ŽILINSKÁ UNIVERZITA V ŽILINE

AUTOREFERÁT
DIZERTAČNEJ PRÁCE

Žilina, apríl 2022

Ing. Milan Ondrašovič

Žilinská univerzita v Žiline
Fakulta riadenia a informatiky

Ing. Milan Ondrašovič

Autoreferát dizertačnej práce

VIZUÁLNE TRASOVANIE OBJEKTOV POUŽITÍM SIAMSKÝCH NEURÓNOVÝCH SIETÍ

na získanie akademického titulu "**philosophiae doctor**" (**PhD.**)
v štúdijnom programe doktorandského štúdia
aplikovaná informatika
v štúdijnom odbore
informatika

Žilina, apríl 2022

Dizertačná práca bola vypracovaná v dennej forme doktorandského štúdia na Katedre matematických metód a operačnej analýzy, Fakulte riadenia a informatiky Žilinskej univerzity v Žiline.

Predkladateľ: *Ing.* ***Milan Ondrašovič***
Katedra matematických metód a operačnej analýzy
Fakulta riadenia a informatiky
Žilinská univerzita v Žiline

Školiteľ: *doc. Mgr.* ***Ondrej Šuch****, PhD.*
Matematický ústav Slovenskej akadémie vied
Banská Bystrica

Školiteľ špecialista: *Ing.* ***Peter Tarábek****, PhD.*
Katedra matematických metód a operačnej analýzy
Fakulta riadenia a informatiky
Žilinská univerzita v Žiline

Oponenti: *prof. Ing.* ***Gregor Rozinaj****, PhD.*
Fakulta elektrotechniky a informatiky
Slovenská technická univerzita v Bratislave

Dr. MSc. ***Lucas Alexandre Ramos***
Department of Computer Vision and Data Science
NHL Stenden - University of Applied Sciences
Leeuwarden - The Netherlands

Autoreferát bol rozoslaný dňa: **.....**
Obhajoba dizertačnej práce sa koná dňa **.....** o **.....** hod. pred komisiou pre obhajobu dizertačnej práce schválenou odborovou komisiou v študijnom odbore **informatika**, v študijnom programe **aplikovaná informatika**, vymenovanou dekanom Fakulty riadenia a informatiky Žilinskej univerzity
v Žiline dňa **.....**

prof. Ing. ***Karol Matiaško****, PhD.*
predseda odborovej komisie
študijného programu **aplikovaná informatika**
v študijnom odbore **informatika**

Fakulta riadenia a informatiky
Žilinská univerzita
Univerzitná 8215/1
010 26 Žilina

Anotácia

Predmetom skumania tejto práce bolo vizuálne trasovanie objektov použitím hlbokého strojového učenia. Cieľom bolo analyzovať dopravné situácie so zameraním na trasovanie vozidiel. Jeden z mnohých problémov, ktoré sú pre túto úlohu typické, je prítomnosť prekrývania sa objektov. Vzhľadom na toto, našou úlohou bolo navrhnúť podporné riešenie na situácie, kedy dochádza k čiastočnému alebo úplnému prekrytiu trasovaného objektu. Naše experimenty boli založené na siamských neurónových sieťach v kombinácií s latentnými priestormi a mechanizmom pozornosti, ktorých inherentné vlastnosti sme využili pre zvýšenie presnosti trasovania za istých podmienok. Dôležitosť siamského prístupu indikuje aj naša rešeršná publikácia so zameraním na túto paradigmu trasovania. Experimenty využívali existujúcu architektúru v rámci ktorej bol identifikovaný priestor na zlepšenie. Táto práca taktiež zahŕňa príspevok v úlohe transformovania obrazu pomocou homografie.

Kľúčové slová: vizuálne trasovanie objektov, hlboké učenie, Siamské neurónové siete, latentné priestory, mechanizmus pozornosti, homografia, analýza dopravy.

Počet strán:	*140*	*Počet použitej literatúry:*	*129*
Počet obrázkov:	*60*	*Počet tabuliek:*	*7*

Annotation

The subject of this study was visual object tracking using deep machine learning. The goal was to analyze traffic situations while focusing on vehicle tracking. One of many typical problems for this task is the presence of object occlusion of varying intensity. In light of this, our objective was to develop a supporting solution for situations in which partial or complete occlusion of the tracked object occurs. We based our experiments on Siamese neural networks in conjunction with latent spaces and an attention mechanism, the inherent properties of which were exploited to increase tracking accuracy under specific conditions. The importance of the Siamese approach is indicated by our survey publication aimed at this tracking paradigm. The experiments exploited an already existing architecture within which we had identified room for improvements. This work also discusses our contribution to the task of image transformation using homography.

Key words: visual object tracking, deep learning, Siamese neural networks, latent spaces, attention mechanism, homography, traffic analysis.

Number of pages:	*140*	*Number of references:*	*129*
Number of figures:	*60*	*Number of tables:*	*7*

1 Úvod

Vizuálne trasovanie objektov, *Visual Object Tracking* (VOT), je jednou zo základných výziev v oblasti počítačového videnia. Cieľom je lokalizovať konkrétny objekt naprieč všetkými snímkami videa za predpokladu, že k dispozícií je len pozícia objektu záujmu v obraze na prvom snímku. Objekt je najskôr detegovaný v snímku a je mu priradený unikátny identifikátor. Následne je snaha korektne priradiť zavedený identifikátor aj v budúcich snímkach.

Trasovanie objektov, či už jedného, *Single-Object Tracking* (SOT), alebo viacerích, *Multi-Object Tracking* (MOT), je často len prostriedkom na dosiahnutie vyšších cieľov v zmysle aplikácií. V tejto práci sa trasovanie vozidiel považuje za primárny praktický prípad použitia nášho aplikovaného výskumu a to aj napriek tomu, že nami navrhnuté metódy sú všeobecne použiteľné pre trasovanie bez ohľadu na doménu.

Najúspešnejšie prístupy k VOT sú založené na hlbokom strojovom učení. Konvolučná neurónová sieť, *Convolutional Neural Network* (CNN), ktorá je považovaná za efektívny a robustný nástroj na extrakciu príznakov z obrazu [5], tvorí základný stavebný pilier nielen tejto oblasti ale aj nášho výskumu.

Vo všeobecnosti je tento problém stále otvorený a najväčšie problémy sú zapríčinené zmenami v osvetlení snímanej scény, polohy či orientácie trasovaného objektu, prípadne pri zmene pohľadu kamery a v neposlednom rade pri čiastočnej alebo úplnej oklúzií (prekrytí). S fenoménom oklúzie sme sa v našom výskume primárne zaoberali kvôli jeho potenciálu veľmi negatívne ovplyvňovať trasovanie, keďže sledovaný objekt sa môže opätovne objaviť po úplnom alebo aspoň veľmi závažnom prekrytí v podstatne inej podobe, a teda môže byť pomýlený za objekt iný.

Kľúčovým elementom, na ktorý je potrebné pri prítomnosti oklúzie klásť dôraz, je správne rozlíšovanie medzi novými a už v histórií trasovanými objektami. Za týmto účelom slúžia latentné resp. vnorené priestory [3] vykazujúce potenciál pre takéto použitie aspoň v niektorých situáciách, ktoré sú generované pomocou podobnostného učenia. Okrem iného, v prípade čiastočnej oklúzie je prvoradej dôležitosti korektne vytýčiť región, v ktorom je trasovaný objekt viditeľný, čo nás motivovalo zvážiť použitie mechanizmu pozornosti [18]. Naviac, v prípade trasovania vozidiel dochádza k vyššej miere ambiguity ako pri ľuďoch, keďže autá vykazujú väčšie množstvo podobných črtov ako ľudia. Na riešenie takýchto situácií je použitie mechanizmu pozornosti opodstatnené kvôli jeho inherentným vlastnostiam.

Vzhľadom na vyššie spomenuté aspekty, v rámci tejto dizertačnej práce bolo naším **cieľom aplikovať prístupy hlbokého strojového učenia využívajúce podobnostné učenie alebo mechanizmy pozornosti so snahou zlepšiť trasovanie objektov v situáciách zaťažených oklúziou objektov, zmenami v pozícií objektu a pohľade naň, ako aj flukuáciami v iluminácií scény.**

2 Teoretické východiská a súčasný stav

2.1 Latentné priestory

Cieľom vytvárania latentných (vnorených) priestorov je naučiť sa funkciu $f_\theta(x) : \mathbb{R}^F \to \mathbb{R}^D$, ktorá mapuje sémanticky podobné body z priestoru $\mathbb{R}^F$ na metricky blízke body v priestore $\mathbb{R}^D$. Analogicky, $f_\theta(\cdot)$ by mala mapovať sémanticky rozdielne body v priestore $\mathbb{R}^F$ na metricky vzdialené body v priestore $\mathbb{R}^D$ [4].

Uvažujme úlohu vizuálnej re-identifikácie, *Re-identification* (ReID), vozidiel. V takomto prípade je korešpondujúci vnorený vektor pre každé vozidlo vytvorený funkciou, ktorá mapuje obrázky vozidiel do vnoreného priestoru tak, aby zábery identického vozidla boli mapované blízko seba narozdiel od záberov iných vozidiel. Ideálne by takéto mapovanie malo byť invariantné voči zmenám v pohľade na dané vozidlo ako aj v osvetlení scény. Existuje mnoho spôsobov akými kvantifikovať mieru podobnosti, napríklad Euklidovská vzdialenosť alebo kosínusová podobnosť.

2.1.1 Dvojicová (kontrastívna) stratová funkcia

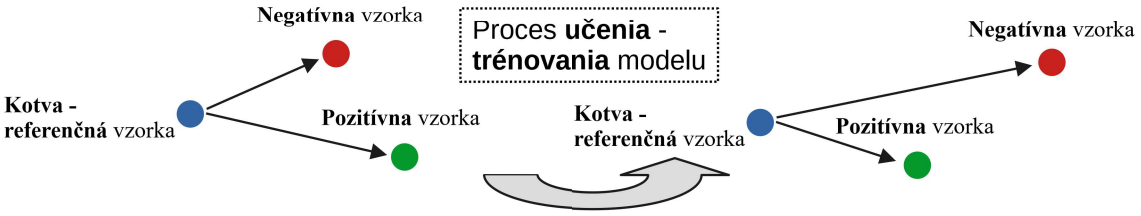
Uvažujme trojicu (x_0, x_1, y), kde x_0 a x_1 reprezentujú vstup, pričom ich kategória $y = 1$ ak x_0 a zároveň x_1 patria do rovnakej triedy, inak $y = 0$. Nech $D(x,y) : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ je metrická funkcia, ktorá meria vzdialenosť vo vnorenom priestore, napríklad L_2 norma, čiže $D(x,y) = \|x-y\|_2$. Nech α je hraničná hodnota reprezentujúca minimálnu vzdialenosť v metrickom priestore, ktorá oddeľuje pozitívne a negatívne vzorky. Kontrastívna stratová funkcia je potom definovaná ako [3]

$$\mathscr{L}_{contr}(\theta) = \frac{1}{2} y D\left(f_\theta(x_0), f_\theta(x_1)\right)^2 + \frac{1}{2}(1-y)\left([\alpha - D\left(f_\theta(x_0), f_\theta(x_1)\right)]_+\right)^2. \quad (1)$$

2.1.2 Trojicová stratová funkcia

V prípade trojicovej stratovej funkcie sú potrebné až tri vzorky, nie len dve. Dôvodom je snaha poskytnúť dodatočný kontext pri formovaní vnoreného priestoru. Prístup využívajúci trojice poskytuje bohatšie informácie o vlastnostiach medzi prvkami, pretože sa počíta s tzv. kotvovou (angl. *anchor*), pozitívnou a negatívnou vzorkou. Tým pádom pozitívna vzorka musí byť mapovaná bližšie ako negatívna.

Nech N je počet všetkých možných validných trojíc $\left(x_a^i, x_p^i, x_n^i\right)$ pre nejaký dataset. Pre každú i-tu trojicu, nech x_a^i je "kotva" pre daný objekt (osoba, vozidlo, atď.) s kategóriou $y\left(x_a^i\right)$; x_p^i je pozitívna vzorka s kategóriou $y\left(x_p^i\right)$, taká že $x_a^i \neq x_p^i \wedge y\left(x_a^i\right) = y\left(x_p^i\right)$; x_n^i s kategóriou $y\left(x_n^i\right)$ je vzorka akéhokoľvek iného objektu spĺňajúc podmienku $y\left(x_a^i\right) \neq y\left(x_n^i\right)$, $\forall i = 1, \ldots, N$. Nech α je hraničná hodnota vzdialenosti medzi pozitívnymi a negatívnymi vozrkami. Na základe tohto môžeme



Obr. 1: Cieľom trojicovej stratovej funkcie je vytvoriť vnorený priestor taký, v ktorom je pozitívna vzorka mapovaná bližšie k ukotvujúcej vzorke a zároveň negatívna vzorka je mapovaná minimálne v parametricky určenej hraničnej vzdialenosti.

trojicovú stratovú funkciu formulovať ako

$$\mathscr{L}_{triplet}(\theta) = \sum_{i=1}^{N} \left[\alpha + D\left(f_\theta\left(x_a^i\right), f_\theta\left(x_p^i\right)\right) - D\left(f_\theta\left(x_a^i\right), f_\theta\left(x_n^i\right)\right)\right]_+, \tag{2}$$

ktorej očakávaný účinok v rámci procesu učenia je uvedený na Obr. 1.

2.2 Siamské vizuálne trasovanie jedného objektu

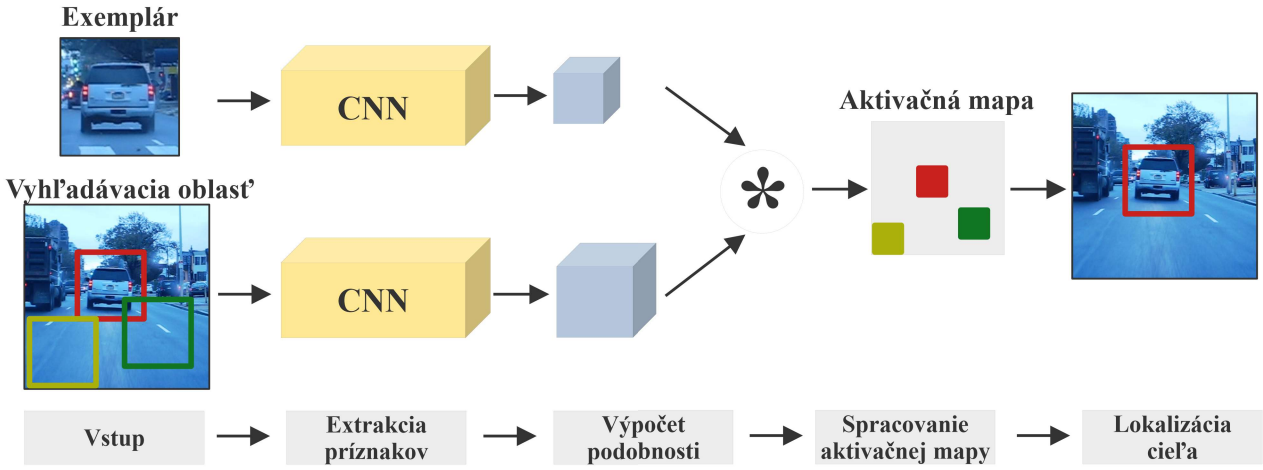
2.2.1 Motivácia

CNN slúži na robustnú extrakciu príznakov z obrazu, preto prínos konvolučných sietí je pre vizuálne trasovanie nepostrádateľný. Avšak, trénovať takéto modely na všeobecnejšiu úlohu podobnostného učenia ako len na samotnú klasifikáciu, ktorá je bežným prípadom použitia pre CNN, je pre trasovanie kľúčové. Tieto myšlienky viedli k vytvoreniu plne konvolučnej siamskej neurónovej siete určenej na VOT, známej ako *Siamese Fully Convolutional Network* (SiamFC) [1]. Jej základným princípom je pracovať s dvojicou snímok. Prvá reprezentuje exemplár, napríklad záznam prvého výskytu objektu záujmu, a druhá reprezentuje vyhľadávaciu oblasť v rámci ktorej je snaha exemplár lokalizovať (Obr. 2). Tento prístup sa stal základom pre mnoho ďalších prác, ktoré uviedli vetvu Siamského trasovania do popredia, napríklad [6, 19]. Táto oblasť tvorila aj základ nášho výskumu.

2.2.2 Základné princípy

Nech γ je transformácia, ktorá extrahuje príznaky zo vstupu, a g je funkcia, ktorá kombinuje dve reprezentácie príznakov vygenerované funkciou γ. Siamské siete aplikujú identickú transformáciu γ na obidva vstupy, vyhľadávací obrázok x a exemplár z, a následne ich kombinujú ako

$$f(x,z) = g\left(\gamma(x), \gamma(z)\right). \tag{3}$$



Obr. 2: Ukážka trasovania pomocou plne konvolučnej siamskej architektúry. Takýto trasovač produkuje aktivačnú mapu, ktorá odzrkadľuje mieru podobnosti vyhľadávaného exempláru (stav z histórie, napríklad z prvého snímku od ktorého trasovanie začalo) v jednotlivých oblastiach v rámci vyhľadávacieho (aktuálneho) snímku. Takýto výpočet je efektívny nakoľko je založený na operácií štandardnej 2D konvolúcie v obraze. Červené, zelené a žlté pixely prislúchajú korešpondujúcim oblastiam na vstupe. Na základe maximálnej odozvy je vykonaná predikcia aktuálnej pozície trasovanéo objektu.

Ak funkcia g používa Euklidovskú vzdialenosť alebo kosínusovú podobnosť, potom hovoríme o spomínaných vnorených priestoroch. V siamskom trasovaní je funkcia g založená na aplikácií krížovej korelácie, výstupom ktorej je aktivačná mapa (Obr. 2).

2.2.3 Závery z nášho prehľadového článku

Táto oblasť trasovačov nás zaujala natoľko, že sme sa rozhodli napísať prehľadový článok, neskôr publikovaný v žurnále [11], ktorý pokrýval moderné prístupy k siamskému trasovaniu a detailnejšie rozoberal výhody a nevýhody týchto architektúr. Identifikovali sme, že od roku 2017 [12] nebola publikovaná žiadna iná rešeršná práca, ktorá by sa explicitne zaoberala touto vetvou prístupov k VOT. V našej publikácií sme sa snažili elaborovať do detailov akým výzvam siamské trasovanie čelí. Ide o špecializovanú hĺbkovú rozpravu o siamskom trasovaní jedného objektu.

Vo všeobecnosti môžeme tvrdiť, že siamské trasovače vykazujú najlepší pomer medzi rýchlosťou a presnosťou. Výsoká rýchlosť je ich inherentná vlastnosť a preto sa hodia pre spracovanie obrazu v reálnom čase. Z hľadiska implementácie a samotného tréningu patria tieto modely medzi jednoduchšie. Ich architektúry sú často tzv. *end-to-end*. Takýto jednotný dizajn je menej náročný na použitie a otvára možnosti pri mnohé inkrementálne zlepšenia, ktoré sú v tejto oblasti všadeprítomné.

Avšak, napriek výhodám, stále sú tu fundamentálne prekážky a tou najzávažnejšou je prítomnosť sémantického pozadia resp. podobnostnej interferencie. Laicky

povedané, jedná sa o objekty inej identity ktoré sú vizuálne veľmi podobné, tzv. distraktory. Naša analýza ukázala, že trasovače, ktoré boli navrhované s cieľom adresovať konsekvencie z prítomnosti týchto "rušivých" objektov vykazovali najlepšie výsledky. Dôvod je ten, že výpočet podobnostnej aktivačnej mapy nie je vo svojej pôvodnej formulácií robustný voči iným vizuálne podobným objektom. Niektoré z riešení, aj keď nie definitívnych, sú implementácia špeciálneho modulu, ktorý si uvedomuje prítomnosť distraktorov [23], využitie stratégie pre rozlišovanie medzi popredím a pozadím [7], prípadne zahrnutie podmienenej re-detekcie objektu [8].

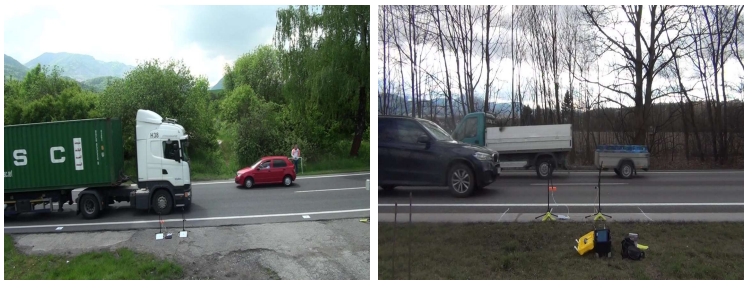
Z hľadiska architektúry sa ukázalo, že použitie mechanizmu pre generovanie oblasti návrhov, v ktorých sa môže potenciálne objekt nachádzať, známeho ako *Region Proposal Network* (RPN) [13], prináša konzistentne výborné výsledky, v dôsledku čoho existujú mnohé úspešné varianty ktorých RPN je súčasťou, napríklad [6, 23].

Paradoxne problematickým aspektom siamského trasovania je použitie krížovej korelácie, ktorá tvorí stavebný kameň. Vo svojej podstate sa jedná o vyhodnocovacie zhody nejakej šablóny objektu, ktorú reprezentuje exemplár, voči aktuálnemu snímku, teda vyhľadávacej oblasti. Otázkou však zostáva, ako často túto šablónu aktualizovať. Príliš časté zmeny môžu spôsobiť, že nepresné zábery objektu sa stanu šablónou. Naopak, raritné alebo žiadne zmeny (prevládajúci prístup) exemplára nereflektujú dynamickú povahu trasovania. Za týmto účelom bolo navrhnutých niekoľko stratégií [9]. Každopádne, samotná operácia sa dá formulovať aj inak a v dnešnej dobe prevažuje použitie hĺbkovej krížovej korelácie, ktorej výstupom je aktivačná s viacerými kanálmi, a teda bohatšími informáciami. Táto operácia sa stala terčom aj pre nasadenie rôznych mechanizmov pozornosti [18] so snahou zlepšiť schopnosť siete vyberať relevantné a zároveň potláčať nepodstatné príznaky. Ako už bolo avizované, mechanizmus pozornosti sme aplikovali v našom výskume aj my.

2.3 Siamské vizuálne trasovanie viacerých objektov

Náš výskum bol pôvodne zameraný na SOT, lenže kvôli zamýšľaným dopravným aplikáciám sme náš obzor rozšírili do oblasti MOT. Ako sa ukázalo, siamské trasovanie úspešne pretravilo svoje prínosy aj do tejto rozšírenej vetvy trasovania.

V práci [16] bol navrhnutý siamský prístup, ktorý simultánne adresoval trasovanie a detekciu objektov, dokonca aj ReID. Unifkácia týchto aspektov do jedného modelu je veľkou výhodou. Naviac, všeobecná formulácia tohto prístupu dovoľuje využitie akéhokoľvek siamského trasovača. Ďalší triviálny, efektívny a zároveň intuitívny prístup bol publikovaný v [17], v ktorom SiamFC trasovač využíval nie jeden ale n exemplárov, generoval n aktivačných máp, a teda bol schopný trasovať n objektov simultánne. Motivácia bola vyhnúť sa výpočtovo náročnému detektoru objektov. Naša experimentálna činnosť v rámci siamského MOT bola založená na práci *Siamese Multi-Object Tracker* (SiamMOT) [15], ktorú neskôr bližšie popíšeme.



(a) **(b)**

Obr. 3: Ukážka záznamov premávky s akými sme pracovali na projekte *Interreg SK-CZ*.

3 Vyhodnocovanie kvality reprojekcie homografie

Naša práca v počiatočných fázach zahŕňala participáciu na projekte *Interreg SK-CZ*. Úlohou bolo poskytovať validačný mechanizmus na báze softvéru pre hardvérový magnetometer s účelom merania rýchlosti vozidiel a určovania ich početnosti v bežnej cestnej premávke. V takýchto situáciách sú z pohľadu spracovania obrazu prístupy fotogrametrie kľúčové a mnohé z nich zahrňajú transformáciu pomocou tzv. homografie. V tejto sekcii popíšeme náš originálny vedecký príspevok publikovaný v žurnále [10], ktorý sme pôvodne zamýšľali využiť aj pre trasovanie objektov, avšak z dôvodu nedostatočného množstva dostupných dátových množín sme kombináciu VOT a homografie nemohli dotiahnúť ďalej.

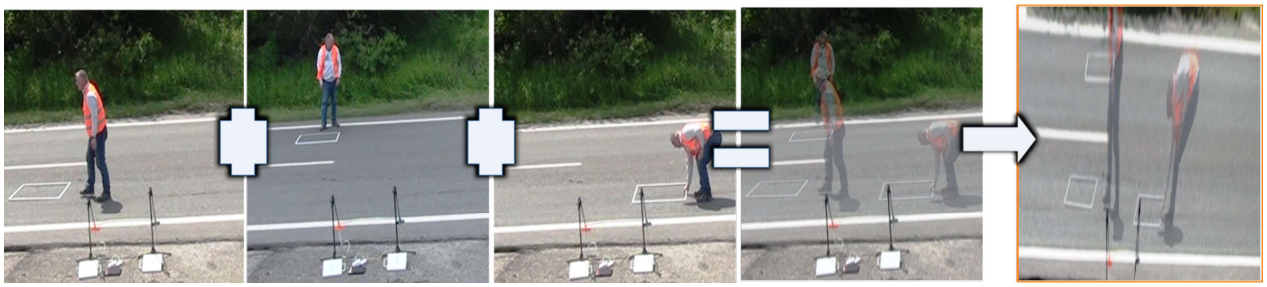
3.1 Motivácia a všeobecný popis prípadu použitia

Počítačové videnie sa často musí vysporiadať s rôznymi transformáciami obrazu pred ďalším spracovaním. Jedna z transformácií, ktorej dopad môže byť veľmi závažný, je perspektívna deformácia. Naším cieľom sa stalo perspektívnu deformáciu z obrazu odstrániť, nakoľko videozáznamy vozovky boli zachytené z bočného pohľadu (Obr. 3). My sme požadovali pohľad taký, ako keby bola kamera umiestnená kolmo na vozovku (Obr. 4). Po takejto transformácií je možné presne merať o akú vzdialenosť sa objekt pohol, okrem iného. Táto úloha je prakticky vyriešená a spomínaná homografia je jeden z nástrojov ako takýto pohľad synteticky vyprodukovať. Avšak, objavili sme prípad použitia, kedy štandardné metódy postačujúce neboli.

Homografia definuje vzťah medzi rôznymi pohľadmi kamery (perspektívnu projekciu) na jednu 2D rovinu pomocou ktorej je možné dosiahnuť zmenu pohľadu kamery na túto rovinu. Homografia je 3×3 matica s 8 stupňami voľnosti

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}. \tag{4}$$

Obr. 4: Demonštrácia syntetickej transformácie pohľadu na vozovku pomocou homografie.



Obr. 5: Demonštrácia pokladania *markeru* do rôznych častí vozovky po nastavení kamery do fixnej polohy s cieľom získať čo najviac kvalitných záznamov. Obrázok pozostávajúci z niekoľkých transparentných vrstiev znázorňuje ako predpoklad statickej kamery umožňuje umelo vytvoriť scénu s použitím viacerých totožných *markerov*. Táto vlastnosť bola dôležitá pre našu metódu kvantifikovania kvality reprojekcie homografií, ktorá predpokladá, že homografia je tým lepšia, čím viacej *markerov* po odsránenní perspektívnej deformácie dokáže transformovať čo najbližšie k ich ideálnemu tvaru (pretože to sú jediné známe objekty).

Vektor $\mathbf{u}^T = [u_x, u_y, 1]$, perspektívne deformovaný bod špecifikovaný v homogénnych súradniciach je mapovaný na svoj rektifikovaný obraz, vektor $\tilde{\mathbf{u}}^T = [\tilde{u}_x, \tilde{u}_y, 1]$, využitím homografie $\mathbf{H}$ pomocou transformácie $s\tilde{\mathbf{u}} \approx \mathbf{Hu}$, kde s je škálovací faktor.

Bežný prístup k odhadovaniu matice homografie je založený na použití aspoň štyroch korešpondencií 2D nezávislých bodov. Ak obraz pre nás známeho útvaru umiestneného na nejakej rovine podlieha perspektívnej deformácií, potom jeho tvar je skreslený vďaka perspektívnej projekcií. Avšak, pri znalosti skutočného tvaru takéhoto objektu je možné zostaviť systém rovníc definujúcich vzťah medzi deformovanými a rektifikovanými bodmi, ktorého riešením je homografia. Po aplikácií získanej transformácie na celý obraz dosiahneme zmenu pohľadu na snímanú rovinu, pričom body mimo túto rovinu budú mapované nesprávne (skreslene, deformovane), čo ale nie je nevýhodou ale len vlastnosťou toho, že dochádza k mapovaniu z 2D do 2D. Objekt, ktorého rozmery sú známe a na scéne sa nachádza buď prirodzene alebo kvôli zámernému umiestneniu, sa nazýva angl. *marker*. Často sa využíva šachovnica alebo iný, ľahko detegovateľný objekt s vhodnými geometrickými vlastnosťami.

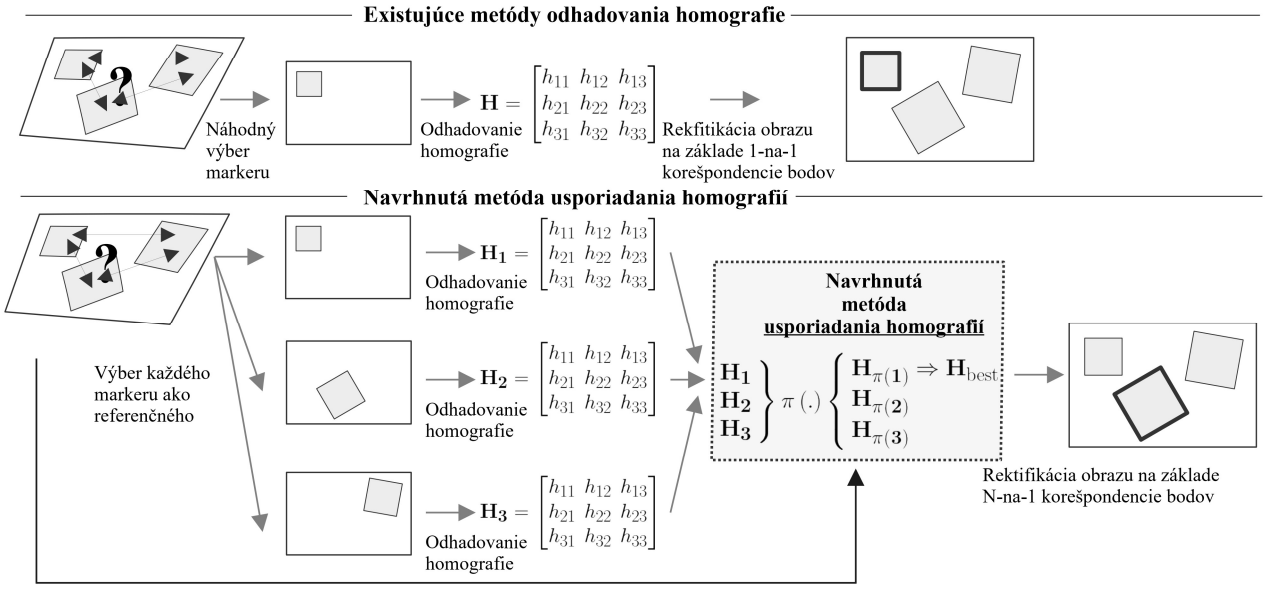
Pre získanie vtáčej perspektívy nad vozovkou sme použili štvorcový *marker* veľ-

kosti 1m $\times$ 1m, ktorý bol umiestnený a snímaný statickou kamerou v rôznych častiach vozovky (roviny, nad ktorou bola snaha získať pohľad z vtáčej perspektívy). Následne sme subjektívne vybrali taký záznam, kde *marker* bol viditeľný najlepšie (Obr. 5). Na základe určenia bodovej korešpondencie sme odhadli homografiu štandardnými metódami a obraz rektifikovali.

Problém nastával v tom, že nie všetky homografie vykazovali rovnakú kvalitu reprojekcie. Našou ideou bolo tieto *markery* spojiť pre vyššiu robustnosť, čo však možné nie je z dôvodu absencie informácie o ich pozícií a to aj napriek tomu, že boli umiestnené na identickej rovine, čo je nutný predpoklad vo všeobecnosti pre prácu s homografiou. Takéto obmedzenie a jeho následné čiastočné obídenie sa stalo prínosom našej metódy. Pre objasnenie uvažujme situáciu, že na scéne je *markerov* prítomných viacej, pričom ich pozície známe nie sú. V takomto prípade nie je možné určiť korešpondenciu medzi bodmi v snímanom a žiadúcom obraze globálne pre všetky *markery*, nakoľko pozície medzi týmito bodmi musia byť dodržané proporčne. Napríklad, ak máme na scéne obdĺžnik reálne veľký 20cm $\times$ 30cm, v obraze jeho pomer strán v pixeloch musí byť rovnako $2/3$, inak by dochádzalo k skresleniu.

Vzhľadom na aktívnu cestnú premávku do ktorej sme zasiahnúť nemohli, sme boli nútení využiť miesta nižšej intenzity prechodu vozidiel a *marker* umiestniť do rôznych oblastí vozovky. Na základe tohto ale relatívne pozície medzi *markermi* nemohli byť určené, keďže sa jednalo o jeden *marker*. Nakoľko bola použitá statická kamera, tak jednotlivé zábery mohli byť považované za viacej *markerov* (Obr. 5). Naviac, aj v prípade kedy by identických *markerov* bolo k dispozícií viacej, z organizačného hľadiska, ale aj kvôli obmedzeniam meracích prístrojov, by sme pozície neboli schopní určiť s dostatočnou presnosťou, aj keď sme sa o to pokúšali. Riešením bolo vybrať *marker* taký, ktorého homografia vykazuje najmenšiu chybu v rekonštrukcii obrazu. Ak sa na scéne nachádza *m markerov*, o ktorých vieme, že sú napríklad štvorce, ale nemáme informáciu o ich pozícií, potom môžeme korešpondenciu bodov definovať len izolovane pre každý *marker*, nie pre všetky naraz, čo by malo za následok zvýšenie presnosti odhadovania homografie.

V bežnej praxi je snímanie obrazu sprevádzané šumom, a teda každá homografia odhadnutá izolovane pre jednotlivé *markery* vykazuje rôznu kvalitu reprojekcie obrazu. Toto nás viedlo k otázke "Ako je možné systematicky kvantifikovať kvalitu reprojekcie *m* izolovaných homografií patriacich *m* geometricky podobným *markerom*?" Ak teda nie je možné všetky známe body spojiť do jedného veľkého, zloženého *markeru* pre čo najlepší odhad homografie, tak sa našim cieľom stalo poskytnúť prístup ako z týchto *m* homografií vybrať tú najlepšiu. Naviac, urobiť "priemer" homografií je matematicky nemožné, takže je len úlohou buď priamo odhadnúť homografiu jednu, alebo jednu z viacerých vybrať. Je zrejmé, že *marker* umiestnený na okraji obrazu bude vykazovať značne horšiu reprojekciu vo vzdialenejších oblastiach. Toto nás viedlo k intuitívnej myšlienke, že kvalita reprojekcie musí byť



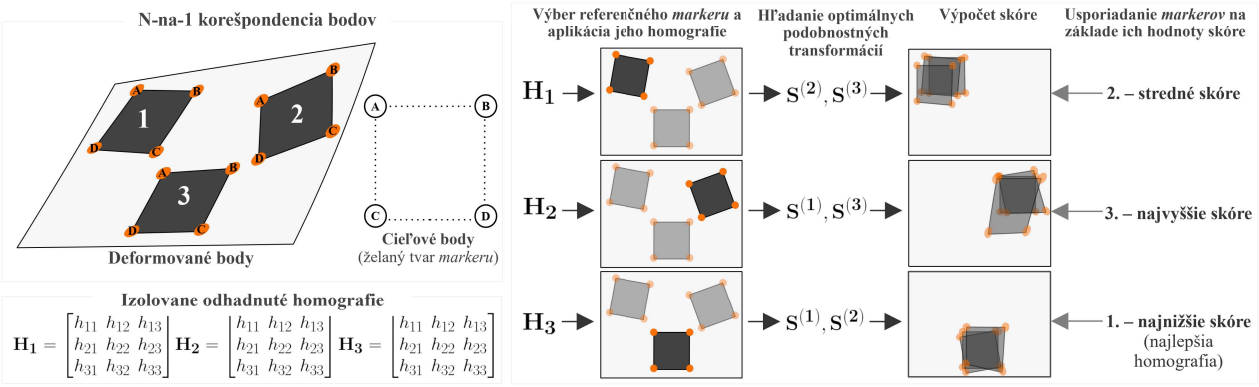
Obr. 6: Ukážka existujúcich metód odhadovania homografie a nášho rozšírenia pre systematický výber, ktorú homografiu použiť na základe chyby reprojekcie meraných na *m markeroch*. Vypočítaná chyba nepriamo slúži ako odhad pre chybu reprojekcie na celom obraze. Predpokladáme, že čím je presnosť vyššia na malej množine pixelov, tým aj transformácia celého obrazu bude presnejšia, hoci to nie je možné určiť v reálnom svete. Algoritmus usporiadania vyskúša všetky homografie (korešpondujúce *markery* sa stanu referenčnými) a každej priradí skóre (chybu reprojekcie), na základe ktorého je vybraná "najlepšia" homografia.

kvantifikovateľná. Odpoveď na túto otázku sme v existujúcej literatúre nenašli, a preto sme sa rozhodli ju po dôkladnej experimentálnej činnosti publikovať.

3.2 Popis navrhnutej metódy

Vzhľadom na nutný predpoklad aby všetky *markery* zdielali rovnakú rovinu v priestore, tak pre každý musí teoreticky existovať valídna homografia, ktorá poskytne pohľad z vtáčej perspektívy na danou rovinou. Avšak, z dôvodu prítomnosti šumu sa kvalita rekonštrukcie obrazu líši. Snahou teda je kvantifikovať, ktorá homografia by potenciálne mohla poskytnúť najlepšiu reprojekciu danej roviny (Obr. 6).

Za týmto účelom sme vyvinuli skórovaciu funkciu založenú na predpoklade, že *m* markerov je geometricky podobných (existuje medzi nimi podobnostná transformácia so 4 stupňami voľnosti) a sú umiestnené na rovnakej rovine v priestore, pričom informácia o ich pozícií absentuje. Pre výpočet skóre sú potrebné pomocné podobnostné transformácie, ale tento krok je v praxi ľahko splniteľný kvôli geometrickej



Obr. 7: Vstupom do našej metódy je niekoľko bodových korešpondencií a informácia o cieľovom tvare *markeru*. Následne sa každý *marker* vyhodnotí ako referenčný, odhadnú sa optimálne podobnostné transformácie zvyšných *markerov*, a vyhodnotí sa skóre. Na základe tohto skaláru je určené poradie homografií pomocou ľubovoľného triediaceho algoritmu.

podobnosti markerov. Skóre je konkrétnu homografiu je vypočítané ako

$$\mathscr{F}\left(\mathbf{H},\mathscr{S}\right)=\frac{1}{m}\sum_{i=1}^{m}\left\|h\left(\mathbf{S}^{(i)}\mathbf{H}\mathbf{W}^{(i)}\right)-\mathbf{T}\right\|_{F}, \tag{5}$$

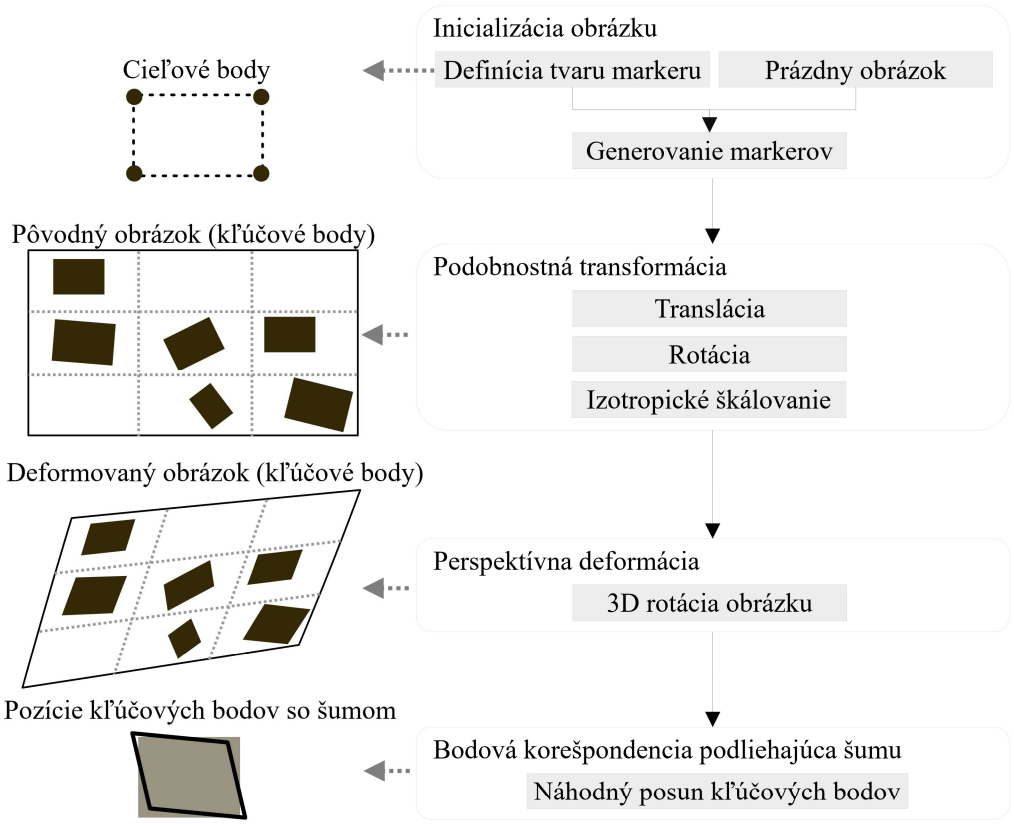
kde $\|\cdot\|_F$ je Frobeniová norma, funkcia $h(\cdot)$ konvertuje body na homogénne koordináty a štyri matice **S**, **H**, **W**, **T** sú podobnostná transformácia, homografia, reprezentácia deformovaných bodov a reprezentácia cieľových bodov, respektíve.

Hlavná myšlienka za touto funkciou je tá, že každá homografia poskytne inú rekonštrukciu obrazu. Lenže nie je možné určiť presnosť pre každý pixel, iba pre tie, o ktorých vieme, že patria známym objektom, čiže *markerom*. Vyhodnocovanie teda zahŕňa rektifikáciu všetkých *markerov*, a následne porovnanie každého jedného s cieľovým (žiadúcim) tvarom (vďaka podobnostnej transformácií je toto možné) a vypočítaní chyby reprojekcie len v kľúčových bodov *markeru* (Obr. 7).

3.3 Experimentálna činnosť

Naše experimenty boli založené výhradne na syntetickej dátovej množine, ktorú sme navrhli pre tento účel. Faktom však zostáva, že reálnu dátovú množinu nie je ani možné vytvoriť. Cieľom bolo simulovať v laboratórnych (ideálnych) podmienkach umiestnenie niekoľko geometricky podobných *markerov* na rovinu. Tieto *markery* mohli byť rôzne otočené, aby sme demonštrovali možné zmeny v translácií, rotácií a izotropickom škálovaní. Následne bol celý takýto obrázok rotovaný náhodne v 3D priestore, čím bola dosiahnutá perspektívna deformácia (Obr. 8).

Nakoľko sme presne vedeli určiť ako má byť deformovaný obraz rekfitikovaný, mali sme možnosť vyhodnocovať aj chybu reprojekcie pre každý jeden pixel, nie len



Obr. 8: Proces tvorby syntetickej dátovej množiny pre vyhodnotenie chyby reprojekcie homografií a schopnosti našej metódy usporiadania homografií priniesť očakávané zlepšenie. V prvej fáze je prázdny obrázok inicializovaný niekoľkými *markermi*. Tieto markery sú modifikované v zmysle valídnej podobnostnej transformácie. Následne je obrázok rotovaný výsledok čoho je perspektívne skreslenie. V poslednej fáze je aplikovaný šum na pozície kľúčových bodov *markeru*, reprezentujúci chybu v detekcií objektu.

pre kľúčové body *markerov*, čoho je schopná naša skórovacia funkcia. Na základe tohto sme korektne vyhodnotili našu metódu a overili predpoklady.

Pri tvorbe syntetickej dátovej množiny sme kládli dôraz na prítomnosť šumu a ukázalo sa, podľa očakávaní, že v prítomnosti šumu je naša metóda schopná poskytnúť konzistentné zlepšenie. Ak nie je šum prítomný tak výber markerov nemá zmysel, ale taká situácia je v praxi nemožná.

Zlepšenie sme porovnávali voči referenčnej metóde založenej na náhodnom výbere *markeru*, ktorou bola jedna z najlepších voľne dostupných implementácií z `OpenCV` knižnice pre počítačové videnie. Myšlienke náhodného vyberu odzrkadľovala nezaujatý výber *markeru* bez ďalších znalostí. Prakticky sme sa snažili zodpovedať na otázku, že ak používateľ využije "odporučanie" našej metódy namiesto náhodného výberu *markeru*, aké percentuálne zlepšenie v rekonštrukcií obrazu môže v priemere očakávať? Veľmi stručne vieme zodpovedať: $\mathbf{60\%}$.

3.4 Vyhodnotenie prínosov

V tejto práci sme navrli metódu, ktorá je rozšírením pre existujúce prístupy k odhadovaniu homografie za predpokladu, že sú využité bodové korešpondencie. Navrhnutá a otestovaná metóda predstavuje systematický spôsob ako zaviesť poradie medzi niekoľkými homografiami za predpokladu, že každá z nich korešponduje nejakému *markeru*, ktorý je umiestnený na rovine nad ktorou chceme získať pohľad z vtáčej perspektívy, pričom všetky *markery* sú geometricky podobné. Zmienené poradie je určené na základe skórovacej funkcia, ktorá tvorí jadro nášho vedeckého prínosu. Vo všeobecnosti môžeme tvrdiť, že naša metóda poskytuje $\mathbf{60\%}$ konzistentné zlepšenie voči bazálnej hladine danej náhodným výberom *markeru*.

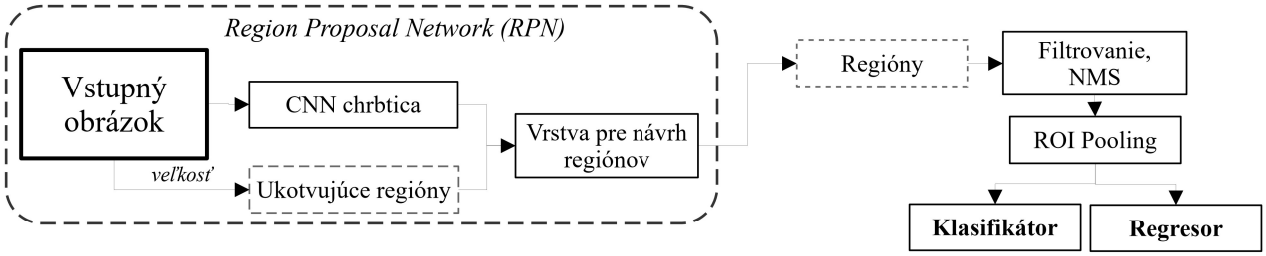
Bez našej metódy by bol používateľ odkázaný na svoje presvedčenie a vizuálne zhodnotenie výstupného obrazu. My sme poskytli nástroj, ktorý je invariantný voči tomu, akým spôsobom bola homografia odhadnutá za predpokladu, že sú k dispozícií korešpondencie kľúčových bodov patriacim geometricky podobným *markerom*.

4 Navrhnuté prístupy k vizuálnemu trasovaniu

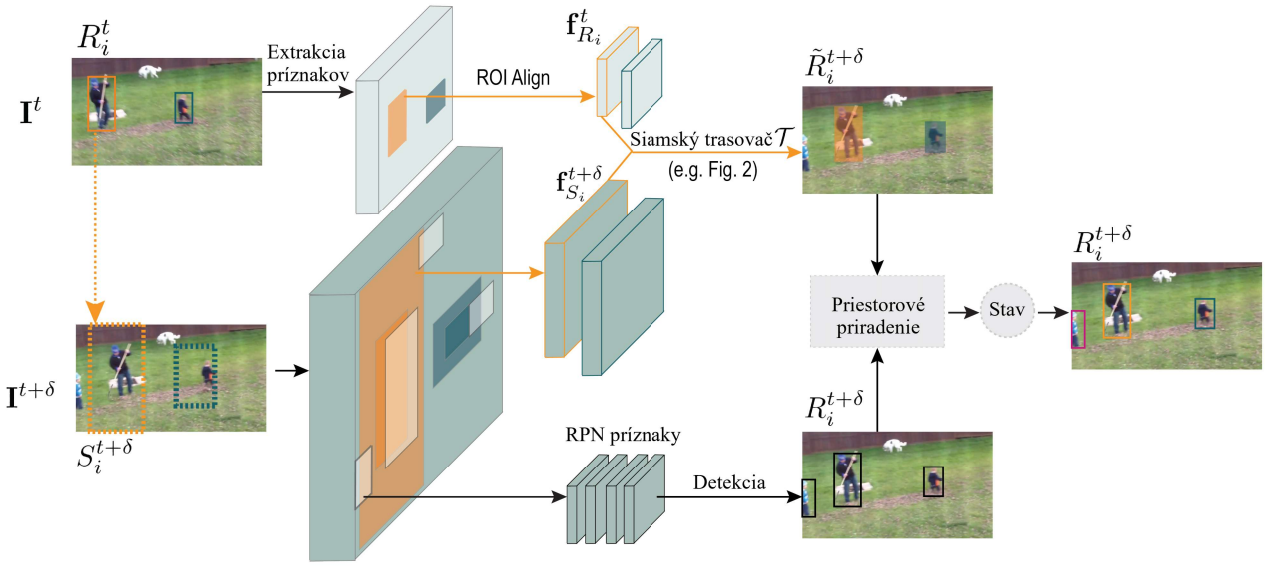
4.1 Architektúra pre Siamské trasovanie viacerých objektov

Architektúra SiamMOT tvorila základ naších experimentov kvôli svojej presnosti a efektívnosti trasovania. Vzhľadom na nadobudnuté znalosti počas písania nášho prehľadového článku o siamskom trasovaní [11] sme usúdili, že takýto trasovač adekvátne demonštruje použitie najlepších prístupov siamského SOT a vhodne ich pretavuje do MOT. Mimo iného, v tejto architektúre sme identifikovali niekoľko slabších miest, na ktoré sme sa rozhodli zamerať. Musíme poznamenať, že daný model je robustný a dosiahnúť zlepšenie dodatočnými modifikáciami architektúry je náročné. Toto na jednej strane potvrdzuje správnosť výberu bazálnej architektúry, avšak na druhej strane to ukazuje, že niektoré naše nápady neboli pre tento model tak vhodné, ako naše hypotézy naznačovali.

Dva kľúčové aspekty modelu SiamMOT sú siamské trasovanie ako také a detekcia objektov pomocou Faster R-CNN [13]. Faster R-CNN detektor objektov pracuje v dvoch fázach, kde prvá generuje návrhy regiónov, ktoré potenciálne môžu obsahovať objekty, a druhá fáza je zodpovedná ako za klasifikáciu objektov v rámci navrhnutých regiónov tak aj za dodatočné spresnenie ohraničujúcich obdlžníkov (Obr. 9). Tento prístup k detekcií patrí medzi najlepšie. V SiamMOT bol použitý ako nezávislý detektor objektov, ktorý využíval extrahované príznaky z rovnakej chrbticovej CNN ako nezávislý siamský trasovač. Tieto dve vetvy produkovali predikcie detekcií objektov a predikcie pohybu ohraničujúcich obdlžníkov trasovaných objektov voči predošlej snímke, respektíve. Predikcie z oboch modelov boli následne spájané vo



Obr. 9: Architektúra dvojfázovej detekcie objektov predstavená v rámci modelu *Faster Region-based Convolutional Neural Network* (Faster R-CNN). Spracovanie vstupu začína extrakciou príznakov pomocou CNN siete tvoriacich vstup pre návrh regiónov, v ktorých by sa mohli objekty nachádzať. Tieto návrhy sú systematicky rozmiestnené v rámci ukotvujúcich regiónov, ktoré predstavujú obdĺžniky rôznych veľkosti a pomerov strán umiestnené na pravidelnej mriežke. Úlohou detektora je predikovať zmenu pozície a veľkosti ukotvujúcich obdĺžnikov tak, aby chyba predikcie bola minimálna. Tieto kroky sú súčasťou ďalšej fázy, ktorá zahŕňa aj klasifikáciu objektov do príslušných tried.



Obr. 10: Vstupom do modelu sú dve snímky ako v bežných siamských trasovačoch. Jedna je exemplár (historická snímka) a druhá je aktuálna vyhľadávacia oblasť. Z obidvoch snímiek sú extrahované príznaky pomocou CNN. Vo vetve siamského trasovača dochádza k predikcií pohybu všetkých trasovaných objektov rovnakým spôsobom, ako pri SOT siamskom trasovaní. Vetva detekcie objektov nezávisle určuje pozície viditeľných objektov bez ohľadu na triedu príslušnosti. Tieto dva prúdy predikcií sú spájané v module, ktorý nie je trénovateľný a je aktívny len počas inferencie s účelom generovať výslednú predikciu celého trasovača.

fáze priestorového priradenia (Obr. 10).

Počas trasovania sa objekt môže nachádzať v troch rôznych stavoch: *nový*, *aktívny* alebo *dormantný*. Každý objekt je inicializovaný ako nový a za predpokladu, že dôveryhodnosť jeho detekcie presahuje parametricky stanovenú prahovú hodnotu, sa

Obr. 11: Ukážka rôznych dopravných situácií v UA-DETRAC [20] dátovej množine.

stane aktívnym. V aktívnom stave pretrváva pokiaľ dôveryhodnosť modelu v existenciu daného objektu je nad špecifickou prahovou hodnotou. V prípade, že táto dôveryhodnosť klesne, tak trasovaný objekt je považovaný za dormantný. V takomto stave existuje len v pamäti a siamský trasovač má snahu opätovne jeho identitu obnoviť neustálym prehľadávaním okolia, v ktorom naposledy zanikol. Ak sa na scéne neobjaví žiadny objekt počas k snímok, ktorý by vykázal dostatočnú zhodu s exemplárom uvažovaného dormantného objektu, potom trasovanie pre tento objekt končí.

4.2 Použité dátové množiny pre tréning a testovanie modelov

Počas nášho výskumu sme využili niekoľko dátových množín, avšak tá najdôležitejšia je UA-DETRAC [20] (Obr. 11) zameraná na trasovanie viacerých vozidiel. Vzhľadom na náš cieľ analyzovať dopravu statickou kamerou je tento projekt ideálny z hľadiska kvality spracovania (nahrávok a anotácie) a objemu dám. K dispozícií je 140 000 snímok zachytených statickou kamerou s 25 *Frames per Second* (FPS) s rozlíšením 960×540 pixelov, ktoré obsahujú až 8 250 manuálne anotovaných vozidiek, čo činí približne 1.21 milióna ohraničujúcich obdĺžnikov.

4.3 Experimenty s pridaním externej re-identifikácie

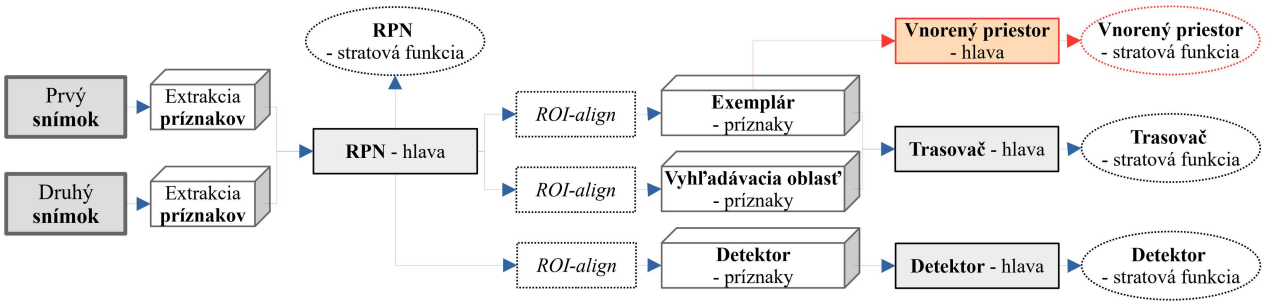
Na počiatku sme sa zamerali na úplnú oklúziu objektu pomocou ReID. Intuícia bola, že ak objekt nie je viditeľný niekoľko konsekutívnych snímok, tak po opätovnom objavení je dôležité jeho predchádzajúci identifikátor korektne priradiť. Nastáva otázka správneho určenia či sa jedná o objekt nový alebo už trasovaný niekedy v histórií, ktorej dĺžka je parametricky určená, zvyčajne aspoň 1 sekundu.

Obr. 12: Čiastočná oklúzia v UA-DETRAC [20] dátovej množine, pri ktorej ohraničujúci obdĺžnik jedného objektu pokrýva veľkú časť objektu druhého, ktorá je pre ReID problematická.

ReID dokáže určiť či dva zábery nejakých objektov reprezentujú objekt identický. Myšlienka bola aby v situácií, kedy SiamMOT trasovač deteguje nový objekt, bol pôvodný mechanizmus siamského prehľadávania okolia nahradený naším algoritmom, ktorý by dokázal pomocou vyhodnotenia vzdialenosti medzi vektormi v nejakom vytvorenom vnorenom priestore určiť úroveň podobnosti dvoch objektov. Vo všetkých ostatných prípadoch by model pracoval podľa pôvodného návrhu. Bazálna formulácia prehľadáva len fixné okolie, spravidla štvornásobne väčšiu oblasť ako pokrýva exemplár. Avšak, úplna oklúzia môže spôsobiť to, že sa objekt opätovne objaví v distálnej časti obrazu, ktorú siamský trasovač ani zvažovať nebude. Autori sa pôvodne zamerali len na krátkodobú oklúziu, no my sme mali cieľ implementovať rozšírenie, ktoré by zvládalo aj oklúziu dlhodobú.

Tento experiment nevykázal zlepšenie. Vo všeobecnosti sa dá povedať, že buď sa nám podarilo presnosť trasovania udržať alebo sa nám ju podarilo mierne zhoršiť. Naviac, rýchlosť inferencie trasovača bola miestami až 5-násobne znížená, a takýto zásah do rýchlosti by nebol ospravedlniteľný minimálnym zlepšením. Každopádne, vďaka tomuto experimentu sme odhalili niekoľko slabín ReID. Tá najdôležitejšia je, že ak dochádza k čiastočnej oklúzií najmä pod uhlom pri použití osovo zarovnaných ohraničujúcich obdlžníkov, potom vo veľkej miere ReID obdrží na vstupe regióny s veľkým prienikom a teda ich vektory podobnosti v rámci vnoreného priestoru budú vykazovať vysokú podobnosť. Pri vozidlách tento aspekt môže byť dodatočne zhoršený podobným výzorom vozidiel (Obr. 12).

ReID je výborné pri záznamoch z viacerých kamier, kde sú zábery vozidla oddelené v priestore a často aj v čase. Vtedy sú vizuálne zmeny významné. V našom prípade bola snaha aplikovať ReID na veľmi podobné objekty, ktoré nemuseli vizuálne zmeny vôbec podstúpiť. Tieto situácie, ktoré sú mimochodom v doprave časté, spôsobovali pre náš navrhnutý prístup problémy. I napriek zlepšeniam v niektorých situáciách, v konečnom dôsledku je toto rozšírenie pre túto architektúru škodlivé. Dá sa tvrdiť, že sa nám podarilo zlepšiť prípady, ktoré sú raritné za cenu toho, že sme mierne zhoršili prípady, ktoré sú časté, a teda vo výsledku bol efekt negatívny.



Obr. 13: Pridanie "hlavy" do pôvodnej SiamMOT architektúry zameranej na tvorbu vnorených priestorov pre následne ReID objektov počas inferencie. Diagram znázorňuje proces počas tréningu. Samotný ReID modul tvorený CNN je so zvyškom modelu previazaný pomocou stratovej funkcie. Celý model je trénovaný *end-to-end* štýlom.

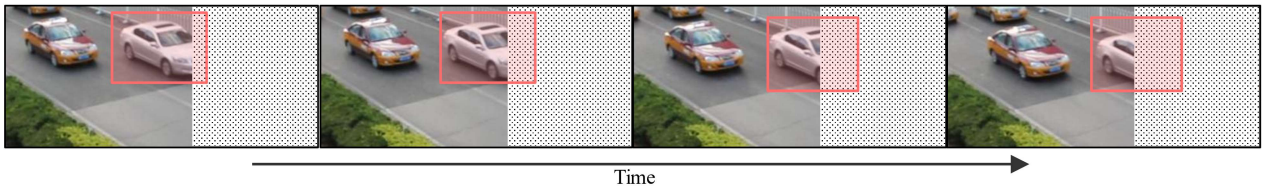
4.4 Experimenty s pridaním hlavy pre vnorené priestory

V ďalšom z naších experimentov bola snaha zjednotiť tréning tvorby vnorených priestorov a trasovania. Rozhodli sme sa architektúru obohatiť o dodatočnú "hlavu", ktorá mala za úlohu naučiť sa tvoriť vnorené priestory pre ReID počas *end-to-end* tréningu celého modelu. Snažili sme sa o jednoduchosť bez značných zásahov do pôvodnej architektúry (Obr. 13). Naviac, takýto spôsob rozširovania sme vo vedeckej literatúre často zahliadli s úspešným prínosom.

V takomto prípade je chrbticová CNN nútená naučiť sa extrahovať príznaky, ktoré sú potrebné nielen pre trasovanie a detekciu, ale aj pre ReID. Príznaky siamského trasovača sú schopné čiastočne realizovať ReID, keďže dochádza k vyhľadávaniu exempláru v budúcej snímke. Počas tréningu sme rozšírili stravovú funkciu pôvodného modelu o pričítanie ďalšieho člena, ktorý reprezentoval dvojicovú (rovnica 1) a aj trojicovú (rovnica 2) stratovú funkciu v závislosti od experimentu. Musíme poznamenať, že trénovanie bolo veľmi nestabilné kvôli explodujúcim gradientom. Za účelom mitigácie takéhoto efektu sme boli nutení aplikovať rôzne techniky na stabilizáciu tréningu. Trojicové stratové funkcie sú známe svojou náročnosťou na tréning. Okrem mnoho iných modifikácií, v rámci tohto experimentu sme vyvinuli aj upravený *Non-Maximum Suppression* (NMS) algoritmus, ktorý zahrňal aj prácu s vnorenými priestormi. Rovnaký spôsob sme neskôr našli aj v literatúre [14].

Toto rozšírenie neprinieslo očakávané benefity. I napriek tomu, že čas inferencie nebol výrazne ovplyvnený vďaka unifikácií, naše výsledky vykazovali značné zníženie presnosti trasovania. Po bližšej investigácií sme zistili, že dochádza k výraznému súpereniu medzi príznakmi. Naše zistenia boli nedávno nezávisle potvrdené vo veľmi dôležitej publikácií [22], ktorá adresuje problémy vyplývajúce zo snahy inkorporovať ReID mechanizmy do trasovačov objektov, ktoré používajú RPN.

V skratke zhrnieme zásadné problémy. Každá z troch hlavných úloh celej architektúry, čiže trasovania, detekcie a v našom prípade aj ReID, vyžaduje iné príznaky



Obr. 14: Demonštrácia postupne narastajúcej umelo vytvorenej oklúzie objektu. V takomto prípade trasovač nesprávne zahŕňa aj objekt, ktorý prekrýva trasovaný objekt. Toto je výhoda pre dočastnú oklúziu. Avšak, ReID následne pracuje s posledným "viditeľným" regiónom v obraze, ktorý bol ešte priradený objektu záujmu, a v ňom už nemusí byť objekt skoro vôbec viditeľný. Aby sme mohli ReID korentne použiť, museli by sme byť schopný detegovať oklúziu ešte skôr ako nastanie veľmi závažné prekrytie objektu, čo je ale náročný problém.

od chrbticovej CNN. I napriek existencií rôznych techník ako tento efekt minimalizovať, ktoré boli vo veľkej miere v tejto architektúre využité, stále dochádza ku konfliktu. Každopádne, najzávažnejšia systematická príčina zlyhania je použitie ukotvujúcich oblastí v RPN za účelom návrhu regiónov pre potenciálny výskyt objektov. Pri architektúrach takéhoto typu je priorita kladená na RPN modul, nakoľko trénovať správne ReID na nesprávnej oblasti v obraze je kontraproduktívne.

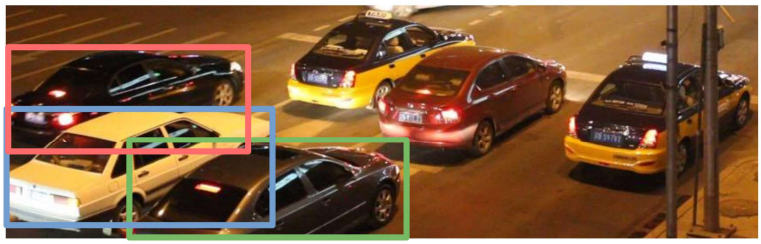
Ďalší dôvod prečo je explicitné použitie ReID problematické je že dochádza k závažnej degradácií viditeľných exemplárov objektu pri oklúzií, ktorú by bolo vhodné konzistentne a presne detegovať. Snaha vyhodnotiť podobnosť výsekov z obrazu, ktoré nemusia obsahovať žiadny objekt značne destabilizuje trasovací proces a dochádza k zamieňaniu identifikátorov objektov (Obr. 14). Tento problém je omnoho náročnejší ako sa zdá a preto sme sa touto cestou nevydali.

4.5 Experimenty s pridaním mechanizmu pozornosti

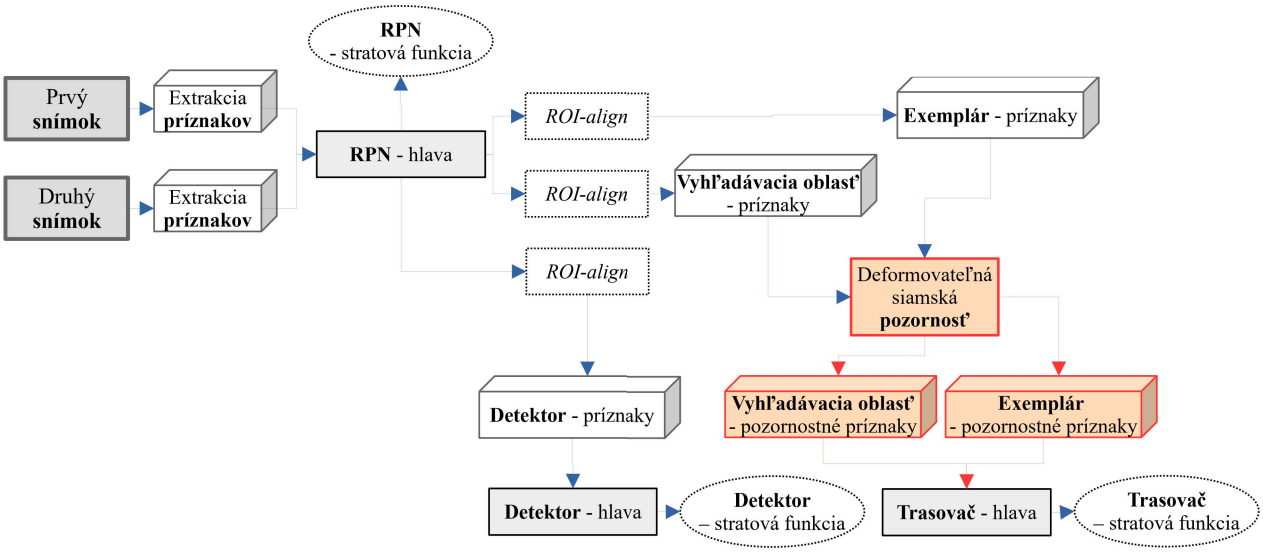
Počas inšpekcie inferencie trasovača sme si všimni často prítomný vzor. Pripomíname, že scény pomocou ktorých sme trénovali a testovali náš model, boli zachytené statickou kamerou. V dôsledku tohto isté video sekvencie obsahovali situácie, kedy niekoľko vozidiel stálo na mieste v dôsledku dopravnej zápchy alebo červeného svetla, avšak boli zachytené kamerou pod uhlom v rozsahu $30 - 60°$ (Obr. 15). V takýchto prípadoch sa oblasti záujmu príslušného objektu prelínala s inými, čo značne zvyšuje šancu pre zámenu trasovaného objektu za sémantické pozadie.

S cieľom adresovať vyššie uvedený problém sme sa rozhodli využiť mechanizmus pozornosti [18], najmä jeho priestorovú formuláciu, o ktorej efektívnosti sme sa presvedčili aj v rámci nášho prehľadového článku [11]. Okrem samotnej pozornosti sme použili aj *Deformable Convolutional Neural Network* (DCNN) [2] namiesto štandardnej CNN pre zvýšenie robustnosti a diskriminačnej schopnosti trasovača.

V literatúre sme sa stretli s *Deformable Siamese Attention* (DSA) [21] modulom,



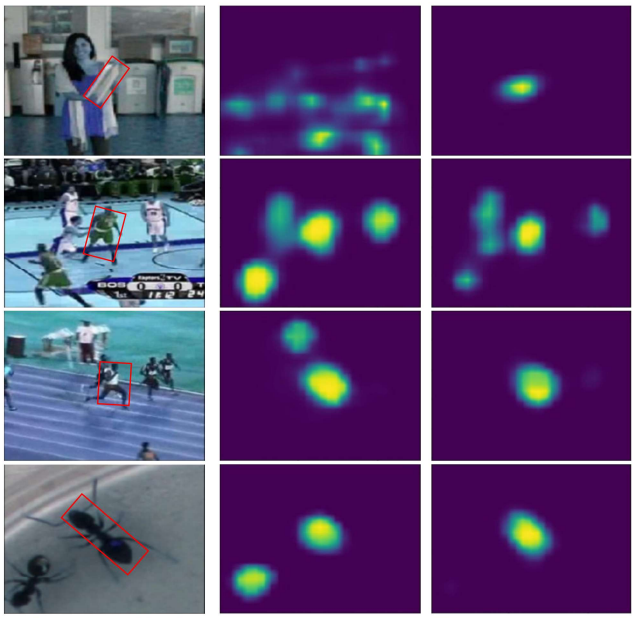
Obr. 15: Ukážka, ako použitie osovo zarovnaného ohraničujúceho obdĺžnika spôsobuje zachytenie blízkych objektov aj bez prítomnosti závažného prekrytia.



Obr. 16: Rozšírenie pôvodnej SiamMOT architektúry pomocou mechanizmu pozornosti. Cieľom bolo naučiť model "upriami5 pozornosť" na relevnatné časti obrazu súvisiace ako s exemplárom tak aj s vyhľadávacou oblasťou. Toto rozšírenie si vyžaduje minimálne modifikácie do pôvodnej architektúry vďaka tomu, že nedochádza k zmenám rozmerov príznakov. Diagram znázorňuje proces počas tréningu. Celý model je trénovaný *end-to-end* prístupom.

ktorý zahŕňal naše nápady a naviac prispel aj myšlienkou krížovej pozornosti. Na základe výsledkov v rámci siamského SOT sme sa rozhodli toto rozšírenie aplikovať pre SiamMOT model, nami označovaný ako DSA-extended (Obr. 16, Obr. 17).

Pridanie DSA modulu do SiamMOT architektúry značne zvýšilo nároky na *Graphics Processing Unit* (GPU) *Video Random Access Memory* (VRAM), nakoľko je nutné počas tréningu vytvoriť aj pozornostné príznaky. Avšak, tieto nároky sú výrazné hlavne počas tréningu. Inferencia modelu nie je zreteľne zaťažená. Každopádne, DSA-extended vykazuje konzistentné zlepšenie vo všetkých relevantných metrikách zameraných na MOT voči pôvodnému modelu (Tabuľka 1). Odhliadnúc od vyšších pamäťových nárokov na tréning, tak rýchlosť inferencie stále môže byť považovaná za vhodnú pre spracovanie v reálnom čase. Oproti pôvodnému modelu DSA-extended verzia dosahuje približne 9% zníženie rýchlosti (Tabuľka 2).



Obr. 17: Vizualizácia aktivačných máp. Prvý stĺpec reprezentuje vyhľadávacie oblasti, druhý demonštruje absenciu DSA modulu, pričom tretí naznačuje zlepšenie v diskriminačnej schopnosti medzi trasovaným objektom a distraktormi pri použití pozornosti. (*zdroj*: [21])

model	**MOTA**	**MOTP**	**precision**	**recall**
DSA-extended	0.7625	0.1548	0.9260	0.8315
original	0.7429	0.1533	0.9137	0.8230

Tabuľka 1: Porovnanie najlepších modelov, ktoré boli trénované a testované na UA-DETRAC dátovej množine. Tieto modely boli vybrané na základe ich *Multiple Object Tracking Accuracy* (MOTA) a *Multiple Object Tracking Precision* (MOTP) páru.

model	**rýchlosť inferencie [FPS]**				
	min.	**max.**	**mean**	**stdev.**	**median**
original	22.82	29.84	26.49	1.61	26.67
DSA-extended	19.06	29.61	24.16	2.67	24.20

Tabuľka 2: Porovnanie rýchlosti inferencie originálneho SiamMOT a DSA-extended verziou. Štatistiky sú založené na priemernej hodnote FPS pre jednu sekvenciu použitím 56 340 snímok naprieč 40 sekvenciami poskytnutými UA-DETRAC validačnou dátovou množinou, s priemernou dĺžkou sekvencie 1408.5 snímok. Naše hardvérové špecifikácie boli `NVIDIA RTX 2080Ti` GPU a `AMD Ryzen Threadripper 2920X 12-Core` CPU.

5 Záver a zhrnutie prínosov

Hlavným cieľom tejto dizertačnej práce bolo prispieť do oblasti VOT použitím hlbokého strojového učenia. Náš vecný vedecký príspevok, či už po stránke teoretickej alebo praktickej, sa skladá z troch hlavných častí.

1. Aktuálne pokroky v rámci siamského trasovania sme pokryli v našom prehľadovom článku publikovanom v žurnále [11], ktorý doplnil chýbajúce miesto v existujúcej literatúre. Tento článok poskytuje kvalitatívnu a kvantitatívnu diskusiu ohľadom fundamentálnych vlastností siamských trasovačov a aktuálnych výzvach, ktorým táto paradigma trasovania aktuálne čelí.

2. Doména trasovania vozidiel podnietila potrebu odstraňovania perspektívnej deformácie nakoľko sme vyžadovali merať rýchlosť a rozmery vozidiel. Za týmto účelom sme využili homografiu. Napriek naším plánom zahrnúť homografiu do trasovania, v dôsledku nedostatočného množstva dátových množín sme túto vetvu výskumu zanechali. Každopádne, naše výsledky boli originálne a publikované v žurnále [10].

3. Náš praktický príspevok v siamskom trasovaní sa skladá z troch častí, pričom ten posledný je najdôležitejší kvôli dosiahnutých zlepšeniam. Cieľom bolo vylepšiť SiamMOT [15] trasovač použitím UA-DETRAC [20] dátovej množiny.

    - V prvom experimente sme ukázali negatívne účinky ReID na trasovač kvôli zníženiu výpovednej hodnoty vnorených vektorov v dôsledku prítomnosti oklúzie. Naše pozorovania a závery sú relevantné pre všeobecnú diskusiu ohľadom použitie ReID v kombinácií so siamským trasovaním.
    - Druhý experiment zahŕňal použitie rozširujúcej hlavy v rámci SiamMOT architektúry, ktorá produkovala príznakové vnorené vektory. Napriek negatívnym výsledkom bol náš výskum prínosný nakoľko sme ukázali konsekvencie simultánneho tréningu formovania vnorených priestorov a trasovača používajúceho RPN. Naše experimenty poskytujú validáciu pre relevantný článok [22], ktorý sa zaoberá "neférovosťou" ReID v MOT.
    - V rámci tretieho experimentu bola snaha zlepšiť diskriminačnú schopnosť trasovača zvládať scény s prítomnosťou čiastočnej oklúzie objektov. Vyvinuli sme prístup založený na mechanizme pozornosti, ktorý dosahuje výborné výsledky. Adoptovali sme publikovanú a úspešnú architektúru DSA [21] spoločne s našími modifikáciami. V dôsledku týchto zmien sme získali 2.6% zlepšenie merané pomocou MOTA metriky.

Vo všeobecnosti môžeme tvrdiť, že dva z troch naších príspevkov sú úspešne ukončené a publikované, konkrétne prehľadový článok o siamskom trasovaní a vy-

vinutá metóda na vyhodnocovanie kvality reprojekcie homografie. Čo sa týka nášho prístupu využívajúceho pozornosť, tak tu je priestor na zlepšenie, najmä v zmysle dodatočnej validácie na iných dátových množinách. Každopádne, analýza dopravy bola prvoradej dôležitosti, a tento účel sme splnili. Naviac, naším cieľom bolo porovnať sa v relatívnych hodnotách vzhľadom na originálny SiamMOT model, čo sa nám rovnako podarilo. Vzhľadom na vyššie pamäťové nároky počas tréningu by bolo vhodné navrhnúť modifikácie, ktoré by tieto nároky redukovali a teda šancu pre použitie tohto rozšírenia ešte zvýšili.

Literatúra

[1] Bertinetto, L., Valmadre, J., et al.: 'Fully-convolutional siamese networks for object tracking', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9914 LNCS, pp. 850–865, 2016, ISSN 16113349, doi: 10.1007/978-3-319-48881-3_56

[2] Dai, J., Qi, H., et al.: 'Deformable convolutional networks', *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 764–773, 2017, doi:10.1109/ICCV.2017.89

[3] Hadsell, R., Chopra, S., LeCun, Y.: 'Dimensionality reduction by learning an invariant mapping', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1735–1742, 2006, ISSN 10636919, doi:10.1109/CVPR.2006.100

[4] Hermans, A., Beyer, L., Leibe, B.: 'In Defense of the Triplet Loss for Person Re-Identification', 2017

[5] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'ImageNet classification with deep convolutional neural networks', *Advances in Neural Information Processing Systems*, vol. 2, pp. 1097–1105, 2012, ISSN 10495258

[6] Li, B., Yan, J., et al.: 'High Performance Visual Tracking with Siamese Region Proposal Network', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8971–8980, 2018, ISSN 10636919, doi:10.1109/CVPR.2018.00935

[7] Li, D., Yu, Y.: 'Foreground information guidance for siamese visual tracking', *IEEE Access*, vol. 8, pp. 55905–55914, 2020, doi:10.1109/ACCESS.2020.2982261

[8] Li, D., Yu, Y., Chen, X.: 'Object tracking framework with siamese network and re-detection mechanism', *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, p. 261, 2019, ISSN 1687-1499, doi:10.1186/s13638-019-1579-x

[9] Liang, Z., Shen, J.: 'Local semantic siamese networks for fast tracking', *IEEE Transactions on Image Processing*, vol. 29, pp. 3351–3364, 2020, doi:10.1109/TIP.2019.2959256

[10] Ondrašovič, M., Tarábek, P.: 'Homography ranking based on multiple groups of point correspondences', *Sensors*, vol. 21, no. 17, 2021, ISSN 1424-8220, doi:10.3390/s21175752

[11] Ondrašovič, M., Tarábek, P.: 'Siamese Visual Object Tracking: A Survey', *IEEE Access*, vol. 9, pp. 110149–110172, 2021, doi:10.1109/ACCESS.2021.3101988

[12] Pflugfelder, R.: 'An in-depth analysis of visual tracking with siamese neural networks', 2018

[13] Ren, S., He, K., et al.: 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017, ISSN 01628828, doi:10.1109/TPAMI.2016.2577031

[14] Salscheider, N.O.: 'Featurenms: Non-maximum suppression by learning feature embeddings', 2020

[15] Shuai, B., Berneshawi, A., et al.: 'Siammot: Siamese multi-object tracking', 2021

[16] Shuai, B., Berneshawi, A.G., et al.: 'Multi-object tracking with siamese track-rcnn', *arXiv preprint arXiv:2004.07786*, 2020

[17] Vaquero, L., Mucientes, M., Brea, V.M.: 'Siammt: Real-time arbitrary multi-object tracking', *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 707–714, 2021, doi:10.1109/ICPR48806.2021.9412625

[18] Vaswani, A., Shazeer, N., et al.: 'Attention is all you need', 2017

[19] Wang, Q., Zhang, L., et al.: 'Fast online object tracking and segmentation: A unifying approach', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 1328–1338, 2019, ISSN 10636919, doi:10.1109/CVPR.2019.00142

[20] Wen, L., Du, D., et al.: ' UA-DETRAC: A New Benchmark and Protocol for Multi-Object Detection and Tracking', *Computer Vision and Image Understanding*, 2020

[21] Yu, Y., Xiong, Y., et al.: 'Deformable siamese attention networks for visual object tracking', 2021

[22] Zhang, Y., Wang, C., et al.: 'Fairmot: On the fairness of detection and re-identification in multiple object tracking', *International Journal of Computer Vision*, vol. 129, no. 11, p. 3069–3087, 2021, ISSN 1573-1405, doi:10.1007/s11263-021-01513-4

[23] Zhu, Z., Wang, Q., et al.: 'Distractor-aware siamese networks for visual object tracking', 2018

6 Zoznam vlastných publikácií autora

- Ondrašovič, Milan, and Peter Tarábek. "*Homography Ranking Based on Multiple Groups of Point Correspondences.*" Sensors 21.17 (2021): 5752.
- Ondrašovič, Milan, and Peter Tarábek. "*Siamese Visual Object Tracking: A Survey.*" IEEE Access 9 (2021): 110149 – 110172.
- Ondrašovič, Milan, Peter Tarábek, and Ondrej Šuch. "*Object Position Estimation from a Single Moving Camera.*" 2021 International Conference on Information and Digital Technologies (IDT). IEEE, 2021.
- Ondrašovič, Milan, and Peter Tarábek. "*Foundations for homography estimation in presence of redundant point correspondencies.*" Mathematics in science and technologies, proceedings of the MIST conference 2020.