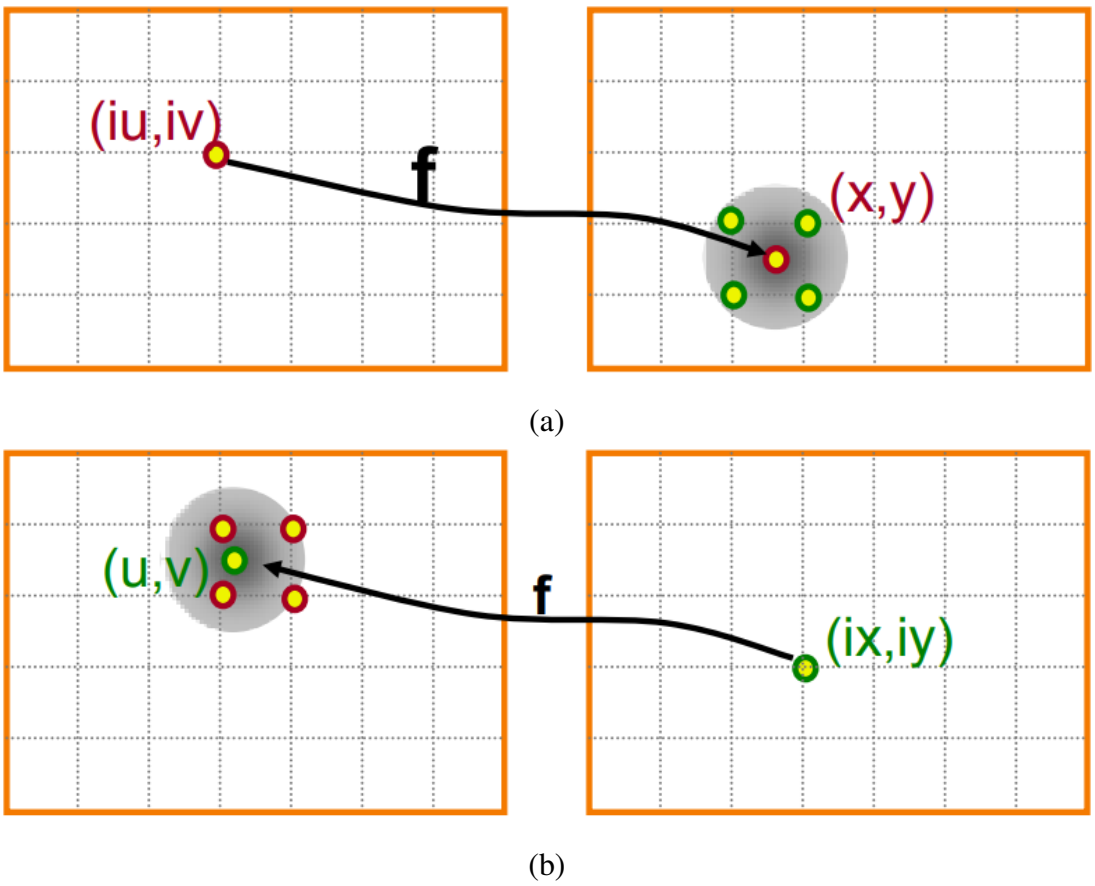
Optički tok između para slika moguće je iskoristiti za procjenu jedne od njih deformiranjem druge. To se često koristi prilikom nenadziranog učenja optičkog toka [41] u videu. Jedan korak učenja započinje procjenom optičkog toka između dvaju ulaznih slika dubokim modelom. Zatim se tim tokom deformira odgovarajuća ulazna (izvorišna) sliku. Gubitak modela tada je funkcija razlike (npr. MSE) između preostale ulazne (odredišne) slike i procijenjene slike dobivene deformiranjem. Napredniji postupci na temelju toka procjenjuju i mapu zaklanjanja [42]. Ona se koristi za relaksaciju funkcije gubitka u dijelovima scene koji nemaju korespondenciju zbog zaklanjanja ili otkrivanja.

Optički tok može biti definiran u unaprijednom ili unatražnom smjeru. Iako se optički tok može odrediti i između dvaju slika nastalih u istom trenutku (npr. lijeva i desna slika u stereoskopskom paru), u ovome razmatranju vrijedit će pretpostavka da su na ulazu dvije slike I_t i I_{t+1} nastale u trenutcima t i $t+1$. Vektori unaprijednog optičkog toka $\mathbf{f}_t^{t+1} = \mathrm{flow}(\mathrm{I}_t, \mathrm{I}_{t+1})$ izviru iz lokacije piksela u trenutku t te završavaju u trenutku $t+1$. Za unatražni tok $\mathbf{f}_{t+1}^{t} = \mathrm{flow}(\mathrm{I}_{t+1}, \mathrm{I}_t)$ vrijedi obrnuto. Budući slikovni okvir I_{t+1} može biti aproksimiran unaprijednom deformacijom [44] prethodnog okvira I_t unaprijednim tokom $\mathbf{f}_t^{t+1} = \mathrm{flow}(\mathrm{I}_t, \mathrm{I}_{t+1})$, ili unatražnom deformacijom okvira I_t i unatražnim tokom $\mathbf{f}_{t+1}^{t} = \mathrm{flow}(\mathrm{I}_{t+1}, \mathrm{I}_t)$:

$$\mathrm{I}_{t+1} \approx \mathrm{warp\_fw}(\mathrm{I}_t, \mathbf{f}_t^{t+1}) \approx \mathrm{warp\_bw}(\mathrm{I}_t, \mathbf{f}_{t+1}^{t}) \tag{2.1}$$

Približna jednakost u (2.1) podsjeća nas da bijektivno mapiranje između dva sukcesivna okvira često ne može biti napravljeno zbog zaklanjanja ili otkrivanja i promjena u perspektivi.

Razlika između ovih dviju deformacija nije zanemariva. U oba slučaja problemi nastaju zbog diskretne slikovne rešetke i nediskretnih vektora pomaka. Unaprijedna deformacija u takvim slučajevima razmazuje jedan izvorišni piksel na više odredišnih piksela. Uslijed toga mogu



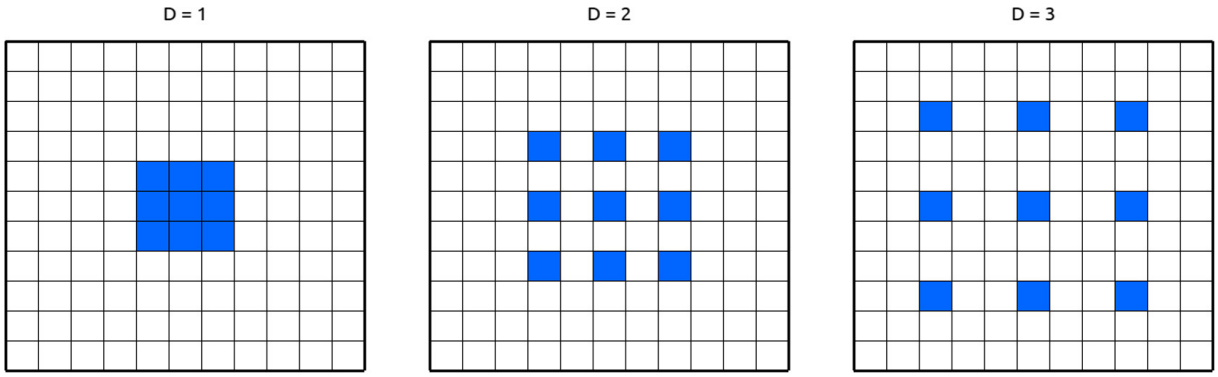
Slika 2.3: Ilustracija dvaju pristupa deformiranju: (a) unaprijedni i (b) unatražni. Na obje slike se lijevo nalazi izvorišna slika, a desno odredišna. Slika preuzeta iz [43].

nastati neželjeni efekti poput zamućenosti i neprikladnog uzorkovanja. Kod unaprijedne deformacije dodatno postoji mogućnost nastanka rupa u odredišnoj slici. Uz to postoji i problem višestrukih pogodaka iste odredišne lokacije. Unatražna deformacija nema problem s razmazivanjem ili višestrukim pogocima jer su vektori toka poravnati sa rešetkom odredišne slike, ali postoji problem s odabirom piksela u izvorišnoj slici. U slučaju nediskretnih pomaka, vrijednost odredišnog piksela se određuje bilinearnom interpolacijom četiri okolna piksela u izvorišnoj slici. Slika 2.3 ilustrira dvije navedene vrste deformiranja, te različite probleme rješavanja nediskretnih pomaka. U praksi se češće koristi unatražna deformacija, ali postoje i problemi u kojima su svojstva unaprijedne deformacije poželjna.

2.3 Deformabilne konvolucije

Jedna jezgra klasičnog 2D konvolucijskog sloja u računalnom vidu ima oblik kvadra. Dubinu toga kvadra određuje broj kanala ulaza, dok prostorne dimenzije jezgre - širinu i visinu zadaje korisnik. Prostorne dimenzije jezgri unutar modela važan su hiperparameter jer izravno utječu na receptivno polje modela. Modeli sa nedovoljnom veličinom receptivnog polja ne mogu is-

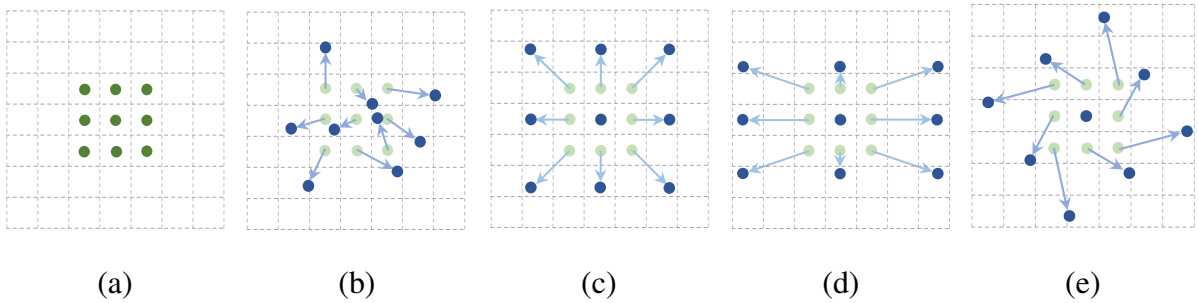
pravno raspoznavati velike objekte. Međutim, konvolucije sa velikim dimenzijama jezgre imaju više parametara i veliku računsku složenost. Jasno je da je odabir dimenzije jezgre delikatan i mora optimirati više kriterija. Povećanje receptivnog polja bez povećanja broja parametara moguće je napraviti dilatacijom jezgre. Dilatacija prostorno proširuje standardnu jezgru umetanjem jedne ili više praznina između elemenata jezgre ovisno o faktoru dilatacije. Slika 2.4 prikazuje konvolucijsku jezgru 3×3 s jednim kanalom na ulazu, dilatiranu s faktorima dilatacije $1, 2$ i 3. Iako povećava receptivno polje, dilatacija i dalje uzorkuje ulaze na pravilnoj rešetci. To limitira mogućnost modela za obavljanje kompleksnijih geometrijskih transformacija. Ovi nedostatci regularnih konvolucija bili su glavna motivacija za otkriće deformabilnih konvolucija [45].



Slika 2.4: Ilustracija utjecaja dilatacijskog faktora na konvolucijsku jezgre dimenzija 3×3. Slika preuzeta iz [46].

Deformabilne konvolucije nemaju predodređen oblik rešetke za uzorkovanje, nego je on prilagođen sadržaju ulaznog tenzora. Lokacija uzorkovanja za svaki prostorni element jezgre pomaknuta je od svojega podrazumijevanog mjesta na diskretnoj rešetci za predviđeni nediskretni pomak. Slika 2.5 ilustrira različite mogućnosti deformacije za konvolucijsku jezgru 3×3. Zeleni i plavi krugovi prikazuju lokacije uzorkovanja za običnu i deformiranu konvoluciju, a plave strelice predviđene pomake, neovisne za svaki element jezgre. Slika (a) prikazuje standardnu konvolucijsku jezgru, a slika (b) deformiranu konvolucijsku jezgru sa potpuno slobodnim pomacima. Preostale slike zapravo su specijalni slučajevi od (b), što pokazuje mogućnost da se deformabilnom konvolucijom modeliraju i neka pravilnija uzorkovanja poput dilatacije.

Važno je naglasiti da su prostorni pomaci lokacija uzorkovanja regresirani na temelju prethodne reprezentacije dubokog modela i da ovise o položaju konvolucijske jezgre u ulaznom tenzoru. Moguće je, dakle, da se u jednom prolazu konvolucijske jezgre po ulaznoj mapi značajki, na nekim lokacijama provede regularna konvolucija, na drugima dilatirana, a na trećima proizvoljno izobličena. Predviđene pomake na temelju ulaznog tenzora računa dodatni konvolucijski sloj koji prethodi deformabilnom prolazu. U standardnom slučaju, konvolucijski sloj za pomake ima jednaku veličinu jezgre (nije nužno) kao i deformabilna konvolucija $k \times k$, a broj izlaznih mapa značajki jednak je $2 \cdot k \cdot k$. Dakle, izlazni tenzor na lokaciji i, j sadrži vektor s



(a) (b) (c) (d) (e)

Slika 2.5: Ilustracija lokacija uzorkovanja kod deformabilnih konvolucija. Slika preuzeta iz [45].

$2 \cdot k \cdot k$ elemenata koji predstavljaju pomak po x i y osi (op.a. 2) za svaki element jezgre kojih ukupno ima $k \cdot k$.

Prilikom deformabilnog prolaza, potrebno je dakle na svakom položaju jezgre očitati predviđeni pomak te u skladu s tim uzorkovati ulazni tenzor, a zatim provesti uobičajenu konvoluciju nad uzorkovanim vrijednostima. Uzorkovanje ulaznih vrijednosti na nediskretnim lokacijama izvedeno je bilinearnom interpolacijom, slično kao kod unatražne deformacije optičkim tokom. U praksi se koristi posebna inačica algoritma *im2col*, koja prilikom preslagivanja ulaza u obzir uzima predviđene offsete i provodi opisano uzorkovanje [47].

Deformabilne konvolucije omogućuju prilagodbu rešetke uzorkovanja sadržaju slike unutar jednoga sloja. Primjerice, ako model sluti da se na nekoj lokaciji nalazi veliki objekt poput kamiona, deformabilna konvolucija ima priliku tu lokaciju pogledati sa velikim receptivnim poljem, i obratno. U ovoj disertaciji, između ostalog, ispituje se i prikladnost deformabilnih konvolucija za problem prognoziranja konvolucijskih značajki. Hipoteza je da bi, slično kao u slučaju različitih veličina objekata, deformabilne konvolucije mogle modelirati gibanja različitih objekata unutar jednoga sloja. Detalji primjene deformabilnih konvolucija u modelu F2MF za prognoziranje značajki opisani su u poglavlju 5, a eksperimentalni rezultati njihovog utjecaja na točnost prognoziranja u poglavlju 6.

Poglavlje 3

Pregled literature

Ovo poglavlje donosi pregled literature tematski povezane sa semantičkim prognoziranjem budućnosti. Predložena metoda prognoziranja F2MF najviše je povezana sa prethodnim radovima na temu gustog semantičkog prognoziranja, i to metodom prognoziranja pomaka u pomak (M2M, (eng. *motion to motion*)) [48] te metodom prognoziranja iz značajki u značajke (F2F, (eng. *feautures to features*) [6]. Osim radova koji se bave gustim semantičkim prognoziranjem, dodirnih točaka sa temom ove disertacije imaju i radovi koji obrađuju prognoziranja RGB piksela budućeg okvira [49] ili efikasnog raspoznavanja u videu [50]. Za našu metodu također su važni i duboki modeli za gustu predikciju u jednoj slici jer ih koristimo kao integralne dijelove sustava za semantičko prognoziranje. Ovaj pregled započinjemo opisom recentnih pristupa zadacima gustog raspoznavanja poput semantičke segmentacije, segmentacije instanci i panoptičke segmentacije.

3.1 Duboki modeli za gusto raspoznavanje u jednoj slici

Gusto raspoznavanje podrazumijeva klasifikaciju na razini piksela u jedan od semantičkih razreda. Semantičke razrede dijelimo na prebrojive (eng. *things*) poput čovjeka, automobila, kamiona, i neprebrojive (eng. *stuff*) poput ceste, neba, zgrade i sl. Za prebrojive razrede postoje oznake koje omogućuju razlikovanje instanci istoga razreda. Skup prebrojivih razreda određen je standardom označavanja koji vrijedi za konkretni skup podataka. Nije rijetkost da se razredi koji označavaju iste vizualne koncepte u nekim podatkovnim skupovima smatraju prebrojivima, a u nekima neprebrojivima. Neki se razredi, iako su u stvarnosti prebrojivi, gotovo u pravilu svrstavaju u neprebrojive. Razlog za to je jeftinije označavanje i ograničena praktična korist u raspoznavanju primjeraka toga razreda. Kao očiti primjer takve prakse ističe se razred koji označava zgradu (građevinu), koja se u više skupova sa prometnim scenama [8, 51, 52] svrstava u neprebrojive razrede.

Nekoliko zadataka računalnog vida bavi se razumijevanjem scene na razini piksela. Se-

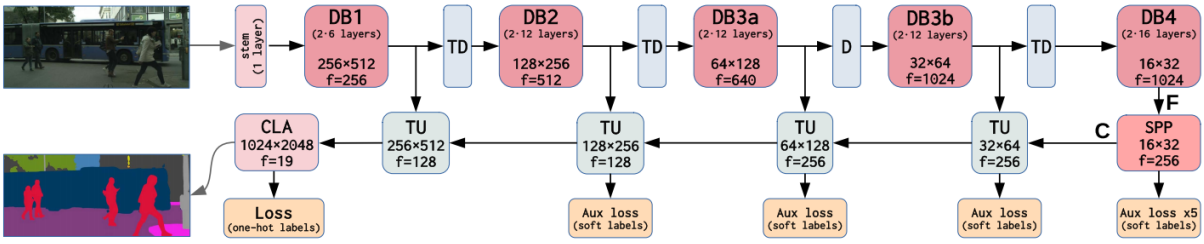
mantička segmentacija [53] svakome pikselu ulazne slike pridjeljuje jedan semantički razred. Ona ne zahtijeva razlikovanje primjeraka istog razreda. Segmentacija instanci [20] detektira instance prebrojivih razreda i povezuje ih s odgovarajućim dijelovima slike. Panoptička segmentacija [54] objedinjuje ta dva zadatka jer svakom pikselu pridjeljuje semantičku kategoriju i indeks primjerka. Svi pikseli koji pripadaju neprebrojivim razredima imaju isti indeks primjerka. Raspoznavanje prebrojivih razreda kod segmentacije instanci i panoptičke segmentacije nije istovjetno. Razlika je u tome što zadatak segmentacije instanci dopušta preklapanja maski između različitih instanci, dok kod panoptike to nije slučaj.

Donedavno je dominantan pristup rješavanju navedenih zadataka bila primjena potpuno konvolucijskih dubokih modela. Adaptacijom transformerskih arhitektura za klasifikaciju slika [55] situacija se počela mijenjati i u zadacima gustog raspoznavanja. Na nekim skupovima podataka najbolje rezultate postižu transformerske arhitekture [56], ili pak hibridne arhitekture koje kombiniraju konvolucijske modele s transformerskim slojevima [57]. U ovoj disertaciji korišteni su potpuno konvolucijski modeli za gusto raspoznavanje u jednoj slici. U nastavku teksta nalazi se kratak pregled najvažnijih konvolucijskih arhitektura za gusto raspoznavanje.

Tipična moderna konvolucijska arhitektura za gusto raspoznavanje sastoji se od kodera i dekodera. Koder obično čini prilagođeni klasifikacijski model predtreniran na ImageNetu (okosnica). Prilagodba klasifikacijskog modela za gustu predikciju obično uključuje odbacivanje posljednjeg potpunog povezanog sloja i globalnog sažimanja. Takav koder proizvodi komprimiranu i sažetu apstraktnu reprezentaciju slike čija je prostorna rezolucija nekoliko puta poduzorkovana. Tu reprezentaciju dekoder prevodi u guste predikcije na finoj rezoluciji ulazne slike. Ovakve arhitekture često slijede oblik pješčanog sata jer se rezolucija skrivene reprezentacije kroz koder postupno smanjuje, a kroz dekoder postupno povećava. Korištenje predtreniranih klasifikacijskih arhitektura na ImageNetu drastično smanjuje potrebu za velikom količinom slika sa gustim oznakama [30, 58], ali sa sobom nosi i neke izazove. Za točnu gustu predikciju potrebno je rekonstruirati detalje o prostornom razmještaju značajki koji su izgubljeni zbog poduzorkovanja. Dodatno, zbog predtreniranja na ImageNet slikama manje rezolucije, modeli imaju ograničeno efektivno receptivno polje [59]. Dizajn konvolucijskih arhitektura za gustu predikciju usmjeren je na rješavanje dva prethodno navedena problema.

Preteča modernih konvolucijskih arhitektura za semantičku segmentaciju, a i ostale zadatke gustog raspoznavanja je FCN (eng. *fully convolutional network*). FCN minimalno adaptira konvolucijske klasifikacijske modele [17, 25, 60] za gustu predikciju [19] i dodaje put naduzorkovanja koji se uči s ostatkom modela. Adaptacija se sastoji od zamjene potpuno povezanih slojeva, tipičnih za klasifikaciju, konvolucijskima. Rezolucija predikcija se postupno oporavlja kroz više modula za naduzorkovanje temeljenih na transponiranim konvolucijama. Model za semantičku segmentaciju, DeepLab [61], izbacuje korak konvolucijskih slojeva u posljednjim procesnim blokovima okosnice kako bi se reduciralo poduzorkovanje. Taj pristup nije efikasan

jer povećava količinu procesiranja na finoj rezoluciji i zahtijeva uvođenje dilatiranih konvolucija. PSPNet [62] se temelji na istoj ideji, ali dodaje modul za piramidalno sažimanje koji povećava receptivno polje i značajke obogaćuje globalnim kontekstom. Efikasniji pristupi usmjereni na precizniju rekonstrukciju prostornih detalja temeljeni su na ideji preskočnih veza između kodera i dekodera. Ideja se prvi put pojavila kod autoenkodera [63], model specijaliziran za obradu medicinskih slika U-Net [64] popularizirao ju je u zadacima gustog raspoznavanja. U-Net ima simetričnu koder-dekoder arhitekturu u kojoj su izlazi procesnih blokova na istim rezolucijama kodera i dekodera povezani preskočnim vezama. Shematski prikaz ovakvih modela često podsjeća na ljestve, pa se često nazivaju modelima sa ljestvičastim naduzorkovanjem. Preskočne veze omogućuju miješanje ranih prostorno preciznih reprezentacija sa kasnijim semantički bogatim reprezentacijama. Na taj način se pospješuje ispravna lokalizacija značajki u putu za naduzorkovanje. Daljnji razvoj te ideje uključivao je korištenje kapacitirane predtrenirane okosnice i lakog puta naduzorkovanja [33, 35]. Uspješnost ovakvog dizajna upućuje na činjenicu da raspoznavanje zahtjeva puno više kapaciteta od određivanja granica segmenata kada je njihova semantika otprilike poznata. Ista ideja još uvijek je prisutna prilikom oblikovanja učinkovitih modela za semantičku segmentaciju [34, 65]. Inačica takve arhitekture [33] producira piramidu značajki koja pospješuje detekciju objekata različitih veličina. Najpoznatija inačica modela Mask R-CNN [20] za segmentaciju instanci također koristi piramidu značajki nastalu ljestvičastim naduzorkovanjem. Model LDN [35], prikazan na slici 3.1, baziran je na okosnici DenseNet i modulu za piramidalno sažimanje prije ljestvičastog naduzorkovanja. Ovakva arhitektura postiže povoljan omjer efikasnosti i točnosti. Sličnu ideju slijedi i jednorazinska varijanta modela SwiftNet [30]. Model za panoptičku segmentaciju Panoptic Deeplab [21] ima dvije odvojene instance ljestvičastog naduzorkovanja. Jednu za semantičku segmentaciju, a drugu za segmentaciju instanci bez određivanja semantičkih kategorija.



Slika 3.1: Model za semantičku segmentaciju sa ljestvičastim naduzorkovanjem i okosnicom DenseNet. Slika preuzeta iz [35].

3.2 Duboko učenje u videu

Duboko učenje u videu dotaknulo je širok spektar zadataka poput: raspoznavanja akcija [66], klasifikacije videa [67], segmentacije objekata u videu [68], interpolacije okvira [69] a na

posljetku i prognoziranja [5, 70]. Osim zadataka usmjerenih na samu obradu, video je korišten i u svrhu nenadziranog učenja dubokih modela za ekstrakciju kvalitetnijih reprezentacija slika/videa [71, 72]. Postojanje vremenske dimenzije i redoslijeda okvira u videu može se iskoristiti kao besplatan signal za učenje. Jedan od načina je da model optimira neki surogat zadatak, poput vremenskog sortiranja slučajno ispremiješane sekvence slikovnih okvira [71].

Najrelevantnija disciplina iz obrade videa za ovu disertaciju je svakako prognoziranje, međutim neki slični koncepti pojavljuju se i u radovima koji se bave drugim zadacima. Radovima iz područja prognoziranja posvećeno je posebno potpoglavlje 3.3, a sljedeći paragraf posvećen je ostalim radovima iz obrade videa, povezanih s predloženom metodom prognoziranja F2MF.

U modelima za obradu videa česta je upotreba optičkog toka i operacije prostornog deformiranja. Moguće je uvećati skup podataka za semantičku segmentaciju propagacijom gustih oznaka na susjedne okvire u videu uz pomoć deformacije optičkim tokom [73]. Modeli za efikasnu gustu predikciju u videu koriste propagaciju značajki deformacijom iz prethodnih ključnih okvira u trenutni [50]. Propagacija značajki iz prošlosti u trenutnu reprezentaciju koristi se i za poticanje vremenske konzistencije gustih predikcija u videu [74]. Posebna varijanta unaprijedne deformacije korištena je za interpolaciju okvira u videu [69]. Njihov pristup razrješava višestruke pogotke u odredišnoj slici na temelju zadanog kriterija (npr. dispariteta u izvorišnoj slici).

3.3 Prognoziranje budućnosti u videu

Tema prognoziranja u računalnom vidu poznata je dugo vremena [3, 66, 75]. Pod prognoziranjem podrazumijevamo sve zadatke u kojima je cilj na temelju jedne ili više slika donijeti zaključak o budućnosti. Buduće predikcije mogu biti niske razine apstrakcije poput intenziteta RGB piksela budućeg okvira [70] ili visoke razine apstrakcije poput akcija [66], trajektorija [76] ili panoptičke segmentacije [7]. U nastavku su posebno obrađene metode za prognoziranje budućih slikovnih okvira i gusto semantičko prognoziranje.

3.3.1 Prognoziranje budućih slikovnih okvira

Prognoziranje budućih slikovnih okvira podrazumijeva ispravnu predikciju jednog ili više budućih okvira na razini RGB piksela na temelju promatranih okvira iz prošlosti [4]. Zbog toga se zadatak često zove i samo RGB prognoziranje, a u nekim radovima [70] korišten je i termin predikcija videa (eng. *video prediction*). Zadatak je iznimno popularan [70, 77, 78, 79, 80, 81] zbog prilike za samonadzirano učenje reprezentacija na praktički neograničenim podacima. U odnosu na općenito generiranje slika, zadatak je zahtjevniji jer prediktirani okvir osim što mora biti vizualno ugodan i jasan, također mora očuvati vremensku konzistentnost u videu. RGB

prognoziranje slično je interpolaciji u videu, međutim interpolacija je lakši zadatak jer su poznati i prethodnik i sljedbenik prediktiranog okvira. Kvaliteta RGB predikcija opisuje se i raznim kvantitativnim metrikama. Neke mjere uspoređuju generirani i stvarni slikovni okvir u RGB prostoru kao što su PSNR (eng. *peak signal-to-noise ratio*) i SSIM (eng. *structural similarity index measure*), a neke u prostoru značajki predtreniranog dubokog modela kao npr. LPIPS [82] i FVD [83]. Metrike koje djeluju u prostoru značajki uspoređuju statistike stvarnog i umjetnog skupa podataka (slika ili videa), kako bi ocijenili kvalitetu generiranih podataka. Takve metrike su se pokazale bližim ljudskoj procjeni kvalitete generiranih slika/videa.

Više je različitih pristupa zadatku RGB prognoziranja. Mathieu et al. [70] kombiniraju višerazinsku arhitekturu sa suparničkim učenjem kako bi smanjili zamućenje u predikcijama - čest artefakt korištenja klasičnog gubitka najmanjih kvadrata. Vukotić et al. [77] prognoziraju proizvoljno daleko u budućnost u jednom koraku ugrađivanjem željenog pomaka u latentnu reprezentaciju konvolucijskog dekodera. Kalchbrenner et al. [84] koriste autoregresivne modele kako bi riješili faktorizaciju izglednosti budućeg slikovnog okvira po vremenskoj i prostornoj dimenziji. Reda et al. [78] prognoziraju budući okvir deformacijom prošlih okvira primjenom prediktiranih jezgri na lokacijama pomaknutima u skladu s prognoziranim tokom. Koncept primjene jezgri na pomaknutim lokacijama sličan je deformabilnim konvolucijama. Naš predloženi modul F2M prognozira na sličan način deformiranjem iz više prošlih reprezentacija. Problem s pristupom iz [78] je u smanjenoj mogućnosti zamišljanja u novootkrivenim dijelovima scene, dok je to u našem pristupu omogućeno paralelnim prognoziranjem modulom F2F. Može se reći da naš kombinirani modul F2MF dekomponira prognoziranja na rekonstrukciju iz prošlosti (F2M) i umetanje novonastalih dijelova scene (F2F). Slična ideja prisutna je i u RGB prognoziranju [49, 79, 85]. Hao et al. [85] uz haluciniranu verziju budućeg okvira predviđaju i težine koje služe za linearnu kombinaciju haluciniranog okvira i okvira dobivenog deformiranjem iz prošlosti. Drugi imaju poseban model za ucrtavanje koji na ulazu prima okvir prognoziran deformiranjem i prikladni tok ili mapu okluzija koja signalizira regije koje je potrebno ucrtati [49, 79]. Osim očigledne razlike u ciljevima prognoziranja (RGB slika vs. semantičke predikcije) između navedenih pristupa i našega rada, postoje još neke. Naš F2M modul koristi deformabilne konvolucije koje imaju pristup prostorno-vremenskim korelacijskim značajkama. Dodatno, prognozirane značajke su kombinacija deformacija iz više trenutaka u prošlosti. Važna činjenica je i to da se procesiranje provodi na značajkama koje su značajno poduzorkovane u odnosu na sliku. Zbog toga, naša metoda može prognozirati semantičke predikcija na velikim rezolucijama uz minimalni dodatni računalni trošak u odnosu na predikciju u jednoj slici. Posebna skupina radova u RGB prognoziranju bavi se generiranjem videa iz jedne slike [79, 80]. Pan et al. [80] razmatraju posebnu varijantu zadatka u kojem ulaz u model čini jedna segmentacijska mapa, koja se prvo prevodi u odgovarajuću sliku, a onda se iz te slike generira video. Ovi modeli također koriste deformiranje optičkim tokom za prognoziranje

budućih okvira. Neki radovi prediktiraju više tokova iz izvornog okvira u sve buduće okvire prognoziranog videa [79, 80]. Ovo je na neki način suprotno onome što radi naš modul F2M. Mi također generiramo više tokova za sve slike iz prošlosti, ali svi ciljaju isti budući trenutak. Ovo omogućuje našem modelu da razriješi neka otklanjanja birajući pravi trenutak iz kojega će deformirati značajke u budućnost.

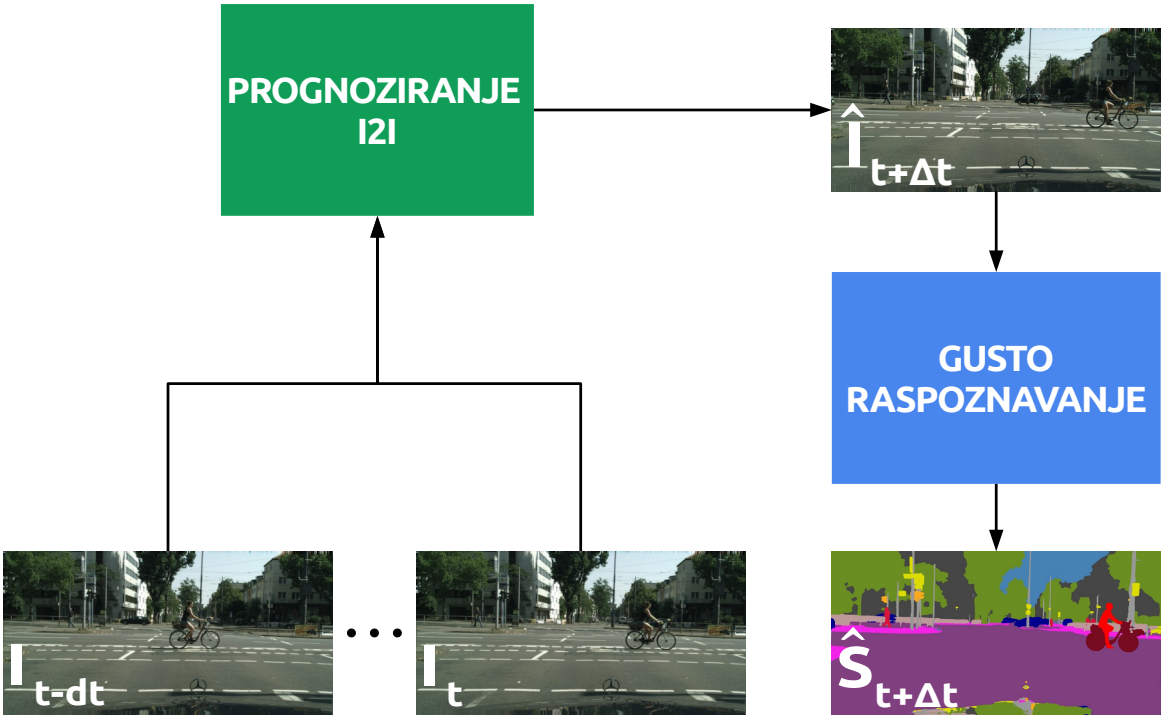
3.3.2 Gusto semantičko prognoziranje

Modeli za gusto semantičko prognoziranje na temelju slika iz prošlosti predviđaju budućnost u vidu semantičkih oznaka na razini piksela. U ovome pregledu literature posebna pažnja dana je metodama čije buduće predikcije odgovaraju semantičkoj segmentaciji, segmentaciji instanci ili panoptičkoj segmentaciji budućeg okvira. Središnji dio metoda za semantičko prognoziranje je prognostički model čija je zadaća prevesti zadanu reprezentaciju iz prošlosti u budućnost. U ovisnosti o vrsti te ulazne odnosno izlazne reprezentacije razlikujemo metode čiji modeli prognoziraju:

- iz slike u sliku, skraćeno I2I (eng. *image to image*),
- iz semantičkih predikcija u semantičke predikcije, skraćeno S2S (eng. *semantics to semantics*),
- iz optičkog toka u optički tok (ili iz pomaka u pomak), skraćemo M2M (eng. *motion to motion*)
- te iz značajki u značajke, skraćeno F2F (eng. *features to features*).

U nastavku se nalazi pregled metoda razvrstanih prema prethodno navedenoj podjeli, a svakome pristupu posvećen je zaseban odlomak.

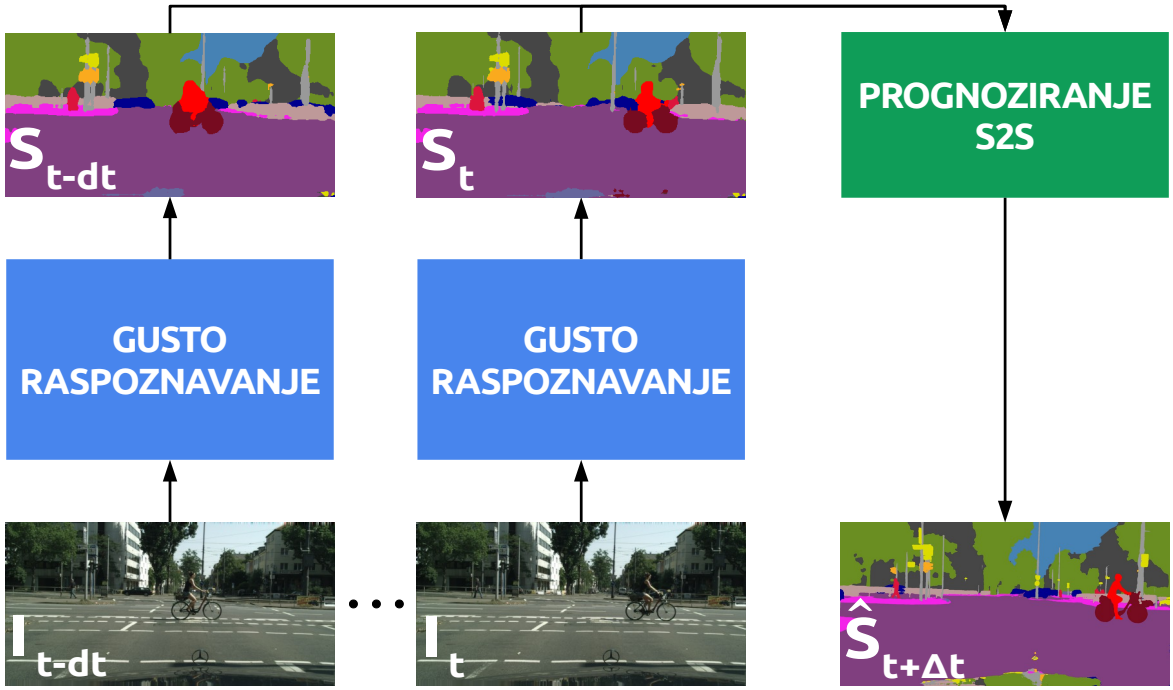
Razmatrajući problem semantičkog prognoziranja, prirodno se nameće ideja da se nekako iskoriste već razvijene metode za RGB prognoziranje iz poglavlja 3.3.1. Primjerice, moguće je prvo promatrajući prošle slikovne okvire prognozirati izgled budućeg, a zatim provesti semantičku segmentaciju prognoziranog okvira prikladnom predtreniranom metodom. Takvi pristupi prognoziraju, dakle, iz slike u sliku i skraćeno ih zovemo I2I (eng. *image to image*). Shematski prikaz takvog sustava prikazan je na slici 3.2. Slika pokazuje dva potrebna koraka za dobivanje gustih semantičkih predikcija budućeg okvira. Prvo, model za RGB prognoziranje (zeleni pravokutnik) predviđa budući slikovni okvir na temelju okvira iz prošlosti. Drugo, model za semantičku segmentaciju (plavi pravokutnik) prima prognozirani okvir i proizvodi semantičke predikcije u budućnosti. Problem ovih pristupa je u tome što se RGB prognoziranje slika visoke rezolucije pokazalo kao težak problem [70]. Primjenom takvih modela za prognoziranje, velika je vjerojatnost za propagaciju pogreške iz RGB prognoziranja u konačne semantičke predikcije budućnosti. Dodatno, upravljačke module većinom zanima semantika buduće scene, a ne doslovan izgled. Ilustrativno govoreći, upravljačkom sustavu je važnija informacija o položaju pješaka u sceni od točne boje njegove odjeće. Zbog toga sustav za percepciju u dodatnom



Slika 3.2: Semantičko prognoziranje bazirano na prognoziranju iz slike u sliku (I2I).

koraku mora obraditi prognozirani slikovni okvir. Iz te perspektive, prognoziranje izgleda budućeg slikovnog okvira čini se nepotrebnim. Štoviše, intuicija nas navodi da bi prognoziranje na većoj razini apstrakcije možda trebalo prethoditi RGB prognoziranju. U tom pristupu prvo bismo odredili položaje i semantičku kategoriju objekata, a zatim detalje izgleda poput tekstura, osvjetljenja ili refleksije.

Takvu motivaciju slijede pristupi koji izravno prognoziraju semantiku budućeg okvira promatrajući semantičke predikcije u prošlosti, poznati pod kraticom S2S (eng. *semantics to semantics*). Slika 3.3 prikazuje dijagram zaključivanja sustava temeljenog na S2S prognoziranju. Slike iz prošlosti se prvo prevode u odgovarajuće semantičke mape primjenom modela za gustu predikciju (plavi pravokutnik) u svakoj od slika posebno. Zatim, prognostički model prevodi semantičke mape iz prošlosti u odgovarajuće semantičke predikcije iz budućnosti. Izravno prognoziranje na semantičkoj razini efikasnije je i točnije od prognoziranja I2I. Eksperimenti iz [5] idu u prilog tome. Pokazalo se da model koji izravno prognozira semantičku segmentaciju na temelju segmentacije prethodnih slika postiže bolju točnost od modela koji prvo prognozira slikovni okvir, a onda ga segmentira. Prvi pristup gustom semantičkom prognoziranju metodom S2S mapirao je semantičku segmentaciju prošlih okvira u buduću semantičku segmentaciju [5]. Za semantičku segmentaciju jedne slike korišten je Dilation10 [86], dok je prognostički model izražen kao višerazinski model s dilatiranim konvolucijama. Kao funkcija gubitka korištena je mjera L1, kao i varijante gradijentnog i suparničkog gubitka. Jin et al. [87] pospješuju točnost

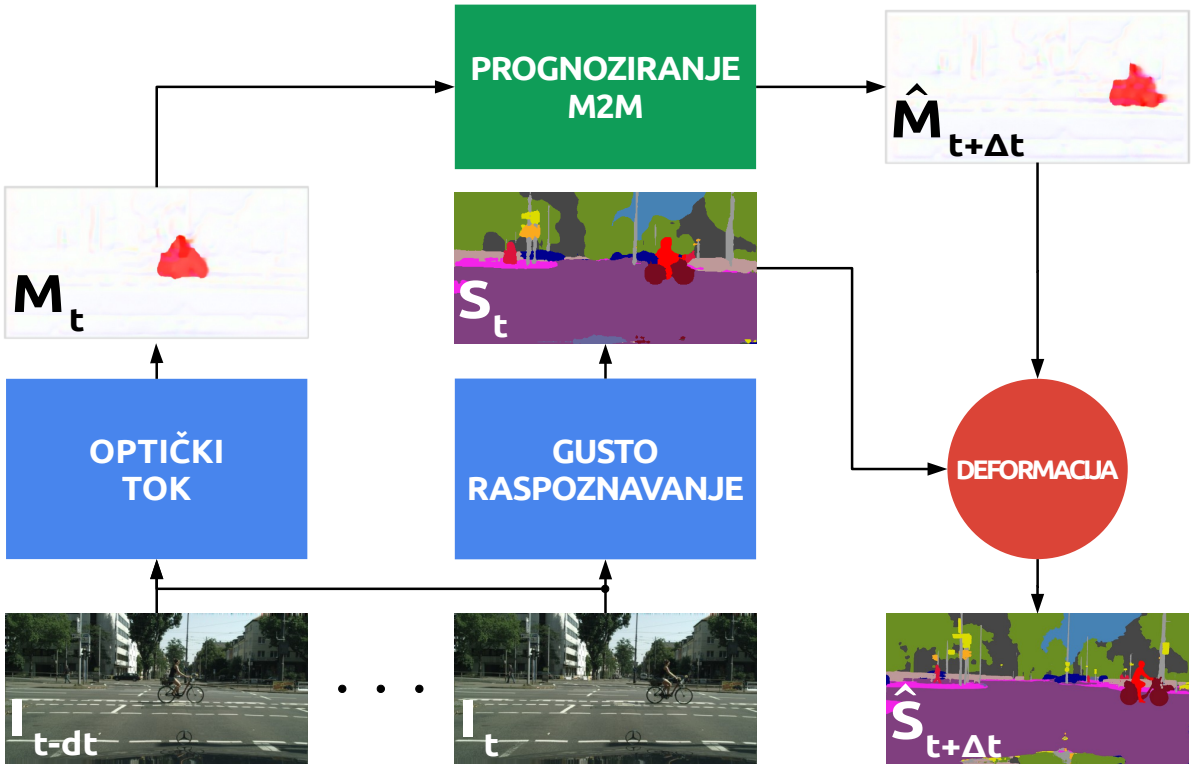


Slika 3.3: Skica sustava za prognoziranje iz semantičkih predikcija u semantičke predikcije (S2S).

buduće semantičke segmentacije paralelnim prognoziranjem budućeg optičkog toka i dijeljenjem značajki između dva modula. Rochan et al. [88] koriste ljestvičasti model za prognoziranje s ugrađenim povratnim ćelijama u preskočnim vezama što omogućuje prognoziranje prozivoljnog broja koraka unaprijed. Chen et al. [89] unaprjeđuju njihov rad korištenjem deformabilnih konvolucija i slojeva pažnje. Neki radovi sa S2S prognoziranjem su se fokusirali na multimodalnu prirodu budućnosti [90, 91]. Naime, budućnost je nepredvidiva te iz jednog te istog skupa događaja u prošlosti može nastati beskonačno mnogo različitih događaja u budućnosti. Neki od njih su vjerojatniji od drugih, te bi, idealno, sustav za prognoziranje uz predikciju više različitih ishoda, svakome od njih trebao dodijeliti i izglednost. Bhattacharyya et al. [90] koriste varijacijsko zaključivanje bazirano na Monte-Carlo isključivanju neurona, kako bi predvidjeli više mogućih semantičkih segmentacija u budućem trenutku. Makansi et al. [91] multimodalno prognoziraju buduće lokacije objekata u sceni i njihove opisujuće pravokutnike. Osim toga, njihova metoda prepoznaje i regije scene u kojima bi se iznenada mogao pojaviti neki do sada nedetektirani objekt i time stvoriti opasnu situaciju. Yao et al. [92] prognozira buduće gibanje kamere što pospješuje točnost lokalizacije objekata u budućnosti. Graber et al. [7] odvojeno prognoziraju prebrojive i neprebrojive razrede kako bi dobili buduću panoptičku segmentaciju. Prognoziranje neprebrojivih razreda sastoji se od procjene parametara 3D transformacije kojom se predikcije deformiraju iz prošlosti u budućnost. Prognoziranje prebrojivih razreda sastoji se od uspostave vremenske korespondencije među detekcijama korištenjem algoritma praćenja

[93] i prognoziranja budućih lokacija svakog objekta. Naš pristup prognoziranju panoptičke segmentacije je jednostavniji. Uspostavu korespondencije implicitno provodi model za prognoziranje značajki, a iz prognoziranih značajki se rekonstruiraju predikcije i za prebrojive i za neprebrojive razrede. Dodatno, naš prognostički model razmatra samo značajke iz prošlosti, dok metoda iz [7] koristi i informacije o odometriji i dubini scene. Odabir gotovih semantičkih predikcija iz prošlosti za ulaz prognostičkog modela donosi posebne izazove. Jedan od tih rizika je propagacija pogrešaka modela za semantičku obradu jednog slikovnog okvira. Izlazi takvih modela su (približno) diskretni i pogreške prilikom donošenja tih odluka propagiraju se i kroz prognoziranje, kada više nema mogućnosti oporavka. Guste predikcije na ulazu stvaraju probleme i s obzirom na računsku složenost postupka. Prostorna rezolucija predikcija je velika jer mora odgovarati rezoluciji ulazne slike. Procesiranje na velikoj rezoluciji je računalno zahtjevno i takvi modeli nisu prikladni za primjenu u stvarnom vremenu. Još jedan od problema je otežana uspostava korespondencije u vremenskoj dimenziji. Naime, za prognoziranje je nužno, barem implicitno, procijeniti gibanje i položaj objekata u različitim vremenskim trenucima. Zato je potrebno prepoznati isti objekt kroz vrijeme, odnosno uspostaviti korespondenciju. To može biti posebno teško napraviti isključivo na temelju prošlih semantičkih predikcija. Primjerice, cilj semantičke segmentacije je prepoznati sve aute u sceni kao jedinstvenu semantičku kategoriju. Unifikacija reprezentacije za različite instance nije dobra za uspostavu korespondencije, jer ona zahtjeva postojanje diskriminativne značajke na razini instance.

Problem uspostave korespondencije izbjegavaju pristupi koji na ulazu izravno primaju informacije o gibanju objekata. U tu kategoriju ubrajaju se metode M2M (eng. *motion to motion*) koje prognoziraju budući optički tok na temelju toka između prošlih okvira. Prognozirani optički tok koristi se za deformiranje (eng. *warp*) semantičkih predikcija iz posljednjeg promatranog okvira te se na taj način dobiju semantičke predikcije koje odgovaraju ciljanom budućem trenutku. Slika 3.4 prikazuje skicu jednog takvog sustava. Pored prognostičkog modela, ovakvi sustavi zahtjevaju još dva neovisna modela. Jedan za procjenu optičkog toka (plavi pravokutnik lijevo), te drugi model za gustu semantičku predikciju u jednoj slici (plavi pravokutnik desno). Temeljna pretpostavka ovih modela jest da se buduća scena može u potpunosti rekonstruirati iz prošlosti, što nije uvijek točno. Budućnost je stohastična i izgledna je pojava dijelova scene koji nisu opaženi u prošlosti. U takvim situacijama prognostički model mora zamišljati, što modeli M2M ne mogu jer se oslanjaju isključivo na rekonstrukciju iz prošlosti. Dodatan problem je što se prilikom samog prognoziranja u potpunosti zanemaruje semantika scene. To je suboptimalno jer se u semantici krije i informacija o uzorcima gibanja određenih objekata. Primjerice, za automobile znamo da se gibaju rigidno, dok se udovi čovjeka gibaju drugačije od njegova trupa. Prognoziranje na razini optičkog toka korišteno je za procjenu buduće semantičke segmentacije [48]. Njihova metoda je u tome trenutku postizala najbolje rezultate u prognoziranju na skupu Cityscapes. Za procjenu optičkog toka korišten je Flownet [23], a za

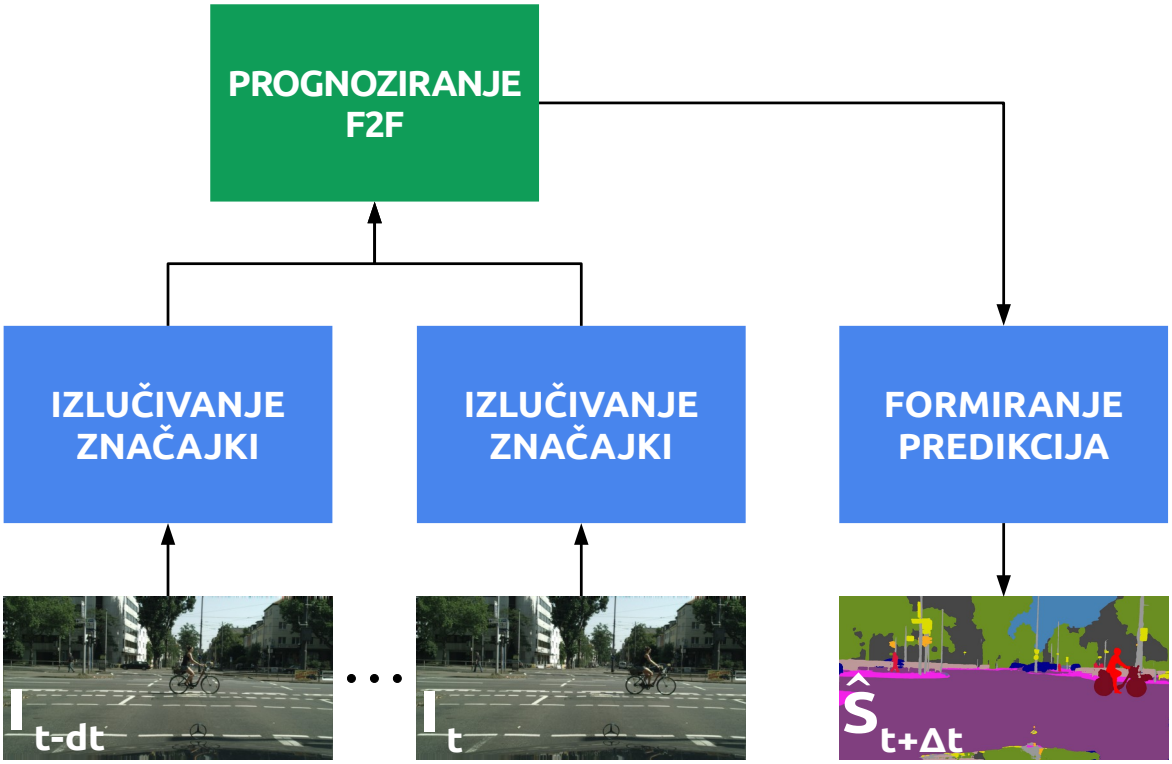


Slika 3.4: Metoda prognoziranja iz pomaka u pomak (M2M).

semantičku segmentaciju posljednjeg okvira PSPNet [62]. Kao prognostički model korišten je konvolucijski LSTM [94]. U ovu skupinu pristupa baziranim na gibanju može se ubrojiti i rad koji je već spomenut prilikom pregleda S2S pristupa [92]. Naime, predložena metoda [92] također promatra optički tok iz prošlosti na temelju čega kodira gustu reprezentaciju uzorka gibanja za svaki objekt, koja se koristi za precizniju predikciju opisujućih okvira u budućnosti. Prognoziranje na razini optičkog toka slično je našem predloženom modulu F2M, koji također prognozira budućnost deformiranjem prošlih reprezentacija. Ipak, postoje neke važne razlike. Modul F2M sam prognozira tok na temelju značajki iz prošlosti, te ne zahtjeva neki vanjski model za procjenu toka, niti dodatne oznake prilikom učenja. Modul F2M uči prediktirati tok koji će deformacijom značajki iz prošlosti najbolje opisati značajke iz budućnosti. Osim toga, naš pristup smanjuje mogućnost propagacije pogreške zbog treniranja s kraja na kraj i značajno je efikasniji zbog primjene na poduzorkovanim značajkama koje su uz to dijeljene između modula za rekonstrukciju gibanja i gusto raspoznavanje. Također, skrivena reprezentacija iz koje se prognozira buduće gibanje sadrži i semantičku informaciju o objektima koja korelira s njihovim budućim gibanjem. Dodatno, naš prognostički pristup ne oslanja se samo na najrecentniju sliku. Umjesto toga, naš pristup miješa doprinose iz svih prošlih slika što omogućuje razrješavanje nekih kompleksnih uzoraka zaklanjanja i otkrivanja i jednostavan odabir najpogodnijeg trenutka za prognoziranje određenog dijela scene. Uz to, naš modul F2M je kombiniran s modulom F2F koji izravno regresira buduće značajke i tako omogućuje točnije prognoziranje u

dijelovima scene koji prethodno nisu bili viđeni.

Trenutno stanje tehnike [95] u gustom semantičkom prognoziranju postižu modeli za prognoziranje na razini skrivene reprezentacije dubokog modela. Pristupi F2F (eng. *features to features*) preslikavaju skrivene značajke dubokog modela iz prošlosti u budućnost. Slika 3.5 prikazuje osnovnu organizaciju sustava za prognoziranje temeljenog na prognoziranju iz značajki u značajke (F2F). Sustav možemo podijeliti na tri dijela: model za prognoziranje značajki



Slika 3.5: Osnovna organizacija sustava za gusto semantičko prognoziranje utemeljenog na preslikavanju značajki (F2F).

(zeleni pravokutnik), modul za ekstrakciju značajki (plavi pravokutnik lijevo) i modul za formiranje semantičkih predikcija (plavi pravokutnik desno). Ovakva podjela posebno je zanimljiva zato što ekstrakcija značajki i formiranje semantičkih predikcija izravno odgovaraju modulima modela za gustu semantiku u jednoj slici. Ekstrakcija značajki odgovara okosnici odnosno putu za poduzorkovanje dok formiranje semantičkih predikcija odgovara putu za naduzorkovanje [13]. Prilikom zaključivanja, prvo se dakle ekstrahiraju značajke iz svih promatranih slikovnih okvira uz pomoć prednjeg kraja semantičkog modela. Zatim, te značajke prognostički model pretvara u značajke koje odgovaraju ciljanom budućem trenutku. Konačno, stražnji kraj semantičkog modela interpretira prognozirane značajke i predviđa guste buduće semantičke oznake. Prognostički model nije ovisan o karakteru semantičkih predikcija i zato ga je lako uklopiti u različite modele za gustu predikciju u jednoj slici. Prednost je, dakle, prognoziranja značajki u tome što jednostavno generalizira na različite ciljne zadatke računalnog vida [15]. Pažljiva ana-

liza cjevovoda zaključivanja otkriva i potencijal ovih metoda za izvođenje u stvarnom vremenu. Prvo, ekstrakcija značajki iz N slika (t, $t - \Delta T$, ...) može se amortizirati jer se svaki tenzor značajki koristi pri N prognostičkih zaključivanja. Drugo, značajke iz tekuće slike mogu se koristiti i za gustu predikciju u tekućem trenutku i za gusto semantičko prognoziranje. Nadalje, samo prognoziranje provodi se na značajkama koje su uobičajeno nekoliko puta poduzorkovane u odnosu na ulaznu sliku. Procesiranje takvih značajki puno je efikasnije od procesiranja semantičkih predikcija na finoj rezoluciji. Pored toga, uspostava korespondencije je lakša jer skrivene značajke u sebi mogu sadržavati diskriminativne informacije o izgledu objekta. Tome svjedoče eksperimenti iz [96] koji uspješno rekonstruiraju ulaznu sliku iz značajki dubokog modela. To omogućava prognoziranje na nekoliko puta poduzorkovanoj rezoluciji što rezultira značajno efikasnijim modelima u odnosu na prethodno spomenute pristupe. Prvi F2F pristup prognoziranju ciljao je vektor značajki iz potpuno povezanog sloja modela AlexNet u svrhu predikcije budućih akcija [97]. Proširenje na prognoziranje gustih konvolucijskih značajki nastalo je s ciljem predviđanja buduće segmentacije instanci [6]. Predložena metoda prognozirala je piramidu značajki iz modula FPN [33], a svaka razina imala je zaseban prognostički model koji je sadržavao niz dilatiranih konvolucija. Budući razvoj te ideje [16, 98] izražava modele za prognoziranje konvolucijskim LSTM-om i uvodi posebne veze između razina piramide koje služe dijeljenju konteksta prilikom prognoziranja. Unaprjeđenje je postiglo značajno bolju točnost prognoziranja, međutim nije se riješilo utega prognoziranja piramide značajki. Naime, najfinija rezolucija značajki u piramidi FPN-a samo je četiri puta poduzorkovana u odnosu na sliku, a prognoziranje značajki fine rezolucije je računalno skupo [99]. Stoga, naše predložene metode [13, 15] uvijek koriste jednorazinsko prognoziranje i ciljaju prostorno najsažetiju moguću reprezentaciju modela za gustu predikciju u jednoj slici. Takav pristup značajno smanjuje zahtjeve za računalnim resursima [15]. Dodatno, sažeta reprezentacija je pogodna za konvolucijsko prognoziranje zbog manjih pomaka značajki kroz vrijeme, a također sadrži bogatu kontekstualnu informaciju. Nedostatak ovakve reprezentacije svakako je mogućnost gubitka informacije za raspoznavanje malih objekata. Vora et al. [100] prognoziraju buduće značajke reprojekcijom 3D rekonstruiranih značajki u skladu s budućom prognoziranom pozicijom kamere. Problem s ovim pristupom je što ne radi dobro u prisutnosti zaklanjanja i otkrivanja te velikih promjena u perspektivi. Također nije u mogućnosti modelirati neovisne pomake ostalih gibajućih objekata. Visoka prognostička točnost naše metode sugerira da precizno prognoziranje zahtjeva balans između rekonstrukcije i raspoznavanja, kao i to da poznavanje 3D scene donosi samo minimalne benefite. Chiu et al. [101] progoziraju značajke na jednoj razini uz pomoć 3D konvolucija. Najnoviji pristup prognoziranju značajki koji ujedno postiže i stanje tehnike na skupu Cityscapes također koristi jednorazinsko prognoziranje [95]. Ipak, njihova metoda ne koristi posebno dizajnirane modele za gustu predikciju u jednoj slici, nego dodaje još jedan korak procesiranja. Naime, piramida značajki izlučena iz FPN modula se prije prog-

noziranja kodira u sažetiju jednorazinsku reprezentaciju uz pomoć varijacijskog autoenkodera [102]. Jednako tako se prognozirana reprezentacija dekodira u piramidu značajki uz pomoć drugog dijela (dekodera) istog varijacijskog autoenkodera. Varijacijski autoenkoder filtrira visoke frekvencije koje je teško prognozirati te stvara reprezentaciju koja je prikladnija za prognoziranje konvolucijskim LSTM-om. Ipak, kodiranje i dekodiranje piramide značajki unosi dodatnu složenost u postupak prognoziranja.

Višemodalno prognoziranje značajki. Mogućnost prognoziranja više različitih ishoda budućnosti nužna je za stvarne primjene prognostičkih sustava. Višemodalno predviđanje budućnosti moguće je ostvariti i modelima koji prognoziraju iz značajki u značajke. Ovaj odjeljak opisuje proširenje našeg prognostičkog modela [13] za višemodalno prognoziranje. Rezultati ovog istraživanja objavljeni su na međunarodnoj konferenciji [9] te u diplomskom radu autora Kristijana Fugošića. Naše proširenje utemeljeno je na uvjetnom generativnom modeliranju te standardnom suparničkom gubitku. Kako bismo izbjegli kolapsiranje modova budućnosti uvjetni dio gubitka umjesto rekonstrukcije piksela [103] zahtijeva rekonstrukciju momenata slike za učenje [104]. U našem višemodalnom sustavu prognostički model ima ulogu generatora uvjetovanog promatranim značajkama iz prošlosti i slučajno uzorkovanim latentnim vektorom. Budućnost je dominantno određena promatranim značajkama iz prošlosti, a zadaća latentnog vektor je kodirati jedan od mogućih ishoda iz danih opažanja. Prilikom treniranja, svaki primjer za učenje kombinira se s više različitih latentnih vektora te se unaprijednim prolazom dobije više prognoziranih tenzora značajki. Iz prognoziranih tenzora računa se srednja vrijednost $\hat{\mu}$ i varijanca $\hat{\sigma}^2$ (prvi i drugi moment razdiobe) koji se koriste u rekonstrukcijskom gubitku momenata:

$$\mathscr{L} = \mathbb{E}_{x,y}\left[\frac{(y-\hat{\mu})^2}{2\hat{\sigma}^2} + \frac{1}{2}\log\hat{\sigma}^2\right], \text{ gdje } (\hat{\mu},\hat{\sigma}^2) = f_\theta(x) \tag{3.1}$$

Kvadrirana razlika usrednjenih prognoza i ciljne buduće reprezentacije osigurava da prognoze u prosjeku odgovaraju stvarnom budućem ishodu. Varijanca u nazivniku izraza sprječava kolaps varijance u nulu i tjera model da razmatra uzorkovani latentni vektor i na osnovu njega prediktira više različitih budućnosti. Spomenutom gubitku dodajemo i standardni gubitak generativnih suparničkih modela koji se temelji na predikciji suparničkog modula (diskriminatora). U ovakvom sustavu na ulaz diskriminatora dovode se značajke, a njegova zadaća je razlikovati prognozirane značajke od budućih značajki koje smo dobili standardnim predikcijskim modelom koji na ulaz dobiva buduću sliku (takav model nazivamo i prorokom). Kao i inače, cilj generatora je zavarati diskriminator. Da bi to postigao prognostički model (generator) mora prognozirati značajke koje su slične stvarnim značajkama izlučenima iz budućeg slikovnog okvira. Rezultati su pokazali da je višemodalnost ostvarena po cijenu točnosti. Očekivano modeli koji na izlazu daju raznolikije predikcije u prosjeku postižu manju točnost prognoziranja.

Poglavlje 4

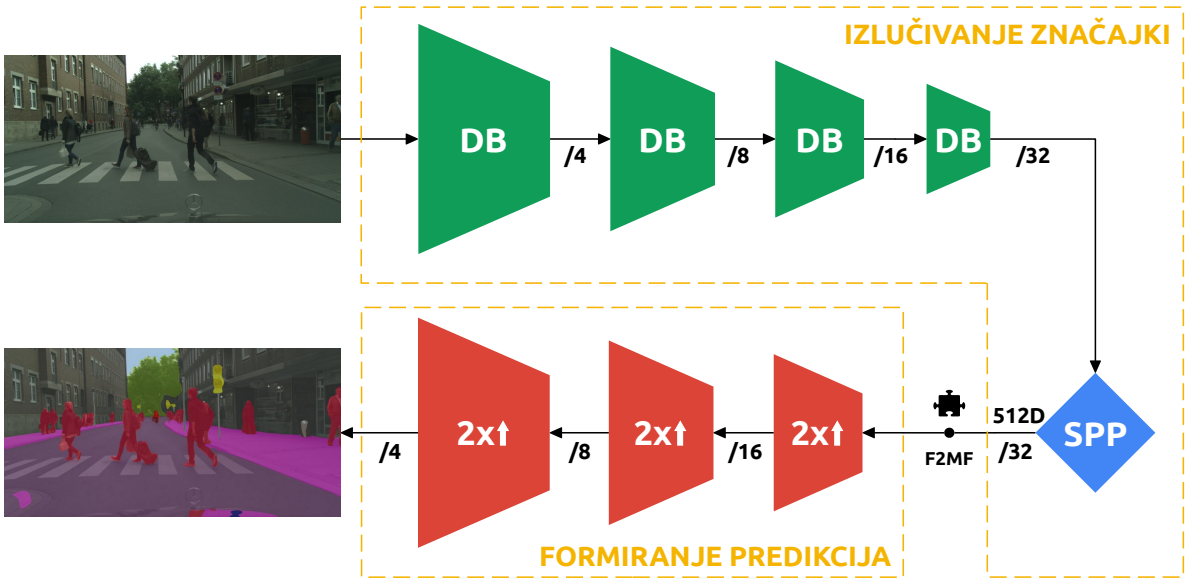
Gusto raspoznavanje u jednoj slici

Sasvim općenito, semantičko prognoziranje na razini značajki trebalo bi biti primjenljivo na različite zadatke guste predikcije. Ova disertacija provjerava točnost te hipoteze integracijom našeg prognostičkog modela u sustave za prognoziranje semantičke segmentacije, segmentacije instanci i panoptičke segmentacije. Svaki od navedena tri sustava zasnovan je na odgovarajućem modelu za gusto raspoznavanje u jednoj slici, koji se specijalizira za ciljani zadatak. Jedna od osnovnih pretpostavki naše metode jest jednorazinsko prognoziranje na sažetoj rezoluciji. Višerazinsko prognoziranje zahtijeva znatno više računskih resursa, a u našim ranim eksperimentima nije podiglo generalizacijsku točnost prognoziranja. U nekim slučajevima, to onemogućuje izravno korištenje gustih semantičkih modela iz literature, ali problem možemo zaobići uz neke jednostavne adaptacije. U nastavku su opisani tako prilagođeni modeli za gusto raspoznavanje u jednoj slici: SwiftNet bez preskočnih veza za semantičku segmentaciju, Mask R-CNN C4 za segmentaciju instanci, te Panoptic Deeplab bez preskočnih veza na rezolucijama R/8 i R/4 za panoptičku segmentaciju.

4.1 SwiftNet bez preskočnih veza

Arhitektura SwiftNet dizajnirana je s ciljem semantičke segmentacije u realnom vremenu, te se stoga temelji na laganim okosnicama i tankom ljestvičastom naduzorkovanju. Izlazne značajke kod SwiftNeta su $4\times$ poduzorkovane u odnosu na sliku. Iz njih se jednom konvolucijom računaju logiti koji se $4\times$ bilinearno naduzorkuju kako bi se dobila semantička segmentacije slike na punoj rezoluciji. Integriranje originalnog SwiftNeta u F2MF prognoziranje semantičke segmentacije zahtijevalo bi prognoziranje izlaznih značajki ili piramide značajki koju čine izlaz okosnice i sve preskočne veze. Međutim, takav sustav bio bi neučinkovit zbog prognoziranja značajki visoke rezolucije. Takvo prognoziranje bilo bi i teško naučiti zbog većih apsolutnih iznosa pomaka objekata na visokoj rezoluciji. Stoga ova disertacija predlaže inačicu arhitekture SwiftNet koja uklanja preskočne veze iz ljestvičastog naduzorkovanja, čime ono zapravo pres-

taje postojati. Ovakva organizacija omogućuje efikasno prognoziranje sažetog tenzora značajki koji je 32× poduzorkovan u odnosu na ulaznu sliku. Koristimo dvije varijante modela: jednu manjeg kapaciteta sa okosnicom ResNet-18, te drugu većeg kapaciteta sa okosnicom DenseNet-121. Slika 4.1 prikazuje inačicu SwiftNeta bez preskočnih veza s okosnicom DenseNet-121. Prilikom ugradnje u prognostički sustav, prikazani model dijelimo u dvije komponente. Prva služi izlučivanju značajki koje se koriste za prognoziranje. Nju čine četiri gusto povezana bloka (zeleni trapezi) iz okosnice te modul za prostorno piramidalno sažimanje (plavi romb). Izlučene značajke su 32× poduzorkovane u odnosu na ulaznu sliku i imaju 512 kanala. Upravo njih naš model F2MF mapira iz prošlosti u budućnost, što je na slici naznačeno crnim simbolom slagalice. Druga komponenta modela u sustavu prognoziranja služi pretvaranju budućih značajki u buduću semantičku segmentaciju. Ona zapravo odgovara putu naduzorkovanja kojeg čine tri modula za naduzorkovanje i konvolucijski sloj za izračun tenzora logita koji predstavlja ulaz u gusti softmaks. Jedan modul za naduzorkovanje sastoji se od jednog slijeda BN-ReLU-CONV i bilinearnog naduzorkovanja za udvostručenje rezolucije. Pri tome BN označava normalizaciju nad grupom, ReLU - aktivacijsku funkciju zglobnice, a CONV - konvolucijski sloj sa jezgrom 3×3. Važan hiperparametar ove komponente jest debljina puta naduzorkovanja odnosno broj mapa značajki u svim modulima. Ako je okosnica modela DenseNet-121, onda debljinu puta naduzorkovanja postavljamo na 256. Inačica modela manjeg kapaciteta koristi slabiju okosnicu ResNet-18, modul za prostorno piramidalno sažimanje na izlazu daje značajke sa 128 kanala, te je debljina puta naduzorkovanja jednaka 128. Model slabijeg kapaciteta postiže manju točnost,

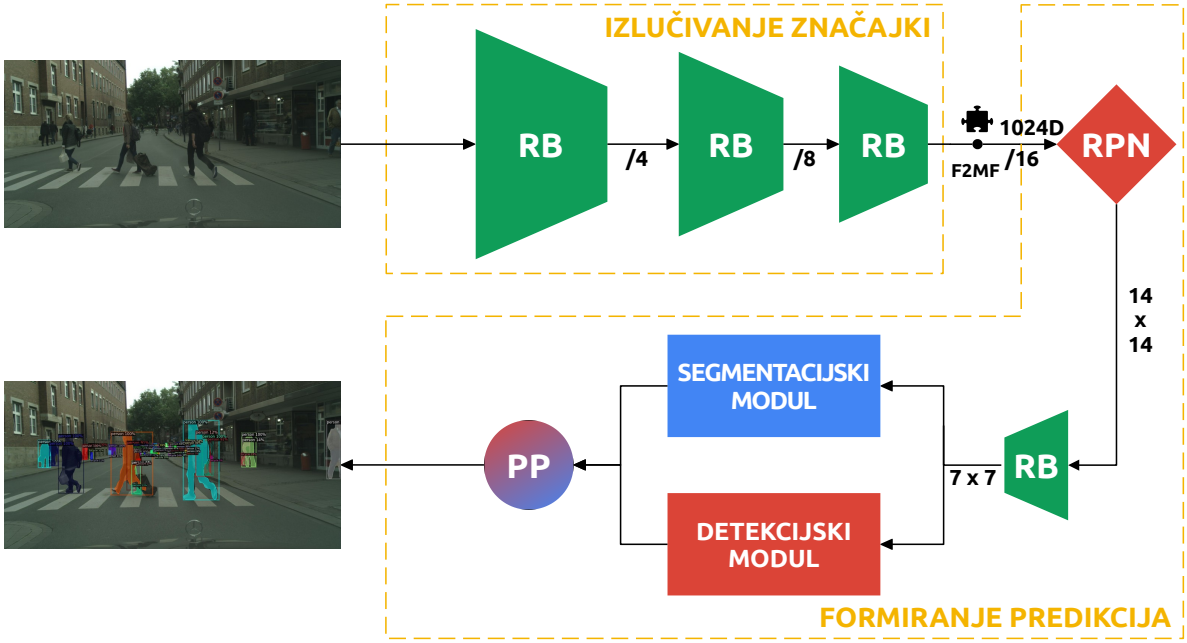


Slika 4.1: Predloženi model za semantičku segmentaciju u jednoj slici temeljen na arhitekturi SwiftNet bez preskočnih veza sa okosnicom DenseNet-121. Prednji dio modela (eng. *feature extraction*) izlučuje značajke, dok ih stražnji dio modela (eng. *semantic formation*) pretvara semantičke predikcije. Crna slagalica označava značajke na kojima ćemo primijenjivati naš prognostički model (F2MF).

ali je zato učinkovitiji. Učinkovitost se u nekoj mjeri prenosi i na proces prognoziranja značajki, zbog razlike u broju kanala.

4.2 Mask R-CNN C4

Najčešće korištena inačica modela Mask R-CNN [20] koristi ljestvičasto naduzorkovanje (eng. *feature pyramid network*, FPN). Regresija okvira objekata provodi se dijeljenim detekcijskim modulom koji se primjenjuje na svim razinama uzduž puta naduzorkovanja. Korištenje takvog modela u sustavima za prognoziranje na razini značajki zahtijeva primjenu višerazinskog prognoziranja što je iznimno računalno skupo [6]. Zato se u našim eksperimentima na prognoziranju segmentacije instanci koristi inačica modela Mask R-CNN C4 [20]. Implementacija ove inačice dostupna je u biblioteci detectron2 [105]. Ta inačica omogućuje jednorazinsko prognoziranje značajki jer se detekcijski modul primjenjuje na jedan tenzor koji odgovara izlazu trećeg rezidualnog bloka okosnice. Ovakav model zapravo odgovara proširenju modela Faster R-CNN [18] dodatnim modulom za regresiju segmentacijske maske detektiranog primjerka. Skica modela prikazana je na slici 4.2. Slično kao kod semantičke segmentacije, model je podijeljen na dva dijela. Izlučivanje značajki uključuje tri rezidualna bloka okosnice ResNet-50 (zeleni trapezi). To implicira da su značajke za prognoziranje u ovome slučaju poduzorkovane 16× u odnosu



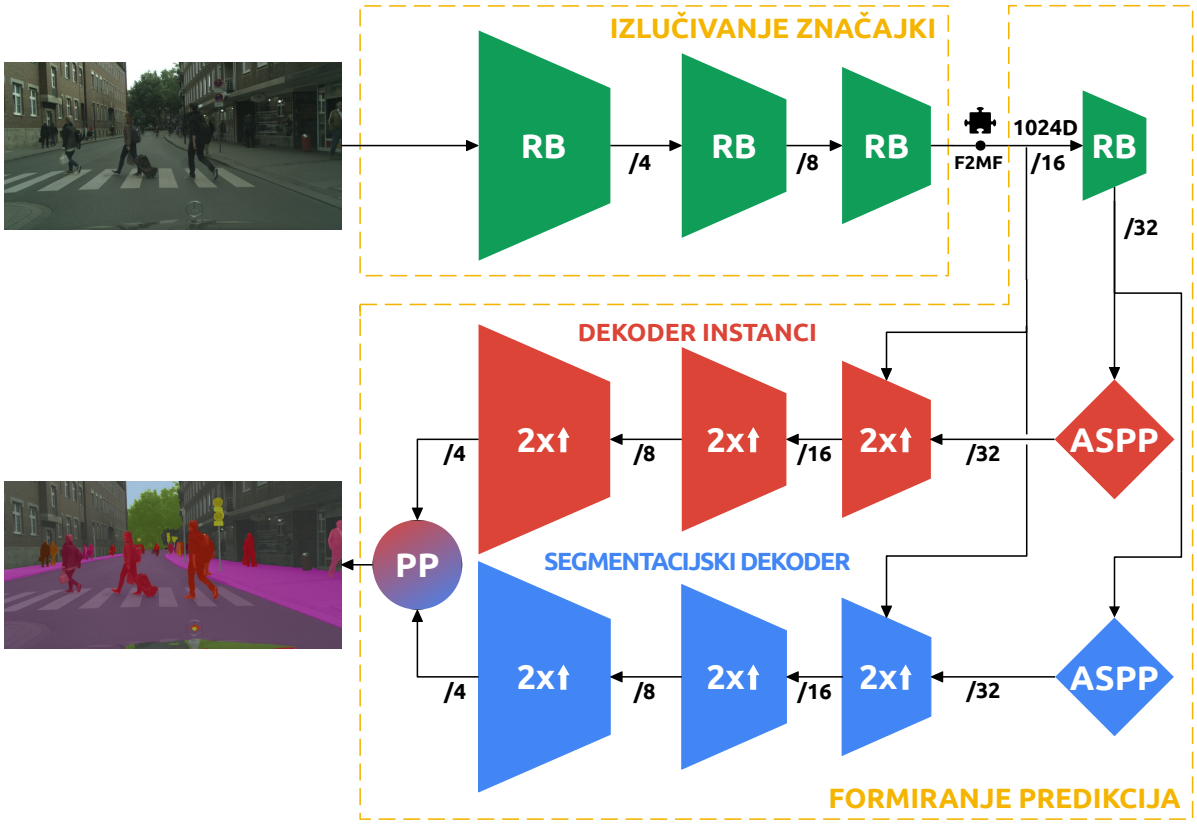
Slika 4.2: Model za segmentaciju instanci Mask R-CNN C4 sa okosnicom ResNet-50. Prednji dio modela (eng. *feature extraction*) izlučuje apstraktne konvolucijske značajke, dok stražnji dio modela (eng. *semantic formation*) predlaže kandidate primjeraka (RPN) u kojima primijenjujemo interpolacijsko sažimanje te odvojene glave za detekciju okvira i regresiju segmentacijske maske. Crna slagalica označava značajke na kojima ćemo primijenjivati naš prognostički model (F2MF).

na ulaznu sliku i imaju 1024 mape značajki. Prognoziranje takvih značajki je nešto izazovnije zbog većih apsolutnih iznosa pomaka, ali i računalno zahtjevnije zbog većeg broja mapa značajki. Točka u kojoj se priključuje model F2MF u sustavu za prognoziranje na slici je naznačena crnom slagalicom. Modul za formiranje predikcija u sustavu za prognoziranje odgovara detekcijskom modulu za segmentaciju instanci. Detekcijski modul započinje konvolucijskim slojem za detekciju kandidata (RPN). On modificira veličinu i položaj pretpostavljenih sidrenih okvira i odabire kandidate u kojima bi se mogao nalaziti neki objekt. Preostala obrada provodi se nezavisno u svakom kandidatu kako slijedi. Prvo se interpolacijskim sažimanjem (eng. *roi align*) izluči reprezentacija kandidata rezolucije 14×14. Na nju se primjenjuje četvrti konvolucijski blok okosnice, potpuno povezana glava za klasifikaciju i poboljšanje okvira primjerka (crveni pravokutnik) te konvolucijska glava za regresiju segmentacijske maske instance (plavi pravokutnik). Konačno se u koraku postprocesiranja fiksna segmentacijska maska rasterizira u prostoru slike unutar detektiranog pravokutnog okvira na odgovarajućoj lokaciji.

4.3 Modificirani Panoptički DeepLab

Panoptički DeepLab [21] je recentni model za panoptičku segmentaciju koji je po svojoj strukturi vrlo sličan SwiftNetu. Sastoji se od konvolucijske okosnice i ljestvičastog puta za naduzorkovanje. Okosnica se tipično predtrenira na ImageNetu. Kao što smo naveli ranije, ljestvičasto naduzorkovanje nije prikladno za jednorazinsko prognoziranje. Međutim, preskočne veze su ovdje važnije nego kod SwiftNeta jer one značajno poboljšavaju točnost pronalaženja malenih objekata koji su važniji za panoptičku nego za segmentacijsku performansu. Zbog toga ovdje predlažemo modifikaciju panoptičkog DeepLaba koja zadržava samo jednu preskočnu vezu, onu na rezoluciji R/16 kao što je ilustrirano na slici 4.3. Prednji dio modela za izlučivanje značajki uključuje prva tri rezidualna bloka (zeleno). Slično kao kod segmentacije instanci, to znači da ciljne značajke za prognoziranje imaju 1024 kanala (pretpostavljamo ResNet-50) i $16\times$ manju rezoluciju u odnosu na ulaznu sliku. Stražnji dio predloženog modela zadužen je za formiranje semantičkih predikcija. Značajke se obrađuju četvrtim rezidualnim blokom te prosljeđuju u dva odvojena dekodera. Prvi dekoder zadužen je za semantičku segmentaciju (plavo), a drugi za određivanje segmentacije instanci nepoznatih semantičkih razreda (crveno). Oba dekodera imaju istu strukturu. Započinju modulom za prostorno piramidalno sažimanje te nakon toga primjenjuju tri modula za naduzorkovanje koji dižu rezoluciju značajki na $4\times$ manju od ulazne slike. Završni konvolucijski sloj u semantičkom dekoderu računa klasifikacijske logite na razini piksela. U dekoderu za segmentaciju instanci postoje dva završna sloja: jedan koji računa toplinsku mapu (eng. *heatmap*) koja indicira centre objekata, te drugi koji za svaki piksel računa dvodimenzionalni vektor pomaka prema odgovarajućem centru objekta. Izlazi svih završnih slojeva se u koraku postprocesiranja kombiniraju u panoptičku segmentaciju. Prvo se

nad toplinskom mapom centara provodi potiskivanje nemaksimalnih odziva čime se dobiva diskretni skup centara objekata. Time je već određen broj detektiranih primjeraka u sceni. Zatim se svi pikseli koji pripadaju prebrojivim razredima pridjeljuju jednom od detektiranih centara. To se radi na način da se apsolutnoj poziciji piksela doda predviđeni vektor pomaka, a zatim se pronađe centar koji je po euklidskoj udaljenosti najbliži toj pomaknutoj poziciji. Indeks svakog piksela instance određen je indeksom pridijeljenog mu centra, čime je završena segmentacija instanci bez poznavanja semantičkog razreda. Semantički razred svake instance određuje se pronalaskom najčešće klase unutar granica instance prema predviđenoj semantičkoj segmentaciji. Svim pikselima instance pridjeljuje se taj isti najčešći semantički razred. Panoptičku predikciju dovršavamo kopiranjem semantičke segmentacije za neprebrojive razrede.



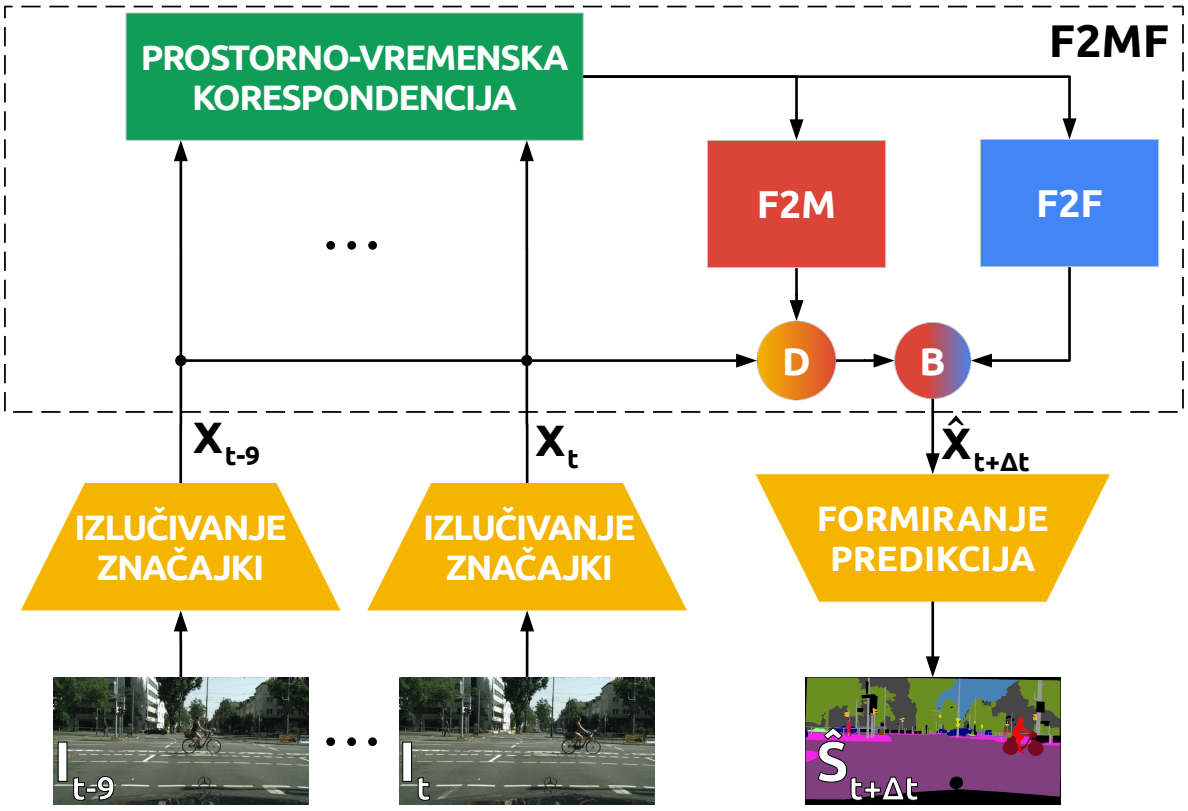
Slika 4.3: Modificirana inačica modela za panoptičku segmentaciju Panoptic Deeplab s jednom preskočnom vezom. Prednji dio modela (eng. *feature extraction*) izlučuje apstraktne konvolucijske značajke, dok stražnji dio modela (eng. *semantic formation*) provodi prožimanje značajki i konteksta (ASPP) te naduzorkovanje s jednom preskočnom vezom. Crna slagalica označava značajke na kojima ćemo primijenjivati naš prognostički model (F2MF).

Poglavlje 5

Gusto semantičko prognoziranje združenom regresijom pomaka i značajki

Ovo poglavlje predstavlja predloženi sustav za gusto semantičko prognoziranje temeljen na regresiji pomaka i značajki. Fokus je na novom modelu za prognoziranje značajki koji čini jezgru cijelog sustava. Model kombinira standardno izravno prognoziranje značajki (F2F) s regulariziranom inačicom (F2M) koja predviđa tok kojim se deformiraju značajke iz prošlosti u budućnost. Prije uranjanja u detalje samog modela za prognoziranje, dobro je podsjetiti se šire slike cjelokupnog sustava. Slika 5.1 donosi grubi pregled sustava od samih ulaznih slikovnih okvira do gustih semantičkih predikcija budućnosti. Zaključivanje započinje izračunom značajki $\mathbf{X}_\tau$ iz korespondentnih slika I_τ, $\tau \in \{t-9, t-6, t-3, t\}$. Značajke nastaju primjenom prvog dijela predtreniranog semantičkog modela (odnosno modula za izlučivanje značajki) u svakoj od promatranih slika iz prošlosti zasebno. Korelacijski modul obogaćuje ekstrahirane značajke korelacijskim koeficijentima koje se nakon daljnje obrade predaju modulima F2M i F2F. Modul F2M predviđa pomake značajki između prošlosti i budućnosti. Modul F2F izravno prognozira buduće značajke. Paralelno s tim, prognostički model računa težinske faktore miješanja na razini piksela. Prognozirane značajke $\hat{\mathbf{X}}_{t+\Delta t}$ nastaju miješanjem prognoza F2M i F2F u skladu s previđenim težinama. Konačno, te značajke interpretira preostali dio semantičkog modela (modul za formiranje semantičkih predikcija) i predviđa konačne semantičke predikcije $\hat{\mathbf{S}}_{t+\Delta t}$ za ciljani budući okvir.

Može se reći da je sustav za prognoziranje građen od dva dijela: modela za gusto raspoznavanje u jednoj slici i modela za prognoziranje značajki. Podsjetimo se da su modul za izlučivanje značajki i modul za formiranje semantičkih predikcija inače dio jednog modela koji obavlja gusto raspoznavanje u jednoj slici. Detalji tih modela objašnjeni su u poglavlju 4. Najveći doprinosi ove disertacije zapravo su ostvareni kroz oblikovanje prognostičkog modela koji skrivene međureprezentacije gustog semantičkog modela mapira iz prošlosti u budućnost. U nastavku ćemo razmatrati predloženi model za prognoziranje značajki F2MF i detaljno opisati

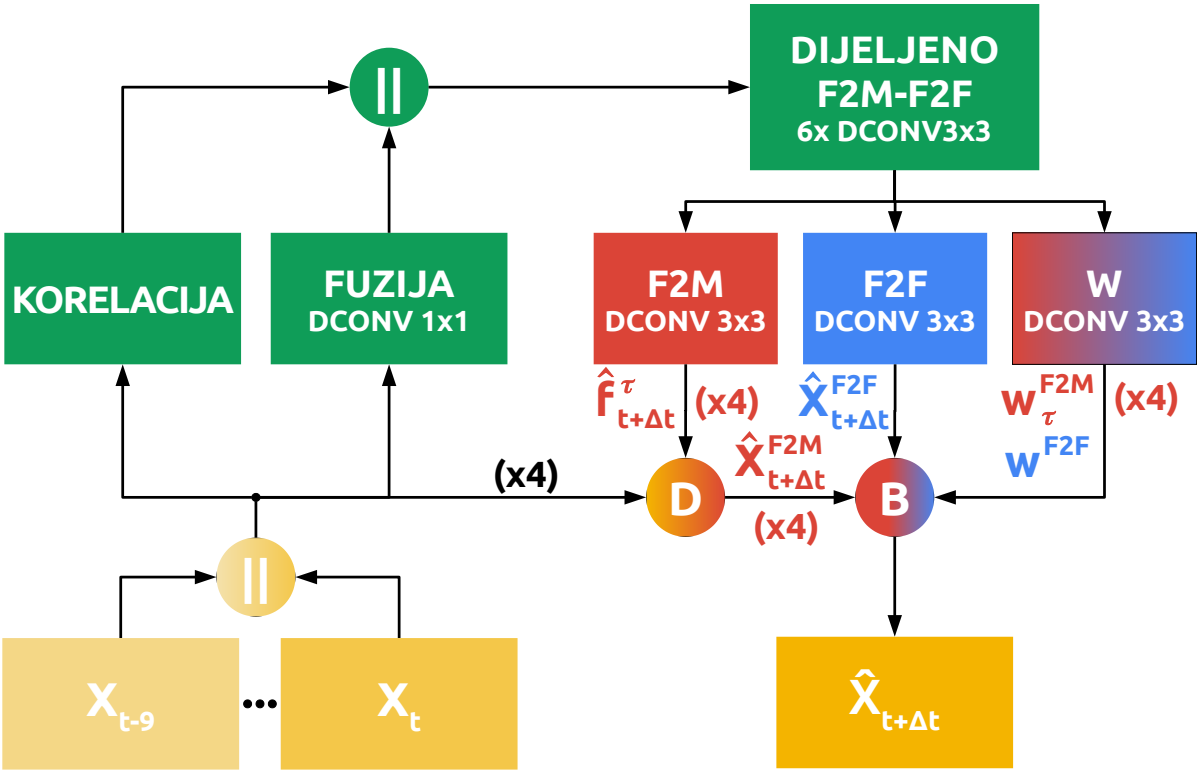


Slika 5.1: Pregled sustava za prognoziranje temeljenog na metodi F2MF. Ulaz u model za prognoziranje čine sažete značajke niske rezolucije $\mathbf{X}_\tau$ ekstrahirane prednaučenim modulom za raspoznavanje iz korespondentnih slika I_τ, $\tau \in \{t-9, t-6, t-3, t\}$. Značajke obogaćene prostorno-vremenskim korelacijskim koeficijentama su procesirane i dovedene na ulaz modulima F2F i F2M. Modul F2F izravno prognozira značajke i specijalizira se za novootkrivene dijelove scene. Modul F2M prognozira deformiranjem značajki iz prošlosti i specijalizira se za prethodno viđene dijelove scene. Prognozirane značajke $\hat{\mathbf{X}}_{t+\Delta t}$ su mješavina izlaza modula F2M i F2F. Konačno, guste semantičke predikcije $\hat{\mathbf{S}}_{t+\Delta t}$ za buduću slikovni okvir predviđene su iz prognoziranih značajki predtreniranim modulom za naduzorkovanje.

njegove ključne dijelove.

Slika 5.2 ilustrira strukturu predloženog prognostičkog modela. Ulaz modela čine $T = 4$ tenzora značajki $\mathbf{X}_{t-9}$, $\mathbf{X}_{t-6}$, $\mathbf{X}_{t-3}$, $\mathbf{X}_t$ (skraćeno ćemo pisati $\mathbf{X}_{\mathtt{t-9:t:3}}$). Ta četiri tenzora proizveo je modul za izlučivanje značajki odgovarajućeg modela za gusto raspoznavanje u jednoj slici. Zadaća modela F2MF je tenzore značajki iz prošlosti $\mathbf{X}_{\mathtt{t-9:t:3}}$ pretvoriti u budući tenzor značajki $\hat{\mathbf{X}}_{t+\Delta t}$. Slika 5.1 pokazuje da se regresirane značajke $\hat{\mathbf{X}}_{t+\Delta t}$ dovode na stražnji dio modela za obradu jedne slike koji ih pretvara u buduće guste semantičke predikcije $\hat{\mathbf{S}}_{t+\Delta t}$. Napomenimo da Δt označava vremenski razmak između posljedneg promatranog okvira i ciljanog okvira u budućnosti. To ćemo obično izražavati u broju sličica u videu između odgovarajućih okvira, a lako se pretvori i u proteklo vrijeme množenjem s recipročnom vrijednosti broja sličica u sekundi za konkretni snimljeni video.

Model F2MF svoju obradu započinje konkatenacijom ulaznih tenzora značajki po semantičkoj dimenziji. Rezultat konkatenacije obrađuje se konvolucijskim slojem kako bi se smanjio



Slika 5.2: Struktura predloženog modela za prognoziranje zasnovanog na kombinaciji izravnog prognoziranja značajki i prognoziranja značajki deformacijom prošlih reprezentacija u skladu s predviđenim tokom.

broj kanala. Nakon ovoga koraka miješanja nemoguće je jednostavno restaurirati podatak o vremenu u kojem je nastala neka značajka. Vjerujemo kako prognostički model implicitno u svoju reprezentaciju prema potrebi ugrađuje i tu informaciju. Paralelno s fuzijom, korelacijski modul računa korelacijske koeficijente u lokalnom susjedstvu za susjedne parove ulaznih tenzora značajki. Fuzionirana reprezentacija i izlazi korelacijskog sloja se kombiniraju konkatenacijom. Rezultat te konkatenacije se obrađuje nizom od šest konvolucijskih slojeva čime nastaje dijeljena reprezentacija za prognoziranje modulima F2M i F2F. Dodatno se iz te dijeljene reprezentacije jednim konvolucijskim slojem predviđaju i težine za miješanje prognoziranih značajki. Težinski modul (označen slovom W u slici 5.2) na izlazu daje pet mapa značajki koje odgovaraju težinama $w^{\mathrm{F2M}}_{t-9}, w^{\mathrm{F2M}}_{t-6}, w^{\mathrm{F2M}}_{t-3}, w^{\mathrm{F2M}}_{t}$ i w^{F2F}. Četiri mape $[w^{\mathrm{F2M}}_{\tau}]$, $\tau \in \{t-9:t:3\}$ predstavljaju vrijednost doprinosa svakog od četiri prošla tenzora u budućoj reprezentaciji prilikom prognoziranja modulom F2M. Peta mapa w^{F2F} predstavlja doprinos modula F2F prilikom združenog prognoziranja. Težine se prije miješanja aktiviraju gustim softmaksom preko pet navedenih prognoza.

Svi konvolucijski slojevi u predloženom prognostičkom modelu izraženi su blokom BN-ReLU-DCONV, gdje BN stoji za normalizaciju nad grupom, ReLU za aktivacijsku funkciju zglobnicom, te DCONV za deformabilnu konvoluciju. Deformabilne konvolucije su prikladne za ovu primjenu jer je zadatak prognoziranja značajki više geometrijski, nego semantički in-

tenzivan. Naime, ulazi i izlazi modela za prognoziranje iste su semantičke složenosti. Obične konvolucije imaju manje šanse naučiti potrebne geometrijske transformacije zbog pravilne rešetke i fiksnih pozicija uzorkovanja. Podsjetimo se da su lokacije uzorkovanja kod deformabilnih konvolucija pomaknute u skladu s pomacima prediktiranima na osnovu ulaznog tenzora značajki. To je prikladno za prognoziranje jer ima mogućnost modelirati dinamiku na razini objekta uzimajući u obzir dostupnu semantičku informaciju. Vjerujemo da deformabilne konvolucije u ovome kontekstu rade posebno dobro zbog korelacijskog modula. Njegovi izlazi otkrivaju informacije o položanju i gibanju značajke što je snažan prediktor za parametre deformacije rešetke kod deformabilne konvolucije.

Nastavak ovog poglavlja donosi detaljan opis preostalih dijelova prognostičkog modela: korelacijskog modula, modula F2M, modula F2F te također opis združenog prognoziranja F2MF.

5.1 Korelacijski modul

Zadaća korelacijskog modula je odrediti prostorno-vremensku korespondenciju između susjednih okvira. Na ulazu, naš korelacijski modul prima izlučene konvolucijske značajke $\mathbf{X}_{\mathtt{t-9:t:3}}$ u obliku tenzora četvrtog reda dimenzija $\mathrm{T} \times \mathrm{C} \times \mathrm{H} \times \mathrm{W}$, gdje T označava broj ulaznih okvira i obično vrijedi $T = 4$. Na izlazu daje prostorno-vremenske korelacijske koeficijente unutar lokalnog susjedstva veličine $d \times d$ za svaki od $T-1$ parova susjednih okvira. Modul prvo ugrađuje značajke iz svakog vremenskog trenutka neovisno u prostor s naglašenim metričkim svojstvima jednom dijeljenom konvolucijom s jezgrom veličine 3×3 i C' kanala (C'=128). Zadaća ove konvolucije je pokušati rekonstruirati diskriminativne značajke koje modelu za raspoznavanje nisu potrebne. Takve značajke primjerice sadrže informacije o izgledu objekta koje nisu bitne za raspoznavanje, ali jesu za uparivanje iste značajke u različitim vremenskim trenucima. Reprezentacija bogatija takvim značajkama bila bi prikladnija za uspostavu korespondencije. U sljedećem koraku se svi C'-dimenzionalni vektori značajki normaliziraju na jediničnu udaljenost čime značajke ugrađujemo u metrički prostor gdje kosinusna sličnost odgovara skalarnom umnošku. Rezultat normalizacije odgovara metričkom ugrađivanju $\mathbf{F}$ dimenzije $\mathrm{T}\times\mathrm{C'}\times\mathrm{H}\times\mathrm{W}$. Konačno, na temelju te reprezentacije određuje se d^2 korespondencijskih mapa između odgovarajućih značajki iz trenutaka τ i $\tau-3$ unutar lokalnog prostornog susjedstva $d \times d$ za svaki $\tau \in \{t-6, t-3, t\}$. Rezultantni korelacijski tenzor $\mathbf{C}^{\tau}$ dimenzija je $d^2 \times H \times W$. Uobičajena veličina prostornog susjedstva određena je s $d = 9$, a ovaj hiperparametar potrebno je prilagoditi rezoluciji prognoziranja.

Označimo s $\mathbf{C}^{\tau}_{ud+v,\mathbf{q}}$ vrijednost korelacijskog tenzora $\mathbf{C}^{\tau}$ na prostornoj lokaciji $\mathbf{q}$ i mapi značajki $ud+v$. Ta vrijednost odgovara skalarnom umnošku između vektora metričkih značajki u trenutku τ na lokaciji $\mathbf{q} \in \mathscr{D}(\mathbf{F})$ i odgovarajućeg vektora značajki u trenutku $\tau-3$ na lokaciji

$\mathbf{q} + [u - \frac{d}{2}, v - \frac{d}{2}]$ gdje su $u, v \in 0..d-1$:

$$\mathbf{C}^{\tau}_{ud+v,\mathbf{q}} = \mathbf{F}^{\tau\top}_{\mathbf{q}} \mathbf{F}^{\tau-3}_{\mathbf{q}+[u-\frac{d}{2},v-\frac{d}{2}]}, \text{za sve } u,v \in [0..d). \tag{5.1}$$

Svaka od d^2 mapa značajki tenzora $\mathbf{C}^{\tau}$ može se izračunati množenjem po elementima tenzora $\mathbf{F}^{\tau}$ i posmaknutog tenzora $\mathbf{F}^{\tau-3}$, te zbrajanjem po kanalima. Izračun ovih korelacijskih koeficijanata prepušten je korelacijskom sloju koji se često pojavljuju u modelima za optički tok [23]. Napomenimo još da $\mathscr{D}(\mathbf{F})$ označava skup svih prostornih lokacija u tenzoru $\mathbf{F}$: $\mathscr{D}(\mathbf{F})$= $\{1..\mathrm{H}\}\times\{1..\mathrm{W}\}$.

Konačni izlaz našeg korelacijskog modula odgovara tenzoru dimenzija $(T-1)\cdot d^2 \times H \times W$. On nastaje konkatenacijom svih rezultantnih korelacijskih tenzora po semantičkoj dimenziji.

5.2 Prognoziranje regresijom pomaka značajki - modul F2M

Pretpostavka modula F2M je da se budućnost u potpunosti može objasniti geometrijskom transformacijom iz prošlosti. U praksi, tu transformaciju možemo izraziti unatražnom deformacijom tenzora značajki iz prošlosti u skladu s predviđenim tokom. Modul F2M zato predviđa gusto polje vektora pomaka $\hat{\mathbf{f}}^{\tau}_{t+\Delta t}$ za svaki od T=4 ulazna tenzora značajki $\mathbf{X}_{\tau}$, $\tau \in \{t-9, t-6, t-3, t\}$. Vektore pomaka određujemo slijedom BN-RELU-DCONV s $T \cdot 2$ izlaznih kanala (množimo s 2 jer svaki vektor ima dvije komponente: pomak po x i y osi). Budući tenzori $\hat{X}^{(\tau)}_{t+\Delta t}$ dobiju se deformacijom odgovarajućih ulaznih značajki $\mathbf{X}_{\tau}$ regresiranim tokom $\hat{\mathbf{f}}^{\tau}_{t+\Delta t}$. Dakle, jedan te isti trenutak iz budućnosti pokušava se objasniti deformacijom značajki iz različitih trenutaka u prošlosti. Na taj način se zapravo dobiju četiri procjene budućeg tenzora značajki. Procjene se miješaju u skladu s predviđenim težinama na razini piksela aktiviranima funkcijom softmax. Konačno, prognozirani tenzor značajki $\hat{\mathbf{X}}^{\mathrm{F2M}}_{t+\Delta t}$ možemo izraziti kao težinsku sumu deformiranih značajki iz prošlosti:

$$\hat{\mathbf{X}}^{(\tau)}_{t+\Delta t} = \mathrm{warp\_bw}(\mathbf{X}_{\tau}, \hat{\mathbf{f}}^{\tau}_{t+\Delta t}) \tag{5.2}$$

$$\hat{\mathbf{X}}^{\mathrm{F2M}}_{t+\Delta t} = \sum_{\tau} \alpha_{\tau} \cdot \hat{\mathbf{X}}^{(\tau)}_{t+\Delta t} \tag{5.3}$$

$$\boldsymbol{\alpha} = \mathrm{softmax}([w^{\mathrm{F2M}}_{\tau}]_{\tau\in\{t-9:t:3\}}) \,. \tag{5.4}$$

Kombiniranje prognoza iz različitih trenutaka u prošlosti omogućuje odabir najprikladnijeg prošlog okvira za objašnjavanje nekog dijela buduće scene. Ovo može biti posebno korisno u nekim slučajevima zaklanjanja kao što je ilustrirano u slici 6.9.

U ovoj disertaciji razmatrali smo i inačicu modula F2M koja umjesto unatražnog prognozira

unaprijedni tok u trenucima $\tau \in \{t-9, t-6, t-3, t\}$:

$$\hat{\mathbf{f}}_{\tau}^{t+\Delta t} = \mathrm{F2M}_{\mathrm{fw}}^{(\tau)}(\mathbf{X}_{\mathtt{t-9:t:3}}) \tag{5.5}$$

U tom slučaju na sva četiri prošla tenzora značajki primjenjuje se unaprijedno deformiranje odgovarajućim tokom:

$$\hat{\mathbf{X}}_{t+\Delta t}^{(\tau)} = \mathrm{warp\_fw}(\mathbf{X}_{\tau}, \hat{\mathbf{f}}_{\tau}^{t+\Delta t}) \tag{5.6}$$

Unaprijedna deformacija stvara buduće aktivacije razmazivanjem [44, 69] značajki iz prošlosti na nediskretnim lokacijama određenim s prognoziranim unaprijednim tokom. Operacija unaprijedne deformacije nije podržana u popularnim okvirima za automatsku diferencijaciju, te stoga unatražna deformacija predstavlja naš podrazumijevani odabir. Ipak, za potrebe istraživanja predlažemo naivnu implementaciju zasnovanu na matričnom množenju. Predložena implementacija može se jednostavno izraziti u bilo kojem okviru za automatsku diferencijaciju. Naša implementacija pretpostavlja da svaki izvorišni piksel doprinosi svim odredišnim pikselima s nekom težinom izraženom jezgrenom funkcijom između odredišne i posmaknute izvorišne lokacije:

$$\hat{\mathbf{X}}_{t+\Delta t}^{(\tau)}[\mathbf{q}] = \frac{1}{N_{\mathbf{q}}} \sum_{\mathbf{p} \in \mathscr{D}(\mathbf{X})} k(\mathbf{p} + \hat{\mathbf{f}}_{\tau}^{t+\Delta t}[\mathbf{p}], \mathbf{q}) \cdot \mathbf{X}_{\tau}[\mathbf{p}]. \tag{5.7}$$

Jezgrena funkcija se brine da minimizira utjecaj jako dalekih pogodaka i naglasi utjecaj bliskih. U jednadžbi iznad, k predstavlja jezgru RBF, a $N_{\mathbf{q}}$ je normalizirajući faktor koji osigurava da norma prognoziranih značajki ostane unutar uobičajenog intervala:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right), \tag{5.8}$$

$$N_{\mathbf{q}} = \sum_{\mathbf{p} \in \mathscr{D}(\mathbf{X})} k(\mathbf{p} + \hat{\mathbf{f}}_{\tau}^{t+\Delta t}[\mathbf{p}], \mathbf{q}). \tag{5.9}$$

Predložena implementacija unaprijednog deformiranja je jako memorijski neefikasna zbog pretpostavke globalnog utjecaja, koja je pak omogućila jednostavnu implementaciju matričnim množenjem. Kasnije su se pojavile i efikasnije implementacije unaprijedne deformacije, optimizirane za grafičke procesore [69]. Ipak, zbog male rezolucije značajki i ova implementacija bila je dovoljna za razmatranje razlike između unaprijedne i unatražne varijante modula F2M.

Spomenute formulacije su prilično različite. Prilikom prognoziranja budućeg toka, konvolucijski sloj je kod unaprijedne varijante poravnat sa pozicijom značajke u prošlosti, dok je kod unatražne varijante poravnat s lokacijom značajke u budućnosti. Budući da se tok prognozira na temelju reprezentacije iz prošlosti, čini se da je unatražnoj varijanti potreban model s većim

receptivnim poljem jer mora biti u mogućnosti pronaći lokaciju odgovarajuće značajke na posmaknutoj lokaciji. Mogli bismo reći da unatražna varijanta adresira *efekte* gibanja: ona donosi odluku pretražujući sve moguće aktivacije koje bi mogle upasti u neku konkretnu lokaciju u budućem tenzoru. Stoga, unatražna varijanta uz dovoljno receptivno polje ima veće mogućnosti razriješiti problem zaklanjanja. S druge strane, unaprijedna varijanta adresira *uzroke* gibanja: ona donosi odluke razmatrajući opaženo gibanje značajke. Stoga, čini se da bi unaprijedni F2M mogao modelirati probabilističku distribuciju mogućih budućih lokacija, što bi bilo prikladno za dugoročno ili multimodalno prognoziranje [9].

Ipak, obje varijante pretpostavljaju da se budućnost može objasniti "preslagivanjem" prošlosti, što je samo djelomično točno, jer budućnost je često nepredvidiva i otvara do tada neviđene dijelove scene. Dodatno, u nekim slučajevima je teško uspostaviti korespondenciju zbog velikog gibanja kamere i promjene u perspektivi. Kao posljedica navedenog, neki dijelovi scene su posebno teško objašnjivi jednostavnim deformiranjem iz prošlosti. Ispravno prognoziranje u takvim slučajevima zahtijeva više slobode, a možda čak i mogućnost zamišljanja. Zato, predložena metoda prognoziranja značajki kombinira modul F2M s modulom F2F kojeg opisujemo u nastavku.

5.3 Izravno prognoziranje značajki - modul F2F

Modul F2F izravno prognozira buduće značajke $\hat{\mathbf{X}}^{\text{F2F}}_{t+\Delta t}$ na temelju zajedničke reprezentacije dijeljene s modulom F2M i težinskim modulom. Modul F2F sadrži jednu BN-ReLU-DCONV jedinicu gdje je broj izlaznih kanala jednak broju kanala jednog ulaznog tenzora značajki iz prošlosti. On nije ograničen na rekonstrukciju iz prošlosti, te zbog toga ima mogućnost razviti sposobnost zamišljanja u novootkrivenim dijelovima scene. Ovaj modul je zapravo sličan modulima za prognoziranje značajki kakvi se koriste u literaturi [6, 16, 101], ali postoje bitne razlike. Prvo, naše izravno prognoziranje značajki uvijek cilja jednorazinsku reprezentaciju, za razliku od primjena iz literature koji prognoziraju piramidu značajki [6, 16]. Drugo, umjesto dilatiranih ili regularnih konvolucija mi koristimo geometrijski moćnije deformabilne konvolucije. Treća razlika je u tome što naš modul F2F ima pristup prostorno-vremenski korelacijskim značajkama, koje smanjuju pritisak učenja korespondencije iz nule kroz konvolucijske slojeve. Eksperimenti potvrđuju jasnu prednost ovih predstavljenih noviteta.

5.4 Združeno prognoziranje regresijom pomaka i značajki - F2MF

Naša je hipoteza da su izravno prognoziranje značajki (F2F) i prognoziranje te deformacija pomakom (F2M) komplementarni, te da bi kombinacija dvaju pristupa trebala unaprijediti točnost. Stoga, konačne prognozirane značajke predloženog modela F2MF zapravo su težinska suma prognoza modula F2M i F2F (kao i prije, vrijedi $\tau \in \{t-9:t:3\}$):

$$\hat{\mathbf{X}}^{\text{F2MF}}_{t+\Delta t} = \beta^{\text{F2F}} \cdot \hat{\mathbf{X}}^{\text{F2F}}_{t+\Delta t} + \sum_{\tau} \beta^{\text{F2M}}_{\tau} \cdot \hat{\mathbf{X}}^{(\tau)}_{t+\Delta t} \tag{5.10}$$

$$\boldsymbol{\beta} = [\beta^{\text{F2F}}] \parallel [\beta^{\text{F2M}}_{\tau}] = \text{softmax}([w^{\text{F2F}}] \parallel [w^{\text{F2M}}_{\tau}]) \,. \tag{5.11}$$

Primijetite da je sada funkcija softmax primijenjena preko pet težinskih faktora: četiri za F2M prognoze iz četiri tenzora prošlosti i jedan za F2F prognozu. To za posljedicu omogućuje podjelu odgovornosti i potiče specijalizaciju modula F2F za prognoziranje u novootkrivenim dijelovima scene. To pak utječe i na modul F2M jer smanjuje njegovu kaznu u takvim regijama i područjima gdje je onemogućena uspostava korespondencije, a istovremeno ga specijalizira za prognoziranje u regijama gdje je to moguće napraviti. Ipak, ovakva podjela odgovornosti može oslabiti i prorijediti signal za učenje ova dva individualna modula. Kako bi se to izbjeglo, prilikom učenja združenog modela za prognoziranje značajki F2MF, funkcija gubitka ima tri komponente. Glavna komponenta $\mathscr{L}_{\text{F2MF}}$ odnosi se na združenu prognozu našom metodom F2MF (5.10). Dvije dodatne komponente $\mathscr{L}_{\text{F2F}}$ i $\mathscr{L}_{\text{F2M}}$ primjenjuju se na odgovarajuće izlaze $\hat{\mathbf{X}}^{\text{F2F}}_{t+\Delta t}$ i $\hat{\mathbf{X}}^{\text{F2M}}_{t+\Delta t}$ individualnih modula. Jedinstvena funckija gubitka dobije se težinskim zbrajanjem navedenih komponenti:

$$\mathscr{L} = \mathscr{L}_{\text{F2MF}} + \lambda_{\text{F2F}}\mathscr{L}_{\text{F2F}} + \lambda_{\text{F2M}}\mathscr{L}_{\text{F2M}} \tag{5.12}$$

Sva tri gubitka izražena su kao prosječna L2 udaljenost između odgovarajućih prognoziranih značajki i značajki izlučenih primjenom odgovarajućeg modela za gusto raspoznavanje u budućem ciljnom okviru. Takav oblik stvaranja točnih predikcija omogućuje samonadzirano učenje prognostičkog modela na neoznačenom videu. Dakle, prilikom treniranja sustava za prognoziranje ljudske anotacije su potrebne samo za učenje modela za gusto raspoznavanje u jednoj slici. To svojstvo koristimo koristimo i u praksi. Modele za gusto raspoznavanje učimo na označenim slikama, a prognostički model na neoznačenom videu.

Poglavlje 6

Eksperimenti

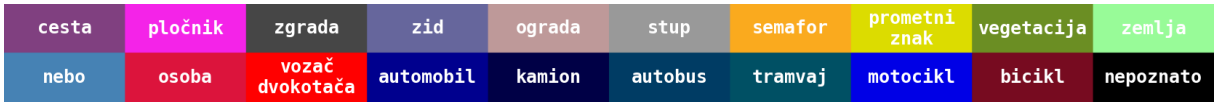
Za treniranje i evaluaciju sustava za gusto semantičko prognoziranje potreban je skup podataka koji sadrži videoisječke s barem jednim označenim okvirom. Treniranje našeg sustava za gusto semantičko prognoziranje odvija se u nekoliko faza. U prvoj fazi se okosnica koja odgovara klasfikacijskoj konvolucijskoj arhitekturi predtrenira na ImageNetu. U drugoj i trećoj fazi redom se treniraju model za gusto raspoznavanje u jednoj slici i prognostički model na istom skupu podataka. Jedan primjerak za učenje modela za gusto raspoznavanje sastoji se od ulazne slike i odgovarajućih gustih oznaka. Guste oznake su obično nastale ljudskim anotiranjem i potrebne su samo u ovoj fazi treniranja. Detalji procesa treniranja modela za gusto raspoznavanje ovise o ciljnom zadatku. Jedan primjerak za učenje prognostičkog modela sastoji se od nekoliko (obično 4) pravilno raspoređenih slikovnih okvira iz prošlosti, te odgovarajućeg slikovnog okvira iz budućnosti. Sustav za prognoziranje obično se testira za različite vremenske udaljenosti budućeg okvira, a tome se onda prilagođava i uzorkovanje primjeraka za učenje. Prognostički model može se učiti na neoznačenim slikama, jer je njegova zadaća mapirati izlučene značajke iz prošlosti u budućnost. U ovoj fazi su parametri modela za raspoznavanje u jednoj slici zaključani i služe samo izlučivanju značajki iz ulaznih slika. Zato je ovu fazu moguće značajno ubrzati spremanjem izlučenih značajki na brzi SSD disk, čime se izbjegava potreba za opetovanim izlučivanjem istih značajki iz istih ulaznih slika. Ipak, spremanje značajki u trajnu memoriju onemogućuje primjenu tehnika rastresanja podataka. Napomenimo još jedanput kako se ovo istraživanje usredotočuje na gusto semantičko prognoziranje na temelju ulaznih slika, te se zbog toga ne koriste dodatni ulazi poput odometrije vozila ili dubine scene.

Točnost gustog semantičkog prognoziranja mjeri se usporedbom prognoziranih predikcija i točnih oznaka u budućem neopaženom slikovnom okviru. Evaluacijske metrike odgovaraju uobičajenim metrikama za ciljni zadatak koje se koriste prilikom evaluacije gustog raspoznavanja u jednoj slici. Za semantičku segmentaciju to je usrednjeni omjer presjeka i unije (eng. *mean intersection over union*, mIoU). Za segmentaciju instanci mjeri se prosječna preciznost (eng. *average precision*, AP) za različite pragove preklapanja između točnih i prediktiranih ma-

ski. Ta mjera poznata je i kao COCO AP, jer je prvi put predstavljena kao evaluacijska metrika za podatkovni skup COCO [106]. Dodatno se mjeri i prosječna preciznost za prag preklapanja jednak 0.5 (AP50). Točnost panoptičke segmentacije mjeri se panoptičkom kvalitetom (eng. *panoptic quality*, PQ) [54]. Panoptička kvaliteta se dodatno često faktorizira na kvalitetu segmentacije (eng. *segmentation quality*, SQ) i kvalitetu raspoznavanja (eng. *recognition quality*, RQ). Gornju granicu točnosti prognoziranja kod svih eksperimenata određuje model prorok. Model prorok odgovara primjeni odgovarajućeg modela za gusto raspoznavanje u budućoj slici. To je, dakle, model koji vidi budućnost te otuda i ime - prorok. Njegova performansa zapravo predstavlja gornju granicu točnosti koju možemo postići unaprijeđenjem metode prognoziranja značajki. Kada bi prognoziranje značajki radilo savršeno, točnost prognoziranja bila bi jednaka točnosti modela proroka. Minimalnu točnost prognoziranja koju očekujemo postići određuje osnovni prognostički model koji kopira semantičke predikcije iz posljednjeg opaženog okvira. Pretpostavka toga modela je dakle da je scena u potpunosti statična i da se od posljednjeg opažanja do ciljanog budućeg trenutka ništa neće promijeniti.

6.1 Izvedbeni detalji prognoziranja na podatkovnom skupu Cityscapes

Podatkovni skup Cityscapes [8] sadrži 2975 fino označenih slika za treniranje, 500 za validaciju i 1525 za testiranje. Slike su rezolucije 1024×2048 piksela i dolaze u paru s oznakama za semantičku segmentaciju, segmentaciju instanci i panoptičku segmentaciju. Svaka označena slika odgovara 20. slikovnom okviru u kratkom video isječku koji sadrži ukupno 30 sličica. Video je sniman frekvencijom od 17Hz što znači da je trajanje jednog video isječka približno jednako 1.8 sekundi. Činjenica da su označene slike dio video isječka čini Cityscapes prikladnim skupom podataka za prognostičke eksperimente. Semantičko taksonomija na Cityscapesu sadrži ukupno 33 razreda, ali službeni evaluacijski protokol sužava taj skup na 19 semantičkih razreda. Slika 6.1 prikazuje imena svih razreda u pravokutniku odgovarajuće boje koja se koristi prilikom vizualizacije gustih predikcija na ovome skupu. Ne brojeći razred "nepoznato", posljednjih 8 razreda sa slike su prebrojivi, dok su ostali razredi neprebrojivi.



Slika 6.1: Semantički razredi podatkovnog skupa Cityscapes i njihove odgovarajuće kodne boje.

U literaturi je uobičajeno evaluirati prognoziranje na Cityscapesu obzirom na vremenski razmak Δt između posljednjeg opaženog i budućeg neopaženog slikovnog okvira. Razlikujemo kratkoročno prognoziranje (3 sličice unaprijed, odnosno 0.18 sekundi) i srednjoročno prognozi-

ranje (9 sličica unaprijed, odnosno 0.54 sekunde). U oba slučaja, uobičajeno je da ulaz u model čine četiri slikovna okvira međusobno razmaknuta po tri sličice.

Naši modeli za semantičku segmentaciju jedne slike treniraju se 250 epoha postupkom optimizacije ADAM [107] na minigrupama veličine 12 koje sadrže isječke slika veličine 768×768 piksela. Stopa učenja se kosinusnim kaljenjem smanjuje od $4 \cdot 10^{-4}$ do 10^{-7}. Za ažuriranje parametara okosnice koristi se $4\times$ manja stopa učenja. Podatkovni skup se umjetno uvećava horizontalnim zrcaljenjem i slučajnim skaliranjem veličine slike.

Model za segmentaciju instanci Mask R-CNN C4 trenira se 30000 iteracija s minigrupom veličine 6 koja sadrži slike pune rezolucije. Težine se ažuriraju stohastičkim gradijentim spustom sa stopom učenja 10^{-2}, koja se nakon 21000 iteracija smanji na 10^{-3}. Za rastresanje podataka koristi se horizontalno zrcaljenje i slučajno skaliranje ulazne slike.

Modificirani Panoptički DeepLab trenira se 90000 iteracija na minigrupama veličine 4 koje sadrže slike pune rezolucije. Za optimizaciju težina koristi se ADAM sa stopom učenja $5 \cdot 10^{-5}$ koja se polinomijalno smanjuje do 10^{-7}. Za rastresanje podataka koristi se horizontalno zrcaljenje i slučajno skaliranje ulazne slike.

Model za prognoziranje F2MF trenira se 160 epoha optimizatorom ADAM [107] i ranim zaustavljanjem. Stopa učenja se kosinusnim kaljenjem smanjuje od $4 \cdot 10^{-4}$ do 10^{-7}. Doprinosi svih komponenti gubitka postavljeni su na 1. Model je implementiran u radnom okviru za automatsku diferencijaciju Pytorch [108]. Treniranje modela sa unaprijed izračunatim značajkama i spremljenim na brzi disk traje 12 sati na jednoj grafičkoj kartici GTX1080Ti. Treniranje modela sa rastresanjem podataka traje 48 sati na tri grafičke kartice. Korištena tehnika rastresanja podataka uključuje horizontalno zrcaljenje i vremensko pomicanje ulaznog primjera po odgovarajućem video isječku. Važno je napomenuti da se prilikom treniranja koriste značajke koje su normalizirane srednjom vrijednosti i standardnom devijacijom izračunatom na cijelom skupu za učenje. Normalizacija ubrzava proces treniranja i pospješuje generalizaciju. Prilikom evaluacije, prognozirane značajke treba denormalizirati kako bi odgovarale distribuciji koju očekuje modul za formiranje semantičkih predikcija.

6.2 Prognoziranje semantičke segmentacije na skupu Cityscapes

Prikladnost modela F2MF za prognoziranje semantičke segmentacije provjeravamo u kombinaciji s dva različita modela za raspoznavanje u jednoj slici. Modeli su detaljno opisani u poglavlju 4, a ovdje se podsjetimo da se radi o modelima različitog kapaciteta. Snažniji model sa okosnicom DenseNet-121 postiže 75.8 mIoU bodova na validacijskom skupu Cityscapesa, dok slabiji model sa okosnicom ResNet-18 postiže 72.5 mIoU boda.

Tablica 6.1 evaluira točnost prognoziranja na validacijskom skupu Cityscapesa. Tablica je

Tablica 6.1: Evaluacija modula F2MF za prognoziranje semantičke segmentacije na skupu za validaciju podatkovnog skupa Cityscapes. *Svi* označava sve razrede, *PO* — razrede pokretnih objekata, *r.p.* — rastresanje podataka, i †— testni skup.

	Kratkoročno: Δt=3		Srednjoročno: Δt=9	
Točnost (mIoU)	Svi	PO	Svi	PO
Prorok-DN121	75.8	75.2	75.8	75.2
Prorok-RN18	72.5	71.5	72.5	71.5
Kopiranje posljednje segmentacije (DN121)	53.3	48.7	39.1	29.7
3Dconv-F2F [101]	57.0	/	40.8	/
Dil10-S2S [5]	59.4	55.3	47.8	40.8
LSTM S2S [88]	60.1	/	/	/
Mask-F2F [6]	/	61.2	/	41.2
FeatReproj3D [100]	61.5	/	45.4	/
Bayesian S2S [90]	65.1	/	51.2	/
DeformF2F [13]	65.5	63.8	53.6	49.9
LSTM AM S2S [89]	65.8	/	51.3	/
LSTM M2M [48]	67.1	65.1	51.5	46.3
LSTM-VAE-F2F [95]	71.1	69.2	60.3	56.7
F2MF-RN18 bez r.p.	66.9	65.6	55.9	52.4
F2MF-DN121 bez r.p.	68.7	66.8	56.8	53.1
F2MF-DN121 s r.p.	69.6	67.7	57.9	54.6
F2MF-DN121 s r.p. †	70.2	68.7	59.1	56.3

podijeljena u tri segmenta. Prvi segment predstavlja rezultate modela proroka te naivnog modela za prognoziranje koji jednostavno kopira segmentaciju posljedneg opaženog slikovnog okvira. Drugi segment predstavlja rezultate iz literature. Treći segment predstavlja naše rezultate sa različitim segmentacijskim modelima i tehnikama rastresanja podataka.

Očekivano, točnost prognoziranja opada s vremenskom udaljenošću budućeg trenutka, pa su rezultati u kratkoročnom prognoziranju znatno bolji, nego u srednjoročnom. U svim slučajevima, točnost prognoziranja pokretnih objekata manja je od prosjeka. Taj efekt posebno je naglašen u srednjoročnom prognoziranju. U vrijeme objavljivanja na konferenciji CVPR 2020 naša metoda F2MF-DN121 s rastresanjem podataka postizala je najbolje rezultate prog-

noziranja semantičke segmentacije na Cityscapesu. Trenutno najbolje rezultate postiže metoda LSTM-VAE-F2F [95] koja je predstavljena na konferenciji ICCV 2021. Ta metoda prognozira 16× poduzorkovane značajke kodirane varijacijskim autoenkoderom. Ipak, naša metoda je efikasnija jer prognozira značajke manje rezolucije i ne koristi nikakav oblik dodatnog kodiranja značajki. Iz tablice se također može primijetiti kako bolja točnost modela za raspoznavanje u jednoj slici vodi ka boljem prognoziranju. Ipak, pri prognoziranju je razlika manja nego pri segmentaciji jedne slike. Razlika između naših modela za predikciju u jednoj slici je 3.3 mIoU boda, dok je razlika u odgovarajućim modelima za kratkoročno prognoziranje 1.8 mIoU bodova, a za srednjoročno 0.9 bodova. Utjecaj rastresanja podataka je također jasno vidljiv. Model s rastresanjem podataka postiže oko 1 mIoU bod bolju točnost prognoziranja. Posljednji redak u tablici otkriva performansu našega modela na skupu za testiranje. Model postiže 70.2 i 59.1 mIoU boda u kratkoročnom i srednjoročnom prognoziranju. Takvi rezultati sugeriraju da predstavljeni rezultati na validacijskom skupu nisu rezultat prenaučenosti ili pristranosti.

Tablica 6.2 uspoređuje točnost po razredima modela proroka, kratkoročnog i srednjoročnog prognoziranja metodom F2MF. Može se primijetiti kako je najteže prognozirati razred stup. Razlog za to leži u činjenici da su stupovi izduženi u smjeru okomitom na gibanje. Zbog toga čak i najmanje greške pri lokalizaciji za posljedicu imaju pogrešnu klasifikaciju cijelog objekta. Stoga se prognostički modeli često odlučuju propustiti prognozirati stupove kako bi izbjegli dvostruku kaznu (mIoU broji i lažne pozitive i lažne negative). Ovo pokazuje i usporedba incidencije razreda stup u predikcijama modela proroka (1.14%) s odgovarajućim statistikama u kratkoročnom (1.00%) i srednjoročnom prognoziranju (0.69%). Od razreda koji predstavljaju pokretne objekte, razred osoba uzrokuje najveću degradaciju točnosti prilikom prognoziranja: 13.5 bodova prilikom kratkoročnog i 32.3 boda prilikom srednjoročnog prognoziranja. Ovi rezultati pokazuju da su gibanja ljudi prognostičkom modelu manje predvidljiva od gibanja vozila. Gibanje svakog čovjeka je specifično i može uzrokovati puno zaklanjanja zbog kretanja u grupama. Vertikalno izduženi oblik ljudi dovodi do sličnih problema kao u slučaju stupova. Ove činjenice pokazuju da je prognoziranje buduće poze i lokacije kod ljudi posebno izazovno.

Tablica 6.2: Točnost po razredima (IoU) na validacijskom skupu podatkovnog skupa Cityscapes. Tablica pokazuje rezultate modela proroka te prognostičkog modela F2MF na kratkoročnom i srednjoročnom prognoziranju.

	cesta	pločnik	zgrada	zid	ograda	stup	semafor	prometni znak	vegetacija	zemlja	nebo	osoba	vozač dvokotača	auto	kamion	bus	tramvaj	motocikl	bicikl	prosjek
Model prorok (DN121)	**97.8**	**82.9**	**91.8**	**60.1**	**59.4**	**59.8**	**65.0**	**74.2**	**91.4**	**62.0**	**93.5**	**78.2**	**58.4**	**94.2**	**80.8**	**85.0**	**68.9**	**61.6**	**73.8**	**75.8**
F2MF-DN121 Kratkoročno	96.7	76.5	89.0	57.8	56.5	44.2	57.5	63.9	88.5	59.0	90.4	64.7	49.8	88.8	77.5	81.3	63.2	50.5	65.2	69.6
F2MF-DN121 Srednjoročno	94.6	66.4	83.0	50.6	49.9	19.2	38.4	42.9	81.9	51.5	83.6	45.9	30.5	78.4	71.1	73.1	47.6	41.0	48.8	57.9

6.3 Prognoziranje segmentacije primjeraka na skupu Cityscapes

Tablica 6.3 predstavlja rezultate prognoziranja segmentacije instanci na validacijskom skupu Cityscapesa. Kao i u prethodnom slučaju tablica je podijeljena u tri dijela: rezultate proroka i naivnog modela, rezultate iz literature te naše rezultate. Naš model za raspoznavanje u jednoj slici (prorok) odgovara inačici C4 modela Mask R-CNN koja omogućava jednorazinsko prognoziranje značajki, kao što je opisano u poglavlju 4. Metode iz literature se oslanjaju na inačicu FPN modela Mask R-CNN koja zahtijeva računski složeno prognoziranje na različitim rezolucijama. Razlika u točnosti između ta dva modela za analizu jedne slike je 1 AP bod. Takav odnos prognostičkim metodama iz literature daje blagu prednost u točnosti nauštrb smanjene efikasnosti. Eksperimenti otkrivaju da naša metoda F2MF uspijeva nadoknaditi taj gubitak i postići najbolju točnost u kratkoročnom prognoziranju. Rezultati na srednjoročnom prognoziranju su usporedivi s literaturom. Dodatno, ablacijski eksperiment (F2F-Corr) u kojem smo uklonili modul F2M otkriva važnost prognoziranja deformiranjem značajki iz prošlosti, odnosno njegove kombinacije sa izravnim F2F prognoziranjem. Model bez F2M modula postiže 2.4 AP boda manju točnost u kratkoročnom prognoziranju i 2.1 bod u srednjoročnom.

Tablica 6.3: Prognoziranje segmentacije instanci na validacijskom skupu podatkovnog skupa Cityscapes. Tablica pokazuje uobičajene osnovice (gore), postupke iz literature (sredina) te naše postupke (dolje).

	Kratkoročno: Δt=3		Srednjoročno: Δt=9	
	AP	AP50	AP	AP50
Prorok (naš)	36.3	63.1	36.3	63.1
Prorok FPN [6, 16, 98]	37.3	65.8	37.3	65.8
Kopiranje posljednje segmentacije	9.6	22.9	2.2	8.1
F2F 4× [6]	19.4	39.9	7.7	19.4
ConvLSTM F2F 4× [16]	22.1	44.3	11.2	25.6
ApaNet [98]	23.2	46.1	**12.9**	**29.2**
F2F-Corr (naš)	21.2	43.3	9.4	19.2
F2MF (naš)	**23.6**	**47.2**	11.5	24.2

6.4 Panoptičko prognoziranje na skupu Cityscapes

Tablica 6.4 prikazuje rezultate prognoziranja panoptičke segmentacije na validacijskom podskupu podatkovnog skupa Cityscapes. Naš sustav za prognoziranje panoptičke segmentacije koristi modificiranu verziju modela Panoptic Deeplab sa samo jednom preskočnom vezom između okosnice i dekodera. Modifikacija omogućuje jednorazinsko prognoziranje značajki koje su 16× poduzorkovane u odnosu na ulaznu sliku kao što smo objasnili u poglavlju 4. Na lanjskom CVPR-u predstavljena je srodna metoda [7] koja se zasniva na originalnom Panoptičkom DeepLabu. Mogućnost jednorazinskog prognoziranja plaćamo lošijom performansom raspoznavanja u jednoj slici i to za 2.3 PQ boda. Naša metoda uvjerljivo pobjeđuje naivnu metodu koja kopira panoptičku segmentaciju posljednjeg viđenog slikovnog okvira. U usporedbi s metodom iz literature [7] naš pristup ima nešto manju točnost. Ipak, važno je naglasiti da metoda iz literature koristi odometriju i informacije o dubini scene. Kao i kod segmentacije instanci, ablacijski eksperiment pokazuje važnost modula F2M za točnost prognoziranja.

Tablica 6.4: Prognoziranje panoptičke segmentacije na validacijskom podskupu Cityscapesa. Tablica pokazuje uobičajene osnovice (gore), postupke iz literature (sredina) te naše postupke (dolje).

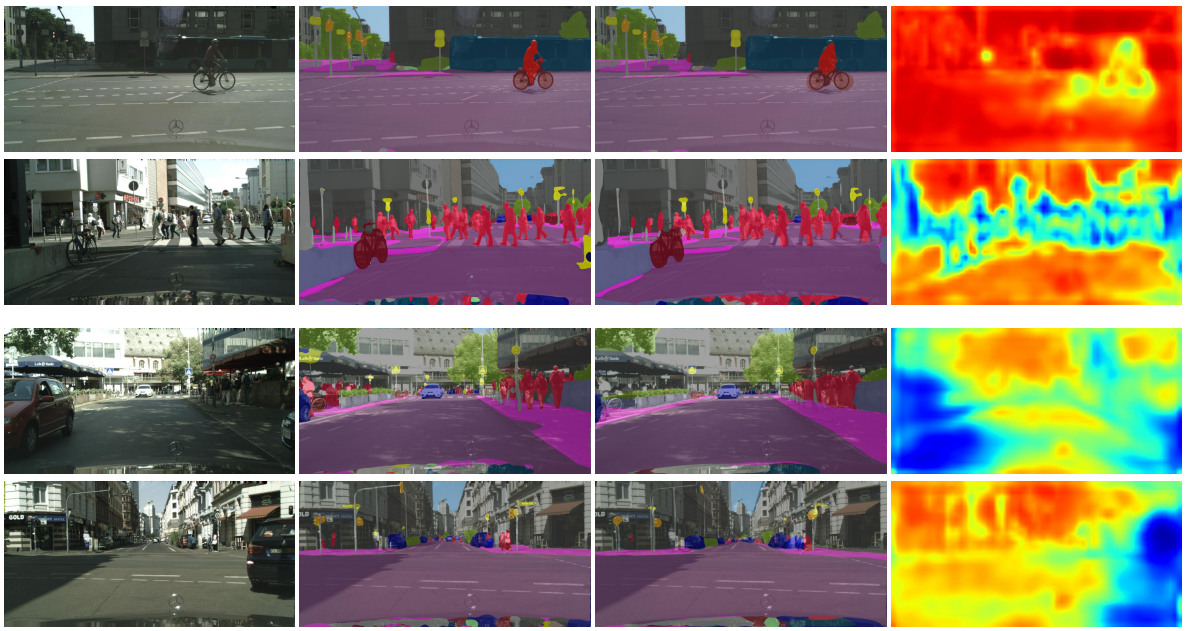
	Kratkoročno: Δt=3			Srednjoročno: Δt=9		
	PQ	SQ	RQ	PQ	SQ	RQ
Prorok (naš)	57.5	79.7	70.8	57.5	79.7	70.8
Prorok (PDL) [7, 21]	59.8	80.0	73.5	59.8	80.0	73.5
Kopiranje posljednje segmentacije	32.3	70.9	42.4	22.3	68.1	30.2
IndRNN-Stack [7]	49.0	74.9	63.3	36.3	71.3	47.8
F2F-Corr (naš)	43.3	74.4	55.5	27.5	69.7	36.0
F2MF (naš)	47.3	75.1	60.6	33.1	71.3	43.3

6.5 Kvalitativni eksperimenti

Rezultate prognoziranja važno je razmatrati i kvalitativno. Slike 6.2, 6.3 i 6.4 prikazuju primjere prognoziranja na Cityscapesu za zadatke semantičke segmentacije, segmentacije instanci te panoptičke segmentacije. Svaka slika prikazuje četiri primjera: po dva za kratkoročno i srednjoročno prognoziranje. Značenje stupaca je također zajedničko za sve tri spomenute slike. Prvi stupac prikazuje posljednji opaženi slikovni okvir iz prošlosti. Drugi i treći stupac prikazuju se-

mantičke predikcije modela proroka i prognostičkog modela. Budući slikovni okviri iscrtani su ispod semantičkih predikcija u svrhu vizualizacije i lakšeg uočavanja kvalitete prognoziranja. Napominjemo kako ova disertacija razmatra samo gusto semantičko prognoziranje, te se ne bavi prognoziranjem izgleda budućih slikovnih okvira. Posljednji stupac prikazuje toplinsku mapu faktora miješanja β^{F2M}. Toplinska mapa otkriva za koje je regije prilikom prognoziranja bio zadužen modul F2M (crvena boja), a za koje modul F2F (plava boja). Regresija faktora miješanja provodi se na $32\times$ ili $16\times$ poduzorkovanoj rezoluciji, pa naduzorkovanjem na rezoluciju slike izgledaju grubo. Ipak, mogu se uočiti neke pravilnosti koje ćemo detaljnije opisati u nastavku.

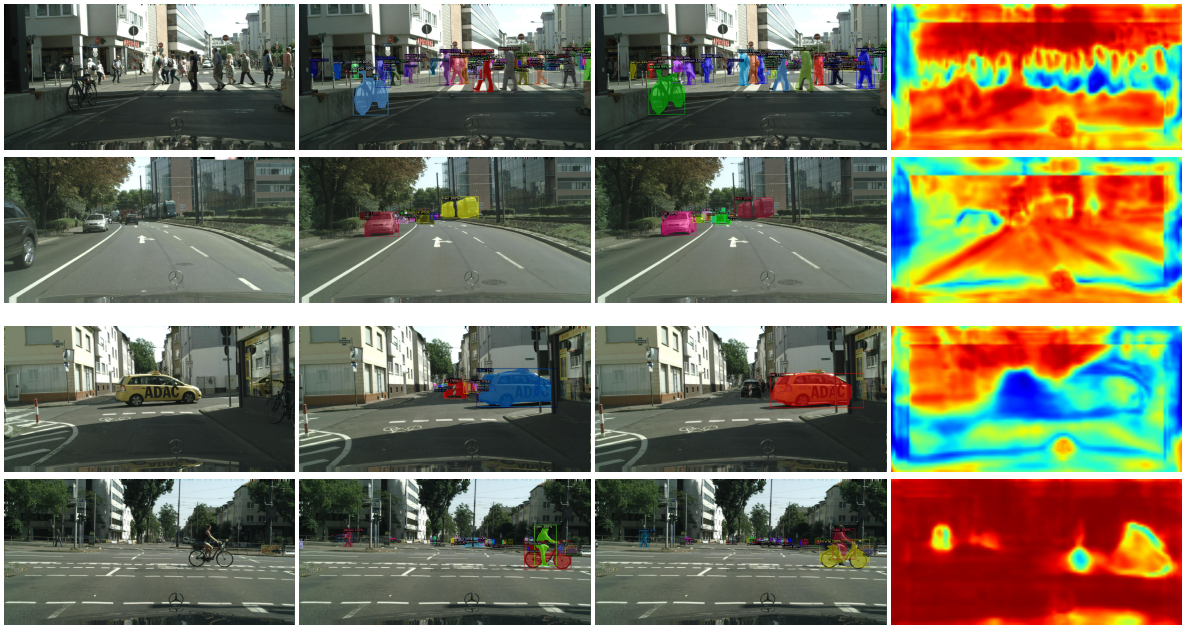
Slika 6.2 prikazuje rezultate prognoziranja semantičke segmentacije. Prvi primjer prikazuje scenu u kojoj se biciklist kreće u desno, dok ostatak scene miruje. Usporedbom predikcija modela proroka i prognostičkog modela može se primijetiti kako je prognoza prilično točna. Dominantno crvena toplinska mapa faktora miješanja otkriva da je za većinu piksela u sceni bio zadužen modul F2M, što je razumljivo obzirom da su pomaci bili mali i bilo je lako uspostaviti korespondenciju. Scena iz drugog retka je bitno drugačija jer prikazuje više pješaka koji prelaze cestu, svaki sa svojim pravcem kretanja. Može se primijetiti jasan pad oštrine i točnosti prognoziranih predikcija. Jasno je da modelu najviše problema stvaraju pješaci jer se svaki ud giba zasebno. Model je u tim dijelovima nesiguran i to se jasno vidi na toplinskoj mapi, jer se prognoziranje u regijama s naglašenim gibanjem prepušta moćnijem modulu F2F. Vjerujemo da su ovi problemi posebno izraženi zbog porijekla značajki koje se prognoziraju. One su izlučene



Slika 6.2: Kvalitativni rezultati prognoziranja semantičke segmentacije na Cityscapesu. Slika prikazuje dva primjera kratkoročnog prognoziranja (gore) i dva primjera srednjoročnog prognoziranja (dolje). Stupci prikazuju posljednji opaženi slikovni okvir, predikcije modela proroka u budućem okviru, našu prognozu te toplinsku mapu faktora miješanja β^{F2M}. Samo pri vizualizaciji, budući slikovni okviri iscrtani su ispod semantičkih predikcija kako bi se lakše pratila točnost naše prognoze.

modelom za semantičku segmentaciju koji ne razlikuje različite primjerke objekata, dok je za prognoziranje to nužno jer svaki od njih ima svoj uzorak kretanja. Na scenama iz srednjoročnog gibanja vidljiv je veći pomak objekata. Oba primjera ilustriraju slučajeve u kojima automobil napušta scenu i iza sebe otkriva novi veliki dio scene, koji model nije vidio u niti jednom od opaženih okvira iz prošlosti. U trećem retku radi se o crvenom automobilu koji se mimoilazi s kamerom, a u četvrtom retku o tamnom automobilu koji skreće udesno. Toplinska mapa otkriva velike regije plave boje koje se u oba slučaja podudaraju s pozicijom nestajućeg automobila iz posljednjeg opaženog okvira. To sugerira da je kombinirani F2MF model prepoznao tu situaciju i prognozu prepustio modulu F2F koji ima mogućnost zamišljanja. U sceni iz trećeg retka zamišljeni dio scene samo je djelomično točan, jer se vidi da je automobil otvorio pogled na bicikl i čovjeka koji su se nalazili iza. Prognostički model je to teško mogao znati, ali ovaj primjer ilustrira važnost prepoznavanja takvih regija. To su regije s visokom nepouzdanosti koje bi prognostički modeli trebali detektirati. Toplinsku mapu faktora miješanja možemo smatrati samo grubom procjenom detekcije takvih regija i kao poticaj za buduće istraživanje.

Slika 6.3 vizualizira prognoziranje segmentacije instanci. Prva scena odgovara sceni iz drugog retka slike 6.2. Može se primijetiti kako su ljudi u ovome slučaju točnije segmentirani. To je posebice vidljivo na primjerima gdje je čak i položaj noge u koraku ispravno prognoziran. Razlog za to može se tražiti u činjenici da se ovdje radi o prognoziranju značajki modela za segmentaciju instanci. Značajke toga modela sadrže diskriminativne informacije koje su važne za razlikovanje primjeraka objekata, pa je olakšano i praćenje gibanja svakog objekta zasebno.



Slika 6.3: Kvalitativni rezultati prognoziranja segmentacije instanci na Cityscapesu. Slika prikazuje dva primjera kratkoročnog prognoziranja (gore) i dva primjera srednjoročnog prognoziranja (dolje). Stupci prikazuju posljednji opaženi slikovni okvir, predikcije modela proroka u budućem okviru, našu prognozu te toplinsku mapu faktora miješanja β^{F2M}. Samo pri vizualizaciji, budući slikovni okviri iscrtani su ispod semantičkih predikcija kako bi se lakše pratila točnost naše prognoze.

Druga scena prikazuje slučaj nestajućeg objekta u kratkoročnom prognoziranju, a što se može primijetiti i prema plavoj regiji na lijevoj strani toplinske mape. Treći redak prikazuje prvu scenu srednjoročnog prognoziranja segmentacije instanci u kojoj se taksi giba s lijeve na desnu stranu. Može se primijetiti kako se taksi pomaknuo za približno cijelu svoju dužinu, što je model ispravno prognozirao. Tom prilikom na sredini scene otvorio se pogled na ulicu koja je do tada bila zaklonjena. To sugerira i duboko plava regija u toplinskoj mapi na sredini scene. Posljednja scena prikazuje bicklista koji se giba s lijeva na desno. Model je ispravno detektirao i prognozirao posebne segmentacijske mape za bicikl i vozača.

Slika 6.4 prikazuje prognoziranje panoptičke segmentacije. Iako semantičke predikcije izgledaju slično kao kod semantičke segmentacije, može se primijetiti kako su instance istog razreda obojani u različite nijanse boje odgovarajućeg semantičkog razreda. Kratkoročno prognoziranje je prilično točno i uspješno razlikuje instance objekata. Treći redak prikazuje već viđenu scenu prilikom srednjoročnog prognoziranja. Automobil na desnoj strani u budućnosti napušta scenu i otvara pogled na cestu i zgradu u pozadini. Model povjerava zadatak prognoziranja modulu F2F koji ispravno prognozira položaje ceste, pločnika i zgrade. Posljednji redak prikazuje statičnu scenu u kojoj prognoziranje za većinu piksela povjereno modulu F2M. Ipak, u nekim manjim dijelovima scene model ima problema s uspostavom korespondencije pa se mogu primijetiti i plave regije u toplinskoj mapi.

Slika 6.4: Kvalitativni rezultati prognoziranja panoptičke segmentacije na Cityscapesu. Slika prikazuje dva primjera kratkoročnog prognoziranja (gore) i dva primjera srednjoročnog prognoziranja (dolje). Stupci prikazuju posljednji opaženi slikovni okvir, predikcije modela proroka u budućem okviru, našu prognozu te toplinsku mapu faktora miješanja β^{F2M}. Samo pri vizualizaciji, budući slikovni okviri iscrtani su ispod semantičkih predikcija kako bi se lakše pratila točnost naše prognoze.

6.6 Validacijski eksperimenti

Ovaj odjeljak donosi skup validacijskih eksperimenata koji će pobliže opisati doprinose različitih elemenata predloženog pristupa prognoziranju značajki F2MF. Zbog brzine učenja svi eksperimenti provedeni su na zadatku semantičke segmentacije u kombinaciji s gustim modelom manjeg kapaciteta temeljenim na okosnici ResNet-18. Nisu korištene tehnike uvećanja podataka kako bi se značajke prilikom treniranja mogle spremiti na brzi SSD disk.

Tablica 6.5 predstavlja rezultate ablacije korelacijskog modula te modula F2M i F2F. Svaki redak u tablici predstavlja jednu konfiguraciju prognostičkog modela. Svaka konfiguracija modela je odvojeno trenirana i to za kratkoročno i srednjoročno prognoziranje. Usporedimo prvo individualno prognoziranje modulima F2F i F2M. Eksperimenti bez korelacijskog sloja (redci 1 i 2) pokazuju da izravno F2F prognoziranje postiže 0.6 mIoU bodova bolju točnost. Kod eksperimenata s prisutnim korelacijskim slojem model F2F bolji je za 0.7 mIoU bodova u kratkoročnom prognoziranje, dok u srednjoročnom prognoziranju rade jednako dobro. Ovi rezultati su očekivani s obzirom na to da je F2F pristup ekspresivniji. Modul F2M budućnost mora objasniti isključivo deformiranjem prošlosti, dok je modul F2F u tom smislu neograničen i može zamisliti bilo što. Ipak, kombinirani model F2MF u oba slučaja nadjačava individualne modele. U kratkoročnom prognoziranju prednost modela F2MF nad modelom F2M je 1.0 odnosno 1.3 mIoU boda sa isključenim odnosno uključenim korelacijskim slojem. U srednjoročnom prognoziranju prednost kombiniranog modela je veća i iznosi 1.2 boda bez odnosno 1.4 boda s korelacijskim slojem. Prednost kombiniranog modela u odnosu na F2F je 0.4 boda u kratkoročnom, odnosno 0.6 bodova u srednjoročnom prognoziranju bez korelacijskog sloja. U suprotnom slučaju, prednosti iznose 0.6 i 1.4 mIoU bod u kratkoročnom i srednjoročnom prognoziranju.

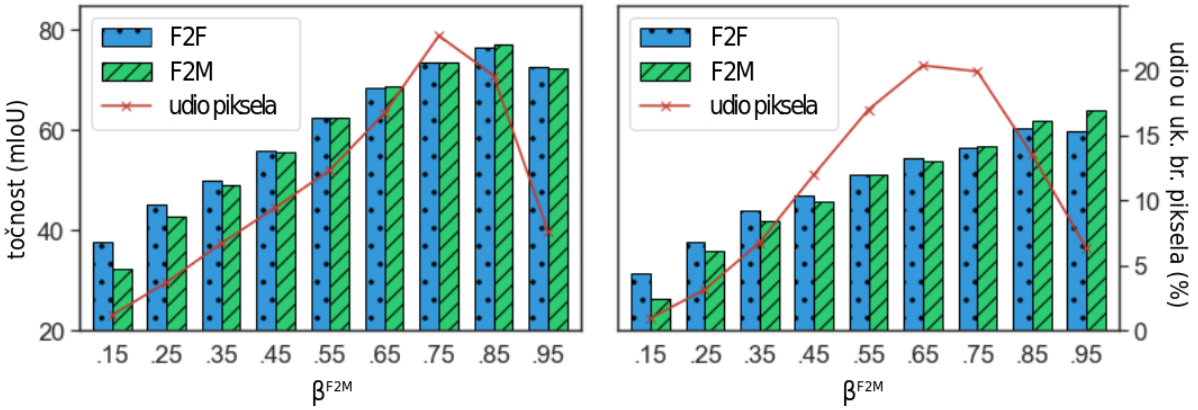
Tablica 6.5: Validacija korelacijskog sloja, te modula F2F i F2M na Cityscapes val podskupu. Modeli sa individualnim modulima F2F i F2M su trenirani zasebno.

Konfiguracija (F2MF-RN18)			Kratkoročno (mIoU)		Srednjoročno (mIoU)	
F2F	F2M	Korelacijski modul	Svi	PO	Svi	PO
	✓		64.8	63.4	52.2	47.6
✓			65.4	64.0	52.8	48.6
✓	✓		65.8	64.7	53.4	49.7
	✓	✓	65.6	64.4	54.5	50.7
✓		✓	66.3	64.9	54.5	50.8
✓	✓	✓	**66.9**	**65.6**	**55.9**	**52.4**

Nadmoć kombiniranog modela sugerira da su pristupi izravnom prognoziranju značajki (F2F) i prognoziranje deformiranjem značajki iz prošlosti (F2M) zaista komplementarni.

Ista tablica otkriva i jasan doprinos korelacijskog modula. Prisutnost prostorno-vremenskih korelacijskih značajki unaprjeđuje točnost modela F2M za 0.8, modela F2F za 0.9, te kombiniranog modela F2MF za 1.1 mIoU bod u kratkoročnom prognoziranju. Doprinosi u srednjoročnom prognoziranju još su i veći: 2.3 mIoU boda za model F2M, 1.7 mIoU bodova za model F2F te 2.5 mIoU boda za kombinirani model F2MF.

U nastavku opisujemo eksperimente koji dodatno istražuju komplementarnost pristupa F2M i F2F evaluacijom neovisnih jednoglavih modela u stratificiranim grupama piksela. Prethodni rezultati daju naslutiti da F2F pristup prognoziranju generalno postiže malo bolje rezultate. Međutim, poznato je da je pristupu F2M posebno teško prognozirati u novootkrivenim dijelovima scene, što navodi na misao da bi F2M mogao biti bolji u prethodno viđenim dijelovima scene. Tu hipotezu testiramo grupiranjem piksela u skupine prema vrijednostima faktora miješanja β^{F2M} izračunatih uz pomoć naučenog kombiniranog modela F2MF. Za svaku skupinu piksela zasebno računamo točnost prognoziranja neovisno naučenih modela koji se oslanjaju na samo jedan modul za prognoziranje (F2F ili F2M). Rezultate eksperimenta prikazuje slika 6.5. Lijeva y-os mjeri točnost prognoziranja, a desna postotak piksela koji pripadaju određenoj skupini. Na x-osi su prikazane vrijednosti faktora miješanja β^{F2M} za svaku skupinu. Slika lijevo prikazuje rezultate kratkoročnog prognoziranja, a slika desno srednjoročnog. Može se primijetiti kako histogram teži desnoj strani u kojoj se nalaze grupe s većom vrijednosti faktora β^{F2M} što sugerira da kombinirani model većinu piksela povjerava predikciji modulom F2M. Usporedbom lijeve i desne slike može se uočiti da je ovaj efekt manje izražen kod srednjoročnog prognoziranja. To je očekivano obzirom da su kod srednjoročnog prognoziranja pomaci veći i dolazi



Slika 6.5: Histogrami točnosti individualnih modela F2F i F2M i incidencije piksela po grupama nastalim prema vrijednosti faktora miješanja β^{F2M} predviđenim kombiniranim modelom F2MF. Lijeva slika prikazuje rezultate za kraktoročno prognoziranje, a desna za srednjoročno prognoziranje na validacijskom skupu Cityscapesa.

do većih otkrivanja. Slika također pokazuje da individualni model F2M radi bolje od modela F2F u pikselima sa visokom vrijednosti faktora β^{F2M}, što znači da kombinirani model ispravno delegira posao prognoziranja. Opisani eksperimenti podupiru hipotezu o komplementarnosti dvaju pristupa prognoziranju značajki.

Tablica 6.6 validira metodu miješanja značajki prognoziranih modulima F2M i F2F kod kombiniranog modela F2MF. Predložena implementacija koristi faktore miješanja na razini piksela koji su predviđeni posebnim težinskim modulom. Težinski modul na izlazu daje pet mapa značajki koje određuju doprinose četiri prognoze modula F2M i jedne prognoze modula F2F konačnim predikcijama budućih značajki. Predložena metoda se uspoređuje s dvije jednostavnije metode. Prva jednostavno uprosječava sve izlaze spomenutih modula. Druga razmatra težinski modul koji umjesto pet mapa značajki i faktora miješanja na razini piksela prediktira jedan vektor sa pet elemenata koji predstavlja faktore miješanja na razini cijelog tenzora značajki. Rezultati pokazuju da faktori miješanja na razini tenzora postižu bolju točnost prognoziranja od jednostavnog uprosječivanja: 1.4 mIoU boda prilikom kratkoročnog prognoziranja i 0.7 mIoU bodova prilikom srednjoročnog prognoziranja. Korištenje faktora miješanja na razini piksela dodatno unaprijeđuje točnost i to za 0.2 mIoU boda kod kratkoročnog i za 0.6 mIoU bodova kod srednjoročnog prognoziranja. Ovaj eksperiment podupire predloženu funkcionalnost i ulogu težinskog modula.

Tablica 6.6: Validacija tri različite inačice težinskog modula za miješanje izlaza modula F2F i F2M na validacijskom podskupu skupa Cityscapes za slučaj prognoziranja semantičke segmentacije.

Metoda miješanja	Kratkoročno (mIoU)		Srednjoročno (mIoU)	
(F2MF-RN18)	Svi	PO	Svi	PO
miješanje prosjekom	65.3	63.7	54.6	50.4
težine na razini slike	66.7	65.1	55.3	51.7
težine na razini piksela	**66.9**	**65.6**	**55.9**	**52.4**

Tablica 6.7 istražuje utjecaj broja ulaznih slikovnih okvira na točnost prognoziranja. Prvi stupac tablice navodi indekse ulaznih okvira u video isječku. Kao i ranije, t označava indeks najrecentnijeg opaženog okvira, a okviri se uzorkuju sa razmakom od 3 vremenska trenutka. Modeli koji prognoziraju budućnost na temelju samo jednog slikovnog okvira rade loše jer ne mogu procijeniti dinamiku scene na temelju jedne referentne točke. Modeli iz prvog retka tablice nemaju korelacijski sloj jer su za izračun prostorno-vremenski korelacijski koeficijanata potrebni tenzori iz barem dva vremenska trenutka. Točnost kratkoročnog prognoziranja je otprilike jednaka za modele koji na ulazu primaju 2-5 ulaznih okvira. Kod srednjoročnog

prognoziranja izražena je prednost modela koji na ulazu prima četiri ulazna okvira. Ovi modeli općenito postižu bolju točnost što potvrđuje ispravnost standardne konfiguracije iz literature. Primijetite da nedostaju rezultati za srednjoročno prognoziranje sa 5 ulaznih slikovnih okvira. Razlog tome je ograničena duljina videoisječaka na Cityscapesu. Naime, zbog evaluacije, ciljani budući slikovni okvir uvijek mora odgovarati 20. okviru u videoisječku jer samo za njega postoje točne oznake. To znači da kod srednjoročnog prognoziranja posljednji opaženi ulaz ima indeks $t = 11$, što uzorkovanje okvira $t - 12$ čini nemogućim.

Tablica 6.7: Validacija brojnosti ulaznih slikovnih okvira na validacijskom podskupa podatkovnog skupa Cityscapes za slučaj prognoziranja semantičke segmentacije.

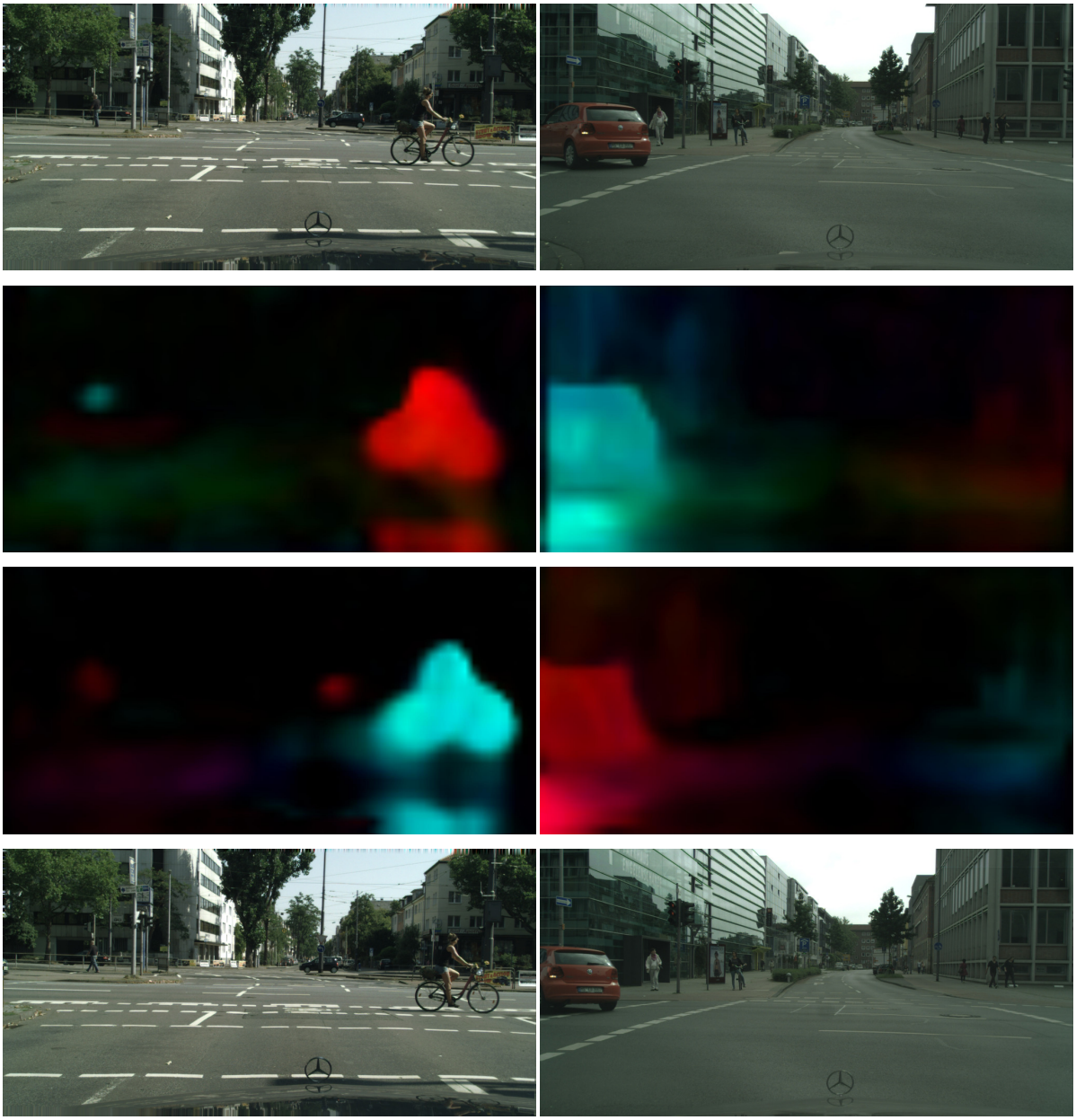
Ulazni okviri	Kratkoročno (mIoU)		Srednjoročno (mIoU)	
(F2MF-RN18)	Svi	PO	Svi	PO
$\{t\}$	57.9	55.5	45.7	39.6
$\{t-3,t\}$	66.4	65.3	54.9	51.2
$\{t-6,t-3,t\}$	66.9	65.6	55.1	51.0
$\{t-9,t-6,t-3,t\}$	**66.9**	**65.6**	**55.9**	**52.4**
$\{t-12,t-9,t-6,t-3,t\}$	66.7	65.1	/	/

Tablica 6.8 uspoređuje individualne modele tipa F2M sa unaprijednom ili unatražnom deformacijom značajki iz prošlosti. Za unaprijednu deformaciju korištena je naivna implementacija opisana u poglavlju 5.2 sa jezgrom RBF i parametrom $\sigma^2 = 0.125$. Rezultati iz prvog odjeljka tablice sugeriraju da dvije formulacije prognoziranja deformiranjem postižu otprilike jednaku

Tablica 6.8: Usporedba točnosti F2M prognoziranja unatražnom (BW) i unaprijednom (FW) deformacijom na validacijskom skupu Cityscapes. Oznaka r.p. označava receptivno polje.

	Kratkoročno		Srednjoročno	
Točnost (mIoU)	Svi	PO	Svi	PO
F2M-BW	64.8	63.4	52.2	47.6
F2M-FW	64.6	63.2	52.2	47.3
F2M-BW (ograničeno r.p.)	60.4	58.1	45.4	37.8
F2M-FW (ograničeno r.p.)	61.2	59.1	47.6	41.1

točnost. Zbog toga u svim ostalim eksperimentima koristimo učinkovito unatražno prognoziranje. Drugi odjeljak tablice uspoređuje ova dva pristupa u režimu ograničenog receptivnog polja. Model F2M koji prognozira tok značajki u ovome slučaju ima ukupno tri (umjesto osam) konvolucijskih slojeva i koristi obične konvolucije umjesto deformabilnih. Rezultati pokazuju jasnu prednost unaprijedne deformacije u režimu rada s ograničenim receptivnim poljem. Ovi eksperimentalni rezultati potvrđuju našu hipotezu da prognoziranje unatražnog toka zahtijeva veće receptivno polje.



Slika 6.6: Vizualizacija toka značajki prognoziranog od strane unaprijednog i unatražnog F2M modela na scenama iz validacijskog skupa Cityscapes. Redci prikazuju posljednju opaženu sliku, unaprijedni tok, unatražni tok te buduću neopaženu sliku. Vizualizacija toka nastala je prema [109], gdje različite boje kodiraju različite smjerove vektora pomaka, a zasićenost odgovara njegovoj apsolutnoj vrijednosti.

Razliku između varijanti modula F2M sa unatražnim odnosno unaprijednim deformiranjem ilustrira i slika 6.6. Slika vizualizira predviđeni tok značajki za dvije scene iz skupa Cityscapes. Redci slike prikazuju posljednji opaženi slikovni okvir, tok unaprijedne inačice modula F2M, tok unatražne inačice modula F2M, te budući slikovni okvir. Tok značajki prikazan je kodiranjem bojom prema [109] gdje svijetlo plava boja označava pomake ulijevo, žuta prema dolje, crvena udesno i tamno-plava prema gore. Zasićenost je proporcionalna apsolutnoj vrijednosti vektora pomaka. Može se primijetiti kako su vektori pomaka dvaju varijanti suprotnih orijentacija. Također, vidljivo je da su vektori unatražnog toka poravnati s budućom lokacijom objekta, a vektori unaprijednog toka s lokacijom objekta u prošlosti. Ova opažanja u skladu su s jednadžbama i diskusijom iz poglavlja 5.2.

Tablica 6.9 validira korištenje deformabilnih konvolucija u prognostičkom modelu F2MF. Uspoređujemo točnost prognoziranja predloženog modela F2MF i istovjetnog modela koji umjesto deformabilnih koristi obične konvolucije. Rezultati pokazuju da deformabilne konvoluciju značajno povećavaju točnost prognoziranja. Model bez deformabilnih konvolucija postiže 1.3 mIoU boda lošiju točnost u kratkoročnom, a 4.2 mIoU boda u srednjoročnom prognoziranju.

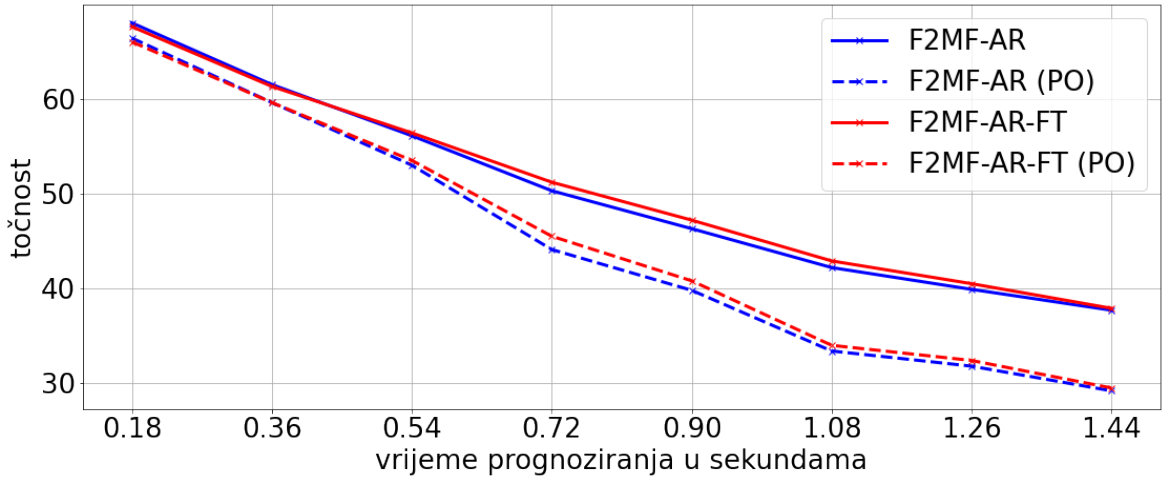
Tablica 6.9: Usporedba točnosti prognoziranja inačicama modela F2MF s običnim odnosno deformabilnim konvolucijama na validacijskom skupu Cityscapes.

	Kratkoročno		Srednjoročno	
Točnost (mIoU)	Svi	PO	Svi	PO
F2MF s običnim konvolucijama	65.6	64.1	51.7	46.9
F2MF s deformabilnim konvolucijama	**66.9**	**65.6**	**55.9**	**52.4**

6.7 Dugoročno prognoziranje autoregresijom

Prognoziranje dalje u budućnost moguće je postići autoregresijskom primjenom modula F2MF. Pod tim podrazumijevamo uzastopnu primjenu našeg najboljeg modela za kratkoročno prognoziranje zamjenjujući svaki put najstariji ulazni tenzor značajki s prognoziranim tenzorom iz prethodne iteracije. Takva primjena omogućuje prognoziranje proizvoljan broj koraka unaprijed. Dugoročno prognoziranje moguće je evaluirati na scenama validacijskog podskupa na Cityscapesu snimljenima u Frankfurtu. Spomenuti videoisječci iz Frankfurta povezani su u jednu veću cjelinu, pa je moguće uzorkovati ulazne okvire koji su jako udaljeni od odgovarajućeg okvira sa točnim oznakama. Evaluiramo dva modela za semantičku segmentaciju: jedan sa finim ugađanjem za autoregresivno prognoziranje, te drugi koji je treniran za kratkoročno

prognoziranje te samo autoregresivno primijenjen. Ugađanje kratkoročnog modela za autoregresivno prognoziranje akumulira gubitak izračunat u trenucima $t+3$, $t+6$ i $t+9$ i propagira gradijent unatrag kroz vrijeme. Rezultati su prikazani na slici 6.7. Na x-osi se nalazi vremenska udaljenost do ciljanog budućeg okvira, a na y-osi točnost prognoziranja. Crvenom bojom je prikazana točnost modela ugađanog za autoregresivnu primjenu (F2MF-AR-FT), a plavom modela bez ugađanja (F2MF-AR). Isprekidana linija prikazuje mIoU uprosječen samo preko razreda koji predstavljaju pokretne objekte. Slika prikazuje pad točnosti prognoziranja s povećanjem vremenske udaljenosti budućeg okvira. Pad točnosti je očekivano veći kod razreda koji predstavljaju pokretne objekte. Model koji je fino ugađan za autoregresivno prognoziranje postiže nešto bolju točnost.



Slika 6.7: Ovisnost točnosti prognoziranja semantičke segmentacije autoregresivno primjenjenih modela o vremenskoj udaljenosti budućeg okvira. Evaluacija je provedena na videoisječcima iz Cityscapesa snimljenima u Frankfurtu. Kratica FT označava model fino ugađan za autoregresivno prognoziranje.

6.8 Generalizacija na podatkovnom skupu CamVid

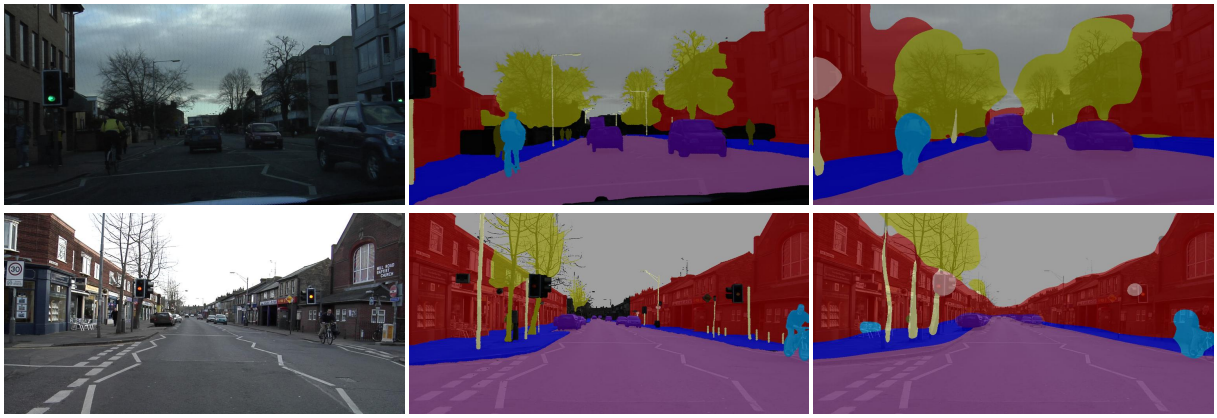
Sposobnost generalizacije u prisustvu pomaka domene prognostičkih modela treniranih na Cityscapesu moguće je provjeriti evaluacijom na podatkovnom skupu CamVid [110]. Takav postupak evaluacije prvi je put primijenjen u [5]. Podatkovni skup CamVid razlikuje se od Cityscapesa prema rezoluciji slike, omjeru stranica slike (slike iz CamVida blago su izdužene po vertikalnoj osi), skupu semantičkih razreda i broju sličica u sekundi snimljenog videa. Obzirom da se radi o potpuno konvolucijskom pristupu prognoziranju, eksperiment nije zahtijevao nikakve adaptacije što se tiče rezolucije slike. Razliku u semantičkim taksonomijama moguće je riješiti jednostavnim mapiranjem iz 19 Cityscapes razreda u 11 CamVid razreda. Ulazne slike CamVida uzorkovane su s razmakom od pet vremenskih trenutaka, što otprilike odgovara tri vremenska trenutka na Cityscapesu.

Tablica 6.10 prikazuje rezultate srednjoročnog prognoziranja semantičke segmentacije na podatkovnom skupu CamVid. Svi prikazani rezultati u tablici odgovaraju modelima treniranima na podatkovnom skupu Cityscapes. Prvi stupac prikazuje rezultate odgovarajućeg modela proroka. Drugi stupac prikazuje točnost prognoziranja. Treći stupac prikazuje relativnu performansu prognostičkog modela u odnosu na proroka. Posljednji stupac ističe relativni pad točnosti prilikom prognoziranja u odnosu na proroka. Prvi odjeljak tablice prikazuje rezultat iz literature [5]. Drugi odjeljak prikazuje rezultate modela F2MF sa snažnijim modelom za raspoznavanje u jednoj slici. Osim osnovnog modela, razmatraju se i autoregresijski modeli iz prethodnog odjeljka. Rezultati pokazuju da autoregresijski model sa finim ugađanjem postiže najbolju točnost i ujedno i najmanji gubitak točnosti u odnosu na odgovarajući prorok.

Tablica 6.10: Točnost prognoziranja semantičke segmentacije na podatkovnom skupu CamVid primjenom modela koje smo naučili na podatkovnom skupu Cityscapes.

	Prorok	Prognoza	Rel. perf.	Pad perf.
Luc ar. ft. [5]	55.4	46.8	84.5%	-15.5%
F2MF-DN121	62.8	51.3	81.7%	-18.3%
F2MF-DN121 ar.	62.8	53.4	85.0%	-15.0%
F2MF-DN121 ar. ft.	62.8	**54.5**	**86.8%**	**-13.2%**

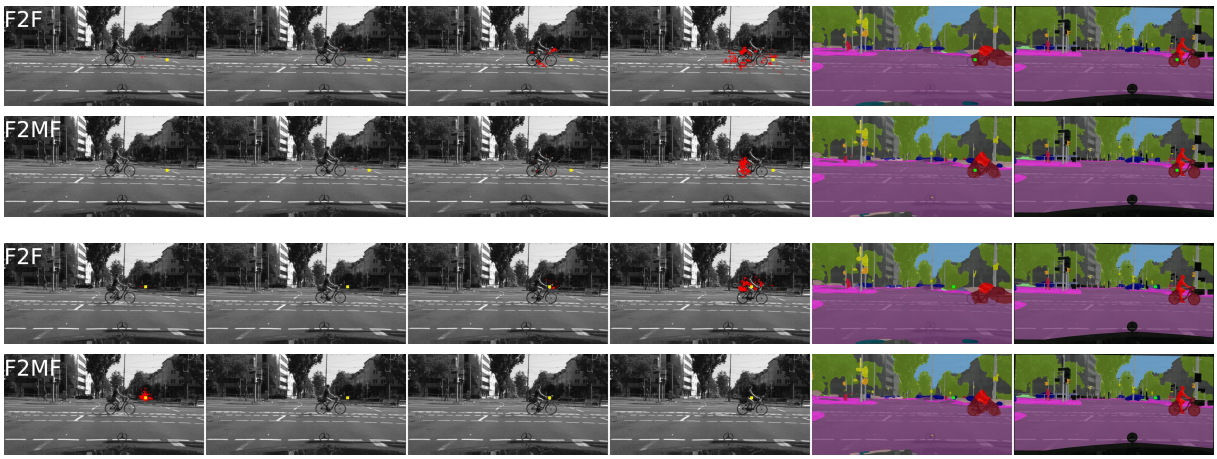
Slika 6.8 vizualizira rezultate prognoziranja najboljeg modela iz tablice 6.10 na dvije scene iz podatkovnog skupa CamVid. Stupci prikazuju posljednji opaženi slikovni okvir, točne oznake i prognoziranu semantičku segmentaciju. Prognoze su većinom točne, što sugerira da predloženi sustav za prognoziranje može generalizirati i uz prisutnost pomaka domene.



Slika 6.8: Prikaz rezultata srednjoročnog prognoziranja na podatkovnom skupu CamVid modelom naučenim na Cityscapesu. Stupci prikazuju posljednji opaženi slikovni okvir, točne oznake i prognoziranu semantičku segmentaciju budućeg slikovnog okvira.

6.9 Interpretacija načina rada modela

Ovaj odjeljak istražuje razliku između kombiniranog prognoziranja modulom F2MF i izravnog prognoziranja modulom F2F. Slika 6.9 uspoređuje raspored gradijenata funkcije log-max-softmax jednog izlaznog piksela po pikselima ulaznih slika za modele F2MF i F2F. Prva četiri stupca prikazuju četiri ulazne slike iz prošlosti na kojima je crvenim točkicama istaknuto 1% magnitudom najvećih gradijenata. Posljednja dva stupca prikazuju prognoziranu semantičku segmentaciju i točne oznake. Zelenim kvadratom je naznačen izlazni piksel iz kojeg se računa gradijent spomenute funkcije, a žutim kvadratom je naznačena ista lokacija u slikama iz prošlosti. Ideja ove vizualizacije je provjeriti ispravnost odnosno objašnjivost odluka prognostičkog modela. Budući da gradijenti s najvećom apsolutnom vrijednošću pokazuju ulazne piksele koji najviše utječu na izabrani izlazni piksel, oni bi se trebali poklapati s korespodentnim lokacijama izabranog piksela u prošlosti. U prvom slučaju (gornja dva retka) lokacija izabranog piksela odgovara stražnjem kotaču bicikla u budućoj slici. Može se primijetiti kako su kod oba modela gradijenti dominantno smješteni u najrecentnijoj slici. Uspoređujući dva modela, možemo primijetiti kako su gradijenti modela F2MF preciznije i s manjom varijancom lokalizirani oko lokacije kotača u posljednjem opaženom okviru. U drugom razmatranom slučaju (donja dva retka) izabrana lokacija piksela u budućnost odgovara pozadini, dok je u posljednjem opaženom okviru tamo smješten biciklist. Gradijenti modela F2F raspršeni su oko te lokacije u posljednjem okviru. To sugerira da je se taj model dominantno naučio oslanjati na najrecentniji okvir. U ovome slučaju traženu prognozu pokušava napraviti interpolacijom iz konteksta u posljednjem okviru. Model F2MF ima mogućnost postupiti mudrije. Kod njega su gotovo svi



Slika 6.9: Vizualizacija gradijenata funkcije log-max-softmax u izabranom izlaznom pikselu po ulaznim slikama za modele F2F i F2MF [59]. Prva četiri stupca prikazuju četiri ulazne slike. Posljednja dva stupca prikazuju prognoziranu semantičku segmentaciju i točne oznake za budući slikovni okvir. Crvene točke prikazuju piksele u kojima se nalaze 1% gradijenata s najvećom prosječnom vrijednošću. Zeleni kvadrat prikazuje izabrani izlazni piksel u prognoziranoj segmentaciji, a žuti istu lokaciju u ulaznim slikama.

gradijenti smješteni u vremenski najudaljenijem slikovnom okviru. Izgleda da je model shvatio da je pozadinu koja se nalazi na izabranoj lokaciji u budućnosti već vidio u ulaznoj slici. Prognoziranje deformiranjem iz svih ulaznih tenzora značajki omogućuje mu da semantički sadržaj izabrane lokacije jednostavno kopira iz prvog ulaznog tenzora. Na taj je način izbjegao teški zadatak zamišljanja koji je morao obaviti model s glavom F2F. To pokazuje da kombinirani model ima sposobnost razumijevanja ovakvih složenih uzoraka zaklanjanja i otkrivanja.

6.10 Analiza računske složenosti

Ovaj odjeljak donosi usporedbu računske složenosti predloženog sustava za gusto semantičko prognoziranje sa odabranim metodama iz literature. Pretpostavljamo mogućnost spremanja aktivacija iz prošlosti u privremenu memoriju i zbog toga razmatramo računalni trošak prognoziranja kada na ulaz sustava dođe novi slikovni okvir. Računalnu složenost mjerimo brojem operacija množenje-akumulacija (eng. *multiply-and-accumulate*, MAC) uz pomoć biblioteke `thop` [111].

Tablica 6.11 uspoređuje računalnu složenost prognoziranja semantičke segmentacije za naš najbolji model sa metodom M2M [48] koja je prethodno pobliže objašnjena u pregledu literature (odjeljak 3.3.2, slika 3.4). Može se primijetiti kako metoda M2M zasnovana na prognoziranju optičkog toka zahtijeva evaluaciju posebnog modela za optički tok što značajno utječe na računalnu složenost. Evaluacija korištenog modela FlowNet2-C na slikama veličine 0.5 MPx košta 88.9 GMAC. Prognoziranje reprezentacije optičkog toka na $8\times$ poduzorkovanoj rezoluciji zahtijeva 38.8 GMAC. Rezultati pokazuju da naš model za prognoziranje ima približno $12\times$ manju računsku složenost. Ako se tome doda evaluacija modela za raspoznavanje u jednoj slici prednost našeg modela je još uvijek značajna ($4\times$). Razlog za takve rezultate leži u činjenicama i) da naš pristup ne zahtijeva evaluaciju zasebnog modela za optički tok i ii) da je rezolucija prognozirane reprezentacije u našem slučaju četiri puta manja.

Tablica 6.11: Računalna složenost prognoziranja semantičke segmentacije na slikama iz Cityscapesa izražena brojem operacija množenje-akumulacija (GMAC).

Moduli	M2M [48]	F2MF-DN121
Model za raspoznavanje u jednoj slici	536.4	145.5
Model za optički tok (FlowNet2-C)	88.9	0.0
Model za prognoziranje	38.8	9.8
Prognoziranje ukupno	127.7	9.8
Ukupno	664.1	155.3

Tablica 6.12 uspoređuje računalnu složenost prognoziranja segmentacije instanci predložene metode F2MF i izvornog višerazinskog pristupa F2F iz literature [6]. Tablica prikazuje trošak prognoziranja značajki na različitim rezolucijama i trošak evaluacije modela za segmentaciju instanci. Naša inačica modela za segmentaciju instanci Mask R-CNN C4 zahtijeva oko $1.6\times$ više operacija od FPN inačice koja se koristi u literaturi. Ipak, ukupan trošak prognoziranja segmentacije instanci kod naše je metode približno $3\times$ manji. Razlog leži u tome što FPN inačica modela Mask R-CNN zahtijeva prognoziranje cijele piramide značajki. To uključuje i značajke visoke rezolucije koje su poduzorkovane samo $4\times$ ili $8\times$. Naša metoda slijedi ideju jednorazinskog prognoziranja i prognozira samo značajke koje su $16\times$ poduzorkovane u odnosu na ulaznu sliku. Ovaj slučaj ilustrira učinkovitost jednorazinskog prognoziranja.

Tablica 6.12: Računalna složenost prognoziranja segmentacije instanci (GMAC).

Moduli	F2F-M-FPN-RN50 [6]	F2MF-M-C4-RN50
Model za raspoznavanje u jednoj slici	401.0	668.4
Prognoziranje značajki rezolucije $1/4$	1417.7	0.0
Prognoziranje značajki rezolucije $1/8$	354.4	0.0
Prognoziranje značajki rezolucije $1/16$	88.6	106.2
Prognoziranje značajki rezolucije $1/32$	22.1	0.0
Prognoziranje ukupno	1865.2	106.2
Ukupno	2266.2	774.6

Tablica 6.13 profilira unaprijedni prolaz kroz sustav za gusto semantičko prognoziranje. Mjeri se vrijeme izvođenja za tri faze zaključivanja: izlučivanje značajki iz četiri ulazna okvira, prognoziranje značajki i formiranje semantičkih predikcija. Mjerenja su napravljena bez

Tablica 6.13: Profil izvođenja neoptimiziranih sustava prognoziranja baziranih na modelu F2MF u milisekundama.

Model	$4\times$ izlučivanje značajki	prognoziranje značajki	formiranje semantičkih predikcija
F2MF-RN18	72	7	7
F2MF-DN121	265	12	8
F2MF-Mask-C4-RN50	204	47	248
F2MF-PDL-RN50	230	52	144

ikakvih optimizacija na grafičkoj kartici GTX1080Ti. Tri odjeljka u tablici redom prikazuju rezultate za prognoziranje semantičke segmentacije, segmentacije instanci i panoptičke segmentacije. Rezultati pokazuju da je prognoziranje značajki znatno brže od evaluacije modela za raspoznavanje u jednoj slici. U realnoj situaciji u kojoj novi slikovni okviri konstantno dolaze na ulaz modela, prognoziranje stvara minimalni dodatni trošak. Prognoziranje semantičke segmentacija sa efikasnim modelom F2MF-RN18 može se izvoditi u stvarnom vremenu pod pretpostavkom spremanja značajki iz prošlosti u privremenu memoriju. U takvoj izvedbi, potrebno je značajke izlučiti samo u novopridošlom okviru.

Poglavlje 7

Zaključak

Anticipacija budućnosti sastavni je dio svakog inteligentnog ponašanja. Prognoziranje semantike buduće scene preduvjet je za pravovremeno i inteligentno odlučivanje o trenutnim akcijama autonomnog agenta. U novije vrijeme, istraživačka zajednica pokušava pristupiti ovom problemu implicitnim postavljanjem zakona dinamike scene primjenom dubokog učenja u videu. Ipak, niti jedan od postojećih pristupa nije sposoban razlikovati novootkrivene od već opaženih dijelova scene. To je suboptimalno jer je za objašnjenje prvoga potrebna mogućnost zamišljanja, dok se drugo može objasniti ekstrapoliranjem prethodnih opažanja. Za razliku od ranijih pristupa, naša metoda može detektirati dijelove slike u kojima će se pojaviti novootkriveni dijelovi scene. Tu informaciju iskorištavamo za razjašnjavanje koji dio dinamike scene nastaje uslijed gibanja objekta, a koji uslijed pojave noviteta u sceni.

Ova disertacija predlaže metodu za gusto semantičko prognoziranje na razini apstraktnih vizualnih značajki dubokog modela za raspoznavanje. Naša metoda kombinira izravno prognoziranje značajki modulom F2F (eng. *features to features*) sa predloženim modulom F2M (eng. *features to motion*) koji deformira značajke iz prošlosti regresiranim poljem pomaka. Konačnu prognozu dobivamo kao linearnu kombinaciju nezavisnih prognoza u skladu s gustim poljem faktora miješanja koje regresira zasebna glava našeg modela. Empirijski smo potvrdili da kombinirani model F2MF (eng. *features to motion and features*) favorizira izravno prognoziranje u novootkrivenim dijelovima scene te prognoziranje defomiranjem tamo gdje se može uspostaviti korespondencija prema prošlim okvirima u videu. Predložena metoda F2MF postiže stanje tehnike u prognoziranju semantičke segmentacije na podatkovnom skupu Cityscapes.

Razvijena metoda uvodi i korelacijski modul koji obogaćuje skrivenu reprezentaciju modela za prognoziranje značajki prostorno-vremenskim korelacijskim koeficijentima. Naša metoda također uvodi i deformabilne konvolucije koje bolje odgovaraju geometrijskoj prirodi ovoga zadatka. Ova dva doprinosa značajno unaprjeđuju točnost prognoziranja u svakom od tri spomenuta pristupa prognoziranju (F2M, F2F i F2MF). Ova disertacija pokazuje da se najučinkovitija i najtočnija prognostička funkcionalnost postiže jednorazinskim prognoziranjem sažetih

značajki. Posljedično, naša metoda prikladna je za primjene u stvarnom vremenu zasnovane na ugradbenom sklopovlju umjerene računske snage. Predložena metoda za prognoziranje značajki može se integrirati s modelima koji obavljaju različite zadatke gustog raspoznavanja. To smo pokazali eksperimentima prognoziranja semantičke segmentacije, segmentacije instanci i panoptičke segmentacije.

Ova disertacija otvara puno mogućnosti za budući rad. Naša metoda ne razmatra multimodalnu prirodu budućnosti, što je ključ za dugoročnije prognoziranje i razmatranje iscrpnog skupa potencijalnih opasnosti koje se mogu pojaviti u budućnosti. Također, vrlo bi zanimljivo bilo osigurati bolju konvergenciju pri treniranju s kraja na kraj te proširiti metodu na prognoziranje RGB piksela budućeg slikovnog okvira.

Literatura

[1]Brš čić, D., Ikeda, T., Kanda, T., "Do you need help? a robot providing information to people who behave atypically", IEEE Transactions on Robotics, Vol. 33, No. 2, 2017, str. 500-506.

[2]Yeong, D. J., Velasco-Hernandez, G., Barry, J., Walsh, J. *et al.*, "Sensor and sensor fusion technology in autonomous vehicles: A review", Sensors, Vol. 21, No. 6, 2021, str. 2140.

[3]Davison, A. J., Reid, I. D., Molton, N. D., Stasse, O., "Monoslam: Real-time single camera slam", IEEE transactions on pattern analysis and machine intelligence, Vol. 29, No. 6, 2007, str. 1052–1067.

[4]Oprea, S., Martinez-Gonzalez, P., Garcia-Garcia, A., Castro-Vargas, J. A., Orts-Escolano, S., Garcia-Rodriguez, J., Argyros, A., "A review on deep learning techniques for video prediction", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[5]Luc, P., Neverova, N., Couprie, C., Verbeek, J., LeCun, Y., "Predicting deeper into the future of semantic segmentation", in Proceedings of the IEEE International Conference on Computer Vision, 2017, str. 648–657.

[6]Luc, P., Couprie, C., Lecun, Y., Verbeek, J., "Predicting future instance segmentation by forecasting convolutional features", in Proceedings of the european conference on computer vision (ECCV), 2018, str. 584–599.

[7]Graber, C., Tsai, G., Firman, M., Brostow, G., Schwing, A. G., "Panoptic segmentation forecasting", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, str. 12 517–12 526.

[8]Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., "The cityscapes dataset for semantic urban scene understanding", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, str. 3213–3223.

[9]Fugoši ć, K., Šarić, J., Šegvić, S., "Multimodal semantic forecasting based on conditional generation of future features", in DAGM German Conference on Pattern Recognition. Springer, 2020, str. 474–487.

[10]Hu, A., Cotter, F., Mohan, N., Gurau, C., Kendall, A., "Probabilistic future prediction for video scene understanding", in European Conference on Computer Vision. Springer, 2020, str. 767–785.

[11]Bei, X., Yang, Y., Soatto, S., "Learning semantic-aware dynamics for video prediction", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, str. 902–912.

[12]Courdier, E., Fleuret, F., "Real-time segmentation networks should be latency aware", in Proceedings of the Asian Conference on Computer Vision, 2020.

[13]Šari ć, J., Oršić, M., Antunović, T., Vražić, S., Šegvić, S., "Single level feature-to-feature forecasting with deformable convolutions", in German Conference on Pattern Recognition. Springer, 2019, str. 189–202.

[14]Saric, J., Orsic, M., Antunovic, T., Vrazic, S., Segvic, S., "Warp to the future: Joint forecasting of features and feature motion", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, str. 10 648–10 657.

[15]Šari ć, J., Vražić, S., Šegvić, S., "Dense semantic forecasting in video by joint regression of features and feature motion", IEEE Transactions on Neural Networks and Learning Systems, 2021.

[16]Sun, J., Xie, J., Hu, J.-F., Lin, Z., Lai, J., Zeng, W., Zheng, W.-S., "Predicting future instance segmentation with contextual pyramid convlstms", in Proceedings of the 27th acm international conference on multimedia, 2019, str. 2043–2051.

[17]Krizhevsky, A., Sutskever, I., Hinton, G. E., "Imagenet classification with deep convolutional neural networks", Advances in neural information processing systems, Vol. 25, 2012, str. 1097–1105.

[18]Ren, S., He, K., Girshick, R., Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks", Advances in neural information processing systems, Vol. 28, 2015, str. 91–99.

[19]Long, J., Shelhamer, E., Darrell, T., "Fully convolutional networks for semantic segmentation", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, str. 3431–3440.

[20]He, K., Gkioxari, G., Dollár, P., Girshick, R., “Mask r-cnn”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, No. 2, 2018, str. 386–397.

[21]Cheng, B., Collins, M. D., Zhu, Y., Liu, T., Huang, T. S., Adam, H., Chen, L.-C., “Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation”, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, str. 12 475–12 485.

[22]Luo, W., Schwing, A. G., Urtasun, R., “Efficient deep learning for stereo matching”, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, str. 5695–5703.

[23]Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., “Flownet: Learning optical flow with convolutional networks”, in Proceedings of the IEEE international conference on computer vision, 2015, str. 2758–2766.

[24]Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., “Imagenet: A large-scale hierarchical image database”, in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, str. 248–255.

[25]Simonyan, K., Zisserman, A., “Very deep convolutional networks for large-scale image recognition”, arXiv preprint arXiv:1409.1556, 2014.

[26]He, K., Zhang, X., Ren, S., Sun, J., “Deep residual learning for image recognition”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[27]Tan, M., Le, Q., “Efficientnet: Rethinking model scaling for convolutional neural networks”, in International conference on machine learning. PMLR, 2019, str. 6105–6114.

[28]Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., “A convnet for the 2020s”, arXiv preprint arXiv:2201.03545, 2022.

[29]Huang, G., Liu, Z., van der Maaten, L., Weinberger, K. Q., “Densely connected convolutional networks”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[30]Orsic, M., Kreso, I., Bevandic, P., Segvic, S., “In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[31]Ioffe, S., Szegedy, C., “Batch normalization: Accelerating deep network training by reducing internal covariate shift”, in International conference on machine learning. PMLR, 2015, str. 448–456.

[32]Santurkar, S., Tsipras, D., Ilyas, A., Madry, A., “How does batch normalization help optimization?”, Advances in neural information processing systems, Vol. 31, 2018.

[33]Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., “Feature pyramid networks for object detection”, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, str. 2117–2125.

[34]Chao, P., Kao, C.-Y., Ruan, Y.-S., Huang, C.-H., Lin, Y.-L., “Hardnet: A low memory traffic network”, in Proceedings of the IEEE/CVF international conference on computer vision, 2019, str. 3552–3561.

[35]Krešo, I., Krapac, J., Šegvi ć, S., “Efficient ladder-style densenets for semantic segmentation of large images”, IEEE Transactions on Intelligent Transportation Systems, Vol. 22, No. 8, 2020, str. 4951–4961.

[36]Pleiss, G., Chen, D., Huang, G., Li, T., Van Der Maaten, L., Weinberger, K. Q., “Memory-efficient implementation of densenets”, arXiv preprint arXiv:1707.06990, 2017.

[37]Muller, P., Savakis, A., “Flowdometry: An optical flow and deep learning based approach to visual odometry”, in 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2017, str. 624–631.

[38]Shin, J., Kim, S., Kang, S., Lee, S.-W., Paik, J., Abidi, B., Abidi, M., “Optical flow-based real-time object tracking using non-prior training active feature model”, Real-time imaging, Vol. 11, No. 3, 2005, str. 204–218.

[39]Sun, D., Yang, X., Liu, M.-Y., Kautz, J., “Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume”, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, str. 8934–8943.

[40]Teed, Z., Deng, J., “Raft: Recurrent all-pairs field transforms for optical flow”, in European conference on computer vision. Springer, 2020, str. 402–419.

[41]Ren, Z., Yan, J., Ni, B., Liu, B., Yang, X., Zha, H., “Unsupervised deep learning for optical flow estimation”, in Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[42]Liu, P., Lyu, M., King, I., Xu, J., “Selflow: Self-supervised learning of optical flow”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, str. 4571–4580.

[43]Rusinkiewicz, S., “Image warping”, Sveu čilišna prezentacija, dostupno na: https://www.cs.princeton.edu/courses/archive/spr11/cos426/notes/cos426_s11_lecture01_intro_color.pdf 2011.

[44]Szeliski, R., Computer vision: algorithms and applications. Springer Science & Business Media, 2010.

[45]Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., “Deformable convolutional networks”, in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[46]Dolz, J., Xu, X., Rony, J., Yuan, J., Liu, Y., Granger, E., Desrosiers, C., Zhang, X., Ben Ayed, I., Lu, H., “Multiregion segmentation of bladder cancer structures in mri with progressive dilated convolutional networks”, Medical physics, Vol. 45, No. 12, 2018, str. 5482–5493.

[47]Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., Lin, D., “MMDetection: Open mmlab detection toolbox and benchmark”, arXiv preprint arXiv:1906.07155, 2019.

[48]Terwilliger, A., Brazil, G., Liu, X., “Recurrent flow-guided semantic forecasting”, in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019, str. 1703–1712.

[49]Gao, H., Xu, H., Cai, Q.-Z., Wang, R., Yu, F., Darrell, T., “Disentangling propagation and generation for video prediction”, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, str. 9006–9015.

[50]Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y., “Deep feature flow for video recognition”, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, str. 2349–2358.

[51]Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P., “The mapillary vistas dataset for semantic understanding of street scenes”, in Proceedings of the IEEE international conference on computer vision, 2017, str. 4990–4999.

[52]Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., Jawahar, C., "Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments", in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019, str. 1743–1751.

[53]Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A., "Semantic understanding of scenes through the ade20k dataset", International Journal of Computer Vision, Vol. 127, No. 3, 2019, str. 302–321.

[54]Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P., "Panoptic segmentation", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, str. 9404–9413.

[55]Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale", arXiv preprint arXiv:2010.11929, 2020.

[56]Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L. *et al.*, "Swin transformer v2: Scaling up capacity and resolution", arXiv preprint arXiv:2111.09883, 2021.

[57]Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.-C., "Max-deeplab: End-to-end panoptic segmentation with mask transformers", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, str. 5463–5474.

[58]He, K., Girshick, R., Dollár, P., "Rethinking imagenet pre-training", in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, str. 4918–4927.

[59]Luo, W., Li, Y., Urtasun, R., Zemel, R., "Understanding the effective receptive field in deep convolutional neural networks", Advances in neural information processing systems, Vol. 29, 2016.

[60]Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., "Going deeper with convolutions", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, str. 1–9.

[61]Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", IEEE transactions on pattern analysis and machine intelligence, Vol. 40, No. 4, 2017, str. 834–848.

[62]Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., “Pyramid scene parsing network”, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, str. 2881–2890.

[63]Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T., “Semi-supervised learning with ladder networks”, Advances in neural information processing systems, Vol. 28, 2015.

[64]Ronneberger, O., Fischer, P., Brox, T., “U-net: Convolutional networks for biomedical image segmentation”, in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, str. 234–241.

[65]Hong, Y., Pan, H., Sun, W., Jia, Y., “Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes”, arXiv preprint arXiv:2101.06085, 2021.

[66]Kitani, K. M., Ziebart, B. D., Bagnell, J. A., Hebert, M., “Activity forecasting”, in European conference on computer vision. Springer, 2012, str. 201–214.

[67]Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., “Large-scale video classification with convolutional neural networks”, in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, str. 1725–1732.

[68]Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A., “A benchmark dataset and evaluation methodology for video object segmentation”, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, str. 724–732.

[69]Niklaus, S., Liu, F., “Softmax splatting for video frame interpolation”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, str. 5437–5446.

[70]Mathieu, M., Couprie, C., LeCun, Y., “Deep multi-scale video prediction beyond mean square error”, arXiv preprint arXiv:1511.05440, 2015.

[71]Lee, H.-Y., Huang, J.-B., Singh, M., Yang, M.-H., “Unsupervised representation learning by sorting sequences”, in Proceedings of the IEEE international conference on computer vision, 2017, str. 667–676.

[72]Srivastava, N., Mansimov, E., Salakhudinov, R., “Unsupervised learning of video representations using lstms”, in International conference on machine learning. PMLR, 2015, str. 843–852.

[73]Zhu, Y., Sapra, K., Reda, F. A., Shih, K. J., Newsam, S., Tao, A., Catanzaro, B., "Improving semantic segmentation via video propagation and label relaxation", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, str. 8856–8865.

[74]Gadde, R., Jampani, V., Gehler, P. V., "Semantic video cnns through representation warping", in Proceedings of the IEEE International Conference on Computer Vision, 2017, str. 4453–4462.

[75]Yuen, J., Torralba, A., "A data-driven approach for event prediction", in European Conference on Computer Vision. Springer, 2010, str. 707–720.

[76]Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S., "Social lstm: Human trajectory prediction in crowded spaces", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[77]Vukoti ć, V., Pintea, S.-L., Raymond, C., Gravier, G., Gemert, J. C. v., "One-step time-dependent future video frame prediction with a convolutional encoder-decoder neural network", in International conference on image analysis and processing. Springer, 2017, str. 140–151.

[78]Reda, F. A., Liu, G., Shih, K. J., Kirby, R., Barker, J., Tarjan, D., Tao, A., Catanzaro, B., "Sdc-net: Video prediction using spatially-displaced convolution", in Proc. ECCV, 2018, str. 718–733.

[79]Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.-H., "Flow-grounded spatial-temporal video prediction from still images", in Proc. ECCV, 2018, str. 600–615.

[80]Pan, J., Wang, C., Jia, X., Shao, J., Sheng, L., Yan, J., Wang, X., "Video generation from single semantic label map", in Proc. CVPR, 2019, str. 3733–3742.

[81]Wu, B., Nair, S., Martin-Martin, R., Fei-Fei, L., Finn, C., "Greedy hierarchical variational autoencoders for large-scale video prediction", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, str. 2318-2328.

[82]Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O., "The unreasonable effectiveness of deep features as a perceptual metric", in CVPR, 2018.

[83]Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S., "Towards accurate generative models of video: A new metric & challenges", arXiv preprint arXiv:1812.01717, 2018.

[84]Kalchbrenner, N., Oord, A., Simonyan, K., Danihelka, I., Vinyals, O., Graves, A., Kavukcuoglu, K., "Video pixel networks", in International Conference on Machine Learning. PMLR, 2017, str. 1771–1779.

[85]Hao, Z., Huang, X., Belongie, S., "Controllable video generation with sparse trajectories", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, str. 7854–7863.

[86]Yu, F., Koltun, V., "Multi-scale context aggregation by dilated convolutions", in 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, Bengio, Y., LeCun, Y., (ur.), 2016, dostupno na: http://arxiv.org/abs/1511.07122

[87]Jin, X., Xiao, H., Shen, X., Yang, J., Lin, Z., Chen, Y., Jie, Z., Feng, J., Yan, S., "Predicting scene parsing and motion dynamics in the future", in NIPS, 2017, str. 6915–6924.

[88]Nabavi, S. S., Rochan, M., Wang, Y., "Future semantic segmentation with convolutional lstm", BMVC, 2018.

[89]Chen, X., Han, Y., "Multi-timescale context encoding for scene parsing prediction", in 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019, str. 1624–1629.

[90]Bhattacharyya, A., Fritz, M., Schiele, B., "Bayesian prediction of future street scenes using synthetic likelihoods", in International Conference on Learning Representations, 2019.

[91]Makansi, O., Cicek, O., Buchicchio, K., Brox, T., "Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[92]Yao, Y., Xu, M., Choi, C., Crandall, D. J., Atkins, E. M., Dariush, B., "Egocentric vision-based future vehicle localization for intelligent driving assistance systems", in 2019 International Conference on Robotics and Automation (ICRA), 2019, str. 9711-9717.

[93]Wojke, N., Bewley, A., Paulus, D., "Simple online and realtime tracking with a deep association metric", in 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017, str. 3645–3649.

[94]Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-c., "Convolutional lstm network: A machine learning approach for precipitation nowcasting", Advances in neural information processing systems, Vol. 28, 2015.

[95]Lin, Z., Sun, J., Hu, J.-F., Yu, Q., Lai, J.-H., Zheng, W.-S., "Predictive feature learning for future segmentation prediction", in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, str. 7365–7374.

[96]Gatys, L. A., Ecker, A. S., Bethge, M., "Image style transfer using convolutional neural networks", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, str. 2414–2423.

[97]Vondrick, C., Pirsiavash, H., Torralba, A., "Anticipating the future by watching unlabeled video".

[98]Hu, J.-F., Sun, J., Lin, Z., Lai, J.-H., Zeng, W., Zheng, W.-S., "Apanet: Auto-path aggregation for future instance segmentation prediction", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[99]Couprie, C., Luc, P., Verbeek, J., "Joint future semantic and instance segmentation prediction", in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, September 2018.

[100]Vora, S., Mahjourian, R., Pirk, S., Angelova, A., "Future segmentation using 3d structure", arXiv preprint arXiv:1811.11358, 2018.

[101]Chiu, H.-k., Adeli, E., Niebles, J. C., "Segmenting the future", IEEE Robotics and Automation Letters, Vol. 5, No. 3, 2020, str. 4202–4209.

[102]Kingma, D. P., Welling, M., "Auto-encoding variational bayes", arXiv preprint arXiv:1312.6114, 2013.

[103]Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A., "Image-to-image translation with conditional adversarial networks", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[104]Lee, S., Ha, J., Kim, G., "Harmonizing maximum likelihood with gans for multimodal conditional generation", in International Conference on Learning Representations, 2018.

[105]Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R., "Detectron2", https://github.com/facebookresearch/detectron2, 2019.

[106]Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., "Microsoft coco: Common objects in context", in European conference on computer vision. Springer, 2014, str. 740–755.

[107]Kingma, D. P., Ba, J., "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980, 2014.

[108]Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., "Automatic differentiation in pytorch", 2017.

[109]Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M. J., Szeliski, R., "A database and evaluation methodology for optical flow", International journal of computer vision, Vol. 92, No. 1, 2011, str. 1–31.

[110]Brostow, G. J., Fauqueur, J., Cipolla, R., "Semantic object classes in video: A high-definition ground truth database", Pattern Recognition Letters, Vol. 30, No. 2, 2009, str. 88–97.

[111]Zhu, L., "Thop: Pytorch-opcounter", https://github.com/Lyken17/pytorch-OpCounter, 2020.

Popis slika

Popis tablica

Životopis

Josip Šarić rođen je 1995. godine u Makarskoj. Osnovno i srednjoškolsko obrazovanje završio je u Kupresu, Bosna i Hercegovina. Preddiplomski i diplomski studij računarstva završio je na Sveučilištu u Zagrebu, Fakultetu Elektrotehnike i Računarstva. Dobitnik je priznanja i brončane plakete Josip Lončar za uspješno savladavanje gradiva prve godine diplomskog studija, odnosno uspješan završetak diplomskog studija. Po završetku diplomskog studija 2018. godine zaposlio se kao zavodski suradnik na Zavodu za elektroniku, mikroelektroniku, računalne i inteligentne sustave Fakulteta elektrotehnike i računarstva u Zagrebu. Sudjelovao je u oblikovanju algoritma strojnog učenja koji je pobijedio na natjecanju RVC 2020. godine. Istraživačka grupa je za taj uspjeh nagrađena fakultetskom nagradom za znanost.

Josip Šarić sudjeluje kao istraživač na projektu "Konvolucijski modeli za semantičko predviđanje razvoja prometnih scena" kojeg financira tvrtka Rimac Technology. Istraživanje je fokusirano na gusto semantičko prognoziranje budućnosti u videu. Njegovi istraživački interesi uključuju učinkovite modele za panoptičku segmentaciju i modele za gusto raspoznavanje u slikama visoke rezolucije. Autor je više radova predstavljenih na međunarodnim konferencijama i časopisima. Josip Šarić tečno govori engleski jezik. Oženjen je i trenutno živi u Zagrebu.

Popis objavljenih djela

Radovi u časopisima

1. **Šarić, J., Vražić, S., Šegvić, S., "Dense Semantic Forecasting in Video by Joint Regression of Features and Feature Motion", IEEE Transactions on Neural Networks and Learning Systems, 2021**

Radovi objavljeni na međunarodnim konferencijama

1. **Šarić, J., Oršić, M., Antunović, T., Vražić, S., Šegvić, S., "Warp to the future: Joint forecasting of features and feature motion", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, lipanj 2020., str. 10648-10657.**

2.Šari ć, J., Oršić, M., Antunović, T., Vražić, S., Šegvić, S., "Single level feature-to-feature forecasting with deformable convolutions", German Conference on Pattern Recognition. Springer, Cham, rujan 2019., str. 189-202.

3.Fugoši ć, K., Šarić, J., Šegvić, S., "Multimodal semantic forecasting based on conditional generation of future features", DAGM German Conference on Pattern Recognition, Springer, Cham., rujan 2020., str. 474-487.

4.Bevandi ć, P., Oršić, M., Grubišić, I., Šarić, J., Šegvić, S., "Multi-domain semantic segmentation with overlapping labels", Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, siječanj 2022., str. 2615-2624.

Ostali radovi

1.Šari ć, J., Oršić, M., Šegvić, S., "Panoptic SwiftNet: Pyramidal Fusion for Real-time Panoptic Segmentation", arXiv preprint arXiv:2203.07908, 2022.

Biography

Josip Šarić was born in 1995 in Makarska, Croatia. He completed elementary school and high school in Kupres, Bosnia and Herzegovina. He received his bachelor's and master's degree in computing from the University of Zagreb, Faculty of Electrical Engineering and Computing. He was a recipient of award and bronze plaque Josip Lončar for successful completion of first graduate year and graduate study respectively. After graduation in 2018., he was employed as research assistant at the Department of Electronics, Microelectronics, Computer and Intelligent Systems of the Faculty of Electrical Engineering and Computing. He participated in machine learning algorithm design which won first place at the international competition Robust Vision Challenge in 2020. The research group later received the faculty science award for that success.

Josip Šarić participates as a researcher in a project called "Convolutional models for semantic forecasting in traffic scenes" which is funded by Rimac Technology. The research is focused on dense semantic forecasting in video. His research interests also involve efficient models for panoptic segmentation, as well as dense recognition models for large resolution images. He is the author of multiple papers published at international conferences and scientific journals. Josip Šarić is proficient in English. He is married and currently lives in Zagreb.