University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Marin Kačan

RECOGNITION OF ROAD INFRASTRUCTURE SAFETY ATTRIBUTES BY COMPUTER VISION

DOCTORAL THESIS

Zagreb, 2025

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Marin Kačan

RECOGNITION OF ROAD INFRASTRUCTURE SAFETY ATTRIBUTES BY COMPUTER VISION

DOCTORAL THESIS

Supervisor: Professor Siniša Šegvić, PhD

Zagreb, 2025

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Marin Kačan

RASPOZNAVANJE OBILJEŽJA SIGURNOSTI CESTOVNE INFRASTRUKTURE RAČUNALNIM VIDOM

DOKTORSKI RAD

Mentor: Prof. dr. sc. Siniša Šegvić

Zagreb, 2025.

The doctoral thesis was written at the University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Electronics, Microelectronics, Computer and Intelligent Systems.

Supervisor: Professor Siniša Šegvić, PhD

Doctoral thesis contains: 79 pages

Doctoral thesis number: _____

About the Supervisor

Siniša Šegvić has received a PhD degree in computer vision and artificial intelligence at UniZg-FER. He was a postdoc researcher at IRISA Rennes (2005-2006) and at TU Graz (2006-2007). Subsequently, he returns to UniZg-FER where he lectures in computer science and performs research in computer vision. He has participated in the introduction of graduate courses Design patterns, Deep learning and Three-dimensional computer vision. He has also participated in the reconstruction of the master course Computer Vision and the introduction of doctoral courses Analysis of dynamic scenes and Models for representing images and video. He mentored seven completed doctoral theses and several hundreds of master and bachelor theses in computer vision and artificial intelligence. His research and professional interests include computer vision, machine learning, scene understanding, recognition of satellite images, and defense from data poisoning attacks. He has published at top conferences (CVPR, ECCV, NeurIPS and AAAI) and scientific journals in computer vision and artificial intelligence (IEEE TPAMI, IJCV, IEEE TNNLS, Patt Recog). He has been a reviewer at top conferences and scientific journals. His research group consists of several postdoctoral and doctoral students that are funded by national projects, European projects and private companies. Together they achieved remarkable results at several competitions in computer vision (ACDC, WildDash, Robust vision challenge, Cityscapes, Fishyscapes and SegmentMeIfYouCan). He has led three research projects funded by Croatian Science Foundation (ADEPT, MultiCLOD, MASTIF), one project from the Croatian RRP programme (VoNoMobil) and several industrial projects funded by Google AI for global goals, P3M, RoMB technology, Rimac Automobiles, MicroBlink and Promet i prostor. He participated in the Center of research excellence DataCross, several projects from the EDF and ERDF programmes (EICACS, A-UNIT, SafeTram) as well as one project from the FP7 programme (ACROSS). He participated in industrial development as a technical consultant. He also led two bilateral research projects in cooperation with researchers from Austria and Germany and organized several bilateral workshops and one international research workshop. Siniša Šegvić speaks english and italian very well, and has basic communication skills in french. He had a six month career break for paternal leave. He is married and has three children.

O mentoru

Siniša Šegvić doktorirao je u području umjetne inteligencije i računalnog vida na zagrebačkom FER-u u 2004. godini. Bio je postdoktorski istraživač na institutu IRISA u Rennesu (2005-2006) te na TU Graz (2006-2007). Nakon toga vraća se na FER gdje predaje u području računarske znanosti i istražuje u području računalnog vida. Sudjelovao je u uvođenju diplomskih kolegija Oblikovni obrasci u programiranju, Duboko učenje te Trodimenzionalni računalni

vid. Također, sudjelovao je i u rekonstrukciji kolegija Računalni vid. Konačno, sudjelovao je i u uvođenju doktorskih kolegija Analiza dinamičkih scena te Modeli za reprezentaciju slike i videa. Mentorirao je sedam obranjenih doktorata te nekoliko stotina diplomskih i završnih radova u području računalnog vida i umjetne inteligencije. Njegovi istraživački i profesionalni interesi uključuju računalni vid, strojno učenje, razumijevanje scena, analizu satelitskih snimaka te obrane od napada putem trovanja podataka. Objavio je radove na vrhunskim konferencijama (CVPR, ECCV, NeurIPS te AAAI) te vrhunskim časopisima iz računalnog vida i umjetne inteligencije (IEEE TPAMI, IJCV, IEEE TNNLS, Patt Recog i IEEE TITS). Recenzent je u vrhunskim konferencijama i znanstvenim časopisima. Njegova istraživačka grupa sastoji se od dva postdoktoranda i šest doktoranada koje financiraju nacionalni projekti, evropski projekti i privatne tvrtke. Zajedno su postigli zapažene rezultate na više natjecanja u računalnom vidu (ACDC, WildDash, Robust vision challenge, Cityscapes, Fishyscapes i SegmentMeIfYouCan). Vodio je tri istraživačka projekta Hrvatske zaklade za znanost (ADEPT, MultiCLOD, MASTIF), jedan projekt iz programa NPOO (VoNoMobil) te industrijska istraživanja koja su financirali Google, P3M, Rimac automobili, RoMB, MicroBlink te Promet i prostor. Sudjelovao je u istraživačkom centru izvrsnosti DataCross, projektima iz programa EDF i ERDF (EICACS, A-UNIT, SafeTram) kao i na jednom projektu iz programa FP7 (ACROSS). Sudjelovao je u industrijskom razvoju kao tehnički konzultant. Vodio je i dva bilateralna istraživačka projekta u suradnji s istraživačima iz Austrije i Njemačke te je organizirao nekoliko bilateralnih i jednu međunarodnu istraživačku radionicu. Siniša Šegvić govori engleski i talijanski jezik te ima osnovne komunikacijske vještine na francuskom jeziku. Bio je na roditeljskom dopustu od šest mjeseci. Oženjen je i ima troje djece.

Dedication

Thank you to my advisor, Siniša Šegvić, for his mentorship, guidance, and patience throughout this journey. His willingness to share knowledge and wisdom has been invaluable.

Thank you also to Marko Ševrović, whose support and advice provided a valuable foundation for my research direction.

I want to thank my friends – Romeo, Stelio, Paolo, Daniel, Natan, Josip, Filip, Ante, Toni, Konrad, Grgur, Entoni, Dino, and Antonio – for the enduring strength of our friendship. Their support was boundless, their laughter a constant lift, and their encouragement a steady push forward, transforming obstacles into cherished memories.

My special thanks go to my colleagues - Ivan M., Ivan S., Anja, Iva, Matej, Jelena, Josip, Ivan G., Petra, and Marin - for bringing energy, camaraderie, and warmth to our daily work. Their openness and intellectual curiosity fostered an environment where ideas flourished alongside friendship, making our shared academic journey both productive and memorable.

I am deeply grateful to my sister Elvira, together with Lukas, Niko and Franko, for their unwavering love, care, and humor that continue to brighten my path.

To the memory of my mother and father, I offer my deepest thanks for their love and guidance, which remain the foundation of who I am.

And finally, thank you to Lorena, for her love, kindness, and unwavering support. She brought joy, order, and serenity into each day, filling our moments together with happiness that sustained me through the challenges of this work. Her presence has been both an inspiration and a source of genuine delight.

Abstract

Road accidents cause over 1.35 million fatalities annually, emphasizing the need for efficient road infrastructure safety assessment. Traditional assessments based on historical accident data are reactive and often limited by incomplete records, especially in regions lacking comprehensive records of accident incidence. In contrast, proactive approaches emphasize regular and systematic evaluations of road infrastructure to identify potential hazards before accidents occur. The International Road Assessment Programme (iRAP) exemplifies this proactive strategy with its Star Rating system, which evaluates road infrastructure safety through a detailed analysis of 52 attributes. However, manual assessment of these attributes is labor-intensive and time-consuming. This thesis proposes a two-stage deep learning approach to automate the recognition of road-safety attributes in monocular video data. Attribute recognition is formulated as a two-stage multi-task multi-class classification problem, where each attribute represents a separate task. The first stage involves local recognition that we formulate as a convolutional neural network with a shared encoder and attribute-specific classification branches. The encoder is initialized by pre-training on semantic segmentation of street scenes. Our problem involves extreme class imbalance due to rarity and variety of road hazards. Under these conditions, straight-forward learning algorithms are bound to suffer from false negatives. While static inverse frequency weighting can alleviate this issue, it has been shown to increase the false positive rate. Furthermore, class weighting with many imbalanced tasks can lead to task interference due to high variance in individual losses. We address these issues by multi-task formulation of dynamic loss weighting, which avoids excessive false positives while maintaining stable magnitudes of individual losses. The second stage enhances local predictions by observing a larger temporal context via per-attribute recurrent models, which capture temporal dependencies. Experiments are conducted on the newly introduced iRAP-BH dataset, comprising over 226,000 labeled images along 2,300 kilometers of roads in Bosnia and Herzegovina. The results confirm the impact of each of our contributions, effectively addressing challenges such as class imbalance, non-orthogonal attribute design, fine-grained classes, and temporal dependencies. Additional evaluations on publicly available datasets demonstrate the model's generalizability and robustness of the proposed machine learning approach, achieving state-of-the-art results on Honda Scenes and performing competitively on FM3m and BDD100k.

Keywords: image classification, road safety, iRAP attributes, deep learning, multi-task learning

Prošireni sažetak

Uvod

Prometne nesreće predstavljaju značajan globalni izazov, uzrokujući preko 1,35 milijuna smrti godišnje. Svjetska zdravstvena organizacija zbog toga ih ubraja među najkritičnije javnozdravstvene probleme na svjetskoj razini. Ovaj zabrinjavajući podatak ukazuje na hitnu potrebu za učinkovitim mjerama za poboljšanje cestovne sigurnosti i smanjenje smrtnosti i ozljeda povezanih s prometom. Problem prometne sigurnosti prepoznala je i organizacija Ujedinjenih naroda inicijativom "Decade of Action for Road Safety".

Tradicionalni pristupi za procjenu cestovne sigurnosti oslanjaju se na reaktivne metode pronalaska visokorizičnih lokacija analizom povijesnih podataka o nesrećama. Iako takvi pristupi mogu otkriti složene čimbenike rizika, oni ovise o prijašnjim incidentima i točnosti dostupnih podataka, te su osobito ograničeni u područjima s nepotpunom evidencijom nesreća. S druge strane, proaktivni pristupi naglašavaju redovite procjene sigurnosti cestovne infrastrukture kako bi se identificirale potencijalne opasnosti prije nego što dođe do nesreća. Takve metode procjenjuju fizičke značajke cestovne infrastrukture - poput izvedbe razdvajanja prometnih traka, opasnosti uz cestu, te objekata namijenjenih ranjivim sudionicima u prometu - kako bi se omogućile ciljane intervencije za umanjivanje rizika.

Udruga International Road Assessment Programme (iRAP) pruža primjer takve proaktivne strategije. Njihov sustav iRAP Star Rating standardizirana je metodologija za procjenu sigurnosti cestovnih dionica na temelju fizičkih karakteristika cestovne infrastrukture. Procjena se provodi analizom 52 obilježja definiranih standardom iRAP. Pri tome se se razina sigurnosti cestovnih dionica zasebno procjenjuje za četiri skupine sudionika u prometu - putnike u vozilima, motocikliste, pješake i bicikliste.

Unatoč svojoj učinkovitosti, ručna procjena iRAP obilježja zahtijeva mnogo vremena i prikladno obučenog osoblja. Ovo istraživanje nastoji automatizirati procjenu tih obilježja u georeferenciranim videozapisima iz perspektive vozača primjenom dubokog učenja i tehnika računalnog vida. Predloženi postupak strojnog učenja teži pružiti učinkovito i skalabilno rješenje za poboljšanje procjene sigurnosti cesta analizom prostorno-vremenskog konteksta cestovnih segmenata.

Sigurnost cestovne infrastukture

Za učinkovitu automatizaciju procjene sigurnosti cesta pomoću dubokog učenja i računalnog vida, ključno je razumjeti strukturu i specifične izazove iRAP obilježja. Obilježja su podijeljeni u sedam skupina: *Road and Context* (metapodaci te broj kolničkih traka prometnice), *Observed Flow* (broj korisnika cestovne dionice u danom trenutku), *Speed Limit* (ograničenja brzine i mjere smirivanja prometa), *Mid-block* (intrinsične značajke ceste poput broja traka, radijusa za-

krivljenosti i kvalitete oznaka), *Roadside* (rizik koji predstavljaju objekti pored ceste, s vozačeve i suvozačeve strane), *Intersection* (tipovi raskrižja po konfiguraciji i signalizaciji), te *Vulnerable Road-User Facilities* (infrastruktura za pješake i bicikliste te karakteristike okolnog područja). Svako obilježje predstavlja specifičnu značajku cestovnog okruženja i za svaku dionicu poprima vrijednost iz specifične taksonomije, pri čemu se broj mogućih razreda razlikuje od obilježja do obilježja.

Analiza obilježja otkriva nekoliko izazova koji utječu na uspješnost modela strojnog učenja za njihovo prepoznavanje. Jedan od glavnih problema je neuravnoteženost razreda, gdje su određeni razredi izrazito zastupljeni, a preostali jako rijetki. Takav disbalans može dovesti do zanemarivanja ključnih, ali rijetkih sigurnosnih značajki, čime se mogu previdjeti situacije visokog rizika. Rješenje ovog problema zahtijeva funkcije gubitka koje daju veću važnost manje zastupljenim razredima i korištenje mjere makro-F1 kao primarne metrike za evaluaciju modela. Dodatni je izazov neortogonalnost određenih obilježja, gdje se više ortogonalnih značajki kombinira u jedinstvenu taksonomiju. Time se povećava broj razreda i povećava neuravnoteženost taksonomije jer se primjeri dijele na više razreda. Također, kapacitet klasifikacijskog modela suvišno se troši na prepoznavanje istih značajki u različitim kombinacijama. Primjerice, obilježje *Skid Resistance* objedinjuje razinu prijanjanja kolnika i tip kolničkog zastora, što rezultira složenom taksonomijom. Pojednostavljivanje takvih složenih taksonomija moglo bi poboljšati učinkovitost modela. Izrazito detaljni i vizualno slični razredi također predstavljaju značajan izazov za raspoznavanje. Obilježja s vrlo specifičnim razredima, poput različitih tipova zaštitnih ograda u obilježju *Roadside severity*, teško je razlikovati zbog suptilnih vizualnih razlika. Takva razina granularnosti u praksi dovodi do pogrešnih klasifikacija i smanjuje učinkovitost modela. Vremenska dinamika obilježja duž slijeda cestovnih segmenata također utječe na učinkovitost modela. Točkasta obilježja bilježe prebrojive pojave infrastrukturnih elemenata koje treba detektirati samo u segmentu koji je najbliži njihovoj lokaciji. Međutim, vizualne značajke koje omogućuju prepoznavanje tih elemenata često se protežu kroz više segmenata, što modelima otežava određivanje točnog segmenta koji treba označiti. S druge strane, intervalna (*smooth*) obilježja rijetko mijenjaju razred duž uzastopnih cestovnih segmenata jer obično opisuju veća područja ili duge kontinuirane infrastrukturne elemente. Lokalni modeli za prepoznavanje mogu imati poteškoća s ovim obrascima zbog ograničenog konteksta, što im otežava internaliziranje inherentne inercije određenih obilježja ili pak pronalazak točnog trenutka pojave obilježja. Razumijevanje ovih vremenskih obrazaca ključno je za razvoj učinkovitih strategija prepoznavanja. Navedeni izazovi potiču slijedno poboljšavanje predikcija lokalnog modela primjenom zasebnih povratnih modela za svako obilježje. To omogućuje učenje vremenskih obrazaca u širem prostorno-vremenskom kontekstu bez značajnog povećanja računalne složenosti.

Ova analiza daje kontekst za razvoj modela koji adresiraju opisane izazove te tako

povećavaju učinkovitost tehnika dubokog učenja za automatsku procjenu sigurnosti cesta u videozapisima uličnih scena.

Klasifikacija slika dubokim učenjem

Duboko učenje donijelo je veliki napredak u području računalnog vida, između ostalog omogućavajući i razvoj algoritama za automatsko raspoznavanje obilježja cestovne infrastrukture iz slika. Dok se tradicionalne metode strojnog učenja oslanjaju na ručno definirane značajke, duboki modeli samostalno uče hijerarhiju značajki izravno iz ulaznih slika. Takvi modeli sastoje se od više slojeva obrade koji koriste ustaljene matematičke operacije poput konvolucije za izdvajanje prostornih značajki, aktivacijskih funkcija za modeliranje nelinearnosti, funkcija sažimanja za smanjenje dimenzionalnosti, te normalizacije za stabilnije učenje i bolju konvergenciju. Zbog sposobnosti prepoznavanja složenih uzoraka ti su modeli pogodni za zadatke koji zahtijevaju detaljnu analizu vizualnih scena, poput prepoznavanja specifičnih obilježja sigurnosti cestovne infrastrukture.

Nagli razvoj dubokog učenja u području klasifikacije slika možemo pripisati nekolicini presudnih čimbenika. Među njima se posebno ističu: pojava velikih označenih skupova podataka, sve veća dostupnost snažnih računalnih resursa, te značajan napredak u metodama učenja dubokih modela. Revolucionarne arhitekture kao što su AlexNet, VGG, ResNet i DenseNet, redom su pomicale granice performansi u ovom području. Posebno je značajan doprinos ResNeta koji uvođenjem rezidualnih veza uspješno rješava problem degradacije u vrlo dubokim mrežama. Time omogućava treniranje znatno dubljih modela nego ranije. DenseNet ide korak dalje uvodeći gustu povezanost slojeva, što je rezultira učinkovitijim protokom informacija kroz mrežu. Moderna konvolucijska arhitektura ConvNeXt integrira spoznaje iz modela temeljenih na pažnji. Značaj navedenih konvolucijskih arhitektura nadilazi njihove rezultate na akademskim skupovima podataka. One su postale temeljni građevni blokovi za složenije primjene računalnog vida, uključujući i procjenu sigurnosti cestovne infrastrukture.

Nadovezujući se na ova dostignuća, prijenos znanja (transfer learning) postao je ključna strategija u primjeni prethodno naučenih modela dubokog učenja za specifične primjene s ograničenim brojem podataka. Predtreniranjem na velikim skupovima podataka možemo naučiti apstraktne reprezentacije koje dobro generaliziraju na nova područja primjene. Ovo je posebno korisno u područjima poput procjene cestovne sigurnosti, gdje je prikupljanje velikih količina označenih podataka zahtjevno. Suvremeni modeli kao što su CLIP i DINOv2 dodatno su unaprijedili mogućnosti prijenosa znanja kroz učenje na još većim skupovima podataka, omogućujući modelima da obuhvate širok spektar vizualnih koncepata relevantnih za prometnu infrastrukturu.

Višezadaćno učenje (multi-task learning) proširuje primjenu dubokog učenja omogućavajući jednom modelu istovremeno obavljanje više povezanih zadataka, pri čemu se reprezentacije

dijele među zadacima kako bi se poboljšala generalizacija i učinkovitost. Međutim, za uspješnu primjenu u prisustvu neuravnoteženih razreda potrebno je osigurati da se prilikom otežavanja rijetkih razreda zadrže stabilne magnitude pojedinih zadataka, kako bi se ublažila inteferencija među njima.

Rješavanje problema neuravnoteženosti razreda ključno je pri treniranju modela na stvarnim skupovima podataka gdje su određeni razredi često podzastupljeni, ali izrazito važni. Otežavanje gubitka prilagođava funkciju gubitka tako što rijetkim razredima pridaje veću važnost, osiguravajući da model ne zanemaruje kritične sigurnosne značajke. Ovo je posebno važno u primjenama kao što je procjena sigurnosti cestovne infrastrukture, gdje propust u prepoznavanju rijetkih, ali opasnih uvjeta može imati ozbiljne posljedice.

U području analize prometnih scena, modeli dubokog učenja uspješno se primjenjuju za detekciju različitih elemenata prometne infrastrukture izravno iz slika ili videozapisa. Neki od ključnih zadataka uključuju detekciju prometne signalizacije, semantičku segmentaciju, kao i sveobuhvatnu procjenu sigurnosti prometnica. Pristupi koji koriste učenje s kraja na kraj (end-to-end) pojednostavljuju proces učenja uklanjanjem potrebe za međukoracima koji mogu unijeti pogreške i iziskuju velik trud za označavanje podataka. Učenjem izravno na zadatku prepoznavanja obilježja model može razviti specijalizirane značajke koje su učinkovitije za tu namjenu, posebice kada je riječ o vizualno sličnim razredima koje zahtijevaju finu granulaciju.

Povratne neuronske mreže, posebice *Long Short-Term Memory* (LSTM) i *Gated Reccurrent Units* (GRU) mreže, omogućuju prepoznavanje vremenskih ovisnosti u slijednim podacima poput videzapisa. Uključivanjem vremenskog konteksta, ovi modeli poboljšavaju prepoznavanje obilježja koji pokazuju specifične vremenske obrasce, rješavajući opisane izazove vezane uz vremensku dinamiku iRAP obilježja. Napredne arhitekture poput LSTM-a riješavaju probleme nestajućeg ili eksplodirajućeg gradijenta, omogućavajući modeliranje dugoročnih ovisnosti.

Opisana dostignuća i pristupi dubokog učenja pružaju sveobuhvatan okvir za rješavanje složenih izazova u automatizaciji prepoznavanja sigurnosnih obilježja prometnica.

Pristup za automatsko raspoznavanje obilježja

Predlažemo model dubokog učenja za procjenu obilježja sigurnosti cestovne infrastrukture na temelju cestovnih videozapisa u dvije faze. Pristup koristi prostorne i vremenske informacije za prepoznavanje obilježja definiranih iRAP standardom, pritom rješavajući izazove višezadaćnog učenja i neuravnoteženosti razreda.

U prvoj fazi provodi se lokalno prepoznavanje konvolucijskim modelom koji obrađuje pojedinačne slike ili kratke nizove slika kako bi izvukao vizualne značajke neposrednog konteksta. Konvolucijska arhitektura sadrži dijeljeni koder i po jednu klasifikacijsku granu za svako obilježje. Zajednički koder koristi okosnicu ResNet-18 nadograđenu modulom prostornog pirami-

dalnog sažimanja, koji objedinjuje kontekstualne informacije na više mjerila kako bi proizveo dijeljenu reprezentaciju fiksne dimenzionalnosti. Koder inicijaliziramo pred-treniranjem na zadatku semantičke segmentacije uličnih scena na skupu podataka Vistas u sklopu segmentacijske arhitekture SwiftNet. Takvo pred-treniranje omogućuje izlučivanje detaljnih prostornih značajki koje doprinose prepoznavanju obilježja sigurnosti cestovne infrastrukture. Klasifikacijska grana svakog obilježja provodi sažimanje pažnjom vođeno vektorom upita (query vector) koji se uči. To omogućuje modelu da se usredotoči na dijelove mape značajki koji su najrelevantniji za određeno obilježje. Tako dobivene sažete vektorske reprezentacije zatim se konkateniraju s dijeljenom reprezentacijom dobivenom prostornim piramidalnim sažimanjem. Model se može proširiti tako da radi s više slikovnih okvira. Svaki okvir obrađuje se zasebno, a njihove se reprezentacije potom konkateniraju. Konkrento, obrađuju se slikovni okviri koji odgovaraju segmentima T, T-1 i T-4, čime se hvata širi vremenski kontekst, uz zadržavanje računalne učinkovitosti. Takav je pristup prikladan za prepoznavanje obilježja koji u pojedinačnim slikama mogu biti samo djelomično vidljivi ili zaklonjeni.

Kako bi se riješio problem izrazite neuravnoteženosti razreda u kontekstu višezadaćnog učenja, uvodimo višezadaćno dinamičko otežavanje funkcije gubitka. Težinski faktori razreda tijekom učenja dinamički se prilagođavaju njihovom odzivu. Pritom se obrnute relativne frekvencije razreda moduliraju stopom lažno negativnih predikcija koja se izračunava nakon svake epohe. Time se povećava utjecaj rijetkih razreda koje model još nije naučio, a istovremeno sprječava prekomjerno otežavanje koje bi moglo dovesti do povećanog broja lažnih pozitiva. Na razini zadatka, gubitak se normalizira s obzirom na zbroj težina primjera tog zadatka, umjesto s obzirom na sam broj primjera. Time se osigurava stabilan iznos gubitka različitih zadataka i sprječava naizmjenična dominacija pojedinih zadatka u ukupnom gubitku kroz iteracije učenja.

Druga faza uključuje slijedno poboljšanje lokalnih predikcija naučenim agregiranjem vremenskih informacija iz dužih slikovnih nizova. Ulaz je oblikovan kao niz lokalnih reprezentacija 21 uzastopnih segmenata koji su centrirani oko trenutnog segmenta, čime se obuhvaća kontekst iz prethodnih i nadolazećih segmenata. Lokalne reprezentacije segmenata su vektori logita i naučenih ugrađivanja predviđenih razreda koje na izlazu daje lokalni konvolucijski model. Četveroslojni dvosmjerni LSTM modeli koriste se za učenje vremenskih obrazaca ponašanja specifičnih za pojedina obilježja. Konačni vektori značajki stvaraju se konkatenacijom završnih skrivenih stanja oba smjera svih povratnih slojeva i skrivenog stanja središnjeg segmenta posljednjeg sloja.

Opisani pristup pruža skalabilno i učinkovito rješenje za automatsko raspoznavanje obilježja cestovne infrastrukture. Gusto semantičko pred-treniranje pospješuje prepoznavanje elemenata infrastrukture, višezadaćno dinamičko otežavanje gubitka adresira disbalans razreda prisutan u mnogim zadacima, dok faza slijednog poboljšanja usklađuje lokalne predikcije s naučenom

vremenskom dinamikom pojedinih obilježja.

Eksperimenti

Učinkovitost predloženog sustava za prepoznavanje obilježja sigurnosti cestovne infrastrukture evaluirali smo na četiri skupa podataka: našem novom skupu iRAP-BH, te tri javno dostupna skupa – Honda Scenes, FM3m i BDD100k. Skup podataka iRAP-BH, razvijen je za ovo istraživanje i sadrži više od 226.000 označenih slika snimljenih duž 2.300 kilometara prometnica u Bosni i Hercegovini. Svaki segment ceste duljine 10 metara ručno je označen vrijednostima svih iRAP obilježja, što čini ovaj skup podataka prikladnim za učenje i vrednovanje modela za raspoznavanje obilježja cestovne sigurnosti.

Mjere vrednovanja modela pažljivo su odabrane kako bi odgovarale prirodi problema koji uključuje više zadataka i više razreda, te kako bi se uspješno nosile s neuravnoteženim taksonomijama. Višerazredne klasifikacijske zadatke na skupovima podataka iRAP-BH i Honda Scenes podataka vrednujemo makro-uprosječenom F1 mjerom, kojom postižemo da svi razredi jednako doprinose ishodu vrednovanja neovisno o njihovoj učestalosti u podacima. Za skup podataka FM3m koji uključuje više zadataka binarne klasifikacije korištena je srednja prosječna preciznost (mAP), dok je za skup BDD100k kao mjera uspješnosti korištena točnost (accuracy), kako bi rezultati bili usporedivi s postojećom literaturom.

Eksperimenti na skupu podataka iRAP-BH pokazuju doprinose komponenti predloženog sustava u adresiranju identificiranih izazova. Pred-treniranje semantičkom segmentacijom poboljšalo je performanse za 1,2 postotna boda u odnosu na klasifikacijsko pred-treniranje na ImageNetu. Time je potvrđena važnost kvalitete lokalnih značajki pri prepoznavanju sigurnosnih obilježja. Višezadaćno dinamičko otežavanje gubitka dodatno poboljšava rezultate, posebno kod prepoznavanja izrazito neuravnoteženih obilježja, za koje su relativna poboljšanja iznosila od 19% do 26,5%. Slijedno poboljšavanje predikcija lokalnog modela povratnim modelima donosi povećanje od 5,1 postotni bod. Rezultati po individualnim obilježjima pokazuju da i točkasta i intervalna obilježja značajno profitiraju od slijednog poboljšavanja.

Usporedne evaluacije na skupovima podataka Honda Scenes, FM3m i BDD100k potvrđuju široku primjenjivost predloženog pristupa. Na skupu podataka Honda Scenes, naš pristup postiže bolje performanse od postojećih pristupa na sva četiri klasifikacijska problema: *Road Place*, *Road Environment*, *Road Surface*, *Weather*. To ukazuje na sposobnost adresiranja raznovrsnih izazova u razumijevanju dinamičkih prometnih scena i generalizaciju na skupove podataka koji nisu specifično dizajnirani za prepoznavanje obilježja sigurnosti cestovne infrastrukture. Na skupu podataka FM3m, naš pristup postiže kompetitivne rezultate, pri čemu višezadaćno dinamičko otežavanje gubitka i pred-treniranje na zadatku semantičke segmentacije donose značajna poboljšanja u odnosu na osnovne modele. Na skupu podataka BDD100k, naš pristup konzistentno nadmašuje postojeće pristupe, kako u standardnoj formu-

laciji tako i kod pomaka domene, što demonstrira robusnost metode.

Kvalitativna analiza pruža intuitivan uvid u način na koji model koristi prostorne i vremenske informacije za predikcije kroz nizove uzastopnih slika. Primjeri ilustriraju učinkovitost slijednog poboljšavanja u ispravljanju pogrešnih lokalnih predikcija za različita obilježja. Mape relevantnih područja slike pokazuju da se lokalni konvolucijski model fokusira na dijelove slike relevantne za određeno obilježje.

Eksperimentalni rezultati potvrđuju učinkovitost opisanog pristupa za raspoznavanje širokog spektra obilježja sigurnosti cestovne infrastrukture. Svaka od opisanih komponenti znanstvenog doprinosa konzistentno donosi porast performansi na različitim skupovima podataka.

Zaključak

Predloženi okvir predstavlja značajan napredak u automatiziranoj procjeni sigurnosti cestovne infrastrukture. Budući koraci uključuju istraživanje primjene velikih transformerskih arhitektura koje su samonadzirano naučene na izrazito velikim skupovima podataka, kao i modela za raspoznavanje u otvorenom svijetu. Kontinuiranim usavršavanjem mogućnosti automatske procjene sigurnosti cestovne infrastrukture, ovo istraživanje može doprinijeti razvoju sigurnijih prometnica, usmjeravati poboljšanja infrastrukture, te u konačnici, smanjiti broj prometnih nesreća i izgubljenih života.

Contents

Chapter 1

Introduction

Road accidents represent a critical global issue, causing over 1.35 million fatalities each year [1]. Therefore they are recognized as one of the most pressing public health challenges worldwide. This alarming statistic underscores the urgent need for effective measures to improve road safety and reduce traffic-related deaths and injuries. The United Nations' Global Plan for the Decade of Action for Road Safety [2] has established key priorities for reducing traffic fatalities worldwide. In alignment with this plan, enhancing road infrastructure safety emerges as a crucial area of focus. The Safe System Approach represents a holistic strategy for eliminating fatal and serious injuries in road traffic [3, 4]. One of the five fundamental pillars of this approach is road infrastructure safety [5, 6].

Traditionally, road infrastructure safety assessments have relied on the analysis of historical accident statistics. This reactive approach involves examining detailed records of past crashes, including information about the time, location, severity, and type of accidents [7]. The data is then used to pinpoint high-risk locations, commonly known as "black spots" [8, 9], which can be targeted for safety improvements. While this approach can uncover complex risk factors and subtle patterns in accident occurrence [10], it has several inherent limitations. It requires accidents to occur before a road section is identified as dangerous [11], potentially allowing hazardous conditions to persist undetected [6]. Additionally, the effectiveness of this approach depends on the accuracy and completeness of historical accident data, which can vary significantly, especially in regions where data collection is less comprehensive [12].

In contrast, proactive approaches to road infrastructure safety emphasize regular and systematic evaluations of the built environment to identify potential hazards before accidents occur. This perspective enables a preventative approach, aiming to mitigate risks inherent in the road infrastructure rather than waiting for adverse events to highlight them. Instead of relying on past events, these approaches leverage a deeper understanding of how specific road infrastructure features can contribute to accidents. This allows for targeted interventions and improvements based on an objective assessment of the inherent safety level of road sections. For example,

assessing the presence and quality of road markings, the adequacy of street lighting, the design of intersections, or the provision of safety features for vulnerable road users like pedestrians and cyclists [13] can reveal deficiencies that, if addressed, may prevent accidents from occurring.

A prominent example of a proactive road infrastructure safety approach is the International Road Assessment Programme (iRAP) Star Rating system [14]. This internationally recognized program provides a standardized method for evaluating and enhancing the safety of roads. The effectiveness of this approach is supported by studies demonstrating significant reductions in traffic fatalities and serious injuries following its implementation across various countries [15, 16].

Central to the iRAP methodology is a comprehensive assessment of road segments based on 52 attributes related to road design, roadside hazards, and provisions for vulnerable road users [13]. These attributes collectively evaluate the inherent safety level of road segments, quantifying the protection offered to different road users, including vehicle occupants, motorcyclists, pedestrians, and cyclists [13]. Each attribute represents a specific feature of the road environment and assumes a value from an attribute-specific taxonomy, with the number of possible values varying across attributes. This granular analysis enables targeted infrastructure enhancements that consider the unique risks faced by each group of road users.

The current iRAP coding manual [13] recommends manual collection of iRAP attributes. However, manual assessment is labor-intensive, time-consuming, and therefore not easily scalable [16, 17]. This thesis proposes a deep learning approach for automatic assessment of road-safety attributes from monocular road-driving video. We formulate attribure recognition as a multi-task multi-class classification problem [18]. Each attribute is treated as a separate task, with the classes corresponding to the possible values of that particular attribute. This approach is designed to enhance the efficiency and scalability of road safety evaluations while maintaining alignment with the iRAP methodology.

The iRAP attribute set presents several challenges for automatic recognition through deep learning. Class imbalance is prevalent throughout the dataset and across attributes, with certain classes appearing far less frequently than others. This imbalance can lead to machine learning models overlooking critical but infrequent safety features, which may lead to undetected high-risk situations. Addressing this challenge requires careful consideration during training and evaluation of learning models. Another challenge arises from the non-orthogonal design of some attributes, where multiple orthogonal features are combined into a single taxonomy. This increases the number of classes and exacerbates class imbalance by distributing examples across more categories. Many attributes exhibit fine-grained distinctions between visually similar classes, making accurate classification particularly challenging. Finally, temporal patterns in attribute behavior across consecutive road segments introduce additional complexity. Models with a limited temporal context may struggle with these patterns, failing to capture the inherent

inertia or the required temporal precision of certain attributes.

We address these challenges by introducing a two-stage visual recognition approach that addresses these complexities by complementing local recognition with attribute-specific sequential enhancement. Local recognition involves a shared encoder and per-attribute classification heads. This design enables efficient feature extraction while maintaining attribute-specific specialization. The encoder is a convolutional neural network that we pre-train for semantic segmentation of street scenes on the Vistas dataset [19]. This pre-training captures detailed spatial features that are relevant to road safety attributes, and thus addresses the challenge of visually similar classes. Per-attribute classification heads are responsible for predicting the classes of their respective attributes based on the shared features [20].

To address the challenge of class imbalance, we propose a multi-task dynamic loss weighting scheme that adjusts class weights during training according to per-class recall [21]. This technique gives higher weights to difficult underrepresented classes, while avoiding excessive false positives. The multi-task formulation ensures stable per-attribute losses, allowing for effective joint learning despite significant class imbalances.

Our detailed analysis has revealed that road safety attributes exhibit distinct temporal patterns. We therefore propose to sequentially enhance initial local predictions by incorporating a broader temporal context. More precisely, we capture and leverage the inherent temporal dependencies [22] with per-attribute recurrent models that we implement with lightweight bidirectional Long Short-Term Memory (Bi-LSTM) cells [23]. These models operate on sequences of local predictions and can thus correct errors that arise from limited spatial context and improve alignment with the annotation conventions of the iRAP standard.

We validate the effectiveness of our approach through experiments on a novel dataset iRAP-BH, which consists of fully labeled video footage along 2,300 kilometers of public roads in Bosnia and Herzegovina. We confirm the impact of each component of our contribution through ablation studies. Moreover, we apply our approach to public road scene classification datasets for the sake of comparison with the related work. Comparative experiments on Honda Scenes [24] show that our model outperforms all concurrent approaches, while competitive performance on FM3m [25] and BDD100k [26] further highlights the generalizability of the proposed approach.

The scientific contributions of this thesis are the following:

1. A technique for improving the generalization performance of visual recognition of road safety attributes by pre-training for semantic segmentation of road-driving images.
2. A technique for dynamic weighting of a multi-task supervised loss based on recall analysis across individual classes of an imbalanced taxonomy.
3. A method for sequential enhancement of local categorical predictions with efficient recurrent models.

This thesis is structured as follows. Chapter 2 provides an overview of road infrastructure safety, detailing the iRAP attribute set and analyzing the inherent challenges in automating attribute recognition. Chapter 3 discusses deep learning techniques relevant to this research, including transfer learning, multi-task learning, methods for addressing class imbalance, and approaches for temporal modeling. Chapter 4 presents the proposed two-stage framework in detail, providing a technical description and the rationale for each component. Chapter 5 describes the experimental setup, datasets, evaluation metrics and results, including ablation studies of each component and illustrative qualitative examples. Finally, Chapter 6 concludes the thesis with a summary of key findings and potential directions for future research.

Chapter 2

Road infrastructure safety

Road infrastructure safety remains a paramount concern in transportation systems globally. The assessment of road infrastructure safety has traditionally relied on historical accident statistics. Detailed data encompassing the time, location, severity, type, and other specifics of past crashes [7] are used to identify hazardous locations known as black spots [8], develop crash-risk maps [27], and construct models to predict future accidents [28]. Such methods are adept at uncovering complex and subtle risk factors that might elude history-agnostic approaches [10].

However, this reactive nature presents a significant drawback: road sections are only deemed unsafe after accidents have occurred [6]. This reliance on prior incidents means that latent dangers may persist unmitigated until sufficient accidents bring them to attention. Additionally, because severe accidents are relatively infrequent events, the historical data available can be sparse. This leads to high-variance predictions, undermining the reliability of risk assessments based solely on past accidents [12].

In contrast, proactive approaches to road infrastructure safety focus on regular inspections of static road-infrastructure features. A prominent example of such an approach is the International Road Assessment Programme (iRAP) Star Rating [14]. Recognized internationally, the iRAP Star Rating provides a standardized framework for evaluating and enhancing the safety of road infrastructure. By focusing on the physical characteristics of the road environment, the iRAP Star Rating enables the identification of potential hazards without the prerequisite of prior accidents. The efficacy of this method is also supported by empirical evidence. Studies show that its implementation can lead to significant reductions in traffic fatalities and serious injuries by identifying high-risk roads and facilitating targeted improvements across various countries [15, 16].

The iRAP methodology assesses the inherent infrastructure safety of road segments by evaluating 52 specific attributes [13]. These attributes cover a wide range of factors, including road design elements, roadside hazards, and the provision of facilities for vulnerable road users like pedestrians and cyclists. We proceed with a detailed description of the iRAP attribute set.

2.1 Road-safety attributes

The iRAP Star Rating system provides a comprehensive quantification of the protection offered by road infrastructure to the four most common types of road users: vehicle occupants, motorcyclists, pedestrians, and bicyclists [13]. The assessment methodology focuses on categorical values of a carefully selected set of 52 attributes related to road-infrastructure elements and roadside objects within each corresponding road segment.

Each attribute represents a specific feature of the road environment and assumes a class from its attribute-specific taxonomy. The number of possible classes varies across different attributes. The attributes with the most numerous classes are those encoding the speed limit (21 classes), roadside severity (17 classes), and intersection type (16 classes). Conversely, the dataset also includes 11 binary attributes. Given this variability, we formulate the problem of attribute recognition as a set of distinct multi-class classification tasks, one for each attribute.

During our study, we found it necessary to exclude certain attributes from our analysis. We discarded four attributes that assume only a single class throughout our entire dataset: *Shoulder rumble strips*, *Centre line rumble strips*, *Motorcycle facility*, and *Pedestrian fencing*. In addition, we excluded five attributes that are not suitable for visual recognition. These include the four speed limit attributes and the *Intersecting road volume* attribute. The speed limit attributes pose a unique challenge as they require knowledge of speed regulations beyond what's visually apparent in a local context. The difficulty lies not in recognizing speed limit signs, but rather in having the knowledge of speed limit signs that may have been placed well outside the immediate visual context captured by our system. Similarly, the *Intersecting road volume* attribute, which captures the average daily traffic from intersecting roads, is more appropriately estimated using data from traffic studies, road counters or aerial analysis rather than from road-driving imagery. Consequently, our experiments focus on the remaining 43 iRAP attributes that are suitable for visual recognition from road-driving imagery. We proceed with a detailed description of these attributes and their respective groups in the following subsections.

2.1.1 iRAP Attribute Groups

The International Road Assessment Programme (iRAP) has established a comprehensive set of attributes to assess road safety [13]. The iRAP standard organized the attributes into seven distinct groups, each focusing on different aspects of road infrastructure and its surroundings. This subsection provides an overview of these attribute groups and highlights their key characteristics.

Road and context attributes (1 attribute) include the attribute *Carriageway label* along with twelve metadata attributes related to data acquisition and annotation processes, such as coder name, coding date, and road name.

Observed flow attributes (5 attributes) quantify the presence of various road users in a given segment. These attributes are determined by counting the occurrences of motorcycles, bicycles, and pedestrians in the recorded segments.

Speed limit attributes (5 attributes) capture the speed limits (four attributes) and the presence of speed-reducing infrastructure such as speed bumps. Although speed limit recognition poses challenges for visual recognition due to the need for broader context, speed-reducing features like speed bumps can often be identified from road-driving imagery.

Mid-block attributes (16 attributes) focus on the intrinsic features of the road rather than its surroundings. These attributes cover various aspects such as road condition and delineation, curvature and sight distance, skid resistance, the presence street lighting, vehicle parking, and the width and number of lanes. The attribute *Median type* stands out as particularly challenging, as it requires distinguishing among fifteen kinds of physical separators or median markings.

Roadside attributes (7 attributes) assess the risk associated with roadside features on both the passenger and driver sides. A notable attribute in this group is *Roadside Severity*, which identifies the most hazardous roadside object based on its type and proximity to the road. The ground truths for this attribute are assigned according to the priority table defined within the iRAP standard [13], which ranks combinations of object types and distances according to risk level. These attributes present significant challenges for classifiers due to two main factors. First, the front dashboard camera captures only a fraction of each roadside, necessitating the use of imagery from previous and subsequent segments. Second, the simultaneous presence of various objects (such as houses, fences, and trees) on the roadside complicates the estimation of severity levels for road users.

Intersection attributes (5 attributes) capture various characteristics of intersections. Most notably, the attribute *Intersection type* has sixteen classes that cover different combinations of intersecting roads, signalization, and special features like roundabouts and railway crossings.

Vulnerable road-user facilities and land use attributes (13 attributes) detail the presence of amenities for pedestrians, cyclists, and motorcyclists, while also capturing characteristics of the surrounding area, such as area type, land use, and the presence of school zones.

2.2 Analysis of road-safety attributes

In this section, we conduct a comprehensive conceptual and empirical analysis of the iRAP attributes present in our dataset. Our objective is to identify inherent challenges associated with these attributes that can significantly impact the performance of machine learning models in road safety assessment.

2.2.1 Class Imbalance

A significant challenge in our dataset is the prevalence of class imbalance among many attributes. Class imbalance occurs when there is a substantial disproportion in the number of examples belonging to different classes within an attribute. For example, certain classes may be overwhelmingly represented while others are rare. This imbalance can have detrimental effects on the performance of classifiers that prioritize overall accuracy, often resulting in the underrepresentation or complete disregard of minority classes [29].

The issue of class imbalance is particularly problematic in the context of road safety assessment, where rare but critical safety features might be overlooked due to their infrequent occurrence in the dataset. Ignoring these features could lead to models that fail to detect high-risk situations, undermining the effectiveness of safety interventions.

To mitigate this challenge, the method described in this work incorporates improved loss functions that assign higher weights to minority classes, encouraging the model to focus more on underrepresented attributes during training. Additionally, macro-F1 is employed as the main evaluation metric, since it accounts for both precision and recall across all classes equally, providing a more balanced assessment of model performance.

2.2.2 Non-orthogonal design

Building upon the issue of class imbalance, another contributing factor is the structure of certain iRAP attributes. Some attributes incorporate multiple, seemingly orthogonal features within a single taxonomy, resulting in classes that are effectively Cartesian products of these features. This design choice not only increases the total number of classes but also exacerbates imbalance by distributing the already rare examples across more categories.

Consider the attribute *Skid resistance*, which is intended to capture both the skidding resistance level and the type of road surface sealing. Specifically, it spans two dimensions: the level of surface grip — categorized as low ("poor"), medium, or adequate — and the type of road surface — either sealed (with a protective layer like asphalt) or unsealed (gravel or dirt road without such a layer). This leads to five combined classes: *Unsealed - poor*, *Unsealed - adequate*, *Sealed - poor*, *Sealed - medium*, and *Sealed - adequate*. While comprehensive, this structure introduces additional complexity that could potentially be simplified.

An alternative, more orthogonal formulation could have separated these concepts into two distinct attributes: *Sealed road*, a binary attribute indicating whether the road is sealed, and *Surface grip*, classifying the grip level as poor, medium, or adequate. Such a decomposition might offer several advantages. Firstly, it would provide a clearer, more intuitive representation of the road characteristics. Secondly, and perhaps more significantly in the context of machine learning, it would reduce the number of classes and increase the sample size per class, thereby

alleviating issues of class imbalance.

2.2.3 Fine-grained and visually similar classes

Some attributes within the iRAP framework are defined with a high degree of specificity, leading to numerous classes that differ only subtly in appearance. Such a nuanced classification scheme, while comprehensive, can lead to difficulties in accurate recognition and exacerbate existing class imbalance issues.

A prominent example is the *Roadside severity* attribute, which encompasses a diverse array of safety barrier types including metal, concrete, wire, and motorcycle-friendly variants. Furthermore, it includes a distinct class for semi-rigid structures such as various fences. These classes are meticulously defined to capture specific safety features, but their visual similarities and the multitude of options can lead to misclassifications in practice.

In light of these challenges, it may be advantageous to reconsider the level of granularity in attribute definitions. Consolidating visually similar classes could improve the overall model accuracy at the cost of some granularity in the output. This trade-off might be particularly beneficial in scenarios where broad categorization is sufficient for decision-making processes related to road safety interventions.

2.2.4 Temporal behaviour

Our dataset comprises videos covering extensive road sections, each composed of sequences of successive 10-meter segments. Along these sequences, our analysis uncovers distinct patterns in the temporal behavior of certain attributes.

According to the iRAP standard [13], some attributes have a default "negative" class. These attributes typically capture countable occurrences of various infrastructure elements, such as intersections or pedestrian crossings. The default class for these attributes is usually *None*, while the "positive" classes correspond to specific instances of the attribute (e.g., *3-leg intersection*). The iRAP standard stipulates that any occurrence of a positive class should be annotated only in the segment closest to its occurrence; with all neighboring segments annotated with the negative class. We refer to such attributes as "single-peak" attributes.

Consider a segment containing an occurrence of a positive class (a peak) and the neighboring segments immediately preceding it. The visual features that a model might leverage to recognize such an attribute are likely present in these neighboring segments as well. For instance, an intersection gradually becomes more visible in the segments leading up to the peak segment. Therefore, a visual recognition model predicting an intersection in a neighboring segment is not entirely incorrect. Furthermore, the model may struggle to discern which exact segment represents the peak.

In contrast to single-peak attributes, there exists another subset of attributes that we denote as "smooth" attributes. These attributes rarely change classes and generally do not oscillate. They describe larger areas, environments, zones, or infrastructure features likely to remain constant across consecutive segments. Examples of smooth attributes include *Area type*, *Road delineation*, and *Carriageway label*.

While many attributes can be categorized as either single-peak or smooth, some attributes exhibit more complex patterns that don't fit neatly into either category. The attribute *Street lighting* exhibits particularly peculiar temporal behavior. It is treated as a single-peak attribute when only a single light post appears in isolation. However, for a sequence of light posts, *Street lighting* should be recognized as present in all segments from the first to the last light post. Given that light posts in such sequences can be up to 100 meters apart, it can be difficult for a vision-based model to distinguish between single occurrences and sequences of street lights.

Figure 5.3 illustrates examples of four iRAP attributes. It presents sequences of five frames from consecutive 10-meter segments, along with the ground truth labels for each corresponding attribute. In row 1, only the third segment is annotated with the positive class of the single-peak attribute *Intersection type*. Conversely, the smooth attributes in rows 3 and 4 maintain constant ground truth labels throughout the sequences. Additionally, the ground truth label of *Street lighting* remains unchanged even when the discriminative visual features are not visible in certain segments.

2.2.5 Motivation for sequential enhancement

Building upon the analysis of temporal behavior patterns presented in the previous section, we now examine class co-occurrence along consecutive segments of the iRAP-BH dataset. This investigation will provide insights into the limitations of our local recognition model and will motivate the need for sequential enhancement.

For a given attribute A and a pair of consecutive segments t and $(t+1)$, we define a co-occurrence as the pair of corresponding classes $(c_{A,t}, c_{A,t+1})$. For an attribute with n classes, we can construct an $n \times n$ co-occurrence matrix where each element (i, j) represents the number of occurrences where $c_{A,t} = i$ and $c_{A,t+1} = j$. We construct two such matrices for each attribute: one using the ground truth labels and another using the predictions produced by our local recognition pipeline described in 4.1.

The expected structure of these matrices varies depending on the attribute type. For single-peak attributes, ground truth transitions can only occur between the default class and a positive class. Consequently, their ground truth co-occurrence matrices have non-zero elements only in the row and column corresponding to the default class. In contrast, smooth attributes typically maintain the same class across consecutive segments, resulting in ground truth matrices with significantly larger diagonal elements compared to off-diagonal elements.

Our analysis reveals consistent discrepancies between the ground truth and the local prediction matrices for both groups of attributes, as shown in examples in Figure 2.1. For single-peak attributes, the local prediction matrices exhibit numerous non-zero diagonal values outside the default class row and column. These discrepancies arise when the model assigns the same positive class to two consecutive segments that are visually very similar (e.g., two segments within an intersection). While this is a reasonable error from a visual recognition standpoint, it indicates that the local recognition model fails to internalize the single-peak annotation convention specified by the iRAP standard [13]. In the case of smooth attributes, we observe significantly larger off-diagonal elements in local prediction matrices compared to ground truth matrices. This indicates the presence of spurious class transitions in local predictions across consecutive segments, suggesting that our local recognition model struggles to capture the inherent inertia of smooth attributes.

These findings motivate the extension of the local recognition pipeline with per-attribute sequential enhancement models. These models are designed to learn temporal behavior patterns across larger context windows without requiring computationally expensive backpropagation through hundreds of video frames. By leveraging these sequential models, we aim to refine local predictions by accounting for temporal context, thereby aligning them more closely with the ground truth annotations.

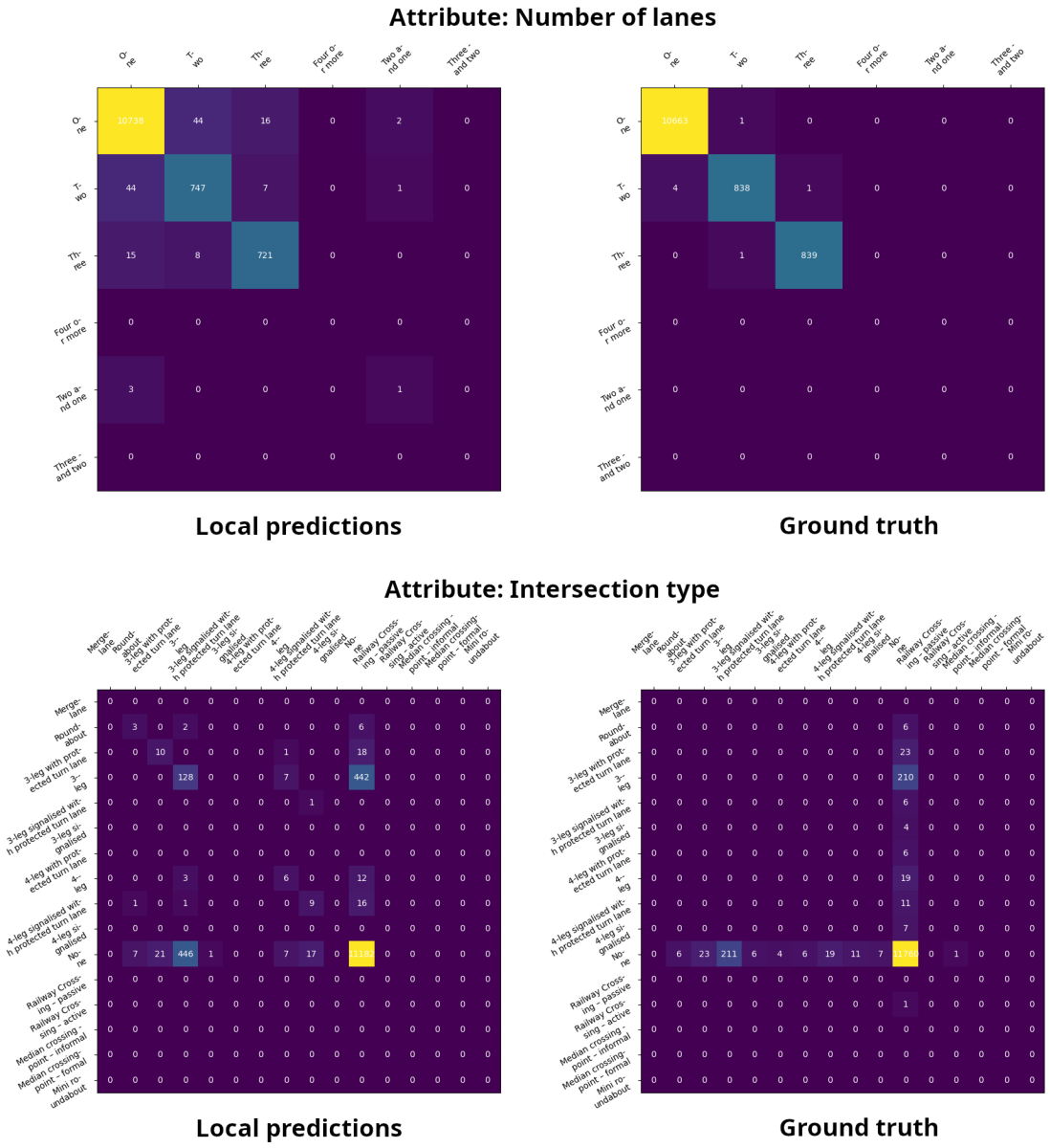

Figure 2.1: Co-occurrence matrices illustrating class transitions between consecutive segments for the smooth attribute "Number of lanes" (top row) and the single-peak attribute "Intersection type" (bottom row). In each row, the local prediction matrix is on the left, and the ground truth matrix is on the right. For smooth attributes, the prediction matrix shows excessive class transitions compared to the strongly diagonal ground truth matrix. For single-peak attributes, the ground truth matrix contains transitions only between the default class and positive classes, while the prediction matrix shows occurrences of consecutive positive classes.

Chapter 3

Deep learning for visual recognition

The field of computer vision encompasses a wide array of tasks and applications, where image classification stands out as a fundamental challenge [30]. Image classification involves assigning a single label to an entire image, categorizing it into one of a predefined set of classes. This task is crucial for numerous applications, including medical imaging [31], autonomous driving [32], and facial recognition [33], where accurate image categorization is critical.

The importance of image classification extends beyond its immediate application. It serves as a cornerstone for more complex computer vision tasks. Deep learning models for classification, particularly those trained on large-scale datasets, have shown remarkable ability to learn rich and robust feature representations directly from raw image data. Thus, these models not only achieve exceptional performance in classification tasks but also act as powerful feature extractors [34] for other problems like semantic segmentation, object detection, and instance segmentation.

3.1 Machine learning

Machine learning is a subfield of artificial intelligence that focuses on developing algorithms capable of learning patterns from data to make predictions or decisions without being explicitly programmed for specific tasks [35]. In the context of pattern recognition, the objective is to learn a function f that maps high-dimensional, unstructured raw input data $\mathbf{x} \in \mathbb{R}^n$ to a semantically meaningful output $\mathbf{y} \in \mathbb{R}^m$ specific to the task domain [36]. For image classification, the input $\mathbf{x}$ represents the pixel values of an image, and the output $\mathbf{y}$ is a label indicating the class to which the image belongs.

Traditionally, machine learning in image classification relied on hand-crafted features [37]. Techniques such as the Bag-of-Visual-Words model [38] and Fisher Vectors [39] were commonly employed.

These supervised learning approaches were characterized by having the mechanisms of pat-

tern recognition encoded implicitly through, and emerging from, several key elements:

- **Feature Extraction**: Manually designing algorithms to extract informative features from raw image data.
- **Data Annotation**: Labeling data to provide supervised learning signals.
- **Algorithm Architecture**: Constructing pipelines or sequences of parameterized operations to process the extracted features.
- **Parameter Selection**: The procedure by which the parameters of the operations are chosen.

In this paradigm, the function f is parameterized by a set of parameters θ, and the goal is to find θ such that $f(\mathbf{x};\theta)$ accurately maps inputs to outputs. The knowledge implicit in the data is encoded into the parameters through the learning process [36]. Machine learning algorithms adjust these parameters by optimizing an objective function, typically using optimization techniques like gradient descent. The loss function $\mathscr{L}(\mathbf{y}, f(\mathbf{x};\theta))$ quantifies the discrepancy between the predicted outputs and the true labels, guiding the parameter updates.

From a high-level perspective, the learning process involves the algorithm adjusting its parameters to:

1. **Identify Relevant Patterns**: Detect and preserve patterns in the input data that are pertinent to the desired output values.
2. **Achieve Invariance**: Learn to abstract away variations in the input data that are irrelevant to the output, such as noise or distortions, thereby achieving invariance to these factors.
3. **Map Patterns to Outputs**: Transform the identified relevant patterns into correct output predictions. This is typically accomplished by funneling the input through a series of parameterized affine transformations, non-linear activation functions, and other similar operations.

Despite the effectiveness of these traditional approaches, manual feature extraction often required extensive domain expertise and could fail to capture the complex patterns inherent in raw data. The transition from manual feature engineering to automated feature learning marks a significant paradigm shift in machine learning. Deep learning leverages large datasets and computational resources to learn hierarchical feature representations directly from raw data [40]. This approach reduces the reliance on domain-specific knowledge for feature extraction, allowing models to automatically discover the optimal features for a given task.

3.2 Development of deep learning

Advancements in machine learning in recent years have propelled the field of computer vision. Specifically, the advent of deep learning has led to a resurgence of neural network-based methods. The success of deep learning methods can be attributed to a confluence of several factors.

The availability of large annotated datasets, such as ImageNet [41, 42], provided the necessary data to train deep neural networks effectively. Advances in GPU hardware and specialized software libraries for parallel computing [43] enabled efficient training of complex models. The development of programming frameworks, like TensorFlow [44] and PyTorch [45], facilitated the implementation and training of neural networks. Finally, novel neural network building blocks and optimization techniques enhanced performance on various tasks.

The superiority of deep learning methods became increasingly apparent as neural networks began to dominate the state-of-the-art in numerous subfields of computer vision [46, 47, 48, 49, 50]. This dominance led to a steep rise in the popularity of deep learning methods, fostering even more advances in the field. Even after more than a decade, the full potential of these advances has yet to be fully exploited and applied across different areas of science and engineering.

Deep learning algorithms are a subset of machine learning algorithms, with some specific characteristics [36]. Their pattern recognition function is composed of multiple processing steps, known as layers, which operate sequentially. Each layer takes as input the output of the preceding layer, except for the first layer, which works on raw input data. This hierarchical composition enables the network to learn complex, high-level representations by building upon simpler, lower-level features extracted by earlier layers.

Formally, a deep neural network can be viewed as a function $f_\theta(\mathbf{x})$ composed of L layers:

$$f_\theta(\mathbf{x}) = f^{(L)}_{\theta_L} \circ f^{(L-1)}_{\theta_{L-1}} \circ \cdots \circ f^{(1)}_{\theta_1}(\mathbf{x}), \tag{3.1}$$

where $\mathbf{x}$ is the input data, $\theta = \bigcup_{l=1}^{L} \theta_l$ represents the set of all learnable parameters, and $f^{(l)}_{\theta_l}$ denotes the function implemented by the l-th layer with its corresponding parameters θ_l.

This hierarchical structure is advantageous because high-level concepts in domains like visual perception and language understanding are often combinations of lower-level features. Deep networks exploit this by reusing and combining features learned in earlier layers to recognize more abstract patterns in later layers.

To successfully determine a set of parameters that enables a deep learning model to recognize patterns, deep learning algorithms must satisfy the following. They need to define a loss function $\mathscr{L}(\mathbf{y}, f(\mathbf{x};\theta))$, which quantifies the discrepancy between the network's predictions and the true labels $\mathbf{y}$ for a given input $\mathbf{x}$. Furthermore, all layers, including the loss function, must be differentiable mathematical operations [51]. This ensures that the entire model function — mapping the input through all layers up to the loss function — is differentiable. Meeting this conditions enables the use of gradient-based optimization methods to adjust the model parameters θ [36].

To efficiently compute the gradients required for optimization, deep learning algorithms

commonly a procedure known as backpropagation [51]. It leverages the chain rule of calculus to propagate gradients through the network layers, with computational complexity linear in the number of layers. Backpropagation shares many of the same operations as forward propagation, which can be expressed as matrix multiplication. This enables deep learning algorithms to leverage the power of efficient general matrix multiplication algorithms and GPUs that support certain general-purpose computing operations.

Neural network parameters are iteratively updated by taking small steps in the direction opposite to the gradient of the loss function with respect to the parameters. This process continues until a satisfactory level of performance - usually measured on a portion of the data not seen during optimization - is achieved. Even with the non-convex nature of loss functions of multi-layer networks and the use of mini-batch stochastic gradient descent - which only approximates the gradient based on small subsets of input data - these optimization procedures have consistently produced models that perform well on various pattern recognition tasks [36].

Initially, machine learning models, including early deep learning approaches, were integrated as components within larger, hand-crafted systems. These systems combined classical algorithms with machine-learned subtasks to achieve the final output. Examples include stereo reconstruction [52], pedestrian tracking [53], and object detection [34]. Over time, there had been a shift toward end-to-end deep learning models that would replace hand-crafted modules entirely, directly mapping raw inputs to desired outputs. Some notable examples of this trend were end-to-end stereo reconstruction [54] and object detection models like Faster R-CNN [55].

Training individual components separately may lead to suboptimal performance, as these components are optimized for proxy sub-tasks that may not align perfectly with the final objective. In contrast, end-to-end training allows the model to learn intermediate representations that are most beneficial for the overall task, since all parameters are optimized jointly. Furthermore, it also alleviates the need to compress intermediate results into human-interpretable forms, which can act as information bottlenecks. End-to-end training also enables the mapping between different data modalities (e.g., images to textual descriptions) without explicitly defining intermediary representations. Additionally, it can reduce data annotation efforts by requiring labels only for the final output, rather than for each subtask.

However, it is worth noting that designing systems with modular components might offer advantages in terms of transparency and interpretability. Such systems may provide insights into internal operations, which may be crucial for safety-critical applications where understanding and controlling the system's behavior is essential to prevent undesirable or dangerous decisions.

3.3 Building blocks of deep learning algorithms

Various mathematical operations are used as layers in a deep neural network.

One fundamental type is the *fully-connected layer*, which performs an affine transformation by multiplying the input vector $\mathbf{x}$ with a weight matrix $\mathbf{W}$ and adding a bias vector $\mathbf{b}$:

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}.$$

This layer does not exploit any knowledge about the topological relationships among the elements within the input vector.

Activation function layers are used to introduce non-linearity to the network. They apply a specified function to each element of the input vector independently. A widely used activation function is the Rectified Linear Unit (ReLU) [56], defined as:

$$\mathrm{ReLU}(x) = \max(0, x).$$

ReLU activation functions allow for better gradient flow during training with backpropagation [57], mitigating the vanishing gradient problem that had been common in deep networks.

Convolutional layers extend convolutional filters - traditionally used in computer vision - with learnable parameters, to learn feature representations from data. These layers can be seen as fully-connected layers with weight sharing and locality constraints.

Specifically, a convolutional layer operates on small, localized regions, applying the same filter across the entire input in a sliding-window approach. This is mathematically represented as:

$$y_{i,j,k} = \sum_c \sum_m \sum_n w_{m,n,c,k} \cdot x_{i+m,j+n,c},$$

where $x_{i,j,c}$ is the input tensor, $w_{m,n,c,k}$ is the filter tensor, and $y_{i,j,k}$ is the output tensor. The convolutional layer is translation equivariant, meaning that a translation in the input results in a corresponding translation in the output.

Pooling operations progressively downsample the spatial dimensions of the feature maps, reducing computational complexity and increasing the receptive field of subsequent layers. For instance, max pooling selects the maximum value within a pooling window, while average pooling computes the mean. These operations introduce a degree of *translation invariance*, making the network less sensitive to small input shifts.

Batch normalization [58, 59] is a technique used to stabilize and accelerate training by normalizing the output of a layer to have zero mean and unit variance. It then re-scales and re-centers the normalized output using learnable parameters γ and β:

$$\hat{x} = \frac{x - \mu}{\sigma}, \quad y = \gamma\hat{x} + \beta,$$

where μ and σ are the mean and standard deviation computed over a mini-batch. Batch normalization acts as a form of regularization and is said to address the problem of internal covariate

shift.

Most image classification architectures consist of a series of blocks comprising convolutional, activation, pooling, and normalization layers, ending in a fully-connected layer that serves as a classifier. The layers preceding the final fully-connected layer can be viewed as a *feature extractor*, transforming the input data into a high-level representation suitable for classification.

To feed the multi-dimensional tensor produced by the feature extractor into the fully-connected classifier, it must be transformed into a one-dimensional vector. One approach is to *flatten* the tensor, but this can result in an impractically large number of parameters in the fully-connected layer, especially for high-resolution inputs. Moreover, such architectures are constrained to fixed input dimensions.

An alternative is to employ *global pooling layers*, which produce fixed-size representations regardless of the spatial dimensions of the input. *Global average pooling* [36, 60] computes the average of each feature map, yielding a vector whose length equals the number of feature maps in the input tensor. This reduces the parameter count and allows the network to handle variable-sized inputs.

Spatial pyramid pooling [61, 62] generalizes global pooling by applying it across multiple scales. It partitions the input tensor into divisions of varying sizes and performs pooling within each division. This approach captures information at different scales and preserves fine-grained spatial details [63].

Attention mechanisms have become increasingly popular in modern deep learning architectures. They allow the network to focus on specific parts of the input data when generating outputs, by computing a weighted sum of input features. The weights are dynamically determined affinity scores between the input features and query vectors, which represent the context. The queries can be obtained from intermediate latent representations (e.g., the hidden state in a recurrent network), randomly initialized learnable vectors, or from input features themselves (self-attention). Learnable query vectors can be used to perform attention-based pooling to selectively aggregate information. For instance, in a multi-task setting, a unique query vector can be learned to emphasize task-relevant features [20].

Recurrent neural networks (RNN) cells are layers specifically designed to process sequential data. They maintain an internal hidden state that captures temporal dependencies in the input sequence. Standard recurrent networks can suffer from issues like vanishing or exploding gradients when dealing with long sequences. More sophisticated recurrent units such as the *Long Short-Term Memory* (LSTM) [23] and *Gated Recurrent Units* (GRUs) [64, 65] address these issues. They incorporate gating mechanisms to control the flow of information, allowing them to learn long-term dependencies more effectively.

An LSTM cell maintains a cell state c_t and hidden state h_t, controlled by three gates. The

operation of an LSTM cell at each time step t is defined by:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{3.2}$$

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{3.3}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{3.4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{3.5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{3.6}$$

where σ is the sigmoid function, W_i, W_f, W_o, W_c are learnable weight matrices, b_i, b_f, b_o, b_c are bias terms, and $\odot$ denotes element-wise multiplication. Recurrent layers using LSTM cells are widely used in tasks involving temporal dependencies, such as video recognition and natural language processing.

3.4 Popular convolutional architectures

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [42] has played a pivotal role in advancing the field of computer vision and deep learning. By providing a large-scale benchmark dataset comprising millions of labeled images across thousands of categories, it enabled researchers to develop and evaluate increasingly sophisticated models. Achieving top results on the ILSVRC became a benchmark for success, and surpassing human-level accuracy was a notable milestone that signified the maturity of deep learning approaches in visual recognition tasks.

In 2012, AlexNet [30] became the first deep learning model to achieve state-of-the-art results on the ILSVRC, with a top-5 error rate of 16.4%. The architecture consisted of five convolutional layers followed by three fully connected layers. Key innovations introduced by AlexNet included the use of the Rectified Linear Unit (ReLU) activation function, which addressed the vanishing gradient problem by allowing for better gradient propagation during training [56]. Additionally, AlexNet leveraged graphics processing units (GPUs) for accelerated computation, enabling the training of deep networks on large datasets. The model also employed data augmentation techniques such as random cropping and horizontal flipping to increase the diversity of training samples and mitigate overfitting.

In 2014, the VGG model [66] further advanced convolutional architectures by systematically investigating the effect of network depth on performance. The 19-layer variant of the model achieved a top-5 error rate of 7.3% on the ILSVRC. A distinctive feature of VGG was the use of small 3×3 convolutional filters throughout the network. This approach allowed for a deeper network architecture while still maintaining a relatively small number of parameters. Despite its simplicity, the VGG architecture demonstrated that network depth is a critical component for

achieving high performance in convolutional neural networks.

The introduction of the ResNet architecture [67, 68] in 2015 marked a significant breakthrough in training very deep neural networks. ResNets surpassed human-level performance on the ILSVRC, achieving a top-5 error rate of 3.6%. The key innovation was the introduction of residual learning through the use of skip (residual) connections. This was motivated by the observation that simply stacking more neural networks layers leads to higher training error - also referred to as the "degradation problem" [67]. This phenomenon was surprising, considering that a deeper network that avoids training performance degradation could be constructed manually merely by inserting identity mapping layers into a shallower network. The fact that not even this, more optimal, solution was found through training pointed to optimization difficulties specific to deeper networks.

ResNets address this by reformulating the mapping between layers as learning residual functions with reference to the layer inputs, denoted as:

$$\mathbf{y} = \mathbf{F}(\mathbf{x}, \{\mathbf{W}_i\}) + \mathbf{x},$$

where $\mathbf{x}$ is the input to the residual block, $\mathbf{F}(\mathbf{x}, \{\mathbf{W}_i\})$ represents the residual function to be learned (typically a series of convolutional layers), and $\mathbf{y}$ is the output. The skip connection adds the input $\mathbf{x}$ directly to the output of the residual function, with the hypothesis that this residual mapping is easier to optimize than the original mapping. The aforementioned identity mappings could now be achieved simply by setting the corresponding residual block parameters to zero. The hypothesis was confirmed by state-of-the-art experimental results and the innovation enabled the training of much deeper networks. ResNet variants range from 18 layers (ResNet-18) up to 152 layers (ResNet-152).

Building upon the concept of residual connections, DenseNets [69] introduced dense connectivity between layers within dense blocks. In a dense block, each layer receives input from all preceding layers within the block through feature concatenation, rather than summation as in ResNet. This design promotes feature reuse and improved information flow throughout the network. When they were introduced, DenseNets achieved state-of-the-art performance on several classification benchmarks while being more parameter-efficient [69, 70].

More recently, ConvNeXt [71] has emerged as a modern reinterpretation of convolutional neural networks inspired by the success of vision transformers [72]. ConvNext incorporates several key architectural modifications: depth-wise separable convolutions [73], inverted bottleneck blocks [74], and Layer Normalization [75] replacing traditional Batch Normalization. Additionally, it features expanded kernel sizes, wider network layers, and adopts the Gaussian Error Linear Unit (GELU) [76] for activation. These design choices result in a convolutional architecture that achieves performance comparable with vision transformers on large-scale image

classification tasks while retaining the advantages of convolutional networks.

3.5 Transfer learning

The introduction of large-scale datasets like ImageNet [41] has facilitated a paradigm shift from isolated learning to *transfer learning* in the field of computer vision.

In many domains, collecting vast amounts of labeled data for every possible specific task is impractical or cost-prohibitive. Transfer learning involves leveraging knowledge acquired from solving a general task with abundant data to enhance performance on a specific task with limited data [77, 78, 79]. It allows models to capitalize on representations learned from large, diverse datasets. This approach is particularly effective when the source and target domains share underlying similarities in their feature spaces [80]. By transferring knowledge from a broader domain to a more specific one, models can achieve improved performance and generalization, often with reduced training time and data requirements.

In the context of deep learning for image recognition, transfer learning often involves pre-training a convolutional neural network on ImageNet and using its learned convolutional layers in the task-specific model [79, 81]. This process leverages the hierarchical feature representations learned from the source task, which are often general and transferable across different visual domains. The sequence of pre-trained convolutional layers can be used as a frozen feature extractor [82] or fine-tuned along with the task-specific output layers [83]. Fine-tuning allows the model to adjust the pre-trained representations to better fit the target domain.

Empirical evidence suggests that this approach improves performance on various image recognition tasks, particularly when the target dataset is small [81, 84]. Pre-training reduces the risk of overfitting, accelerates convergence during training, and enhances generalization capabilities [80]. For instance, Oquab et al. [79] showed that features learned on ImageNet can be effectively transferred to tasks such as object and action recognition with minimal adaptation. Similarly, Oršić et al. [80] demonstrated that pre-trained ImageNet architectures improve real-time semantic segmentation performance in road-driving images.

Moreover, the success of transfer learning is influenced by the similarity between the source and target domains and tasks [81, 85]. When the source and target tasks are closely related, the transferred features are more likely to be beneficial. Conversely, significant domain discrepancies may reduce the effectiveness of transfer learning, necessitating techniques such as domain adaptation to bridge the gap [86, 87, 88].

Recently, the development of modern foundation models has further advanced the capabilities of transfer learning. These models, trained on massive and diverse datasets, yield rich transferable representations. Two notable examples are the vision-language model CLIP [89] and the self-supervised model DINOv2 [90].

CLIP (Contrastive Language-Image Pre-training) learns visual concepts from natural language supervision via contrastive learning. The training objective encourages the model to produce image and text embeddings that are close if they are associated and distant otherwise. Aligning visual and textual representations enables zero-shot transfer to various vision tasks without task-specific fine-tuning, simply by providing textual descriptions.

DINOv2, on the other hand, employs self-supervised learning techniques to extract meaningful representations from unlabeled data at scale. It extends the self-distillation with no labels (DINO) framework, where the model is trained to predict its own representations under different augmentations. The teacher-student setup encourages the student network to match the teacher's output, which is computed on a differently augmented view of the same image. By training on massive unlabeled datasets, DINOv2 captures a wide variety of visual concepts and exhibits strong transfer performance on downstream tasks.

3.6 Multi-task learning

Multi-task learning is a machine learning paradigm where a single model is trained to perform multiple tasks simultaneously [18, 91]. Unlike transfer learning, however, these tasks are learned jointly by leveraging shared representations, with the aim of improving generalization across all tasks.

The typical architecture of a multi-task model consists of shared layers followed by task-specific branches [91]. The shared layers capture common features from the input data. Each task-specific branch processes these shared features to produce task-specific outputs.

Let $\mathbf{x}$ denote the input data, and suppose there are T tasks. The shared representation $\mathbf{h}$ is computed as:

$$\mathbf{h} = f_{\text{shared}}(\mathbf{x}; \theta_{\text{shared}}), \tag{3.7}$$

where f_{shared} represents the shared layers parameterized by θ_{shared}. For each task $t \in \{1, 2, \ldots, T\}$, the task-specific output is obtained via:

$$\hat{\mathbf{y}}^{(t)} = f_{\text{task}}^{(t)}(\mathbf{h}; \theta_{\text{task}}^{(t)}), \tag{3.8}$$

where $f_{\text{task}}^{(t)}$ denotes the task-specific layers with parameters $\theta_{\text{task}}^{(t)}$.

Each task typically has its own loss function $\mathscr{L}^{(t)}$ [92]. The overall loss is often formulated as a weighted sum of these individual losses:

$$\mathscr{L}_{\text{total}} = \sum_{t=1}^{T} \alpha^{(t)} \mathscr{L}^{(t)}, \tag{3.9}$$

where $\alpha^{(t)} \geq 0$ are weighting coefficients that balance the contribution of each task [93]. The

model is trained end-to-end by minimizing $\mathscr{L}_{\text{total}}$ with respect to all parameters:

$$\theta_{\text{shared}}, \theta_{\text{task}}^{(1)}, \ldots, \theta_{\text{task}}^{(T)} = \arg\min_{\theta} \mathscr{L}_{\text{total}}. \tag{3.10}$$

Alternatively, training can involve alternating optimization, where the model intermittently optimizes individual task losses during training [94].

Multi-task learning can offer some advantages over training separate task-specific models. Feature sharing enables the transfer of knowledge across tasks and might lead to more robust features and better generalization [18, 95]. At the same time, training and inference are faster since a large portion of the network is shared among tasks, reducing the need to run separate models [95].

However, it is important to note that multi-task learning does not always guarantee superior or even equal performance compared to individually trained tasks. Not all tasks are mutually beneficial when learned together. Empirical studies have shown that some groups of tasks may not be mutually beneficial [96]. These incompatible tasks can interfere with each other, leading to decreased performance compared to single-task learning [96, 97]. Task interference can occur due to conflicting objectives [98], imbalanced task importance [93], and limited model capacity [99].

3.7 Learning on Imbalanced Datasets

Class imbalance is a prevalent issue in real-world classification tasks, where examples across different classes are very disproportionately distributed [29]. Imbalance can make it difficult for accuracy-oriented classifiers to recognize underrepresented classes.

Addressing this issue becomes particularly important when the underrepresented classes are of equal or greater importance than the majority classes, as is often the case in safety-critical applications. For instance, in the context of road-safety assessment, a dataset may contain mostly safe roadside segments and only a few instances of very dangerous roadside objects. From a purely accuracy-oriented perspective, a classifier might achieve high scores by simply ignoring the underrepresented dangerous class. However, in practice, misclassifying a dangerous roadside as safe one could have far more severe consequences than the reverse scenario.

Various techniques have been developed to address class imbalance issues, both during training and evaluation. Data-level approaches attempt to rebalance the class distribution by oversampling examples of rare classes or undersampling examples of frequent classes [100, 101, 102]. While effective in single-task setups, these methods are less feasible in multi-task scenarios [103, 104] where tasks are uncorrelated and exhibit non-uniform imbalance. In such cases, an image that is rare in one task might be frequent in another, complicating the

direct application of oversampling.

Cost-sensitive learning approaches adapt learning algorithms by assigning larger loss weights to misclassified examples of underrepresented classes [105, 106, 107, 108]. While these cost-sensitive approaches are widely used, standard inverse-frequency loss weighting schemes have been shown to improve recall at the cost of reduced precision [21]. To address this issue, Tian et al. [21] propose dynamic assignment of class weights. During training, class weights are dynamically set according to the current false negative rate of the corresponding class. This approach prevents classes that achieve high recall from suffering excessive false positives. Unfortunately, this method is not directly applicable in multi-task setups, as we discuss in Section 4.2.

Regarding the evaluation of classifiers on imbalanced datasets, naive metrics like accuracy can hide poor classification performance on rare classes, making even trivial classifiers appear deceptively good. One solution is to use performance metrics that give equal importance to each class, regardless of its frequency. An example is the macro-F1 score, calculated as the mean of F1 scores for each class. For a given class in a multi-class problem, its F1 score is calculated by treating the problem as binary classification where the given class is the positive class and all other classes are grouped into the negative class. The F1 score is the harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3.11}$$

If either precision or recall is very low, the F1 score will also be very low. If a model were to learn to ignore a particular class, its recall and - consquently - its F1 score will be very low, significantly decreasing the macro-F1 score of the classifier.

3.8 Image Recognition in Traffic Scenes

Computer vision techniques have been extensively applied to detect and recognize various elements of road infrastructure pertinent to road safety assessment.

Prior research has focused on tasks such as localization of traffic control devices [109, 110], recognition of fleet management attributes [25], detection of traffic signs [109, 111], and identification of road surface markings [110, 112, 113]. Semantic segmentation methods have also been employed to classify pixels into different categories within road scenes, providing detailed scene understanding [80, 114]. Yi et al. [115] leverage active learning techniques to detect road-safety elements such as guardrails and utility poles, improving detection performance with fewer labeled samples.

While these works are related to our task, they typically target only specific subsets of road-safety attributes, focusing on particular elements rather than providing a comprehensive assessment. Some approaches attempt to classify iRAP attributes by utilizing an intermediate

semantic segmentation step, from which they extract the attributes using rule-based systems [116, 117, 118]. For instance, Sanjeewani et al. [116] employ outputs from semantic segmentation models to identify road-safety attributes based on the presence of certain semantic categories.

However, creating a training dataset for semantic segmentation requires dense pixel-level annotations, which entails a significant annotation effort and is considerably more time-consuming than annotating images with image-wide attributes provided by human coders. Moreover, a semantic segmentation model receives learning signals only from segmentation labels, not directly from the attribute class labels crucial for road-safety assessment. Consequently, such models learn features optimized for semantic segmentation, which may not be optimal for the specific task of attribute recognition. Additionally, the use of separate stages for segmentation and attribute extraction can lead to error propagation, where inaccuracies in the segmentation stage adversely affect the attribute recognition performance.

In contrast, deep learning models can be trained to recognize road-safety attributes directly from input data in an end-to-end manner [20]. By predicting directly from images, the approach avoids complexities and potential errors associated with intermediate representations. This allows the model to learn latent representations specifically tailored for attribute recognition, leading to improved performance.

Some approaches have attempted to predict even higher-level road-safety metrics directly from video data without relying on intermediate representations. For example, Song et al. [119] aim to predict the Star Rating Score - a composite metric used in road safety assessment - directly from video sequences, bypassing the need even for explicit attribute recognition.

The authors of the Honda Scenes dataset [24] present a baseline approach for infrastructure-related event detection in road-driving video. They pre-train a ResNet-50 backbone on the Places365 dataset to capture scene-level features and utilize a frozen semantic segmentation model to mask out dynamic objects like traffic participants, focusing on static infrastructure elements. Their pipeline includes recurrent processing of frozen convolutional features followed by standard softmax classification.

Context MTL [103] addresses recognition on Honda Scenes dataset using a multi-task learning architecture. They regularize the loss function with a lower bound of mutual information between the input and latent-space features, computed using the Jensen-Shannon divergence [120], to encourage the model to learn informative representations.

3.9 Recurrent models for video recognition

Long Short-Term Memory (LSTM) networks [23] have been widely employed in video classification and action recognition tasks [121], owing to their ability to model temporal dependencies

in sequential data. In video recognition applications, a common approach involves combining convolutional neural networks (CNNs) with LSTM networks. CNNs are utilized to extract spatial features from individual video frames, capturing the visual content within each frame. Frame-level features are then fed into an LSTM network, which models the temporal dynamics across the sequence of frames [122].

Beyond video recognition, LSTM networks have been utilized to enhance sequential predictions in various domains. Tu et al. [123] applied LSTMs to improve speech recognition systems by modeling temporal dependencies in audio sequences, leading to more accurate transcription of spoken words. Similarly, Kratzert et al. [124] employed LSTMs for rainfall-runoff modeling, capturing the temporal dynamics of hydrological processes and improving the prediction of water flow in river basins.

In the context of traffic scene analysis, Narayanan et al. [24] proposed an LSTM-based architecture inspired by temporal region proposal methods [125, 126]. Their approach involves a two-stage process:

1. **Event Proposal Stage**: Task-agnostic event proposals are generated as video intervals, identifying segments that may contain relevant events. This stage uses temporal region proposal techniques to segment the video into intervals of interest without assigning specific labels.
2. **Classification Stage**: The proposed intervals are classified into specific traffic scene events using features pooled both spatially and temporally. This is achieved by feeding the spatio-temporally pooled features into an LSTM network, which outputs the event classifications.

Since the event proposal and classification stages are decoupled, the second stage treats the problem as a single-task multi-class classification, independently assigning each proposed interval to an event category.

Trabelsi et al. [127] extended the traditional LSTM network by incorporating multi-head attention mechanisms [128], enhancing the model's ability to focus on different aspects of the input sequence. By combining this enhanced LSTM with a CNN, they capture and interpret the complex dynamics of driver behavior in traffic scenes. The attention mechanism allows the model to weigh the importance of different time steps and features, improving its capacity to model long-term dependencies and interactions within the data.

However, recurrent models present challenges of their own, such as vanishing or exploding gradients in very long sequences. While architectures like LSTMs and GRUs aim to mitigate these issues, they do not fully eliminate them. Moreover, the sequential nature of recurrent models limits parallelization, potentially leading to longer training and inference times compared to fully convolutional approaches.

Additionally, the sequential nature of recurrent models can limit parallelization, potentially

leading to longer training and inference times compared to fully convolutional approaches.

In the context of road safety attribute recognition, the use of recurrent models offers the potential to capture the temporal evolution of attributes along a road segment. This temporal information can be crucial for attributes that may change gradually or exhibit patterns over time.

Chapter 4

Automatic road-safety assessment in road-driving video

In this chapter, we present our proposed two-stage deep learning framework for automatic recognition of road-safety attributes from driving video data. Building upon the challenges and insights discussed in Chapter 2.2, our approach aims to efficiently and accurately assess road infrastructure safety by automating the detection of the attributes defined by the iRAP standard.

Our framework comprises three key components. We employ a convolutional neural network (CNN) for *local recognition* (4.1), performing multi-task recognition using the local spatio-temporal context. This stage focuses on extracting rich spatial features from individual frames or short sequences, capturing immediate visual cues relevant to each attribute. Notably, the shared part of our local-recognition model, including the CNN backbone (ResNet-18) and the spatial pyramid pooling module, is pre-trained for semantic segmentation of street scenes on the Vistas dataset [19]. Specifically, we utilize the encoder weights from a SwiftNet semantic segmentation architecture [129]. This pre-training enables our model to better recognize infrastructure elements and details relevant to road-safety attributes.

To further improve accuracy, we preform *sequential enhancement* of the initial predictions by integrating temporal information over longer sequences of frames (4.3). This stage aims to provide context beyond the local scope, leveraging recurrent models to better understand the temporal dependencies inherent in road infrastructure.

Throughout both stages, we address the issue of significant class imbalance through multi-task dynamic loss weighting (4.2). This approach adaptively adjusts the importance of classes within each task, while also maintaining stable values of per-attribute losses. At the class level, all weights are re-calculated each epoch by modulating inverse relative class frequencies with current per-class false negative rates. At the task level, each individual loss is normalized by the sum of the weights of all examples in the current iteration, rather than merely by the number

of the examples. This prevents individual tasks from intermittently dominating the overall loss and hampering the learning of other tasks.

By integrating these components, our framework leverages both spatial and temporal information while effectively addressing the challenges of multi-task learning and class imbalance. The local recognition stage provides initial attribute predictions based on spatial features, while the sequential enhancement stage refines these predictions by considering temporal patterns. Our dynamic loss weighting formulation facilitates balanced learning across all classes in a multi-task setup.

The rest of this chapter is organized as follows. In Section 4.1, we detail the architecture and training methodology of the local recognition stage,. Section 4.2 describes our dynamic loss weighting scheme, explaining how it addresses class imbalance within each task and maintains stable loss magnitudes across tasks. In Section 4.3, we discuss the sequential enhancement stage, outlining how temporal modeling is implemented to enhance attribute recognition.

4.1 Recognition in the local spatio-temporal context

Having outlined the overall framework, we now detail the local recognition stage. Figure 4.1 illustrates our convolutional architecture designed for multi-task visual recognition of road-safety attributes in road-driving imagery. The architecture comprises a shared encoder and multiple task-specific classification heads, enabling efficient feature extraction and prediction across different attributes.

Shared Encoder

The shared encoder begins with a ResNet-18 backbone to extract features from road-driving images. The choice of ResNet-18 balances computational efficiency and representational capacity, making it suitable for applications requiring real-time or near-real-time processing.

The features extracted by the ResNet-18 backbone are then processed by a Spatial Pyramid Pooling (SPP) module with grid dimensions of 6×6, 3×3, 2×2, and 1×1 [62]. The SPP module aggregates multi-scale contextual information, producing a fixed-size representation that captures both local and global spatial cues. This multi-scale representation can be beneficial for recognizing road-safety attributes that may appear at different sizes and positions within the image. It also serves as a common feature space for all attribute-specific classification heads.

Attribute-Specific Classification Branches

The architecture incorporates 43 attribute-specific classification branches, each designated to recognize a particular road-safety attribute. Each branch begins with an attention pooling mod-

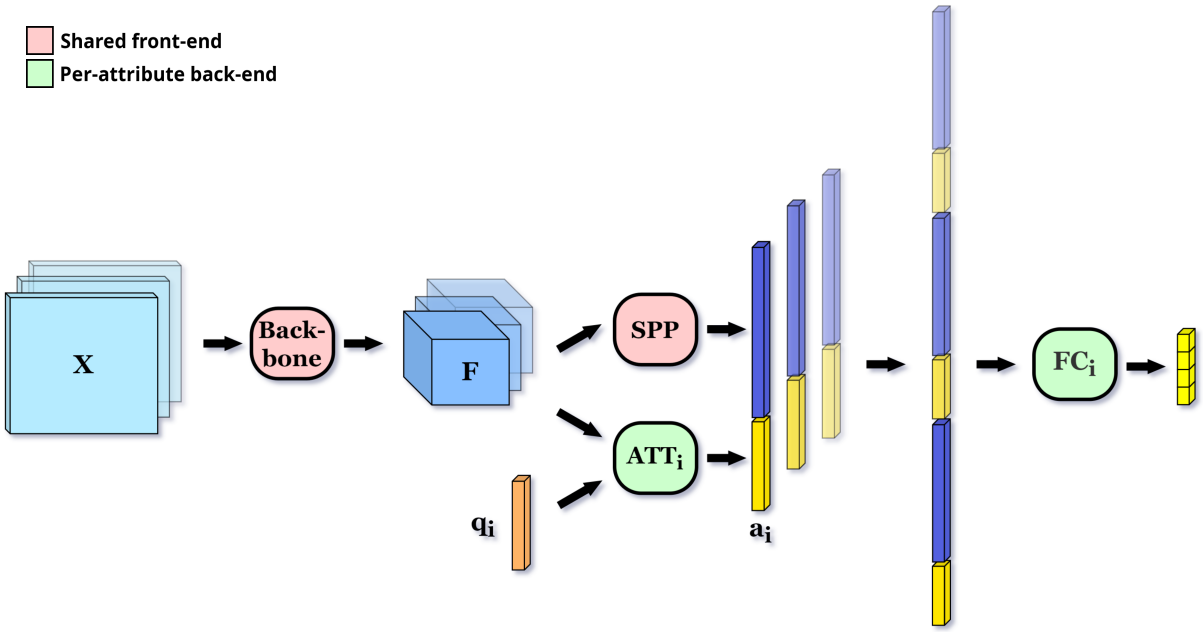

Figure 4.1: Our multi-frame local recognition pipeline recognizes road-safety attributes in multi-frame input **X**. Data tensors are represented as cuboids, while processing modules are shown as rounded rectangles. The shared front-end (red) maps each input frame into convolutional features (**F**) that are subsequently pooled by the **SPP** module. Attribute-specific back-ends (green) produce attention pools (yellow) and concatenate them with the shared spatial pools (blue). Fully-connected layers (FC$_i$) map the concatenated descriptors into attribute-specific logits.

ule ATT_i [20], which operates on the shared convolutional features. The module is guided by a learned attribute-specific query vector q_i, allowing the model to focus on regions of the feature map that are most relevant to the attribute A_i. The output of the attention pooling module is an attribute-specific representation, which is then concatenated with the shared SPP features to form a comprehensive single-frame descriptor for attribute A_i. This descriptor is subsequently processed by the corresponding fully connected layer (prediction head) that predicts the posterior probability $P(A_i|\mathbf{x})$ for attribute A_i.

Extension to Multi-Frame Input

Our architecture can be naturally extended to operate on multi-frame input to capture a slightly larger spatio-temporal context beneficial for recognizing partially visible and occluded attributes. In the multi-frame setup, the convolutional backbones and the two pooling modules process each frame independently. The resulting single-frame descriptors are then concatenated to form a multi-frame attribute descriptor. The per-attribute predictions heads process the concatenated descriptors to produce the final predictions.

It is worth noting that while multi-frame input can provide richer contextual information, there are practical limitations to the number of input frames that can be processed simultaneously. To balance performance and computational efficiency, our multi-frame models are de-

signed to produce predictions for segment T by observing frames of segments T, T-1, and T-4. This configuration allows for a temporal window that captures both immediate and moderately distant context while avoiding excessive memory requirements.

Training Objective

Each per-attribute prediction head is associated with a cross-entropy loss function, measuring the discrepancy between the predicted probabilities and the ground-truth labels. The per-attribute loss for attribute A_i is given by:

$$\mathscr{L}_i = -\sum_{c=1}^{C_i} y_{i,c} \log P(A_i = c \,|\, \mathbf{x}),$$

where C_i is the number of classes for attribute A_i, and $\mathbf{y}_i$ is a one-hot representation of the ground truth class of attribute A_i for the input example $\mathbf{x}$. Following the multi-task learning paradigm [18], the total loss $\mathscr{L}$ is computed as the mean of all per-attribute losses:

$$\mathscr{L} = \frac{1}{K} \sum_{i=1}^{K} \mathscr{L}_i, \tag{4.1}$$

where K is the total number of attributes (tasks). This approach allows for a balanced optimization process that considers the importance of each road-safety attribute equally.

Semantic Segmentation Pre-Training

To enhance the representational power of our shared encoder, we pre-train the ResNet-18 backbone and the spatial pyramid pooling (SPP) module on the task of semantic segmentation using the Vistas dataset [19]. Semantic segmentation requires assigning a semantic label to every pixel in an image, necessitating the extraction of detailed spatial features that capture the identity and boundaries of various elements within street scenes.

In this pre-training step, we employ the SwiftNet architecture [129], leveraging its encoder weights to initialize the ResNet-18 backbone and the SPP module. SwiftNet is designed to achieve a balance between accuracy and computational efficiency, making it suitable for applications that require real-time or near-real-time performance.

The Vistas dataset provides a large and diverse set of street-level images captured in various conditions, including different weather, lighting, and geographic locations. Pre-training on such a rich dataset exposes our encoder to a wide variety of visual patterns and scenarios.

The decision to use semantic segmentation as a pre-training task stems from its similarity to the ultimate objective of recognizing road infrastructure components. While classification tasks often emphasize global image context, segmentation involves understanding the relationships

between smaller scene elements. This capability directly transfers to our attribute classification tasks, where understanding the spatial arrangement of infrastructure elements is crucial (e.g. attribute *Road severity*). By initializing our shared encoder with weights learned from semantic segmentation, we provide it with strong prior knowledge of street scene composition and structure, enabling more robust detection of road-safety attributes.

4.2 Dynamic loss weighting for multi-task learning

In multi-task learning, the challenge of class imbalance is particularly pronounced, as each task may contain classes with vastly different frequencies (cf. 3.7). To address this issue, we aim to increase the influence of rare classes on the training objective, thereby improving the model's performance on underrepresented classes. At the same time, we do not want this to adversely affect task-level learning dynamics.

We denote our training set as $\{(x_n, y_n)\}_{n=1}^{N}$, where $x_n \in \mathbb{R}^d$ represents the n-th input sample, $y_n \in \{1, \ldots, C\}$ is the corresponding class label, with C being the total number of classes. The total number of training samples is N. Let $P_n^c = P(Y = c \mid x_n)$ denote the predicted posterior probability of class c given input x_n.

The standard cross-entropy loss function is commonly used for classification tasks. It can be expressed as the negative logarithm of the geometric mean of the correct class posterior probabilities across all training samples $\overline{P} = \left(\prod_{n=1}^{N} P_n^{y_n}\right)^{1/N}$ over all training examples [21]:

$$\begin{aligned} \mathrm{CE} &= -\frac{1}{N}\sum_{n=1}^{N} \ln P_n^{y_n} && (4.2) \\ &= -\frac{1}{N} \ln\left(\prod_{n=1}^{N} P_n^{y_n}\right) \\ &= -\ln\left(\prod_{n=1}^{N} P_n^{y_n}\right)^{\frac{1}{N}} \\ &= -\ln \overline{P} && (4.3) \end{aligned}$$

This formulation highlights that minimizing the cross-entropy loss is equivalent to maximizing the geometric mean of the predicted probabilities for the true classes.

To further analyze the impact of class imbalance, we can express the cross-entropy loss as a weighted sum over per-class geometric mean posteriors. Let N_c denote the number of samples

belonging to class c, such that $\sum_{c=1}^{C} N_c = N$. Then, we can rewrite the cross-entropy loss as:

$$\begin{aligned} \mathrm{CE} &= -\frac{1}{N}\sum_{c=1}^{C}\sum_{n:y_n=c}\ln P_n^c && (4.4)\\ &= -\sum_{c=1}^{C}\frac{N_c}{N}\ln\left(\prod_{n:y_n=c}P_n^c\right)^{\frac{1}{N_c}} \\ &= -\sum_{c=1}^{C}\frac{N_c}{N}\ln\overline{P^c} && (4.5) \end{aligned}$$

where $\overline{P^c} = \left(\prod_{n:y_n=c} P_n^c\right)^{\frac{1}{N_c}}$ is the geometric mean posterior probability for class c. The notation $\{n : y_n = c\}$ denotes the set of indices of samples belonging to class c.

Equation (4.5) shows that the standard cross-entropy loss gives more weight to classes with higher frequencies, as the term $\frac{N_c}{N}$ represents the relative frequency of class c in the dataset. Consequently, rare classes contribute less to the loss, which can lead to poor performance on these classes.

To mitigate this issue, we can assign higher weights to rare classes by using inverse frequency weighting. To each class we assign a weight that is inversely proportional to its relative frequency: $w_c = \frac{N}{N_c}$. This leads to the inverse-frequency-weighted cross-entropy loss [106, 107]:

$$\begin{aligned} \mathrm{CE}^{\mathrm{IFW}} &= -\frac{1}{N}\sum_{n=1}^{N} w_{y_n}\ln P_n^{y_n} && (4.6)\\ &= -\sum_{c=1}^{C}\frac{w_c N_c}{N}\ln\overline{P^c} \\ &= -\sum_{c=1}^{C}\ln\overline{P^c} && (4.7) \end{aligned}$$

Here, the weighting effectively cancels out the relative frequencies, ensuring that each class contributes equally to the loss function. While this approach balances the influence of each class, it may introduce new issues.

Specifically, the standard cross-entropy loss focuses on maximizing the posterior probability of the correct class, but does not take into account the distribution of the posterior probabilities over the incorrect classes. In essence, cross-entropy focuses solely on penalizing false negatives while disregarding false positives. Consequently, assigning a large weight to a particular class might inadvertently increase the incidence of false positives for that class. Empirical analysis has shown that increasing the class weight indeed often decreases the precision for that class, as the model becomes more likely to over-predict it [21]. At the same time, as classes achieve higher recall scores, class weighting starts to show diminishing returns in terms of additional increase in recall. These observations suggest that overly emphasizing rare classes without

considering model performance can be counterproductive.

To address this issue, we can adapt class weights dynamically based on model performance, specifically in terms of per-class recall [21]. Let $R_{c,t}$ represent the validation recall of class c after epoch $t-1$. Then, recall-balanced class weights $w^R_{c,t}$ can be defined as follows [21]:

$$w^R_{c,t} = \frac{N}{N_c}(1 - R_{c,t}) + \varepsilon \tag{4.8}$$

where $\varepsilon = 10^{-4}$ is a small constant added to prevent the weight from becoming zero in the unlikely event of perfect recall. This formulation ensures that classes with low recall receive higher weights, thereby focusing the learning process on difficult classes that the model struggles with.

When the recall $R_{c,t}$ for a class is close to zero, the weight $w^R_{c,t}$ approaches the inverse relative frequency, similar to the inverse-frequency weighting scheme. As the recall improves, the weight diminishes, reducing the emphasis on classes that the model already predicts well.

However, class weighting schemes based on inverse frequencies, including both static and dynamic approaches, can lead to high variance in loss magnitudes across training iterations. Batches containing more examples from extremely rare classes will tend to have a much larger loss compared to batches with fewer such examples, as these rare examples are assigned correspondingly large weights. The extreme scarcity of these examples results in their sporadic presence across batches, causing significant loss fluctuations between training iterations that can destabilize the learning process.

In multi-task learning, where the total loss is often computed as the arithmetic mean over individual task losses, this issue is exacerbated. In a given training iteration, if an imbalanced task experiences a large loss magnitude due to the presence of extremely rare examples, it can dominate the total loss and impede the learning of other tasks. When there are multiple extremely imbalanced tasks, this scenario can occur frequently, leading to a situation where different tasks intermittently suppress each other's progress.

To address this issue, we seek to stabilize the loss magnitude across tasks by normalizing each task's loss with respect to the sum of the weights of individual examples. This approach stabilizes the loss magnitude across different batches and tasks. Utilizing the recall-balanced weights from Equation (4.8), we express the loss for each individual task as:

$$\mathrm{CE}^{\mathrm{R}}_{\mathrm{MT}} = \frac{-\sum_{n=1}^{N} w^R_{y_n,t} \ln P_n^{y_n}}{\sum_{n=1}^{N} w^R_{y_n,t}} \tag{4.9}$$

This formulation ensures that the contribution of each task to the overall loss remains stable, preventing situations where one task might impede the progress of others due to class imbalance. By dynamically adjusting the class weights based on per-class recall and normalizing the task

losses, our approach balances precision and recall while maintaining stable multi-task training dynamics.

4.3 Sequential enhancement

In the second stage of our recognition approach, we aim to enhance the initial local predictions by aggregating information across a larger temporal context. The motivation behind this strategy is to leverage the inherent temporal patterns and dependencies present in road-safety attributes along a sequence of road segments.

To achieve this, we construct temporal inputs as sequences of $T = 21$ vectors, encompassing consecutive segments from time $t - 10$ to $t + 10$. This forms a temporal window centered at the current segment t, providing context from both preceding and succeeding segments. By incorporating data from a broader context, we can capture the temporal dynamics and improve the robustness of our predictions against transient noise or occlusions that may affect individual frames.

Rather than relying on hand-crafted post-processing rules, we propose to utilize deep recurrent neural networks for sequence classification [22]. Specifically, we employ per-attribute recurrent models that can learn attribute-specific temporal behavior patterns described in Section 2.2.4. This approach allows the model to learn complex temporal patterns directly from the data, potentially capturing nuances that might be difficult to encode in manual rules.

Our recurrent models are based on bidirectional Long Short-Term Memory (Bi-LSTM) networks, consisting of four layers of Bi-LSTM cells. Each Bi-LSTM layer processes the input sequence in both forward and backward directions through separate unidirectional LSTM modules. This bidirectional approach enables the model to capture both past and future temporal dependencies for each attribute, which is particularly useful in identifying patterns that evolve consistently over time.

At each time step i, the input to our recurrent model is constructed by concatenating the local logits $\mathbf{s}_i^a$ and the embedding $\mathbf{e}_{c_i^a}$ of the predicted class $c_i^a = \arg\max \mathbf{s}_i^a$. The local logits provide the model with the raw predictions from the first stage, while class embeddings provide additional context regarding the predicted class through its learned representation.

The matrices of per-attribute class embeddings are jointly learned during the training of these recurrent models, enabling them to adapt to the specific characteristics of each attribute. We set the dimensionality of the class embeddings to $\max(4, C)$, where C denotes the number of classes for a particular attribute. This ensures sufficient representation capacity while keeping the model lightweight.

The hidden states in all Bi-LSTM layers have a dimension of 128. Each Bi-LSTM layer outputs hidden states for each time step, but for our final representation, we focus on the last

hidden states produced by both the forward and backward LSTM modules in each layer. By concatenating these last hidden states across all four layers, we capture hierarchical temporal features from different levels of abstraction. Additionally, we include the hidden state corresponding to the middle element (time t) from the last layer, which emphasizes the representation at the current segment.

The concatenation of these hidden states results in a feature vector of size 1280, which is then passed through a fully-connected softmax classifier to produce the final classification output. This output represents the predicted posterior distribution over the classes for attribute a at time t, denoted as $P(A_i = c^i_j \mid x_{t-T:t+T})$.

The use of Bi-LSTMs in this context allows our model to effectively capture the temporal evolution of road-safety attributes, which is crucial for understanding attributes that may change gradually or exhibit specific patterns over time.

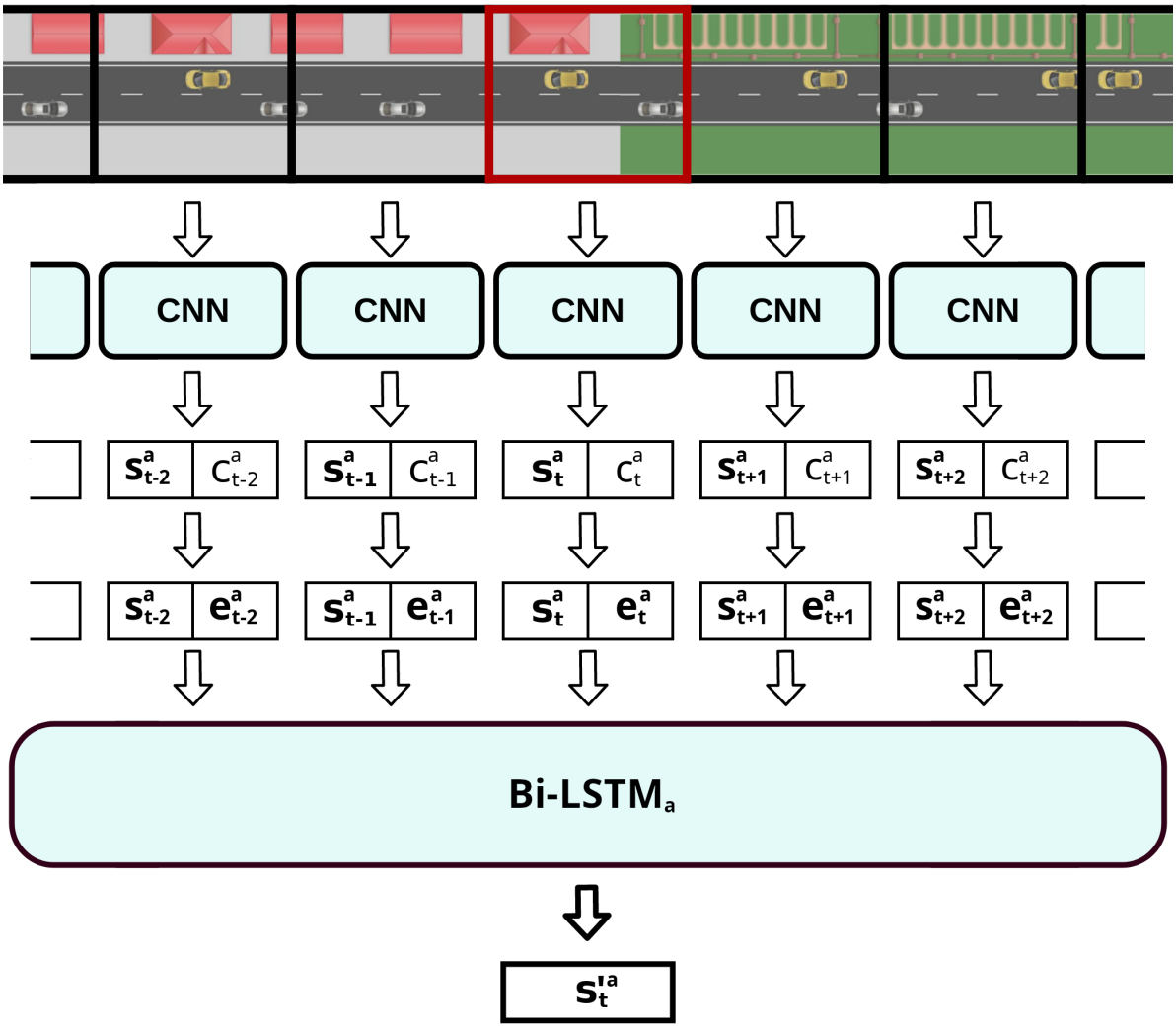

Figure 4.2: Sequential enhancement corrects local predictions with per-attribute Bi-LSTM models. For each attribute a, the model Bi-LSTM$_a$ outputs corrected logits $\mathbf{s}'^{\mathbf{a}}_{\mathbf{t}}$ in segment t by observing $T = 21$ vectors that correspond to segments from *(t-10)* to *(t+10)*. Each of these vectors is a concatenation of the logits $\mathbf{s}^{\mathbf{a}}_{\mathbf{i}}$ and the jointly learned embedding $\mathbf{e}^{\mathbf{a}}_{\mathbf{i}}$ of the the most probable class according to the local model.

Chapter 5

Experiments

This chapter presents a comprehensive experimental evaluation of our approach for automatic road-safety attribute recognition. We begin by introducing the datasets used in our experiments, including the novel iRAP-BH dataset and three established datasets from the literature that enable a broader validation of our method. We then detail our experimental setup, including evaluation metrics and training configurations, before presenting extensive quantitative and qualitative results. Our experiments demonstrate the effectiveness of our approach in capturing spatial and temporal dependencies for road safety attribute recognition. Through ablation studies and comparisons to state-of-the-art methods, we validate the impact of our key contributions: semantic segmentation pre-training, multi-task dynamic loss weighting, and sequential enhancement.

5.1 Datasets

In this section, we present the datasets used to evaluate our approach for automatic road-safety assessment. Our experiments encompass four distinct datasets, each offering unique characteristics and challenges relevant to the task of road-safety attribute recognition.

The cornerstone of our evaluation is the novel iRAP-BH dataset, which we introduce as a comprehensive corpus of georeferenced video data specifically designed for off-line road-safety assessment.

To further validate the efficacy and generalizability of our approach, we extend our experiments to three additional datasets from the literature: Honda Scenes [24], FM3m [25], and BDD100k [26]. These datasets, while not specifically curated for road-safety attribute recognition, offer valuable benchmarks for related tasks in traffic scene understanding and classification.

In the following subsections, we provide detailed descriptions of each dataset, including their composition, annotation schemes, and specific considerations for our experimental setup.

5.1.1 iRAP-BH

Road safety assessment requires high-quality, diverse data that captures the complexity of real-world infrastructure. To meet this need, we present iRAP-BH, a novel dataset of georeferenced video recordings specifically designed for off-line road-safety assessment. It was acquired along 194 public city, inter-city and rural road sections in Bosnia and Herzegovina, spanning a total of 2,300 km. The dataset captures a diverse road infrastructure and varying road conditions, providing a rich dataset for training and evaluating road-safety assessment algorithms.

All videos were recorded in RGB format at a resolution of 2704×2028 pixels and a frame rate of 25 frames per second. The average road section within this dataset consists of 1,175 segments, each segment corresponding to a 10-meter stretch of road, with an average segment spanning approximately 18 frames. This granular segmentation allows for detailed analysis and annotation of road attributes, capturing subtle changes that may occur over short distances. To illustrate the geographical coverage of our dataset, Figure 5.1 presents a map showing the distribution of recorded road sections across Bosnia and Herzegovina.

The iRAP-BH dataset was meticulously annotated with all iRAP attributes by trained human annotators at the Faculty of Transportation and Traffic Sciences, University of Zagreb (UniZG-FTTS). The annotators underwent specialized training to ensure consistency and accuracy in the annotations. The process was facilitated by a manual annotation interface, designed at UniZG-FTTS to streamline the workflow and reduce the potential for human error.

While the iRAP Star Rating Score typically requires 100-meter granularity, iRAP-BH was deliberately annotated at a finer ten-meter level. This decision was made to provide more fine-grained supervision for machine learning algorithms, as the increased resolution in annotations allows models to learn from more detailed and localized features.

For the purposes of model training and evaluation, we divided the dataset into three distinct splits: 214,073 segments for training, 5,813 for validation, and 6,563 for testing. We ensured that all segments belonging to the same road section were allocated to the same split, in order to avoid any risk of data leakage between training and evaluation phases. This strategy also enables the training of sequential and multi-frame models without compromising the integrity of the evaluation process.

Each 10-meter road segment in the dataset is represented by its middle frame, resized to 384×288 pixels to reduce computational complexity while retaining sufficient visual detail. This preprocessing results in a comprehensive multi-task, multi-class video recognition dataset comprising 226,449 images. The iRAP-BH dataset thus provides a rich resource for the development and evaluation of computer vision and machine learning techniques for road-safety attribute recognition.

Figure 5.1: The geographical coverage of the IRAP-BH dataset. The road network captured by the dataset is shown in blue.

5.1.2 Honda Scenes

The Honda Scenes dataset [24] offers a comprehensive collection of annotated road-driving videos intended to advance research in traffic scene understanding and classification. The dataset consists of 80 training videos and 20 evaluation videos. Each frame within these videos has image-wide labels for four distinct traffic scene classification problems: *Road place*, *Road environment*, *Road surface*, and *Weather*. Figure 5.2 illustrates the four problems and their temporal annotations.

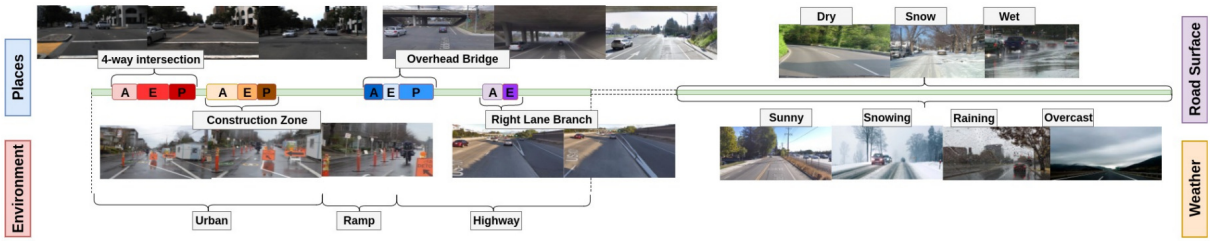

Figure 5.2: Video annotations for four separate problems of the Honda Scenes dataset: *Road Place*, *Road Environment*, *Road Surface* and *Weather*. The figure is reproduced from the original paper that describes the dataset [24].

To generate frame-level data for the *Road place* and *Road environment* classification problems, frames were extracted by subsampling the videos at a rate of 3Hz. This resulted in a substantial dataset consisting of approximately 760,000 training frames and 160,000 evaluation frames, which serve as inputs for our recognition models. In our experimental setup, we treat these consecutive frames as consecutive road segments, analogous to the 10-meter segments in the iRAP-BH dataset. This approach allows us to apply our multi-frame and sequential enhancement models, capturing temporal dependencies and improving classification performance.

The *Road Surface* and *Weather* classification problems involve images sampled from both the Honda Scenes dataset and the BDD100k dataset [26]. Specifically, the *Road Surface* dataset comprises a total of 10,139 images, with 2,676 images from Honda Scenes and 7,463 images from BDD100k. The dataset is split into 9,150 images for training and 898 images for evaluation. The *Weather* dataset, meanwhile, consists exclusively of data from BDD100k, containing 11,781 training images and 1,255 evaluation images.

In the rest of this section, we provide a detailed description of each classification problem, including their relevance to our research and the specific considerations in our experimental approach.

Road place

The *Road place* classification problem is unique among the four as it involves multiple, concurrent multi-class classification tasks. It encompasses a series of fine-grained temporal labels that capture the dynamic relationship between the vehicle and various road elements. These labels

include Approaching (A), Entering (E), and Passing (P), indicating the relative position of the vehicle to a specific place of interest in each frame.

The tasks within this problem cover a wide array of road infrastructure elements:

- *Construction Zone*
- *Intersection (3-way)*
- *Intersection (4-way)*
- *Intersection (5-way & more)*
- *Overhead Bridge*
- *Rail Crossing*
- *Merge - Gore on Left*
- *Merge - Gore on Right*
- *Branch - Gore on Left*
- *Branch - Gore on Right*
- *Background*

These tasks are relevant to our work with the iRAP-BH dataset, as they involve recognizing road infrastructure elements in a sequential manner. The fine-grained temporal labeling allows our models to capture the nuanced temporal dynamics of driving scenes. It is important to note that some of these tasks, such as rail crossings and complex intersections, suffer from significant class imbalance, reflecting the natural distribution of these features in real-world driving scenarios.

Road Environment

The *Road environment* problem classifies scenes into four distinct categories: *Local*, *Highway*, *Ramp*, *Urban*. Unlike *Road Place*, this problem does not utilize temporal labels. However, since it consists entirely of frames from continuous driving videos from the Honda Scenes dataset, we can still apply our multi-frame and sequential models. By incorporating sequential frames, we aim to capture the temporal continuity inherent in the road environment, enhancing the robustness of our classification results.

Road surface

The *Road Surface* problem involves multi-class classification, where each image is categorized into one of three surface conditions: *Wet*, *Dry*, *Snow*. Since this problem involves only non-sequential, individual images, our evaluation is limited to the single-frame version of our model. Additionally, the classes within this subset of data are fairly balanced, which diminishes the need for weighted loss functions, as they do not significantly impact model performance in this particular case.

Weather

The *Weather* classification problem involves categorizing images into one of four weather conditions: *Clear*, *Overcast*, *Rainy*, *Snowy*. Similar to the *Road surface* problem, this dataset also consists of relatively balanced non-sequential frames, thus we only evaluate our single-frame model, without loss balancing or sequential enhancement.

5.1.3 Fleet Management Dataset (FM3)

The third iteration of the Fleet Management Dataset, FM3 [25], offers a diverse collection of 11,448 images depicting various traffic scenes from roads in Croatia. For our experiments, we utilize the main subset of this dataset, referred to as FM3m, which consists of 6,413 images labeled for classification tasks. Each image is annotated with eight binary labels (true/false) corresponding to eight distinct classification attributes. These attributes encompass various aspects of road infrastructure and traffic conditions: *highway*, *road*, *tunnel*, *exit*, *settlement*, *overpass*, *booth*, and *traffic*.

The dataset is partitioned into training, validation, and test sets containing 1,607, 1,600, and 3,206 images, respectively. These subsets we constructed through uniform random frame assignment. As a result, consecutive frames from the same driving sequence are generally distributed across different subsets. This random allocation disrupts the temporal continuity of data within each subset and precludes the use of multi-frame or temporally linked segments as input for training and evaluation. Therefore, in our experiments with FM3m, we restrict our evaluation exclusively to the single-frame version of our approach. This experimental setup serves as an important baseline to assess the effectiveness of our model when operating without the sequential enhancement strategies that play a crucial role in other experiments.

5.1.4 Berkeley Deep Drive (BDD100k)

The Berkeley Deep Drive (BDD100k) dataset [26] is a large-scale collection curated for heterogeneous multi-task visual recognition in road driving scenes. It consists of 100,000 video clips, each 40 seconds in length, capturing a wide variety of driving environments. These environments include both urban and rural areas, various weather conditions, and different times of day, providing a diverse representation of real-world driving situations.

From each video, a single frame has been chosen and annotated with detailed visual features. These annotations encompass object bounding boxes, drivable areas, lane markings, and full-frame panoptic segmentation labels. In addition to the detailed per-pixel annotations, BDD100k incorporates three image-wide classification tasks: *Scene*, *Weather*, and *Time of Day*. The dataset is partitioned into 70,000 training images, 10,000 validation images, and 20,000 test images, respectively.

The *Scene* task includes seven distinct classes: *Tunnel*, *Residential*, *Parking Lot*, *City Street*, *Countryside*, *Gas Station*, and *Highway*. Our experiments concentrate on this task, as it more closely aligns with road-safety attributes pertinent to our study and provides a relevant benchmark for comparison with previous work in the field. The other two tasks, *Weather* and *Time of Day*, have not been the focus of related previous work. Additionally, there is an overlap between the *Weather* task of BDD100k and the homonymous task within the Honda Scenes dataset, which makes the separate evaluation of these tasks redundant within the context of our objectives.

It is important to note that the structure of the BDD100k dataset does not facilitate the use of sequential processing techniques. Each annotated image is extracted from a distinct video sequence, resulting in the absence of consecutive frames within the dataset. Consequently, our experiments on BDD100k exclusively employ single-frame models. This approach mirrors our methodology with the FM3m dataset and the *Road Surface* and *Weather* tasks of Honda Scenes, where sequential data is also unavailable.

5.2 Experimental Results and Analysis

In this section, we present a comprehensive evaluation of our approach for the automatic recognition of road-safety attributes. Our experimental evaluation primarily focuses on the iRAP-BH dataset, complemented by additional experiments on Honda Scenes [24], FM3m [25], and BDD100k [26] datasets that contain related road scene classification tasks.

In the following subsections, we delve into the specifics of our experimental setup, including the evaluation metrics and training hyperparameters. We then present the results obtained on the iRAP-BH dataset, analyzing the impact of various design choices and providing qualitative examples. Subsequently, we report our findings on the Honda Scenes, FM3m, and BDD100k datasets, comparing our results with those of existing methods and highlighting the strengths of our approach.

5.2.1 Evaluation metrics

In this section, we outline the evaluation metrics used to assess the performance of our approach across different datasets. Given the multi-task multi-class nature of our problem and the presence of class imbalance, selecting appropriate metrics is crucial for a fair and informative evaluation.

For binary classification tasks, the performance of a model is commonly evaluated using precision, recall, and the F1 score.

Precision measures the proportion of correctly predicted positive instances among all in-

stances predicted as positive:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}. \tag{5.1}$$

Recall, also known as sensitivity, measures the proportion of correctly predicted positive instances among all actual positive instances:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{5.2}$$

The **F1 score** is the harmonic mean of precision and recall, providing a single metric that balances both:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{5.3}$$

A low value in either precision or recall results in a low F1 score, highlighting any deficiencies in the model's ability to correctly identify positive instances.

In multi-class classification problems, especially those with class imbalance, it's important to evaluate performance across all classes equitably. The macro-F1 score addresses this by computing the F1 score for each class independently, treating the problem as binary classification (one-vs-rest), and then taking the average:

$$\text{Macro-F1} = \frac{1}{C} \sum_{i=1}^{C} \text{F1}_i, \tag{5.4}$$

where C is the total number of classes and F1_i is the F1 score for class i.

Given its equal weighting of all classes, the macro-F1 score can be more informative than accuracy, which can obscure poor classification of minority classes in imbalanced datasets. For instance, in a dataset where a rare class comprises only 1% of the samples, a naive classifier that always predicts the majority class achieves 99% accuracy while completely failing to identify the minority class. Conversely, the macro-F1 score paints a different picture. The classifier's complete failure to identify any minority class instances yields zero recall, and consequently an F1 score of zero. This substantially lowers the macro-averaged F1, revealing the classifier's inability to detect these infrequent instances.

Given these considerations, we evaluate our approaches on the iRAP-BH and Honda Scenes datasets using the mean macro-averaged F1 score [130, 131]. Notably, the creators of Honda Scenes also employ macro-F1 in their experiments, facilitating direct comparison.

For datasets involving multiple binary classification tasks, such as FM3m, we use the mean Average Precision (mAP) as the evaluation metric. The Average Precision (AP) for a single binary task is calculated from the precision values at each recall threshold where the prediction

confidence results in a true positive. It can be expressed as:

$$\mathrm{AP} = \sum_{n} (\mathrm{R}_n - \mathrm{R}_{n-1})\mathrm{P}_n, \tag{5.5}$$

where P_n is the precision and R_n is the recall at the nth threshold, and the sum is taken over all unique recall values.

The mAP is then computed by averaging the AP across all tasks [132]:

$$\mathrm{mAP} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{AP}_t, \tag{5.6}$$

where T is the total number of tasks and AP_t is the Average Precision for task t.

This metric is suitable for FM3m since all tasks involve binary classification, and it effectively captures the model's ability to balance precision and recall across different thresholds [25].

In the case of the BDD100k dataset, we evaluate classification performance using accuracy, which measures the proportion of correct predictions among all predictions made. This metric is standard for this dataset and is used in related works [133, 134].

To ensure clarity and consistency, we present all performance metrics as percentage points (pp) throughout our analysis.

5.2.2 Training setup

As described in Chapter 4, our method comprises two stages: local recognition and sequential enhancement. In this section, we detail the training setup and hyperparameter configurations for both stages, with particular attention to data augmentation and optimization parameters.

To improve the robustness and generalization of our model without significantly altering key image features, we apply data augmentation to the input images. Specifically, we use color jittering by varying the brightness, contrast, saturation, and hue with relative ratios of 0.6, 0.3, 0.2, and 0.02, respectively. We deliberately exclude horizontal flipping since it could disrupt the detection of attributes that are specific to right-hand traffic scenarios (e.g. *Roadside severity - passenger side*). Similarly, we avoid random cropping and related augmentation methods, since they may remove critical visual information from the peripheral regions of the images. which is essential for identifying roadside attributes like *Street lighting*. This peripheral information is crucial for identifying various roadside attributes, such as *Street lighting*.

Both stages of our approach are trained using the Adam optimizer. For the local recognition stage, we set the learning rate to 1×10^{-5}, weight decay to 1×10^{-3}, and use a batch size of 12. The training is conducted for 15 epochs, with the learning rate decreasing according to a multiplicative scheduler using an annealing factor of 0.88 per epoch. For the sequential

enhancement stage, we set the learning rate to 5×10^{-4}, the weight decay to 1×10^{-4}, and train for 10 epochs with a batch size of 32.

To determine the optimal hyperparameter values, we systematically explored multiple configurations through two grid searches on the validation split of the iRAP-BH dataset. The first grid search optimized the parameters of the local recognition stage, followed by a second search focused on the sequential enhancement stage. The resulting optimal hyperparameters identified are consistently applied in all subsequent experiments across all datasets.

5.2.3 iRAP-BH

In this section, we present and analyze the experimental results obtained on the iRAP-BH dataset. We begin our analysis by examining the impact of our proposed contribution on the overall performance. Table 5.1 demonstrates the cumulative impact of each component on the overall Macro-F1 score on the iRAP-BH test set.

Table 5.1: Impact of our contributions on overall Macro-F1 performance on iRAP-BH test. IFW - inverse frequency weighting; R - dynamic recall weighting; SE - sequential enhancement.

Model	Pre-training	Macro-F1
ConvCE	ImageNet	53.79
ConvCE	Vistas	54.96
$\text{ConvCE}^{\text{IFW}}_{\text{MT}}$	Vistas	56.43
$\text{ConvCE}^{\text{R}}_{\text{MT}}$	Vistas	57.77
ConvCE-SE	Vistas	59.68
$\text{ConvCE}^{\text{R}}_{\text{MT}}$-SE	Vistas	**62.86**

Semantic segmentation pre-training on the Vistas dataset improves performance by 1.2 percentage points (pp) in mean Macro-F1 compared to pre-training on ImageNet-1k. This indicates that leveraging prior knowledge of road scene semantics is beneficial for recognizing road-safety attributes.

Next, we examine the impact of different loss weighting strategies. The multi-task formulation of inverse-frequency weighting ($\text{ConvCE}^{\text{IFW}}_{\text{MT}}$) delivers an improvement of 1.5 pp in Macro-F1. Multi-task dynamic recall weighting ($\text{ConvCE}^{\text{R}}_{\text{MT}}$) further enhances the performance by 1.3 pp over IFW.

Finally, we incorporate sequential enhancement ($\text{ConvCE}^{\text{R}}_{\text{MT}}$-SE) through recurrent models that capture a larger temporal context. It further increases performance by 5.1 pp. The combination of semantic segmentation pre-training, dynamic loss weighting and sequential enhancement yields the best performance, achieving a mean Macro-F1 score of 62.86%.

To gain deeper insights into the effectiveness of our approach on individual attributes, we present per-attribute Macro-F1 scores in Table 5.2.

Table 5.2: Per-attribute mean macro-F1 performance of two ablated models and our best model on iRAP-BH test.

Attribute	CNN CE	+LSTM CE	+LSTM CE^{R}_{MT}	Attribute	CNN CE	+LSTM CE	+LSTM CE^{R}_{MT}
Area type	90.49	90.93	94.56	Ped. obs. flow, driver	24.73	27.61	29.37
Bicycle facility	53.72	100.00	100.00	Ped. obs. flow, pass.	26.96	28.74	35.62
Bicycle observed flow	35.77	35.82	34.83	Property access points	56.64	57.28	59.71
Carriageway label	93.88	94.20	97.54	Quality of curve	65.46	65.99	71.24
Curvature	55.22	59.55	64.77	Road condition	72.76	73.32	77.70
Delineation	98.12	99.17	98.94	Roadside severity, driver dist.	56.94	57.05	57.86
Grade	53.60	53.63	51.85	Roadside severity, driver obj.	35.25	35.26	43.31
Intersection channelisation	61.55	63.41	62.74	Roadside severity, pass. dist.	54.90	56.87	59.97
Intersection quality	47.83	51.14	52.02	Roadside severity, pass. obj.	46.94	50.51	51.06
Intersection type	27.39	33.51	34.97	Roadworks	70.61	74.29	78.80
Land use - driver	62.35	64.46	63.85	Sc. zone crossing supervisor	62.84	65.60	66.51
Land use - passenger	62.81	67.08	69.60	Sc. zone warning	62.57	66.90	68.03
Lane width	64.99	71.21	75.60	Service road	55.28	59.44	62.26
Median Type	40.76	44.45	55.62	Sidewalk - driver-side	43.63	46.08	43.72
Motorcycle observed flow	26.04	33.23	36.45	Sidewalk - passenger-side	44.31	45.83	50.19
Number of lanes	71.09	84.32	98.49	Sight distance	70.00	70.85	75.67
Paved shoulder - driver	63.54	64.71	65.02	Skid resistance / grip	38.77	44.52	50.90
Paved shoulder - passenger	46.43	48.52	54.01	Speed management	61.57	100.00	100.00
Ped. crossing - inspected rd.	35.61	35.73	45.03	Street lighting	90.58	91.38	91.53
Ped. crossing - side rd.	39.12	43.78	46.58	Upgrade cost	61.48	62.45	63.28
Ped. crossing quality	48.11	53.11	59.21	Vehicle parking	56.29	60.57	60.28
Ped. obs. flow, across	26.22	33.54	44.31	**Mean**	**54.96**	**59.68**	**62.86**

We compare three models: the convolutional model trained with standard cross-entropy loss (CNN CE), the model with added sequential enhancement (+LSTM CE), and our full model with both sequential enhancement and multi-task dynamic loss weighting (CNN+LSTM CE^{R}_{MT}). All three models incorporate semantic segmentation pre-training.

We observe that attributes affected by severe class imbalance benefit the most from our dynamic loss weighting approach. Notably, the attributes *Pedestrian crossing - inspected road*, *Median type*, *Pedestrian observed flow along the road passenger-side*, *Roadside severity - driver-side object*, and *Bicycle facility* show relative improvements ranging from 19% to 26.5% compared to the standard cross-entropy loss. This confirms that our loss weighting strategy effectively mitigates the negative impact of class imbalance in multi-task learning.

Our analysis further reveals the benefits of sequential enhancement for different attribute types. Single-peak attributes that benefit most from sequential enhancement include *Speed management* and *Intersection type*, with relative improvements of 62.4% and 22.3%, respectively. This improvement suggests that recurrent models are able to learn and accommodate the annotation rule specific for these attributes, which mandates that the positive class be annotated only in the one nearest segment per appearance.

For smooth attributes, which rarely change classes, the sequential model effectively corrects spurious class transitions made by the local model by considering a larger temporal context. The attributes *Bicycle facility*, *Number of lanes*, and *Skid resistance / grip* show the largest relative improvements in this group of attributes.

Figure 5.3 illustrates four examples where sequential enhancement successfully corrects erroneous predictions of the local model. In the first example involving the single-peak attribute *Intersection type*, the local model incorrectly assigns a positive class (*3-way intersection*) to two consecutive segments. Sequential enhancement rectifies this prediction to adhere to the single-peak annotation convention. In the second and third examples, involving smooth attributes *Bicycle facility* and *Number of lanes*, the local model predicts spurious class transitions due to visual ambiguities or momentary occlusions. For instance, the local model mistakes a motorcyclist near a tram rail for a dedicated bicycle lane and fails to predict the correct number of lanes. Sequential enhancement considers a larger temporal context and successfully corrects these local mistakes. The fourth example involves the *Street lighting* attribute, which should be annotated continuously through all segments between two nearby lighting poles. In the example, the upcoming lighting poles are obscured by road curvature and overgrown roadside vegetation, causing the local model to misclassify segments between poles. Sequential enhancement utilizes a larger context to infer the presence of street lighting and corrects the misclassifications.

To evaluate the effect of different pre-training strategies for our model's backbone, we conduct an ablation study summarized in Table 5.3. We compare semantic segmentation pre-training on the Vistas dataset with classification pre-training on ImageNet-1k and two road scene classification datasets: BDD100k and Honda Scenes. Results indicate that semantic segmentation pre-training consistently outperforms classification pre-training. This suggests that detailed spatial understanding of visual concepts, as learned through semantic segmentation, is beneficial for recognizing road-safety attributes in traffic scenes.

Table 5.3: Impact of different pre-training strategies on overall Macro-F1 performance on iRAP-BH test.

Dataset	Task	Macro-F1
ImageNet-1k	Classification	61.26
BDD100k	Classification	61.19
Honda Scenes	Classification	61.35
Vistas	Semantic Segmentation	**62.86**

Overall, our experimental results on the iRAP-BH dataset demonstrate the efficacy of our multi-stage approach and the impact of our contributions. The combination of semantic seg-

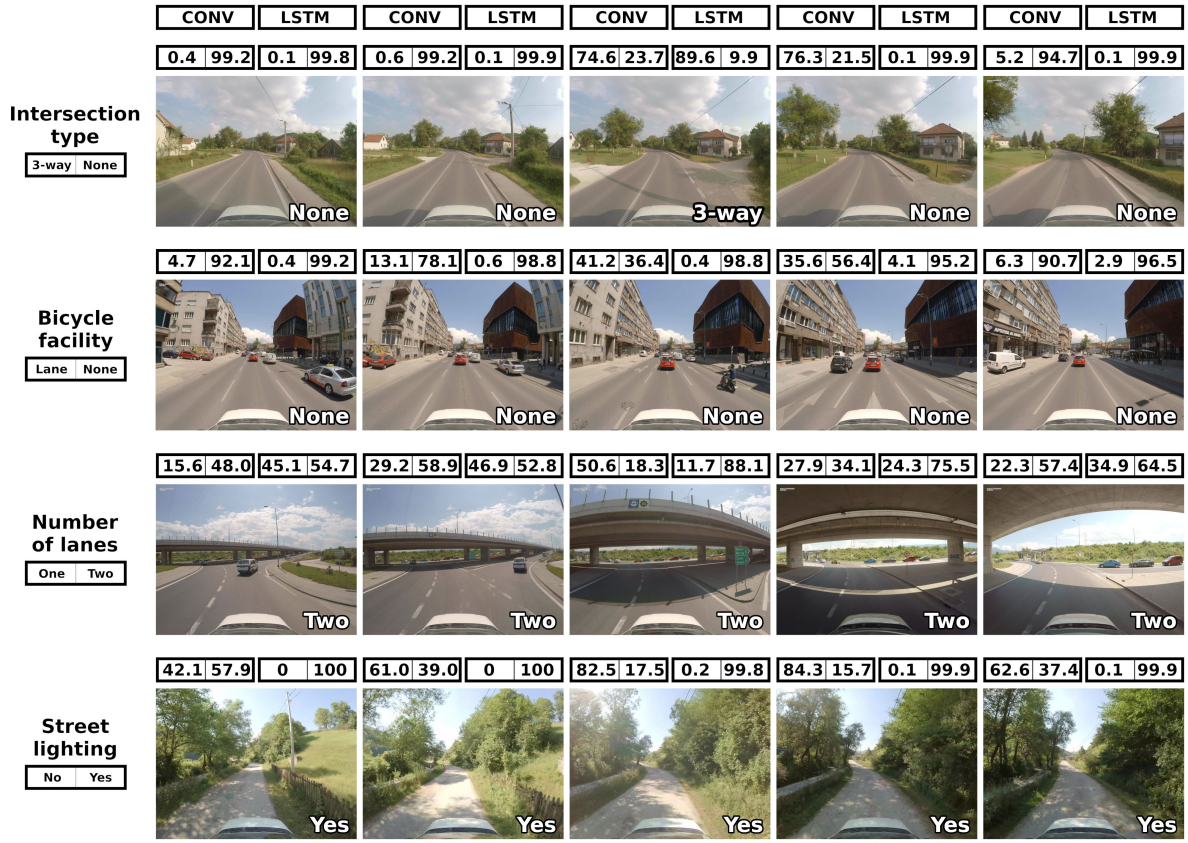

Figure 5.3: Four iRAP-BH examples where sequential enhancement (LSTM) succeeds to correct local visual predictions (CNN). For each of the five consecutive segments (columns), we display the categorical predictions by both models (top) and the ground truth label (bottom right). Row 1 involves a single-peak attribute - *Intersection type*. We observe that the local model incorrectly assigns a positive class (*3-way intersection*) in column 4. Rows 2 and 3 involve smooth attributes - *Bicycle facility* and *Number of lanes*. We observe that the local model mistakes a motorcyclist near a tram rail for a dedicated bicycle lane in column 3 and fails to predict the correct number of lanes again in column 3. Row 4 involves the *Street lighting* attribute. We observe that the upcoming lighting poles are obscured by the road curvature and overgrown roadside bushes. Consequently, the local model fails in columns 2-4. In all cases, the sequential model succeeds to correct the mistakes by leveraging a larger temporal context.

mentation pre-training, multi-task dynamic loss weighting, and sequential enhancement leads to substantial improvements in recognizing road-safety attributes, particularly those affected by class imbalance and temporal dependencies.

5.2.4 Honda Scenes

This subsection compares our method with prominent previous work on the Honda Scenes dataset. We include several methods from the original paper [24], as well as Context MTL [103], MTAN [104], and two of our own ablations that demonstrate the impact of our contributions.

Road Place

Table 5.4 evalutes the overall performance across all tasks of the Road place problem. We refer to the original sequential baseline as Honda BiLSTM and their two-stage sequential approach as Honda Event [24].

Table 5.4: Macro-F1 performance on Honda Scenes - Road-place.

Model	BB	Road place	
		Mean	Mean w/o Background
Honda BiLSTM [24]	rn50	27.56	25.23
Honda Event [24]	rn50	28.36	25.91
Context MTL [103]	rn50	-	27.92
MTAN [104]	wrn28	29.14	26.73
ConvCE (ours)	rn18	34.11	31.96
$\text{ConvCE}^{\text{R}}_{\text{MT}}$ (ours)	rn18	37.00	34.92
$\text{ConvCE}^{\text{R}}_{\text{MT}}$-SE (ours)	rn18	**40.93**	**39.00**

Table 5.5 offers a more granular view by providing per-task results. Despite using a weaker backbone, our baseline model (multi-frame model with standard loss and without sequential enhancement) outperforms all previous approaches. The strong performance of this baseline

Table 5.5: Experimental evaluation on all Road-Place tasks of Honda Scenes (macro-F1, percentage points). Legend: BB - backbone; B-*Background*, A-*Approaching*, E-*Entering*, P-*Passing*.

Model	BB	B	Intersection 5-way				Railway Crossing				Construction				Left Merge			Right Merge		
		-	A	E	P	Mean	A	E	P	Mean	A	E	P	Mean	A	P	Mean	A	P	Mean
Honda BiLSTM [24]	rn50	88	0	0	9	3	24	14	46	28	2	5	29	12	9	28	19	16	23	20
Honda Event [24]	rn50	**92**	0	0	0	0	23	47	46	39	2	6	38	15	5.6	8	7	13	16	15
Context MTL [103]	rn50	-	0	6	0	2	1	35	52	32	0	4	38	14	4	6	5	26	18	22
MTAN [104]	wrn28	92	1	2	5	3	19	27	42	29	3	9	24	12	11	17	14	19	12	16
ConvCE (ours)	rn18	90	19	0	5	8	13	49	52	38	2	11	56	23	22	29	26	29	33	31
$\text{ConvCE}^{\text{R}}_{\text{MT}}$ (ours)	rn18	91	27	0	10	12	15	56	59	43	3	12	63	26	27	36	32	31	35	33
$\text{ConvCE}^{\text{R}}_{\text{MT}}$-SE (ours)	rn18	91	29	0	9	**13**	28	53	71	**51**	11	22	64	**32**	29	43	**36**	34	45	**40**
Model	BB		Overhead Bridge				Intersection 3-way				Intersection 4-way				Left Branch			Right Branch		
			A	E	P	Mean	A	E	P	Mean	A	E	P	Mean	A	P	Mean	A	P	Mean
Honda BiLSTM [24]	rn50		23	55	53	44	3	28	27	19	14	68	66	49	36	22	29	28	28	28
Honda Event [24]	rn50		42	58	59	53	8	16	23	16	31	7	67	56	30	19	25	24	22	23
Context MTL [103]	rn50		47	59	60	**55**	11	38	28	26	14	78	79	57	33	19	26	34	27	31
MTAN [104]	wrn28		31	49	57	46	15	35	31	27	25	75	73	58	38	26	32	30	19	25
ConvCE (ours)	rn18		31	57	60	50	10	33	32	25	25	77	66	56	29	18	24	35	38	37
$\text{ConvCE}^{\text{R}}_{\text{MT}}$ (ours)	rn18		33	58	61	51	10	34	34	26	26	77	68	57	31	23	27	37	42	40
$\text{ConvCE}^{\text{R}}_{\text{MT}}$-SE (ours)	rn18		37	59	64	53	14	32	38	**28**	30	78	73	**60**	42	24	**33**	41	44	**43**

makes the impact of our subsequent contributions more compelling. Incorporating multi-task dynamic loss weighting increases performance by 2.9 percentage points (pp). This is due to significant class imbalance, recognized also by the authors of the dataset [24]. We observe the greatest relative improvements on tasks *Intersection (5-way or more)*, *Railway*, *Left Merge*, and *Left Branch*. Sequential enhancement brings an additional improvement of 3.9 pp.

Road Environment

Table 5.6 compares per-class performance of our multi-frame model with Context MTL, MTAN, and the two original frame-based approaches [24]. The original approaches employ a ResNet-50 backbone pre-trained on Places365 and leverage the DeepLabV2 semantic segmentation model. The latter is utilized either to mask out traffic participants (Honda Frame - Mask) or to augment the input image with its segmentation map (Honda Frame - SemSeg).

Our model achieves superior results across most classes, with the largest improvement occurring on the challenging class *Ramp*. This class is the least frequent in the dataset and benefits the most from multi-task loss weighting. The addition of sequential enhancement further improves overall performance by 1.2 pp.

Table 5.6: Macro-F1 performance on the Road environment problem of Honda Scenes.

Road environment						
Model	BB	Local	Highway	Ramp	Urban	Mean
Honda Frame - Mask [24]	rn50	33.0	91.0	20.0	83.0	56.8
Honda Frame - SemSeg [24]	rn50	34.0	89.0	13.0	81.0	54.3
Context MTL [103]	rn50	36.0	92.0	21.0	81.0	57.5
MTAN [104]	wrn28	**37.2**	91.1	19.7	80.3	57.1
ConvCE (ours)	rn18	29.1	91.2	36.3	82.9	59.9
$\text{ConvCE}^{\text{R}}_{\text{MT}}$ (ours)	rn18	30.8	92.3	42.6	83.7	62.4
$\text{ConvCE}^{\text{R}}_{\text{MT}}$-SE (ours)	rn18	32.8	**93.1**	**44.1**	**84.2**	**63.6**

Road surface

Table 5.7 shows that our single-frame model surpasses previous approaches on the Road surface task. Since the classes in this task are relatively balanced, loss weighting does not not yield additional performance gains. Furthermore, multi-frame and sequential enhancements are not applicable, as this task is restricted to single-frame prediction.

Table 5.8 reveals that pre-training our ResNet-18 on the Vistas semantic segmentation dataset provides better results than using the larger ResNet-50 pre-trained on ImageNet-1k.

Table 5.7: Macro-F1 performance on the Road surface problem of Honda Scenes.

Road surface						
Model	BB	Dry	Wet	Snow	Mean	Mean w/o Snow
Honda Frame - Mask [24]	rn50	93.0	92.0	99.7	94.9	92.5
Honda Frame - SemSeg [24]	rn50	92.2	92.5	99.0	94.6	92.4
Context MTL [103]	rn50	93.0	92.0	-	-	92.5
MTAN [104]	wrn28	94.2	94.1	98.9	95.7	94.2
Conv single (ours)	rn18	**98.5**	**98.5**	**100.0**	**99.0**	**98.5**

Table 5.8: Ablation of segmentation pre-training (Macro-F1) on the Road surface problem of Honda Scenes.

Road surface					
Pre-training	BB	Dry	Wet	Snow	Mean
IN-1k (ours)	rn18	97.2	96.8	99.7	97.9
IN-1k (ours)	rn50	97.7	97.3	99.7	98.2
Vistas (ours)	rn18	**98.5**	**98.5**	**100.0**	**99.0**

Weather

Table 5.9 shows that our single-frame model outperforms previous approaches on the classes *Overcast* and *Snow*, as well as in overall performance, while underperforming on classes *Clear* and *Rain*.

Table 5.9: Macro-F1 performance on Weather problem of Honda Scenes.

Weather							
Model	BB	Clear	Overcast	Rain	Snow	Mean	Mean w/o Snow
Honda Frame - Mask [24]	rn50	92.0	83.0	96.0	94.0	91.3	90.3
Honda Frame - SemSeg [24]	rn50	91.6	83.4	96.0	93.9	91.2	90.3
Context MTL [103]	rn50	**93.2**	84.0	**97.0**	-	-	91.4
MTAN [104]	wrn28	90.4	85.9	91.2	92.5	90.0	89.2
Conv single (ours)	rn18	91.0	**90.9**	95.0	**95.3**	**93.0**	**92.3**

Similar to the Road Surface problem, the balanced distribution of classes means that loss weighting does not provide additional benefits, while sequential enhancement is not applicable since due is a single-frame prediction task. Moreover, ablations in Table 5.10 suggests that task-

specific segmantic segmentation pre-training is more beneficial than classification pre-training, even using a larger model.

Table 5.10: Ablation of segmentation pre-training (Macro-F1) on the Weather problem of Honda Scenes.

Weather						
Pre-training	BB	Clear	Overcast	Rain	Snow	Mean
IN-1k (ours)	rn18	89.5	86.8	94.0	93.0	90.8
IN-1k (ours)	rn50	90.1	87.6	94.3	94.0	91.5
Vistas (ours)	rn18	**91.0**	**90.9**	**95.0**	**95.3**	**93.0**

5.2.5 FM3m

We evaluate our approach on the Fleet Management (FM3m) dataset by comparing against several approaches.

The original dataset authors provided two baselines which leverage SVM classifiers with radial basis function kernels. These approaches leverage SVM classifiers with radial basis function kernels. The classifiers operate on image descriptors extracted by ResNet-50 and DenseNet-121 models, both pre-trained on ImageNet-1k. It is important to note that the backbones were not fine-tuned on the FM3m dataset.

We also evaluate our method against other contemporary traffic scene recognition approaches, namely the two variants of the Honda Frame-based model [24] and the Multi-Task Attention Network (MTAN) [104]. The Honda Frame models use a ResNet-50 backbone pre-trained on the Places365 dataset and improve classification performance by leveraging semantic segmentation masks. One variant masks out traffic participants, while the other adds the segmentation mask as an additional input channel.

The results presented in Table 5.11 demonstrate that our single-frame models achieve competitive performance with all of the aforementioned approaches. Furthermore, multi-task dynamic loss weighting improves performance across nearly all classes.

To further explore the impact of pre-training and backbone architecture on our model's performance, we conduct an ablation study comparing three variants of our single-level model:

1. A model with a ResNet-18 backbone pre-trained on the ImageNet-1k dataset.
2. A model with a ResNet-18 backbone pre-trained on the Vistas dataset.
3. A model with a larger ResNet-50 backbone pre-trained on the ImageNet-1k dataset.

As shown in Table 5.12, semantic segmentation pre-training on the Vistas dataset offers more substantial improvements in classification performance than increasing the capacity of the backbone.

Table 5.11: AP performance on the FM3m dataset: H-*Highway*, R-*Road*, Tu-*Tunnel*, E-*Exit*, S-*Settlement*, O-*Overpass*, B-*Booth*, Tr-*Traffic*.

Model	H	R	Tu	E	S	O	B	Tr	Mean
RN50-SVM [25]	99.8	91.1	**100.0**	97.7	98.3	97.2	98.8	86.8	96.2
DN121-SVM [25]	93.7	**100.0**	98.0	98.3	97.9	98.8	87.9	**96.8**	96.4
Honda Frame - Mask [24]	99.5	91.0	99.2	93.1	97.1	96.5	97.9	80.4	94.4
Honda Frame - SemSeg [24]	96.5	92.9	95.8	92.1	95.6	97.5	94.0	92.3	94.6
MTAN [104]	98.2	92.3	98.1	94.3	98.0	94.1	97.9	91.0	95.5
Conv single, CE (ours)	**100.0**	94.2	99.8	98.0	98.2	98.5	99.4	90.5	97.3
Conv single, CE^{R}_{MT} (ours)	**100.0**	94.6	**100.0**	**98.6**	**98.5**	**98.9**	**99.8**	91.2	**97.7**

Table 5.12: Ablation of pre-training on FM3m (AP): H-*Highway*, R-*Road*, Tu-*Tunnel*, E-*Exit*, S-*Settlement*, O-*Overpass*, B-*Booth*, Tr-*Traffic*.

Model	BB	H	R	Tu	E	S	O	B	Tr	Mean
IN-1k (ours)	rn18	99.2	91.5	99.6	97.3	97.8	97.0	98.2	89.9	96.3
IN-1k (ours)	rn50	99.6	92.5	99.9	98.2	98.0	97.4	98.9	90.5	96.9
Vistas (ours)	rn18	**100.0**	**94.6**	**100.0**	**98.6**	**98.5**	**98.9**	**99.8**	**91.2**	**97.7**

It is important to note that sequential enhancement is not applicable in this context, as the FM3m dataset involves only single-frame predictions.

5.2.6 BDD100k

In this section, we evaluate our approach on the BDD100k dataset, focusing on the *Scene* classification task. We consider both the default and the cross-domain evaluation setup.

In the default setup, shown in table 5.13, we train and test our model on the full dataset without any domain constraints. In the cross-domain setup, shown in table 5.14, we simulate domain shifts by training models exclusively on images captured under sunny weather conditions and evaluating them on images from cloudy, rainy, and snowy conditions. This setup tests generalization capabilities under significant domain shifts caused by varying weather conditions. We present the performance of our method compared to several state-of-the-art approaches. In both setups, we compare against Honda Frame - Mask, Honda Frame - SemSeg [24], and the Multi-Task Attention Network (MTAN) [104].

The default setup additionally includes the Local-Global FCRNN model [133], which employs a two-stream architecture combining local features from Faster R-CNN and global features from InceptionV2.

Table 5.13: Comparison with prior work on the Scene task BDD100k (default setup).

BDD100k default setup	
Model	Accuracy
Honda Frame - Mask [24]	76.8
Honda Frame - SemSeg [24]	76.0
MTAN [104]	73.9
Local-Global FCRNN [133]	76.0
Conv single, CE (ours)	78.4
Conv single, CE^{R}_{MT} (ours)	**78.7**

For the cross-domain setup, we additionally compare against the "source-only" baseline from the Sparse Adversarial Domain Adaptation (SADA) method [134]. This baseline provices a reference point for evaluating generalization under domain shifts.

Table 5.14: Comparison with prior work on the Scene task of BDD100k (cross-domain setup).

BDD100k cross-domain setup			
Model	Cloudy	Rainy	Snowy
Honda Frame - Mask [24]	70.9	63.8	62.2
Honda Frame - SemSeg [24]	72.3	65.1	62.3
MTAN [104]	71.1	66.2	60.7
SADA [134]	70.5	62.7	59.1
Conv single, CE (ours)	75.9	71.4	70.0
Conv single, CE^{R}_{MT} (ours)	**76.6**	**71.5**	**70.9**

Experiments show that our single-frame models consistently outperform all competing methods across both the default and cross-domain setups. The results imply some degree of robustness of our method in handling diverse scenes and weather conditions. Incorporating multi-task dynamic loss weighting increases performance in every configuration. Lastly, our sequential enhancement approach is not applicable in this context due to the single-frame prediction constraint of the BDD100k dataset.

5.3 Qualitative examples from iRAP-BH

In this section, we present qualitative examples that showcase the detection of road-safety attributes on the iRAP-BH dataset. These examples are designed to provide a more intuitive

understanding of how our models perform on individual road segments and how they generate predictions across a sequence of frames. Figure 5.4 displays examples of five different attributes. For each example, we display a sequence of five images corresponding to successive 10-meter road segments.

Each road segment is represented by two aligned images: the original input image and its corresponding saliency map generated by the convolutional model. The saliency maps highlight the image regions that most strongly influence the model's decisions. In the upper-right corner of each image, we provide the ground truth label for the corresponding road-safety attribute. Additionally, the predictions from the convolutional (conv) and sequential (seq) models are displayed in the upper-left and upper-right corners of the corresponding saliency map, respectively. The name of each attribute is shown to the left of the image sequence, along with its associated class labels. The attributes showcased in these examples include *Median type*, *Sidewalk - passenger-side*, *Roadside severity - passenger-side object*, *Roadworks*, and *Pedestrian crossing - inspected road*.

These qualitative results demonstrate how our models capture both spatial and temporal features to make accurate predictions across consecutive road segments.

Figure 5.4: Qualitative experiments on iRAP-BH for 5 road-safety attributes. For each attribute, the top row shows 5 input images with the ground truth label, while the bottom row presents the corresponding saliency maps and the predictions of the two stages of our model (conv, seq).

Chapter 6

Conclusion

This thesis introduces a novel two-stage visual recognition framework for automatic assessment of road infrastructure safety attributes from monocular video data. In contrast to traditional reactive road safety assessment methods that depend on historical accident data, this framework adopts a proactive approach. It attempts to identify potential hazards within the road environment before the incidents occur. The shift towards proactive road safety assessment is crucial, as traffic-related fatalities continue to be a significant global issue. This approach aligns with international efforts to develop safer transportation infrastructure and reduce the loss of life on roads.

The proposed framework consists of two stages - local recognition and sequential enhancement. Both stages leverage deep learning techniques to handle the complexities inherent in real-world road scenes and the challenges posed by the iRAP attribute set, such as class imbalance, visually similar fine-grained classes, and temporal behavior.

The local recognition stage pre-trains a convolutional feature extractor for semantic segmentation on the Vistas dataset. This facilitates detailed representations of road infrastructure elements. This pre-training strategy proves more effective than classification pre-training on ImageNet-1k and demonstrates the importance of detailed scene understanding for road safety attribute recognition.

To address the pervasive challenge of extreme class imbalance in road safety datasets, this work introduces a dynamic multi-task loss weighting strategy. By adjusting class weights based on their recall score during training, the system maintains focus on rare classes without compromising its ability to recognize more common ones. Additionally, by normalizing each per-task loss with the sum of the example weights, we alleviate inter-attribute interference during training. This enables more effective joint learning across many significantly imbalanced tasks.

A key innovation of this research lies in the analysis and treatment of specific temporal patterns of road infrastrucure safety attributes. The sequential enhancement stage, implemented through per-attribute bidirectional LSTM networks, refines the local predictions by incorpo-

rating a larger temporal context from adjacent road segments. This approach is particularly effective for attributes that exhibit characteristic temporal behaviors, such as "single-peak" patterns where infrastructure elements should be predicted only in the closest road segment, or "smooth" patterns where the attributes tend to maintain consistency across consecutive segments. The bidirectional nature of these networks enables the system to leverage both past and future context, resulting in more coherent and accurate predictions across sequences of road segments.

The effectiveness of the proposed framework was validated through experiments across multiple datasets. The novel dataset iRAP-BH encompasses over 226,000 labeled images of 10-meter road segments collected along 2,300 kilometers of diverse road infrastructure in Bosnia and Herzegovina. Each image is annotated with the values of all iRAP attributes, providing a comprehensive resource for training and evaluation of road safety assessment models. Experimental results on the iRAP-BH dataset validate the impact of each contribution - semantic segmentation pre-training, dynamic multi-task loss weighting and sequential enhancement. Qualitative analyses provide intuition and illustrate the model's capacity to leverage spatial and temporal context in order to adhere to the nuanced annotation conventions outlined in the iRAP standard. In addition, the robustness and generalizability of the framework was tested on three established road scene classification datasets: Honda Scenes, FM3m and BDD100k. The framework outperforms all previous work on the Honda Scenes dataset, particularly on the Road place task which provides sequential input and involves significantly imbalanced taxonomies. We also achieve competitive performance on FM3m and BDD100k, even though our approach employs a weaker backbone and the two datasets allow only for single-frame input. The results confirm the framework's ability to accurately and robustly classify complex road scene scenarios across different datasets and environmental conditions.

This work provides a foundation for future advancements in automated road safety assessment. The rapid advancement of transformer architectures for image and video recognition presents opportunities for enhancing both the local and sequential components of the system. Self-supervised and contrastive training of very large foundation models on large-scale datasets enable extraction of general, robust and transferrable image features across various domains.

The presented contributions mark a step towards efficient, scalable, and automated road infrastructure safety assessment. The framework provides valuable tools for infrastructure planning and maintenance by enabling more comprehensive and frequent road safety assessments. The resulting insights can inform policy decisions, guide infrastructure improvements, and contribute to the reduction of traffic-related fatalities worldwide.

Bibliography

[1] World Health Organization, "Global status report on road safety 2018: Summary," 2018, WHO/NMH/NVI/18.20.

[2] United Nations General Assembly, "Improving global road safety," 2020, A/RES/74/299.

[3] L. Mooren, R. Grzebieta, and R. S. Job, "Safe system – comparisons of this approach in australia," in *Australasian College of Road Safety Conference "A Safe System: Making it Happen!", September 1-2, 2011, Melbourne, Australia*, 2011.

[4] B. F. Corben, D. B. Logan, L. Fanciulli, R. Farley, and I. Cameron, "Strengthening road safety strategy development 'towards zero' 2008–2020 – western australia's experience scientific research on road safety management swov workshop 16 and 17 november 2009," *Safety Science*, vol. 48, no. 9, pp. 1085–1097, 2010, scientific Research on Road Safety Management.

[5] M. Green, C. Muir, J. Oxley, and A. Sobhani, "Safe system in road safety public policy: A case study from victoria, australia," *IATSS Research*, vol. 46, no. 2, pp. 171–180, 2022.

[6] H. Stipdonk, S. Job, and B. Turner, "The safe system approach in action," ITF, 2022.

[7] S. Chatterjee and S. Mitra, "Safety assessment of two-lane highway using a combined proactive and reactive approach: Case study from indian national highways," *Transportation Research Record*, vol. 2673, no. 7, pp. 709–721, 2019.

[8] H. Cui, J. Dong, M. Zhu, X. Li, and Q. Wang, "Identifying accident black spots based on the accident spacing distribution," *Journal of Traffic and Transportation Engineering (English Edition)*, 2022.

[9] S. Aziz and S. Ram, "A meta-analysis of the methodologies practiced worldwide for the identification of road accident black spots," *Transportation Research Procedia*, vol. 62, pp. 790–797, 2022, 24th Euro Working Group on Transportation Meeting.

[10] S. Job, "Advantages and disadvantages of reactive (black spot) and proactive (road rating) approaches to road safety engineering treatments: When should each be used?" in

Australasian Road Safety Research Policing Education Conference, 2012, Wellington, New Zealand, 2012.

[11] A. Adedokun, "Application of road infrastructure safety assessment methods at intersections," Ph.D. dissertation, Linköping University, 2016.

[12] S. He, M. A. Sadeghi, S. Chawla, M. Alizadeh, H. Balakrishnan, and S. Madden, "Inferring high-resolution traffic accident risk maps based on satellite imagery and GPS trajectories," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 11 957–11 965.

[13] iRAP - International Road Assesment Programme, "iRAP Coding Manual Version 5.0 – Drive on Right Edition," 2019.

[14] ——, "iRAP Star Rating and Investment Plan Implementation Support Guide," 2017.

[15] Q. Li, J. Bradford, and A. M. Bachani, "Statistical estimation of fatal and serious injuries saved by irap protocols in 74 countries," *PLOS ONE*, vol. 19, no. 4, pp. 1–10, 04 2024. [Online]. Available: https://doi.org/10.1371/journal.pone.0301993

[16] Steve Lawson, "iRAP methodology and safer roadsides," 2017.

[17] Road Safety Foundation, "Engineering Safer Roads: Star Rating roads for in-built safety," 2015.

[18] Y. Bengio, A. C. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[19] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 5000–5009.

[20] M. Kacan, M. Orsic, S. Segvic, and M. Sevrovic, "Multi-task learning for irap attribute classification and road safety assessment," in *23rd IEEE International Conference on Intelligent Transportation Systems, ITSC 2020, Rhodes, Greece, September 20-23, 2020*. IEEE, 2020, pp. 1–6.

[21] J. Tian, N. C. Mithun, Z. Seymour, H. Chiu, and Z. Kira, "Striking the right balance: Recall loss for semantic segmentation," in *2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022*. IEEE, 2022, pp. 5063–5069.

[22] Ö. Yildirim, "A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification," *Comput. Biol. Medicine*, vol. 96, pp. 189–202, 2018.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[24] A. Narayanan, I. Dwivedi, and B. Dariush, "Dynamic traffic scene classification with space-time coherence," in *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*. IEEE, 2019, pp. 5629–5635.

[25] I. Sikiric, K. Brkic, P. Bevandic, I. Kreso, J. Krapac, and S. Segvic, "Traffic scene classification on a representation budget," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 1, pp. 336–345, 2020.

[26] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 2633–2642. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Yu_BDD100K_A_Diverse_Driving_Dataset_for_Heterogeneous_Multitask_Learning_CVPR_2020_paper.html

[27] B. Dimitrijevic, S. Darban Khales, R. Asadi, and J. Lee, "Short-term segment-level crash risk prediction using advanced data modeling with proactive and reactive crash data," *Applied Sciences*, vol. 12, p. 856, 01 2022.

[28] S. M. Gaweesh, M. M. Ahmed, and A. V. Piccorelli, "Developing crash prediction models using parametric and nonparametric approaches for rural mountainous freeways: A case study on wyoming interstate 80," *Accident Analysis & Prevention*, vol. 123, pp. 176–189, 2019.

[29] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Springer, 2018.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1106–1114.

[31] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Anal.*, vol. 42, pp. 60–88, 2017. [Online]. Available: https://doi.org/10.1016/j.media.2017.07.005

[32] S. M. Grigorescu, B. Trasnea, T. T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020. [Online]. Available: https://doi.org/10.1002/rob.21918

[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, X. Xie, M. W. Jones, and G. K. L. Tam, Eds. BMVA Press, 2015, pp. 41.1–41.12. [Online]. Available: https://doi.org/10.5244/C.29.41

[34] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 580–587. [Online]. Available: https://doi.org/10.1109/CVPR.2014.81

[35] T. M. Mitchell, *Machine learning, International Edition*, ser. McGraw-Hill Series in Computer Science. McGraw-Hill, 1997. [Online]. Available: https://www.worldcat.org/oclc/61321007

[36] I. J. Goodfellow, Y. Bengio, and A. C. Courville, *Deep Learning*, ser. Adaptive computation and machine learning. MIT Press, 2016.

[37] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[38] J. Sánchez and F. Perronnin, "High-dimensional signature compression for large-scale image classification," in *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*. IEEE Computer Society, 2011, pp. 1665–1672. [Online]. Available: https://doi.org/10.1109/CVPR.2011.5995504

[39] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013. [Online]. Available: https://doi.org/10.1007/s11263-013-0636-x

[40] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nat.*, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539

[41] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 2009, pp. 248–255. [Online]. Available: https://doi.org/10.1109/CVPR.2009.5206848

[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015. [Online]. Available: https://doi.org/10.1007/s11263-015-0816-y

[43] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "cudnn: Efficient primitives for deep learning," *CoRR*, vol. abs/1410.0759, 2014. [Online]. Available: http://arxiv.org/abs/1410.0759

[44] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *CoRR*, vol. abs/1603.04467, 2016. [Online]. Available: http://arxiv.org/abs/1603.04467

[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 8024–8035. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[46] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information*

Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 91–99. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3431–3440. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298965

[48] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, 2020. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2844175

[49] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 5695–5703. [Online]. Available: https://doi.org/10.1109/CVPR.2016.614

[50] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 2758–2766. [Online]. Available: https://doi.org/10.1109/ICCV.2015.316

[51] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct 1986. [Online]. Available: https://doi.org/10.1038/323533a0

[52] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, pp. 65:1–65:32, 2016. [Online]. Available: http://jmlr.org/papers/v17/15-535.html

[53] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing, ICIP 2017, Beijing, China, September 17-20, 2017*. IEEE, 2017, pp. 3645–3649. [Online]. Available: https://doi.org/10.1109/ICIP.2017.8296962

[54] A. Kendall, H. Martirosyan, S. Dasgupta, and P. Henry, "End-to-end learning of geometry and context for deep stereo regression," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 66–75. [Online]. Available: https://doi.org/10.1109/ICCV.2017.17

[55] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2577031

[56] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, J. Fürnkranz and T. Joachims, Eds. Omnipress, 2010, pp. 807–814. [Online]. Available: https://icml.cc/Conferences/2010/papers/432.pdf

[57] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, ser. JMLR Proceedings, G. J. Gordon, D. B. Dunson, and M. Dudík, Eds., vol. 15. JMLR.org, 2011, pp. 315–323. [Online]. Available: http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf

[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

[59] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 2488–2498.

[60] M. Lin, Q. Chen, and S. Yan, "Network in network," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: http://arxiv.org/abs/1312.4400

[61] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*. IEEE Computer Society, 2006, pp. 2169–2178. [Online]. Available: https://doi.org/10.1109/CVPR.2006.68

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015. [Online]. Available: https://doi.org/10.1109/TPAMI.2015.2389824

[63] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 6230–6239. [Online]. Available: https://doi.org/10.1109/CVPR.2017.660

[64] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, D. Wu, M. Carpuat, X. Carreras, and E. M. Vecchi, Eds. Association for Computational Linguistics, 2014, pp. 103–111. [Online]. Available: https://aclanthology.org/W14-4012/

[65] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014. [Online]. Available: http://arxiv.org/abs/1412.3555

[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[68] ——, "Identity mappings in deep residual networks," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9908. Springer, 2016, pp. 630–645. [Online]. Available: https://doi.org/10.1007/978-3-319-46493-0_38

[69] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017.*

IEEE Computer Society, 2017, pp. 2261–2269. [Online]. Available: https://doi.org/10.1109/CVPR.2017.243

[70] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 6391–6401.

[71] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 11 966–11 976. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01167

[72] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[73] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017. [Online]. Available: http://arxiv.org/abs/1704.04861

[74] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 4510–4520. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html

[75] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, vol. abs/1607.06450, 2016. [Online]. Available: http://arxiv.org/abs/1607.06450

[76] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," *CoRR*, vol. abs/1606.08415, 2016. [Online]. Available: http://arxiv.org/abs/1606.08415

[77] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010. [Online]. Available: https://doi.org/10.1109/TKDE.2009.191

[78] L. Torrey, J. W. Shavlik, T. Walker, and R. Maclin, "Transfer learning via advice taking," in *Advances in Machine Learning I: Dedicated to the Memory of Professor Ryszard S. Michalski*, ser. Studies in Computational Intelligence, J. Koronacki, Z. W. Ras, S. T. Wierzchon, and J. Kacprzyk, Eds. Springer, 2010, vol. 262, pp. 147–170. [Online]. Available: https://doi.org/10.1007/978-3-642-05177-7_7

[79] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 1717–1724. [Online]. Available: https://doi.org/10.1109/CVPR.2014.222

[80] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *Proc. CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 12 607–12 616.

[81] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 3320–3328. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html

[82] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 2014, pp. 512–519. [Online]. Available: https://doi.org/10.1109/CVPRW.2014.131

[83] M. Huh, P. Agrawal, and A. A. Efros, "What makes imagenet good for transfer learning?" *CoRR*, vol. abs/1608.08614, 2016. [Online]. Available: http://arxiv.org/abs/1608.08614

[84] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, ser. JMLR Workshop and

Conference Proceedings, vol. 32. JMLR.org, 2014, pp. 647–655. [Online]. Available: http://proceedings.mlr.press/v32/donahue14.html

[85] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 36–45. [Online]. Available: https://doi.org/10.1109/CVPRW.2015.7301270

[86] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 2962–2971. [Online]. Available: https://doi.org/10.1109/CVPR.2017.316

[87] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018. [Online]. Available: https://doi.org/10.1016/j.neucom.2018.05.083

[88] I. Martinovic, J. Saric, and S. Segvic, "Mc-panda: Mask confidence for panoptic domain adaptation," *CoRR*, vol. abs/2407.14110, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2407.14110

[89] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: http://proceedings.mlr.press/v139/radford21a.html

[90] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P. Huang, S. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, vol. 2024, 2024. [Online]. Available: https://openreview.net/forum?id=a68SUt6zFt

[91] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997. [Online]. Available: https://doi.org/10.1023/A:1007379606734

[92] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine*

Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011, L. Getoor and T. Scheffer, Eds. Omnipress, 2011, pp. 689–696. [Online]. Available: https://icml.cc/2011/papers/399_icmlpaper.pdf

[93] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 7482–7491. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Kendall_Multi-Task_Learning_Using_CVPR_2018_paper.html

[94] R. Goyal, E. Mavroudi, X. Yang, S. Sukhbaatar, L. Sigal, M. Feiszli, L. Torresani, and D. Tran, "MINOTAUR: multi-task video grounding from multimodal queries," *CoRR*, vol. abs/2302.08063, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2302.08063

[95] S. Ruder, "An overview of multi-task learning in deep neural networks," *CoRR*, vol. abs/1706.05098, 2017. [Online]. Available: http://arxiv.org/abs/1706.05098

[96] A. R. Zamir, A. Sax, W. B. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proc. IJCAI*, S. Kraus, Ed. ijcai.org, 2019, pp. 6241–6245.

[97] T. Standley, A. Zamir, D. Chen, L. J. Guibas, J. Malik, and S. Savarese, "Which tasks should be learned together in multi-task learning?" in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 9120–9132. [Online]. Available: http://proceedings.mlr.press/v119/standley20a.html

[98] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html

[99] S. Wu, H. R. Zhang, and C. Ré, "Understanding and improving information transfer in multi-task learning," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SylzhkBtDB

[100] P. Bevandić, M. Oršić, J. Šarić, I. Grubišić, and S. Šegvić, "Weakly supervised training of universal visual concepts for multi-domain semantic segmentation," *International Journal of Computer Vision*, Jan 2024. [Online]. Available: https://doi.org/10.1007/s11263-024-01986-z

[101] S. R. Bulò, L. Porzi, and P. Kontschieder, "In-place activated batchnorm for memory-optimized training of dnns," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 5639–5647.

[102] P. Bevandic, M. Orsic, I. Grubisic, J. Saric, and S. Segvic, "Multi-domain semantic segmentation with overlapping labels *," in *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*. IEEE, 2022, pp. 2422–2431. [Online]. Available: https://doi.org/10.1109/WACV51458.2022.00248

[103] Y. Lee, J. Jeon, J. Yu, and M. Jeon, "Context-aware multi-task learning for traffic scene recognition in autonomous vehicles," in *IEEE Intelligent Vehicles Symposium, IV 2020, Las Vegas, NV, USA, October 19 - November 13, 2020*. IEEE, 2020, pp. 723–730.

[104] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 1871–1880. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Liu_End-To-End_Multi-Task_Learning_With_Attention_CVPR_2019_paper.html

[105] I. Kreso, D. Causevic, J. Krapac, and S. Segvic, "Convolutional scale invariance for semantic segmentation," in *Proc. GCPR*, B. Rosenhahn and B. Andres, Eds., 2016.

[106] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 5375–5384.

[107] Y. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 7029–7039.

[108] Y. Cui, M. Jia, T. Lin, Y. Song, and S. J. Belongie, "Class-balanced loss based on effective number of samples," *CoRR*, vol. abs/1901.05555, 2019.

[109] S. Segvic, K. Brkic, Z. Kalafatic, and A. Pinz, "Exploiting temporal and spatial constraints in traffic sign detection from a moving vehicle," *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 649–665, 2014.

[110] V. Zadrija, J. Krapac, S. Segvic, and J. Verbeek, "Sparse weakly supervised models for object localization in road environment," *Comput. Vis. Image Underst.*, vol. 176-177, pp. 9–21, 2018.

[111] D. Tabernik and D. Skocaj, "Deep learning for large-scale traffic-sign detection and recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1427–1440, 2020.

[112] Q. Zou, H. Jiang, Q. Dai, Y. Yue, L. Chen, and Q. Wang, "Robust lane detection from continuous driving scenes using deep neural networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 41–54, 2020.

[113] J. C. McCall and M. M. Trivedi, "Video-based lane estimation and tracking for driver assistance: survey, system, and evaluation," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 20–37, 2006.

[114] I. Krešo, J. Krapac, and S. Segvic, "Efficient ladder-style densenets for semantic segmentation of large images," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–11, 2020.

[115] H. Yi, C. Bizon, D. Borland, M. Watson, M. Satusky, R. Rittmuller, R. Radwan, R. Srinivasan, and A. Krishnamurthy, "Ai tool with active learning for detection of rural roadside safety features," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 5317–5326.

[116] T. G. P. Sanjeewani and B. K. Verma, "Learning and analysis of ausrap attributes from digital video recording for road safety," in *Proc. IVCNZ*. IEEE, 2019, pp. 1–6.

[117] P. Sanjeewani and B. Verma, "An optimisation technique for the detection of safety attributes using roadside video data," in *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2020, pp. 1–6.

[118] Z. Jan, B. Verma, J. Affum, S. Atabak, and L. Moir, "A convolutional neural network based deep learning technique for identifying road attributes," in *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2018, pp. 1–6.

[119] W. Song, S. Workman, A. Hadzic, X. Zhang, E. Green, M. Chen, R. R. Souleyrette, and N. Jacobs, "FARSA: fully automated roadway safety assessment," *CoRR*, vol. abs/1901.06013, 2019.

[120] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[121] J. Y. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 4694–4702.

[122] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2599174

[123] Y. Tu, J. Du, L. Sun, and C. Lee, "Lstm-based iterative mask estimation and post-processing for multi-channel speech enhancement," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017, Kuala Lumpur, Malaysia, December 12-15, 2017*. IEEE, 2017, pp. 488–491.

[124] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, "Rainfall–runoff modelling using long short-term memory (lstm) networks," *Hydrology and Earth System Sciences*, vol. 22, no. 11, pp. 6005–6022, 2018.

[125] H. Xu, A. Das, and K. Saenko, "R-C3D: region convolutional 3d network for temporal activity detection," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 5794–5803.

[126] Y. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 1130–1139.

[127] R. Trabelsi, R. Khemmar, B. Decoux, J.-Y. Ertaud, and R. Boutteau, "Staf: Spatio-temporal attention framework for understanding road agents behaviors," *IEEE Access*, vol. 10, pp. 55 794–55 804, 2022.

[128] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*

2017, December 4-9, 2017, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[129] M. Orsic and S. Segvic, "Efficient semantic segmentation with pyramidal fusion," *Pattern Recognit.*, vol. 110, p. 107611, 2021.

[130] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, "Optimal thresholding of classifiers to maximize F1 measure," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II*, ser. Lecture Notes in Computer Science, T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, Eds., vol. 8725. Springer, 2014, pp. 225–239.

[131] J. Opitz and S. Burst, "Macro F1 and macro F1," *CoRR*, vol. abs/1911.03347, 2019.

[132] M. Zhu, "Recall, precision and average precision," *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, vol. 2, no. 30, p. 6, 2004.

[133] J. Ni, K. Shen, Y. Chen, W. Cao, and S. X. Yang, "An improved deep network-based scene classification method for self-driving cars," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022. [Online]. Available: https://doi.org/10.1109/TIM.2022.3146923

[134] M. Saffari, M. Khodayar, and S. M. J. Jalali, "Sparse adversarial unsupervised domain adaptation with deep dictionary learning for traffic scene classification," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 7, no. 4, pp. 1139–1150, 2023. [Online]. Available: https://doi.org/10.1109/TETCI.2023.3234548

Biography

Marin Kačan was born in 1995 in Rijeka. He earned his BSc and Msc degrees from UniZg-FER, Faculty of Electrical Engineering and Computing in 2016 and 2019, respectively. During his studies, he interned as a software engineer at Infobip and Amazon, and as a research engineer at Microblink and the TakeLab laboratory at UniZg-FER.

Upon defending his master's thesis in 2019, he was employed at the Faculty of Transport and Traffic Sciences, University of Zagreb, as an expert associate on the project SLAIN: Saving Lives Assessing and Improving TEN-T Road Network Safety. In 2020, he started his PhD at the Faculty of Electrical Engineering and Computing on the topic of automatic road infrastructure safety assessment where he soon gets employed as a research associate. From 2021 to 2023, he participated in the research project DATACROSS. From 2023, he has been involved in two research projects - VoNoMobil: Research, development and production of new mobility vehicles and supporting infrastructure; and Google.org AI for Global goals: iRAP Star Rating for Schools in Vietnam. He also participates in the peer review of the scientific literature as a reviewer for international conferences and scientific journals.

His research interests include multi-task visual recognition of road safety infrastructure, efficient video classification in the wild, long-tail learning and satellite image classification.

List of publications

Journal papers

1. **Kačan M., Ševrović M., Šegvić S., "Dynamic Loss Balancing and Sequential Enhancement for Road-Safety Assessment and Traffic Scene Classification", IEEE Transactions on Intelligent Transportation Systems, Vol. 25, No. 11, November 2024, pp. 15628–40**
2. Kačan, M., Turčinović, F., Bojanjac, D., Bosiljevac, M., "Deep learning approach for object classification on raw and reconstructed GBSAR data", Remote sensing, Vol. 14, No. 22., November 2022, p.5673
3. Turčinović, F., Kačan, M., Bojanjac, D., Bosiljevac, M., Šipuš, Z., "Utilizing Polarization Diversity in GBSAR Data-Based Object Classification", Sensors, Vol. 24, No. 7, April

2024, p.2305

4. Turčinović, F., Kačan, M., Bojanjac, D., Bosiljevac, M., "Near-distance raw and reconstructed ground based SAR data", Data in brief, Vol. 51, December 2023, p.109620

Conference papers

1. **Kačan, M., Oršić, M., Šegvić, S., Ševrović, M., "Multi-task learning for iRAP attribute classification and road safety assessment", in 2020 IEEE 23rd International Conference on Intelligent Transportation Systems, September 2020, pp. 1-6**
2. Turčinović, F., Kačan, M., Bojanjac, D., Bosiljevac, M., "Deep learning approach based on GBSAR data for detection of defects in packed objects", in 2023 17th European Conference on Antennas and Propagation, March 2023, pp. 1-4
3. Turčinović, F., Kačan, M., Bojanjac, D., Bosiljevac, M., "Ground based SAR system for object classification with parameter optimization based on deep learning feedback algorithm", in Proceedings of Image and Signal Processing for Remote Sensing XXIX, Vol. 12733, pp. 244-250
4. Turčinović, F., Kačan, M., Bojanjac, D., Bosiljevac, M., "Impact of Ground Based SAR Parameters on Radar Data Based Object Classification", in 2023 24th International Conference on Applied Electromagnetics and Communications, September 2023, pp. 1-5

Životopis

Marin Kačan rođen je 1995. godine u Rijeci. Na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu 2016. godine završio je preddiplomski, a 2019. godine diplomski studij računarske znanosti. Tijekom studija radio je kao pripravnik na poziciji softverskog inženjera u tvrtkama Infobip i Amazon, te na poziciji istraživačkog inženjera u Microblinku i laboratoriju TakeLab Fakulteta elektrotehnike i računarsta Sveučilišta u Zagrebu.

Nakon završetka diplomskog studija 2019. godine, zaposlio se na Fakultetu prometnih znanosti Sveučilišta u Zagrebu kao stručni suradnik na projektu SLAIN: Saving Lives Assessing and Improving TEN-T Road Network Safety. Godine 2020. upisao je doktorski studij na Fakultetu elektrotehnike i računarstva s temom automatske procjene sigurnosti cestovne infrastrukture gdje se nakon godinu dana i zapošljava. Od 2021. do 2023. bio je zaposlen na Fakultetu elektrotehnike i računarstva, na istraživačkom projektu DATACROSS. Od 2023. godine uključen je u dva istraživačka projekta - VoNoMobil: Istraživanje, razvoj i proizvodnja vozila nove mobilnosti i prateće infrastrukture; te Google.org AI for Global goals: iRAP Star Rating for Schools in Vietnam. Uključen je u istorazinsku procjenu znanstvene literature kao recenzent za međunarodne konferencije i znanstvene časopise.

Njegovi istraživački interesi uključuju višezadaćno vizualno raspoznavanje obilježja sigurnosti cestovne infrastrukture, učinkovitu klasifikaciju videa u stvarnim uvjetima, učenje na velikim i neuravnoteženim taksonomijama i klasifikaciju satelitskih snimaka.