VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS INSTITUTAS
PROGRAMŲ SISTEMŲ STUDIJŲ PROGRAMA

Vaizdo žaidimo eismo vaizdų generavimas naudojant nesuporuotų nuotraukų vertimą

Video game traffic image generation using unpaired image to image translation

Kursinis darbas

Atliko: 4 kurso 5 grupės studentas
Gytis Oksas (parašas)

Darbo vadovas: j. asist. Boleslovas Dapkūnas (parašas)

Vilnius – 2023

TURINYS

Įvadas

Paveikslėlio vertimas kitu paveikslėliu (angl. image-to-image (I2I) translation) yra plačiai taikoma metodika kompiuterių regos ir paveikslų apdorojimo problemoms spręsti [PLQ+21]. Metodai, sprendžiantys šias problemas, gali būti taikomi parinktų specifinių paveikslėlių požymių keitimui, trūkstamų paveikslėlio pikselių užpildymui arba realistinių vaizdų generavimui. Šiame darbe pristatomas naujas įrankis generuoti žaidimo eismo vaizdus iš realistinių eismo vaizdų. Ši taikymo sritis gali būti naudinga kuriant žaidimų meninį turinį (pavyzdžiui, tekstūras, realistinių objektų dizaino atitikmenis žaidimui), kuriant animacijas. Ir nors realistinių vaizdų transformavimas į nerealistinių vaizdų domeną nėra naujas (pavyzdžiui, veido nuotraukas transformuoti į animacinių filmukų stilių), kelių eismo nuotraukų transformavimas į vaizdo žaidimų domeną yra labai mažai nagrinėtas.

Šio tyrimo tikslas ir yra - sukurti modelį, kuris generuotų tam tikro žaidimo vaizdus iš jų norimo realistinio atitikmens, šiame tyrime buvo pasirinkta kurti žaidimo „Grand Theft Auto: Vice City" (toliau trumpinama „GTA:VC") domeną dėl jo išskirtinio meninio stiliaus ir ryškaus domeno požymių skirtumo lyginant su realaus eismo nuotraukomis.

Tyrimo tikslas yra sukurti modelį, kuris pasiektų FID įvertį žemesnį nei 100, vizualiai būtų kuo panašesnis į žaidimų vaizdus ir išsaugotų kuo daugiau tikrų nuotraukų detalių.

1. Literatūra ir kiti tyrimai

1.1. Žaidimų vaizdai į realybės vaizdus

Tyrimas kuris atkreipė daug dirbtinio intelekto ir vaizdo žaidimų mėgėjų dėmesį yra aprašytas Intel dirbtinio intelekto mokslininkų straipsnyje ”Enhancing Photorealism Enhancement” [RAK21], kuriame pasakojama, kaip išnaudojus modernias dirbtinio intelekto technologijas yra sukuriamas nuotraukų transformatoriaus modelis, sugebantis žaidimo GTA V automobilių eismo vaizdus paversti foto realistiniais, lyg įrašytais automobilio registratoriumi. Jis vienas pirmųjų įtikinamai, švariai, be prarastų esminių nuotraukų detalių ir be pridėtinių artefaktų sugebėjo transformuoti nuotrauką į realistiškai atrodantį vaizdą (žr. 1 pav.). Tai davė idėją šiam tyrimui - apsukti puses ir sukurti modelį, kuris pagal tuos pačius atributus sugebėtų transformuoti realybės automobilių eismo vaizdus į vaizdo žaidimo automobilių eismo vaizdus.

1 pav. Vaizdai prieš ir po nuotraukos transformavimo, atitinkamai.[RAK21]

1.2. CycleGAN modelis

Vienas iš trijų panaudotų nuotraukų transformavimo modelių yra CycleGAN modelis [ZPI+17], sukurtas 2017 metais, tačiau išpopuliarėjęs savo galimybe vaizdus transformuoti į abi puses, t.y. modelį išmokius nuotraukas transformuoti iš domeno A į domeną B, jis taip pat sugebės vaizdus transformuoti iš domeno B į domeną A, nors nebūtinai taip pat gerai.

Tyrimo straipsnyje yra vaizduojama, kaip modelis sugeba transformuoti vaizdus tarp realistinių domenų (pavyzdžiui, iš arklio į zebrą ir iš zebro į arklį) ir tarp nerealistinių ir realistinių (pavyzdžiui, iš kraštovaizdžio nuotraukų į Monet paveikslus ir atvirškčiai). Tarp dviejų nerealistinių domenų pavyzdžių nėra, tačiau galima nuspėti, jog su kokybišku apmokymo procesu ir duomenų

rinkiniu tokį uždavinį taip pat nesunkiai įveiktų.

Nors modelis jau lygintinai dirbtinio intelekto pasaulyje yra senas, tačiau dėl jo kodo realizacijos prieinamumo ir architektūros paprastumo yra vertas dėmesio ir laiko.

1.3. CUT modelis

Antras iš trijų naudotų modelių yra CUT (pilnas angliškas pavadinimas - contrastive unpaired translation), išleistas 2020 metais. Pagrindinis jo bruožas yra, jog jis yra pritaikytas mokymui su nesuporuotomis nuotraukomis (t.y. nuotraukai iš domeno A, nėra tiesioginio atitikmens iš domeno B). Kitaip nei CycleGAN, CUT mokymas ir nuotraukų transformavimas vykdomas tik į vieną pusę, tai reiškia, kad norint nuotraukas transformuoti iš domeno A į domeną B ir iš domeno B į domeną A, yra reikalingi du atskirai išmokyti modeliai. Dėl vienpusio nuotraukų transformavimo, mokymo procedūra yra supaprastinama ir paspartinama. Šio modelio tikslas yra transformuojant perimti norimo domeno išvaizdą, bet išlaikyti transformuojamos nuotraukos struktūrą ir esminį turinį. Būtent ši CUT modelio savybė ir pakiša koją transformuojant nuotraukas kai kuriuose domenuose, kadangi jei apmokant modelį yra dažnai pasitaikančių artefaktų, tai jie gali dažnai atsikartoti ir transformuojamose nuotraukose, tai yra pabrėžiama "Enhancing Photorealism Enhancement" [RAK21] straipsnyje, kur transformuojant GTA V vaizdus, CUT modelis dažnai ant žaidėjo automobilio variklio dangčio uždėdavo Mercedes žvaigždę, kuri beveik visados matoma Cityscapes duomenų rinkinyje [COR+16], kuriuo ir buvo mokinti modeliai.

1.4. MSPC modelis

Trečiasis naudotas modelis yra MSPC (pilnas pavadinimas angliškai - Maximum Spatial Perturbation Consistency) [XXW+22]. Jis yra sukurtas CycleGAN [ZPI+17] pagrindu, todėl jo nuotraukų transformavimas taip pat yra komutatyvus. Šis modelis yra sukurtas su tikslu jį naudoti nesuporuotų nuotraukų duomenų rinkiniams, kurie dažnai priveda prie nuotraukos turinio išdarkymo. Būtent šią problemą MSPC modelis ir bando spręsti, bandant geriau išsaugoti turinio bruožus ir jų turinį. Šis ir CycleGAN modeliai gali būtų mokomi suporuotais, nesuporuotais ir mišriais duomenų rinkiniais, tačiau sprendžiant šį uždavinį neįmanoma surinkti kokybiško ir kiekybiško suporuoto duomenų rinktinio, todėl yra parinktas šių modelių nesuporuoto mokymo metodas.

2. Metodologija

2.1. Duomenų rinkiniai

2.1.1. Naudojami duomenų rinkiniai

Vaizdų transformavimui iš realistinių į GTA:VC vaizdus reikėjo dviejų duomenų rinkinių: realaus eismo vaizdų ir žaidimo eismo vaizdų. Realaus eismo vaizdų duomenų rinkinių yra pakankamai daug, todėl jų rinkti nereikėjo ir buvo pasirinktas bdd100k rinkinys, o GTA:VC vaizdams reikėjo kurti savo duomenų rinkinį, kadangi kokybiško ir kiekybiško rinkinio internete rasti nepavyko.

2.1.2. Realistinių vaizdų duomenų rinkinys

Kaip minėta, buvo naudojamas bdd100k duomenų rinkinys. Jis sudarytas iš 100 tūkst. automobilių eismo nuotraukų, padarytų naudojant kameras nukreiptas į automobilio važiavimo kryptį ir yra dažniausiai ant automobilio priekinės panelės. Buvo naudotas jo poaibis sudarytas iš 10 tūkst. nuotraukų, kadangi GTA:VC nuotraukų duomenų rinkinys gavosi lygintinai mažas. 10 tūkst. nuotraukų duomenų rinkinys yra padalintas 70:20:10 santykiu mokymui, testavimui ir validacijai (atitinkamai gaunasi 7 tūkst, 2 tūkst ir 1 tūkst). Kiekviena nuotrauka yra 1280:720 pikselių raiškos.

Papildomo apdorojimo ar modifikacijų duomenų rinkiniui nereikėjo, išskyrus paprasčiausią direktorijų keitimą ir aplankų pervadinimą, jog modelio treniravimo algoritmas žinotų, kur tiksliai rasti šaltinio domeno duomenų rinkinį.

2.1.3. Grand Theft Auto: Vice City duomenų rinkinys

Eksperimentą daryti buvo norima su GTA:VC vaizdais, tačiau jo duomenų rinkinio internete rasti nepavyko, todėl teko jį sukurti. Džiugu, jog kokybišką ir kiekybišką duomenų rinkinį leido sukurti žaidimo grafiniai nustatymai, todėl papildomų modifikacijų į žaidimą diegti nereikėjo, vienintelis dalykas, ko prireikė - ekrano vaizdo įrašymo programinės įrangos. Ekrano vaizdo įrašymui naudojama buvo OBS Studio programa.

OBS Studio programos paruošimo nereikėjo, bet buvo pakeistos kelios parinktys, kad kuriamas duomenų rinkinys būtų struktūriškai kuo panašesnis į šaltinio duomenų rinkinį ir vaizdo įrašai nesigautų be reikalo dideli ir sunkūs apdoroti. Pirma pakeista parinktis buvo pakeista raiška į bdd100k rinkinio nuotraukų raišką (kuri yra 1280:720 pikselių). Antras pakeitimas buvo nustatyti vaizdo įrašų kadrų dažnį kiekį į 1 kadrą per sekundę. Su paskutine parinktimi supaprastiname vaizdo įrašų perdarymą į nuotraukas, nes nereikia išmesti perteklinių nuotraukų nustačius didesnį kadrų dažnį (pvz. nustačius standartinius 60 kadrų per sekundę, gautume žymiai per daug nuotraukų).

GTA:VC žaidimui reikėjo kelių vaizdinių pakeitimų tam, kad būtų švaresni vaizdai ir būtų struktūriškai panašesnis į bdd100k duomenų rinkinį. Naudojant standartinius nustatymus yra rodomas vaizdas trečiuoju asmeniu (t.y. iš žaidėjo galo), apatiniame kairiajame kampe yra rodomas mažas žemėlapis ir viršuj dešinėje yra rodomi žaidimo duomenys: laikas, pinigai, gyvybės taškai,

šarvų taškai, esami ginklai ir aktyvumas, kuriuo policija ieško žaidėjo (žr 2 pav.). Tam, kad vaizdas būtų artimesnis šaltinio domeno duomenų rinkiniui, reikėjo panaikinti žemėlapį ir informacines detales, tą žaidimas leido nustatymuose bei reikėjo nustatyti pirmą asmenį ir automobilio perspektyvos, kad nesimatytų pačio veikėjo ir taip išvengti nenorimų artefaktų (panašiai kaip įprastai lieka naudojant [COR+16] duomenų rinkinį, jame filmuojamas vaizdas iš automobilio Mercedes, o kadangi vaizduose yra matomas automobilio kapotas prie kurio pritvirtinta ikoniška Mercedes žvaigždė, todėl daugelyje transformuotų nuotraukų atsiranda minėtoji Mercedes žvaigždė). Šį pakeitimą žaidimas taip pat leidžia daryti, kaip konfiguracinį nustatymą. Šiuos pakeitimus implementavus, gaunamas kokybiškas ir švarus vaizdas, kuris nepalieka artefaktų ir užfiksuoja esminį turinį (žr. 3 pav.).

2 pav. GTA:VC žaidimo vaizdas su įprastais nustatymais

3 pav. GTA:VC žaidimo vaizdas su pakeistais nustatymais

Duomenų rinkinio nuotraukos buvo kuriamos įrašinėjant GTA:VC žaidimo vaizdą, įrašuo-

se važinėjant po žaidimo erdves keliais, bandant padengti kuo daugiau esamų vaizdų. Tas buvo daryta tiek žaidimo dienos metu, tiek nakties, kad būtų sukuriamas duomenų rinkinio aplinkybių vienodumas ir įvairumas. Žaidimo Pasaulis yra suskirstytas į 8 regionus. Kiekvienas regionas buvo išvažinėtas ir nufilmuotas žaidimo dienos ir nakties metu.

Kitame skyriuje yra minima, jog rezultatuose yra pastebėta spragų - vaizdai turi žymiai mažiau automobilių nuotraukose nei bdd100k duomenų rinkinys, todėl po kelių eksperimentų, reikėjo papildyti surinktą duomenų rinkinį nuotraukomis, kuriuose yra daug automobilių arba jie užima didesnį ekrano plotą. Tokių nuotraukų iš viso buvo padaryta 300 ir jas pridėjus prie kitų nuotraukų buvo sukurta antra duomenų rinkinio versija.

Visus regionus nufilmavus, jie buvo apdoroti Python kalbos skriptu, kuris vaizdo įrašo kadrus išskaidė į atskiras nuotraukas. Taip buvo iš viso sudaryta 2000 nuotraukų, o antra duomenų rinkinio versija buvo sudaryta iš 2300 nuotraukų.

2.2. Mokymo platformos

Mokymo proceso kalibravimui ir testavimui buvo naudojama Google Colab platforma. Tačiau mokamų versijų resursai yra limituoti, vienam mokymo ciklui (100 epochų) prireikė mėnesiui skiriamų skaičiavimo resursų (Google Colab Pro prenumerata). Dėl šios problemos bei Google Colab dėl neaktyvumo išjungiamo programos veikimo greitai buvo pereita prie Vilniaus Universiteto matematikos ir informatikos fakulteto superkompiuterio naudojimo. Mokoma buvo naudojant Linux Ubuntu operacinę sistemą (20.04 versiją) su PyTorch karkasu.

2.3. Mokymo procesas

Mokymui buvo išrinkti trys modeliai: CycleGAN, CUT ir MSPC. Kadangi CUT ir MSPC kodo implementacija yra sukurta CycleGAN modelio realizacijos programinio kodo pagrindu [JW22], tiek duomenų rinkinio, tiek pačio mokymo proceso keisti drastiškai nereikėjo.

Kiekvienas modelis buvo mokomas 100 epochų su vientisu mokymo greičiu (0.0002) ir 35 epochomis tiesišku mokymosi greičio nykimu (angl. decay). Mokymosi greitis yra apskaičiuojama formule pavaizduota 4 paveikslėlyje. Kiekvienas modelis buvo apmokytas du kartus, vieną kartą su pirmąja GTA:VC eismo vaizdų duomenų rinkiniu, o antrą kartą su antrąja jo versija, todėl iš viso yra kiekvieno modelio dvi versijos vadinamos V1 ir V2, pavyzdžiui, CUT V1 ir CUT V2.

$$lr = 2 \times 10^{-4} \times \frac{max(0, ep - ep_n)}{ep_d + 1}$$

4 pav. Epchos mokymo greičio skaičiavimo formulė (čia ep - einamoji epocha, ep_n - epochų kiekis, ep_d - nykimo epochų kiekis)

CycleGAN ir MSPC modelių mokymas vyko žymiai ilgiau nei CUT modelio mokymo, dėl jų transformacijų komutatyvumo požymio, kadangi reikia nuotrauką transformuot tiek iš domeno A į

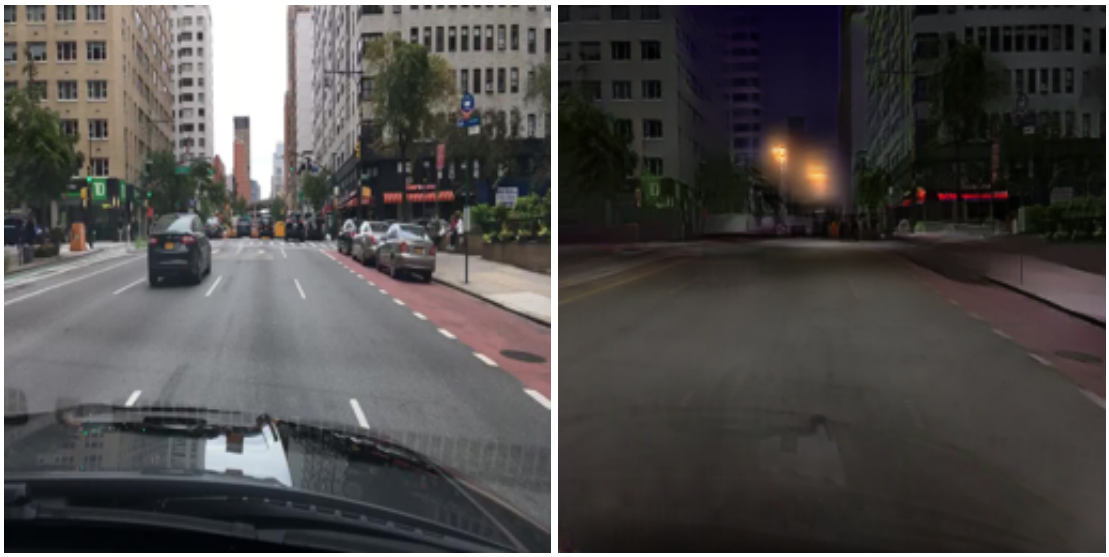
B, tiek iš B į A ir MSPC modelio mokymas truko ilgiau nei CycleGAN dėl mažesnio skaičiavimų kiekio, reikalingų minėtam vaizdo detalių palaikymui.

1 lentelė. Modelių mokymo trukmės

Modelio pavadinimas	Mokymo laikas, minutės
Cut V1	1524
Cut V2	1520
MSPC V1	4136
MSPC V2	3897
CycleGAN V1	2806
CycleGAN V2	2620

2.4. „Dingstančių automobilių" problema

Išmokius CycleGAN modelį buvo pastebėta, jog pradėjo dingti tam tikros stambios detalės iš nuotraukų, svarbiausia iš jų - automobiliai. Išmokyta pirmoji CycleGAN modelio versija linksta trinti automobilius ir uždengti tai aplinkos detalėmis ar fonu. Dėl to nuspręsta buvo išbandyti CUT modelį, kadangi jis pritaikytas nesuporuotų duomenų rinkinių uždaviniams, tačiau pirmoji jo versija yra dar labiau linkusi trinti automobilius iš nuotraukų. Dėl šios problemos taip pat buvo išbandytas MSPC modelis, nes juo yra sprendžiama nuotraukų semantinių detalių palaikymo problema. Nors MSPC modelis pasiekia geresnius rezultatus nei kiti du ties šia problema, tačiau išlaikomų automobilių kiekis vis tiek liko labai mažas (žr. 2 lentelę).



5 pav. CycleGAN modelio transformuotos nuotraukos pavyzdys, kuriame yra panaikinami automobiliai.

2 lentelė. Atpažintų automobilių kiekis duomenų rinkiniuose.

Modelio pavadinimas	Atpažintų automobilių kiekis, vnt.	Atpažintų automobilių dalis, %
Testavimo duomenų rinkinys	426	-
CycleGAN V1	15	3,5 %
CycleGAN V2	14	3,2 %
Cut V1	7	1,6 %
Cut V2	10	2,3 %
MSPC V1	10	2,3 %
MSPC V2	17	3,9 %

Automobilių atpažinties nuotraukose testavimui naudojamas Centernet HG-104 [DBX+22] modelis iš anksto apmokytas su Microsoft CoCo duomenų rinkiniu [LMB+14].

Modeliai išmokyti su patobulintu duomenų rinkiniu, kuriame yra daugiau nuotraukų, kuriose yra automobilių užimančių pakankamai didelį nuotraukos plotą, arba yra nuotraukų su daug automobilių nuotraukoje, demonstruoja geresnį sugebėjimą išlaikyti automobilius (išskyrus su CycleGAN modeliu, kurio statistika minimaliai sumažėjo). Atpažintų automobilių padidėjimas siekia 69% su MSPC modeliu, tačiau toks patobulėjimas yra beveik bereikšmis kadangi bendras atpažintų automobilių procentas lieka labai mažas.

Rezultatai ir išvados

Santrumpos

1. **GTA:VC** - Žaidimas „Grand Theft Auto: Vice City”.
2. **GTA V** - Žaidimas „Grand Theft Auto 5”.
3. **FID** - angliškai Fréchet inception distance, o lietuviškai Frečeto pradžios atstumas, yra metrika naudojama įvertinti generatyvinių modelių kokybę.

Literatūra

[COR+16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth ir Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding, 2016. DOI: 10.48550/ARXIV.1604.01685. URL: https://arxiv.org/abs/1604.01685.

[DBX+22] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang ir Qi Tian. CenterNet++ for Object Detection, 2022. DOI: 10.48550/ARXIV.2204.08394. URL: https://arxiv.org/abs/2204.08394.

[JW22] Taesung Park Jun-Yan Zhu ir Tongzhou Wang. CycleGAN and pix2pix in PyTorch, 2022. URL: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev ir k.t. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312, 2014. arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

[PLQ+21] Yingxue Pang, Jianxin Lin, Tao Qin ir Zhibo Chen. Image-to-Image Translation: Methods and Applications, 2021. DOI: 10.48550/ARXIV.2101.08629. URL: https://arxiv.org/abs/2101.08629.

[RAK21] Stephan R. Richter, Hassan Abu AlHaija ir Vladlen Koltun. Enhancing Photorealism Enhancement. *arXiv:2105.04619*, 2021.

[XXW+22] Yanwu Xu, Shaoan Xie, Wenhao Wu, Kun Zhang, Mingming Gong ir Kayhan Batmanghelich. Maximum Spatial Perturbation Consistency for Unpaired Image-to-Image Translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 18311–18320, 2022-06.

[ZPI+17] Jun-Yan Zhu, Taesung Park, Phillip Isola ir Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.