

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

**Vaizdo žaidimo eismo vaizdų generavimas
naudojant nesuporuotų vaizdų transformacijas**

**Video game traffic image generation using unpaired image
to image translation**

Bakalauro baigiamasis darbas

Atliko:	Gytis Oksas	(parašas)
Darbo vadovas:	j. asist. Boleslovas Dapkūnas	(parašas)
Darbo recenzentas:	doc. dr. Linas Petkevičius	(parašas)

Vilnius – 2025

Santrauka

Pristatomas metodas vaizdo žaidimo eismo nuotraukų generavimui pasitelkiant porą generatyviųjų adverziniai tinklų ir vaizdo segmentavimo modelį, taip sukuriant metodą sugebantį transformuoti nuotrauką į norimą domeną išsaugant norimos domeno klasės detales. Šiame straipsnyje aprašoma, kaip šiuo būdu galima transformuoti realaus eismo nuotrauką į vaizdo žaidimo eismo nuotrauką, atskirai transformuojant svarbiausią objekto klasę - automobilius. Papildomai, rezultatai yra palyginami su naujais iš anksto apmokytais difuzijos modeliais.

Gautas metodas nors ir sugeba tam tikrais atvejais sugeneruoti kokybiškesnę nuotrauką nei pavienis GAN modelis, tačiau yra mažiau stabilus ir gauna prastesnius įverčius. Difuzijos modeliai generuoja švaresnį vaizdą ir daug geriau išsaugo detales, tačiau arba per stipriai pakeičia vaizdą, arba ji pakeičia per mažai.

Raktiniai žodžiai: **Generatyviniai adverziniai tinklai, segmentavimo modelis, difuzijos modeliai, realaus eismo vaizdai, vaizdo žaidimų eismo vaizdai**

Summary

A method of generating video game traffic images is presented, using a pair of generative adversarial networks and an image segmentation model, thus creating a method capable of transforming the image to the desired domain while preserving the details of domain's most important details. In this paper it is described how using this method is it possible to transform images from a real traffic domain into a video game traffic domain while separately transforming its most important image class - cars. Additionally, results are compared against recently released pretrained diffusion models.

The created method, while capable of generating better images than a standalone GAN under certain conditions, is less stable and is worse by evaluation metrics. While diffusion models generate images with much less noise and preserve details much better, they tend to either change the content too much or barely change it at all.

Keywords: **Generative adversarial networks, diffusion models, segmentation model, real traffic images, video game traffic images**

TURINYS

SANTRAUKA	2
SUMMARY	3
PADĖKA UŽ HPC IŠTEKLIUS	6
1. VAIZDŲ TRANSFORMAVIMO MODELIAI	9
1.1. GAN modelis	9
1.1.1. CycleGAN modelis	9
1.1.2. CUT modelis	9
1.1.3. MSPC modelis	10
1.2. Difuzijos modelis	10
1.2.1. ControlNet-seg modelis	11
1.2.2. CycleDiffusion modelis	12
1.2.3. InstructPix2Pix modelis	12
2. EKSPERIMENTAS	13
2.1. Duomenų rinkiniai	13
2.1.1. Realių vaizdų duomenų rinkinys	13
2.1.2. Grand Theft Auto: Vice City duomenų rinkinys	13
2.1.3. Automobilių duomenų rinkinys	15
2.2. Mokymo platformos	16
2.3. Pirminių modelių mokymas	17
2.3.1. „Dingstančių automobilių“ problema	17
2.3.2. Paros meto keitimosi problema	19
2.3.3. Kokybiskai transformuoojamos detalės	19
2.4. Modifikuoto MSPC modelio kūrimas ir mokymas	20
2.5. Panaudoti iš anksto apmokyti difuzijos modeliai	22
2.5.1. ControlNet-Seg modelis	22
2.5.2. InstructPix2Pix modelis	23
2.5.3. CycleDiffusion modelis	25
2.6. Modelių vertinimo metrikos	25
2.6.1. FID	25
2.6.2. SSIM	26
2.7. Gauti įverčiai	26
2.7.1. FID įverčiai	26
2.7.2. Atpažintų automobilių kiekiai	27
2.7.3. SSIM įverčiai	27
REZULTATAI IR IŠVADOS	29
SANTRUMPOS	30
ŠALTINIAI	31
PRIEDAI	32
1 priedas. Darbo saugykla	33
2 priedas. Pirminių modelių (CycleGAN, CUT, MSPC) nuotraukų pavyzdžiai (pirmosios duomenų rinkinio iteracijos)	34
3 priedas. Pirminių modelių (CycleGAN, CUT, MSPC) nuotraukų pavyzdžiai (antrosios duomenų rinkinio iteracijos)	35
4 priedas. Difuzijos modelių nuotraukų pavyzdžiai	36

Padėka už HPC išteklius

Darbo autorius dėkoja Vilniaus universiteto Matematikos ir informatikos fakulteto Informaciinių technologijų atviros prieigos centru už suteiktus HPC išteklius šio darbo tyrimams atlikti.

Ivadas

Vaizdų transformacija (angl. *image-to-image 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. Ši taikymo sritis gali būti naudinga kuriant žaidimų meninį turinį (pavyzdžiu, tekštūras, realistinių objektų dizaino atitinkmenis žaidimui), kuriant animacijas. Ir nors realių vaizdų transformavimas į nerealistinių vaizdų domeną néra naujas (pavyzdžiu, veido nuotraukas transformuoti į animacinių filmukų stilių, kaip yra taikoma MSPC modelio pavyzdžiuose [XXW⁺22]), jis vis dar yra aktyviai analizuojamas, kadaangi yra likusių ir pastoviai iškylančių problemų ir iššūkių (pavyzdžiu, vaizde daug triukšmo, besikartojantys artefaktai, nelogiški piešiniai). Vaizdo žaidimų vaizdų transformavimas į realistišius vaizdus yra spręstas, tačiau atvirkštinis variantas yra gan mažai išnagrinėtas.



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

Tyrimas kuris atkreipė daug dirbtinių neuroninių tinklų ir vaizdo žaidimų mėgėjų dėmesį yra aprašytas Intel dirbtinių neuroninių tinklų mokslininkų straipsnyje „Enhancing Photorealism Enhancement“ [RAK21], Jame pristatomas metodas, kaip išnaudojus modernias dirbtinių neuroninių tinklų technologijas yra sukuriamas nuotraukų transformatoriaus modelis, sugebantis žaidimo GTA V automobilių eismo vaizdus paversti foto realistiniai, 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.

Darbo tikslas

Šio tyrimo tikslas ir yra sukurti metodą, gebantį transformuoti realias eismo nuotraukas į vaizdo žaidimo vaizdus, tačiau neprarandant esminių semantinių objektų (šiuo atveju – automobilių). Pasirinkta buvo transformuoti nuotraukas į 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 uždaviniai:

1. Sukurti vaizdo žaidimo duomenų rinkinį.
2. Sukurti vaizdo žaidimo automobilių duomenų rinkinį.
3. Sukurti metodą sugebantį atskirai transformuoti tam tikrą nuotraukos semantinę klasę ir gautą rezultatą atgal įklijuoti į nuotrauką.
4. Palyginti modelius.

1. Vaizdų transformavimo modeliai

Šiame skyriuje yra apžvelgiami pagrindiniai eksperimente naudoti modeliai bei jų versijos, aprašomi jų veikimo principai, privalumai bei istorija.

1.1. GAN modelis

Pirmasis generatyvinis adversarinis tinklas (angl. *generative adversarial network*) buvo pri-
statytas 2014 metais Montrealio universiteto mokslininkų [GPM⁺14]. Jo principas sudarytas iš
dviejų dalių. Pirmoji dalis yra generatyvinis modelis G , kuris išmoksta nuotraukų domeno požy-
mius ir pagal juos sugeba kurti nuotraukas. Antroji – diskriminacinis modelis D , kurio tikslas yra
gavus nuotrauką nuspėti tikimybę ar nuotrauka yra iš mokomo domeno, ar yra sugeneruota mo-
delio G . Šiuo metodu, generatyvinis modelis G yra pastoviai mokomas geriau kurti nuotraukas,
kurios apgautų diskriminacinių modelių D , o diskriminacinis modelis D yra pastoviai mokomas
geriau atpažinti sugeneruotą nuotrauką nuo tikros.

Šis modelis sugeba tik sukurti naujas nuotraukas, jų netransformuoja. Todėl jam reikia
tik vieno duomenų rinkinio. Toliau šiame skyriuje minimi modeliai kuria nuotraukas jas trans-
formuojant iš pirminės, todėl jiems reikalinga nuotraukos įvestis, jei taip pat reikalauja dviejų
duomenų rinkinių – vieną šaltinio domeną ir vieną tikslo domeną.

1.1.1. CycleGAN modelis

Vienas iš trijų panaudotų nuotraukų transformavimo modelių yra CycleGAN modelis [ZPI⁺17]. Šis modelis sukurtas 2017 metais ir išpopuliarėjo dėl savo galimybės vaizdus trans-
formuoti į abi puses, t.y. modelių 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.
Tai nulemia, jog modelių sudaro du generatyviniai modeliai ir du diskriminaciniai modeliai.

Tyrimo straipsnyje yra vaizduojama, kaip modelis sugeba transformuoti vaizdus tarp realių
domenų (pavyzdžiui, iš arklio į zebrą ir iš zebro į arklių) ir tarp ne nerealistinių ir realių (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 yra senesnis nei dauguma šiuo metu populiarūjų nuotraukų transformavimo
modelių architektūrų, tačiau dėl jo kodo realizacijos prieinamumo ir architektūros paprastumo
yra vertas dėmesio ir laiko.

1.1.2. CUT modelis

Antras iš trijų naudotų modelių yra CUT (angl. *contrastive unpaired translation* – kontrastyvi
neporuota transformacija), išleistas 2020 metais. Pagrindinis jo bruožas yra, jog jis yra pritaikytas
mokymui su nesuporuotais duomenų rinkiniais (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 tuo pačiu yra šio modelio didžiausia stiprioji ir silpnoji vieta transformuojant nuotraukas kai kuriuose domenuose, kadangi jei apmokant modelį yra dažnai pasitaikančių artefaktų, tai jie gali dažnai atsikartoti ir transformuojamose nuotraukose, tai yra pabrėžiamas „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.

1.1.3. MSPC modelis

Trečiasis naudotas modelis yra MSPC (angl. *Maximum Spatial Perturbation Consistency* – maksimalus erdinės perturbacijos pastovumas) [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 ši uždavinį neįmanoma surinkti kokybiško ir kiekybiško suporuoto duomenų rinktinio, todėl yra parinktas šių modelių nesuporuoto mokymo metodas.

1.2. Difuzijos modelis

Difuzijos tikimybinis modelis, bendrai vadinamas difuzijos modeliu, buvo pristatytas 2015 metais Stanfordo universiteto mokslininkų [SWM⁺15], nors tuo metu šis metodas buvo pavaudintas giliuoju neprižiūrimu mokymusi naudojant nepusiausvirą termodinamiką. Šis metodas naudoja Markovo grandinės principą, kad palaipsniui paversti vieną duomenų pasiskirstymą kitu, t.y. įvesti paversti išvestimi. Dėl šio sprendimo, šios architektūros veikimas yra kompleksiškesnis ir daugiau resursų reikalaujantis nei GAN architektūros, dėl reikalingo kiekvieno Markovo žingsnio apskaičiavimo, kurių dažnai būna dešimtys, šimtai ar net daugiau, ir tai yra pagrindinis šių modelių minusas.

Pirmasis difuzijos modelis pritaikytas nuotraukų generavime buvo pristatytas Berklio Kalifornijos universiteto mokslininkų 2020 metais [HJA20]. Kaip ir praetame skyriuje apie GAN modelius, šis modelis sugeba tik generuoti nuotraukas ir jų netransformuoja, tačiau toliau pristatomi modeliai transformuoja įvesties nuotraukas į išvestį.

Pirmasis difuzijos modelis generuojantis nuotraukas pagal tekstą buvo pristatytas OpenAI kompanijos pavadinimu DALL-E. Šis modelis veikia paprastu principu – gauna teksto įvestį ir pagal tai sugeneruoja nuotrauką. Toliau pristatomi modeliai veikia panašiu principu – įvesta nuotrauka yra transformuojama pagal tekstinę įvestį (žr. 2 pav.).

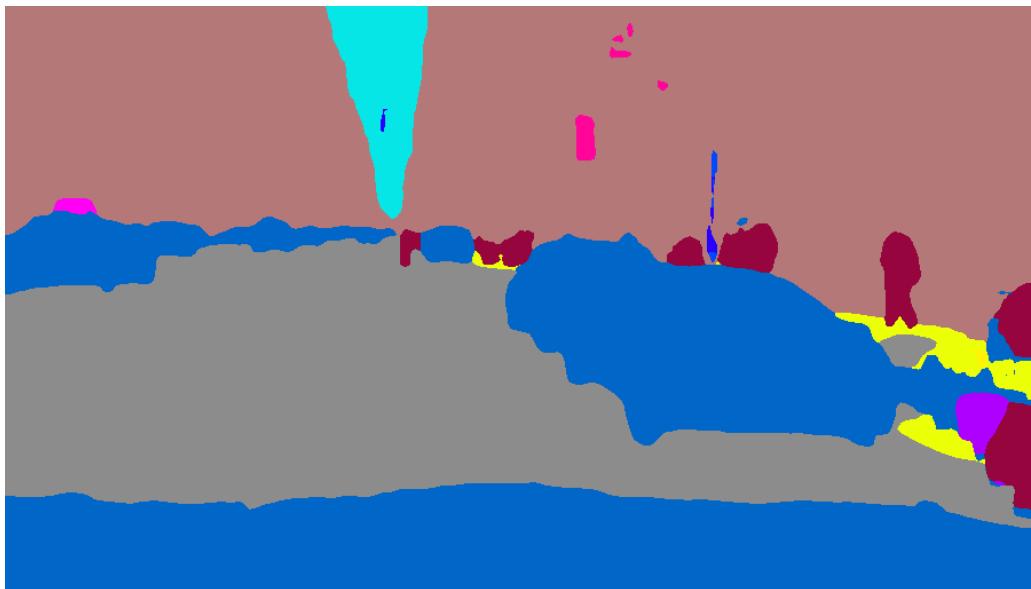


2 pav. Vaizdai prieš ir po nuotraukos transformavimo, naudojant ControlNet-Seg modelį su tekstiniu raginimu „Vice City car traffic“.[ZA23a]

Kol kas šie modeliai pritaikyti mokyti tik suporuotų duomenų rinkiniams, todėl toliau modelių mokymo galimybės ir aspektai nėra minimi.

1.2.1. **ControlNet-seg modelis**

Pirmas panaudotas difuzijos modelis yra ControlNet-Seg modelis, pristatytas 2023 metais ir yra ControlNet modelio [ZA23b] versija, kuri transformuoja nuotrauką remiantis segmentavimo rezultatais, t.y. pradžioje nuotrauka yra segmentuojama ir pagal gautas kaukes, skirtinges vietas yra transformuojamos skirtingai, išlaikant kiekvienos klasės turinį (žr. 3 pav.).



3 pav. Transformuoojamos nuotraukos segmentacijos kaukės.

Šis modelis taip pat leidžia transformuoti atsižvelgiant į turinio kraštus, spalvų šuolio kraštų kaukes, asmens pozą (jei transformuojama žmogaus nuotrauka), nuotraukos turinio gylį.

Šie modeliai taip pat leidžia pridėti papildomus ir vengtinus tekstinius nurodymus, kurie dažnai naudojami kontroliuoti kokybei. Pavyzdžiu, numatytojos reikšmės papildomiem tekstiniams nurodymams yra „best quality, extremely detailed“, o numatyti neigiami nurodymai yra „longbody, lowres, bad anatomy, bad hands, missing fingers, extra digit, fewer digits, cropped, worst quality, low quality“.

1.2.2. CycleDiffusion modelis

Antrasis panaudotas difuzijos modelis yra CycleDiffusion, pristatytas 2022 metų spalį ir jo tikslas yra simetruoti nuotraukos generavimą iš latentinių vektorių (angl. *latent space*) ir vektorių generavimą iš nuotraukos, taip iš esmės yra „sutaisomas“ atsitiktinis sėklos kodas. Tai leidžia modeliui generuoti beveik identiškas nuotraukas iš to pačio sėklos kodo.

Šis modelis kaip įvesti priima tris parametrus: nuotrauką, nuotraukos apibūdinimą ir norimos nuotraukos transformavimą. Papildomas parametras leidžia detaliau apibūdinti turinį arba atkreipti dėmesį į reikalingas detales, taip leidžiant vartotojui pagal jo poreikius modifikuoti kas yra transformuojama ir kaip.

1.2.3. InstructPix2Pix modelis

Trečiasis panaudotas difuzijos modelis yra InstructPix2Pix [BHE22] modelis, pristatytas 2023 metais, sukurtas StableDiffusion modelio pagrindu [RBL⁺21]. Nors StableDiffusion modelio įvestis yra tekstas iš kurio yra sugeneruojama nuotrauka, InstructPix2Pix priima nuotrauką ir tekstą, iš kurių sugeneruoja naują nuotrauką.

Kaip galima nuspėti iš modelio pavadinimo, jis veikia „paliepimo“ principu, kur tekstinė įvestis yra kaip komanda, liepanti pakeisti nuotraukos turinį pagal instrukcijas, pavyzdžiu, „nuspalvink švarką mėlynai“ ar „pakeisk paros metą į naktį“.

2. Eksperimentas

Mokymo procesas vykdomas pagal keturis pagrindinius žingsnius:

1. Duomenų rinkinio paruošimas,
2. Modelių paruošimas;
3. Mokymas;
4. Apmokyti difuzijos modelių paruošimas ir panaudojimas
5. Rezultatų išvedimas.

Jei rezultatai nėra patenkinami, tai ciklą kartojama nuo numanomo problemos taško. Šiame skyriuje aprašomi šio proceso žingsniai.

2.1. Duomenų rinkiniai

Vaizdų transformavimui iš realių į GTA:VC vaizdus reikia dviejų duomenų rinkinių: realių eismo vaizdų ir žaidimo eismo vaizdų. Realaus eismo vaizdų buvo pasirinktas BDD100K [YCW⁺18] rinkinys, o GTA:VC vaizdams buvo kurtas naujas duomenų rinkinys.

2.1.1. Realių vaizdų duomenų rinkinys

Kaip minėta, buvo naudojamas BDD100K duomenų rinkinys. Jis sudarytas iš 100 000 automobilių eismo nuotraukų, padarytų naudojant kameras nukreiptas į automobilio važiavimo kryptį ir yra dažniausiai ant automobilio priekinės panelės. Buvo naudotas jo poaibis sudarytas iš 10 000 nuotraukų (dar vadinamas kaip bdd10k), kadangi GTA:VC nuotraukų duomenų rinkinys susidaro lygintinai mažas. 10 tūkst. nuotraukų duomenų rinkinys yra padalintas 70:20:10 santykliu mokymui, testavimui ir validacijai (atitinkamai gaunasi 7 tūkst, 2 tūkst ir 1 tūkst), nes kitaip rinkinys gautųsi stipriai nesubalansuotas. Kiekviena nuotrauka yra 1280:720 pikselių raiškos.

2.1.2. Grand Theft Auto: Vice City duomenų rinkinys

Eksperimentas buvo vykdomas su GTA:VC vaizdais, kurių duomenų rinkinį teko sukurti. Kokybiską duomenų rinkinį leidžia sukurti žaidimo grafiniai nustatymai, todėl papildomų modifikacijų į žaidimą diegti nereikia, užtenka ekrano vaizdo įrašymo programinės įrangos. Ekrano vaizdo įrašymui buvo naudojama OBS Studio programa.

OBS Studio programos nustatyti parametrai, kad kuriamas duomenų rinkinys būtų struktūriškai kuo panašnis iš šaltinio duomenų rinkinį ir vaizdo įrašai nesigautų per dideli ir sunkiai apdorojami. Igyvendinti du parametrų pakeitimai: įrašo raišką pakeista į BDD100K rinkinio nuotraukų raišką (kuri yra 1280:720 pikselių) ir nustatyta vieno kadro per sekundę įrašo sparta. Su paskutine parametru yra supaprastinamas vaizdo įrašų perdarymas į nuotraukas, nes nereikia išmesti perteklinių nuotraukų nustačius didesnį kadrą dažnį (pvz. nustačius standartinius 60 kadrių per sekundę, gaunama žymiai per daug nuotraukų).

GTA:VC žaidimas reikalauja kelių vaizdinių pakeitimų tam, kad vaizdai būtų švaresni ir struktūriškai panašesni į BDD100K duomenų rinkinį. Naudojant standartinius nustatymus yra rodomas vaizdas trečiuoju asmeniu (t.y. iš žaidėjo galio), kuriame matyti daug interfeiso elementų

(žr. 4 a) 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, tame 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). Ši pakeitimą žaidimas taip pat leidžia daryti, kaip konfiguracinių nustatymų. Šiuos pakeitimus implementavus, gaunamas kokybiškas ir švarus vaizdas, kuris nepaliela artefaktų ir užfiksuoja esminį turinį (žr. 4 b) pav.).



4 pav. GTA:VC žaidimo vaizdas su įprastais nustatymais (a) ir su pakeistais nustatymais (b).

Duomenų rinkinio nuotraukos buvo kuriamos įrašinėjant GTA:VC žaidimo vaizdą, įrašuose 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 vienodus ir įvairumas. Žaidimo erdvė yra suskirstyta į 8 regionus. Kiekvieno regiono aplinka yra nufilmuojama žaidimo dienos ir nakties metu.

Po pirminių eksperimentų 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.

Vaizdo įrašai apdoroti Python kalbos skriptu, konvertuojančiu vaizdo įrašą į atskiras kadru nuotraukas, taip sudarydamas GTA:VC vaizdų duomenų rinkinį sudarytą iš 2000 nuotraukų, o jo antrą versiją iš 2300.

2.1.3. Automobilių duomenų rinkinys

Apmokius pirminius modelius, buvo pastebėta, kad dingsta automobiliai ir nors vaizdo žaidimo duomenų rinkinys buvo papildytas nuotraukomis, kuriose yra automobilių, rezultatai vis dar buvo prasti. Dėl šios priežasties buvo nuspręsta kurti atskirą modelį, transformuojantį tik nuotraukas, kuriose yra automobiliai, t.y. viskas kas nepriklauso automobilių klasei – lieka paverčiama permatoma. Tam pasirinktas *PNG* nuotraukų formatas, kuris sudarytas iš 4 kanalų – raudona, žalia, mėlyna ir permatomumo (dar vadinama *alpha* kanalu).

Šio duomenų rinkinio generavimui buvo pasitelkti atsisiųstas BDD100K ir surinktas GTA:VC duomenų rinkiniai bei nuotraukų segmentavimo modelis. Buvo pasirinktas DeepLabV3 segmentacijos modelis su ResNet-101 pagrindu, prieinamas PyTorch segmentacijos modelių bibliotekoje. Pagal numatytyus nustatymus, šis modelis atsiūlė vieną iš 19 klasės segmentacijos rezultatų, kurie nurodė kiekvienam pikseliui tikimybę, jog Jame yra atvaizduojamas automobilis. Tam, kad pikselis būtų pilnos reikšmės, arba jis būtų visiškai permatomas, reikėjo suapvalinti kiekvieno pikselio tikimybę iš 1 arba 0. 0,5 slenkstinė reikšmė gavosi per maža ir automobilių nuotraukose gavosi dideli tarpių tušumas, dėl to pasirinkta buvo 0,25 reikšmė, kuri suteikia pilnesnį automobilių vaizdą (žr. 5 pav.). Gauta kaukė yra panaudojama kaip *alpha* kanalo reikšmė nuotraukai, taip paverčiant pikselius, kuriuose spėjama, jog nėra automobilio, permatomais.

Tas pats procesas buvo pritaikytas ir žaidimo vaizdams maskuoti. Šis modelis apmokytas realiomis nuotraukomis gan gerai įveikė užduotį ir gavosi patenkinamos kokybės nuotraukos (žr. 6 pav.). Daug nuotraukų gavosi tuščios, t.y. be automobilių ir dėl to tapo juodos, arba labai prasenos kokybės su iškirptomis automobilių dalimis arba nereikalingais artefaktais, dėl šių priežasčių daug nuotraukų teko ištinti ir duomenų rinkinys sumažėjo nuo 6893 iki 2474 nuotraukų realiam domenui ir nuo 2279 iki 410 nuotraukų žaidimo domenui.



5 pav. Automobilių duomenų rinkinio realaus domeno nuotraukos pavyzdys.



6 pav. Automobilių duomenų rinkinio GTA:VC domeno nuotraukos pavyzdys.

2.2. Mokymo platformos

Paruošus modelius, jie buvo mokomi VU MIF superkompiuteriu. Mokoma buvo naudojant Linux Ubuntu operacinę sistemą (20.04 versiją) su PyTorch karkasu, pasitelkiant superkompiutero Nvidia DGX-1 stotį, kurią sudaro keturios Nvidia Tesla V100 vaizdo plokštės.

2.3. Pirminių modelių mokymas

Pirminiam mokymui buvo išrinkti trys modeliai: CycleGAN¹, CUT² ir MSPC³. Kadangi CUT ir MSPC kodo implementacija yra sukurta CycleGAN modelio realizacijos programinio kodo pagrindu, tiek duomenų rinkinio, tiek pačio mokymo proceso keisti drastiškai nereikėjo. Parinkti modeliai nebuvo modifikuoti ar kaip nors kitaip keisti.

Kiekvienas modelis buvo mokomas 100 epochų su vientisu mokymo greičiu (0.0002) ir 35 epochomis tiesišku mokymosi greičio nykimu (angl. *decay*). 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. Mokymosi greitis yra apskaičiuojama pagal šią formulę:

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

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 transformuoti tiek iš domeno A į B, tiek iš B į A ir MSPC modelio mokymas truko ilgiau nei CycleGAN dėl mažesnio skaičiavimų kiekiečių 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

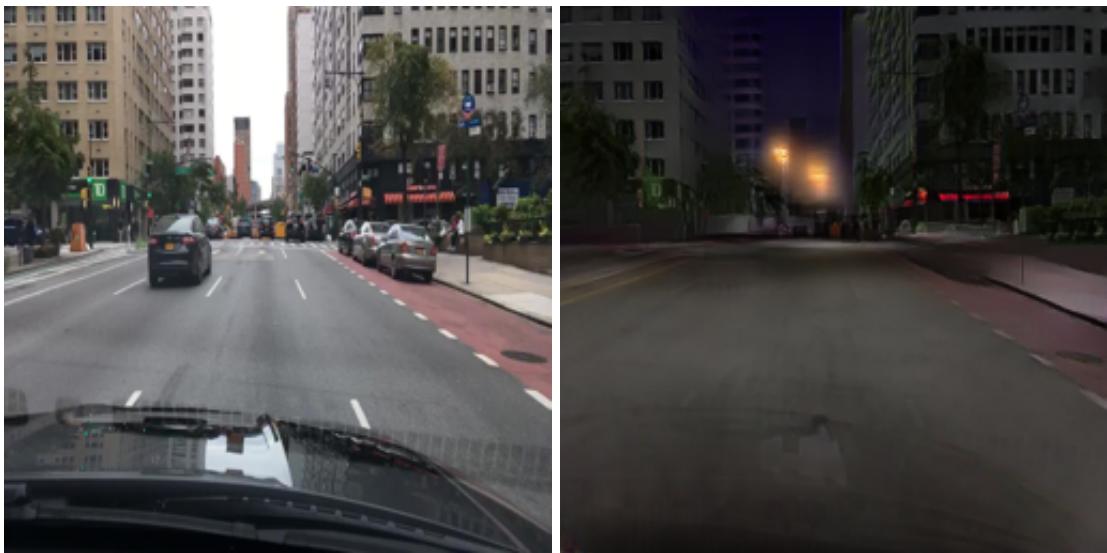
2.3.1. „Dingstančių automobilių“ problema

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 nuspresta 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).

¹CycleGAN kodo realizacija pasiekama nuoroda – <https://github.com/junyanz/PyTorch-CycleGAN-and-pix2pix>

²CUT modelio kodo realizacija pasiekama nuoroda – <https://github.com/taesungp/contrastive-unpaired-translation>

³MSPC modelio kodo realizacija pasiekama nuoroda – <https://github.com/batmanlab/MSPC>



7 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]. Šis modelis sugeba pakankamai gerai atpažinti automobilius net ir vaizdo žaidimuose, tokiuose, kaip GTA:VC.

Modeliai apmokyti 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.

Gali būti daug priežasčių, kodėl modeliai yra taip linkę trinti automobilius iš nuotraukų, tačiau labiausiai tikėtina iš jų yra automobilių kiekių duomenų rinkiniuose – GTA:VC duomenų rinkinyje jų yra žymiai mažiau. 3 lentelėje yra pateikti mokymo duomenų rinkiniuose esantys automobilių kiekių ir kokia dalis nuotraukų turi juose atpažintą bent vieną automobilį.

3 lentelė. Duomenų rinkinių atpažintų automobilių kiekiai ir nuotraukų.

Duomenų rinkinys	Atpažintų automobilių kiekis, vnt.	Dalis nuotraukų kuriuose yra atpažinta automobilių, %
BDD100K	21624	89 %
GTA:VC V1	274	11 %
GTA:VC V2	653	23 %

2.3.2. Paros meto keitimosi problema

Gautose nuotraukose GAN modelių transformuotose nuotraukose pastebėta problema, jog tam tikrais atvejais yra pakeičiamas paros metas iš dienos į naktį. Tas dažniausiai įvyksta kai būna debesuota arba kai yra mažai matoma dangaus, tarkim kai aplinkui yra daug aukštų pastatų arba gamtos (žr. 8 pav.).



8 pav. CycleGAN V2 pavyzdžiai kur modelio dėl aukštų pastatų arba debesų yra diena pakeičiamama į naktį.

2.3.3. Kokybiskai transformuojamos detalės

Modeliai itin gerai sugeba transformuoti gamtinius objektus, kaip žolę ir medžiai, iš realaus domeno į vaizdo žaidimo domeną. GTA:VC duomenų rinkinyje yra daug žaidime paplitusių palmių, kurios atrodo maždaug vienodai ir yra pavaizduotos 6 paveiksluke.



9 pav. GTA:VC domeno palmių pavyzdys.

Modelis medžius sugeba gerai atskirti ir juos transformuoti į GTA:VC domene matomus. Iš tamsiai žalios į šviesiai žalią spalvą pereinantys lapai yra nupiešiami medžiams ir net kamienai yra transformuojami į dryžuotai rudus. Ši elgsena yra būdinga visiems trims modeliams ir abejoms kiekvieno jų versijoms.



10 pav. MSPC V2 pavyzdžiai kur modelio dėl aukštų pastatų arba debesų yra diena pakeičiama į naktį.

Kitos dvi gerai transformuojamos detalės yra kelias ir dangus (kelio ir dangaus pavyzdys matomas 6 pav.). Dangus beveik visada įgauna išprastą saulėtos dienos mėlyną arba giedros nakties juodą spalvą. Nors, kaip prieš tai minėta, dangus kartais pakeičia dienos laiką, tačiau ir naktinis dangus yra tiksliai transformuojamas. Asfalto spalva taip pat yra labai gerai parenkama. Nors šios detalės yra labai paprastai transformuojamos, tačiau jos sudaro didelę dalį bendros transformuotos nuotraukos kokybės ir panašumo į vaizdo žaidimo domeną.

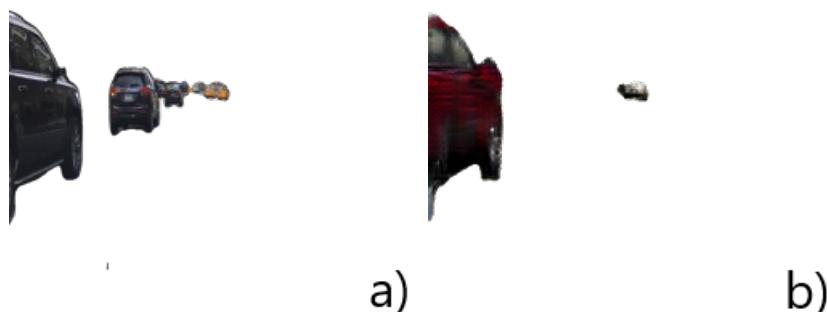
2.4. Modifikuoto MSPC modelio kūrimas ir mokymas

Po pirminių modelių mokymo rezultatų, buvo pastebėta, kad dingsta daug automobilių ir dėl to buvo nuspręsta sukurti metodą, sugebantį išskirti automobilius iš nuotraukos, juos transformuoti atskirai nuo bendro turinio ir po to gautą rezultatą įklijuoti į transformuotą nuotrauką,

Šis modelis toliau vadinamas „MSPC Car“. Kadangi po pirminių mokymų geriausius rezultatus davė MSPC modelis, todėl naujam metodui buvo pasirinktas būtent šis modelis. MSPC modelis transformuojantis bendrą vaizdą buvo pirminiame eksperimente apmokytas MSPC V2 modelis, o modelis transformuojantis tik automobilius buvo minimaliai modifikuotas MSPC modelis. Buvo atliktos dvi modifikacijos: pridėtas ketvirtasis įvesties bei išvesties kanalas ir papildyta paklaidos formulė.

Pagal numatytą kodą, modelis palaiko tik du režimus – spalvotas ir juodai baltas nuotraukas. Ši modelių reikėjo pritaikyti transformacijai nuotraukų su permatomumo sluoksniu, tam kodel reikėjo modifikuoti nuotraukos apdorojimo prieš paduodant modeliui funkcijas bei įvesties ir išvesties sluoksnius, kad palaikytų 4 dimensijų aibes.

Antroji modifikacija buvo atlikta po to, kai buvo pastebėta, jog transformuotos nuotraukos dažnai prarado turinį (t.y. toje vietoje, kur buvo automobilis, staiga ta vieta tapo permatoma) arba atsirado turinio, kur prieš tai buvo tuščia vieta. Ši paklaidos skaičiavimo formulė yra paprasta kvadratinės paklaidos formulė ir jos rezultatas pridedamas prie bendros paklaidos. Ši papildoma paklaida skirta „nubausti“ modelių už matomo turinio vėtės keitimą.



11 pav. Automobilio nuotraukos pavyzdys prieš (a) ir po (b) transformacijos.

Šis modelis apmokytas buvo tuo pačiu principu kaip modeliai praeitame skyriuje, išskyrus buvo sumažintas skaičius epochų iki 50 siekiant išvengti persimokymo.

Galutiniam metodui sukurti, buvo sukurta programa naudojanti pirminiame eksperimente sukurtą MSPC V2 modelį, automobilių transformavimo MSPC modelį ir duomenų rinkinio kūrime naudotą DeepLabV3 segmentavimo modelį. Ši programa veikia tokiu eiliškumu:

1. Bendro MSPC (antrojo duomenų rinkinio versijos) modelio paruošimas;
2. Automobilių transformavimo MSPC modelio paruošimas;
3. Nuotraukos nuskaitymas;
4. Bendros nuotraukos transformavimas;
5. Originalios nuotraukos segmetavimas ir *alpha* kaukės paruošimas naudojant DeepLabV3 modelį;
6. Maskuotos nuotraukos transformavimas antruoju modeliu;
7. Maskuotos nuotraukos sujungimas su transformuota bendra nuotrauka;
8. Bendros nuotraukos išsaugojimas.

2.5. Panaudoti iš anksto apmokyti difuzijos modeliai

Buvo panaudoti 3 difuzijos modeliai: ControlNet-Seg, InstructPix2Pix ir CycleDiffusion. Jie visi trys veikia skirtingai ir duoda unikalius rezultatus. Jų tikslėsnis veikimas aprašytas pirmo skyriaus antrame poskyryje. Šie modeliai nebuvo apmokomi, o buvo naudoti jau apmokyti (angl. *pre-trained*).

Visi šie modeliai buvo panaudoti su testavimo duomenų rinkiniu.

2.5.1. ControlNet-Seg modelis

Šis modelis atspindi tai, koks buvo MSPC Car modelio tikslas (transformuoti remiantis turinio klase), nors rezultatai nėra tokie, kokių buvo siekta su MSPC Car modeliu. Kaip minėta modelių apžvalgos skyriuje, šis modelis transformuoja remiantis segmentacijos kaukėmis, taip žinantis kurią dalį kaip transformuoti. Tai suteikia korektiškesnį turinio išlaikymą ir mažesnį semantinį nuokrypi.

Šis modelis buvo panaudotas naudojant žiniatinklio paslaugą (angl. *API*) prieinamą puslapio nuorodoje <https://replicate.com/jagilley/controlnet-seg>. Joje suteikiamas limituotas kiekis nemokamų užklausų ir vėliau užklausos tampa mokamos.

Buvo pastebėta, jog nuotraukų transformacijos rezultatai stipriai priklauso nuo pasirinkto nuotraukos dydžio. Pasirinkus 512 pikselių kvadratinę nuotrauką, rezultatai yra aiškus ir tvarkingi (žr. 12 pav.), tačiau pasirinkus 256 pikselių nuotraukos dydį, rezultatai tapo abstraktiškesni su suliejamomis detalėmis bei daug kur atsirasdavo nelogiškos spalvos, pavyzdžiui, žalias dangus (žr. 13 pav.).



12 pav. ControlNet-Seg modelio pavyzdinis rezultatas su 512 pikselių nuotraukos dydžio parametru.



13 pav. ControlNet-Seg modelio pavyzdinis rezultatas su 256 pikslelių nuotraukos dydžio parametru.

Nors ir statistikos šiam modeliui nėra geros, tačiau nuotraukų kokybė, subjektyviai žiūrint, yra geriausia. Detalės yra išlaikomos, stilius perteikiamas, nuotraukos nėra konservatyviai transformuojamos ir nuotraukose lieka labai mažai triukšmo.

2.5.2. InstructPix2Pix modelis

Kaip minėta modelių apžvalgos skyriuje, šis modelis veikia „paliepimo“ principu. Pateikiant nuotrauką, pateikta tekstinė įvestis turi būti komandos tipo, pavyzdžiu „paversk kalnus snieguotais“. Šis modelis yra jau apmokytas ir pateiktas viešai prieigai HuggingFace bibliotekoje. Jo panaudojimas labai lengvas, nereikalaujantis ilgo kodo ar daug pasiruošimo. Užtenka trumpo Python kalbos kodo, kuris paremtas oficialia pavyzdine implementacija, prieinama nuoroda <https://huggingface.co/timbrooks/instruct-pix2pix>:

```
import torch

from diffusers import StableDiffusionInstructPix2PixPipeline
import transformers
from PIL import Image
import glob
import os

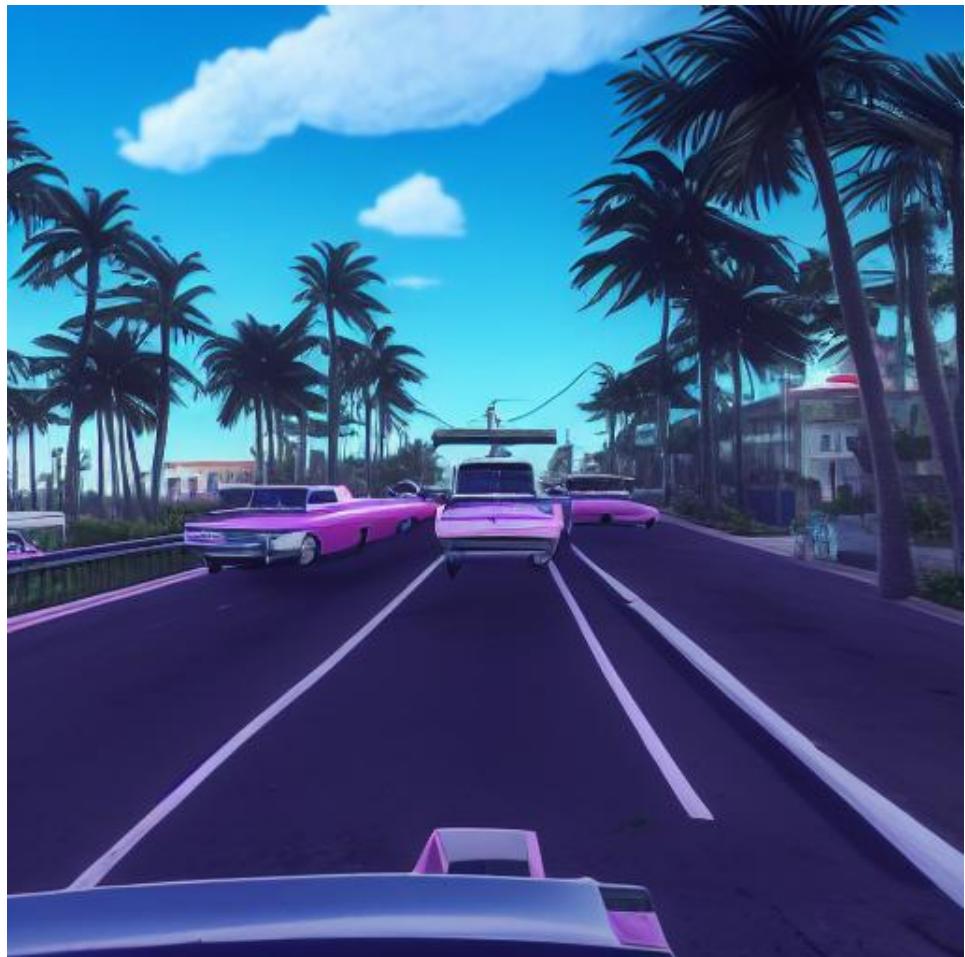
model_id_or_path = "timbrooks/instruct-pix2pix"
pipe = StableDiffusionInstructPix2PixPipeline.from_pretrained(model_id_or_path,
    torch_dtype=torch.float16).to("cuda")
generator = torch.Generator("cuda").manual_seed(0)
```

```
init_image = Image.open("input.jpg").convert("RGB").resize((512, 512))

edited_image = pipe(
    "turn into Vice City style",
    image=init_image,
    num_inference_steps=60,
    image_guidance_scale=1.5,
    guidance_scale=10,
    generator=generator,
).images[0]

edited_image.save("output.jpg")
```

InstructPix2Pix modelis duoda kiek nuvilliančius rezultatus, nors ir švaresnius bei su išliekančiomis detalėmis. Modelio pagrindinės transformacijos yra spalvų pakeitimas į žydrą ir rožinę, tačiau nei kelio, nei pastatų, nei dangau spalvos neatitinka GTA:VC stiliaus (žr. 14 pav.).

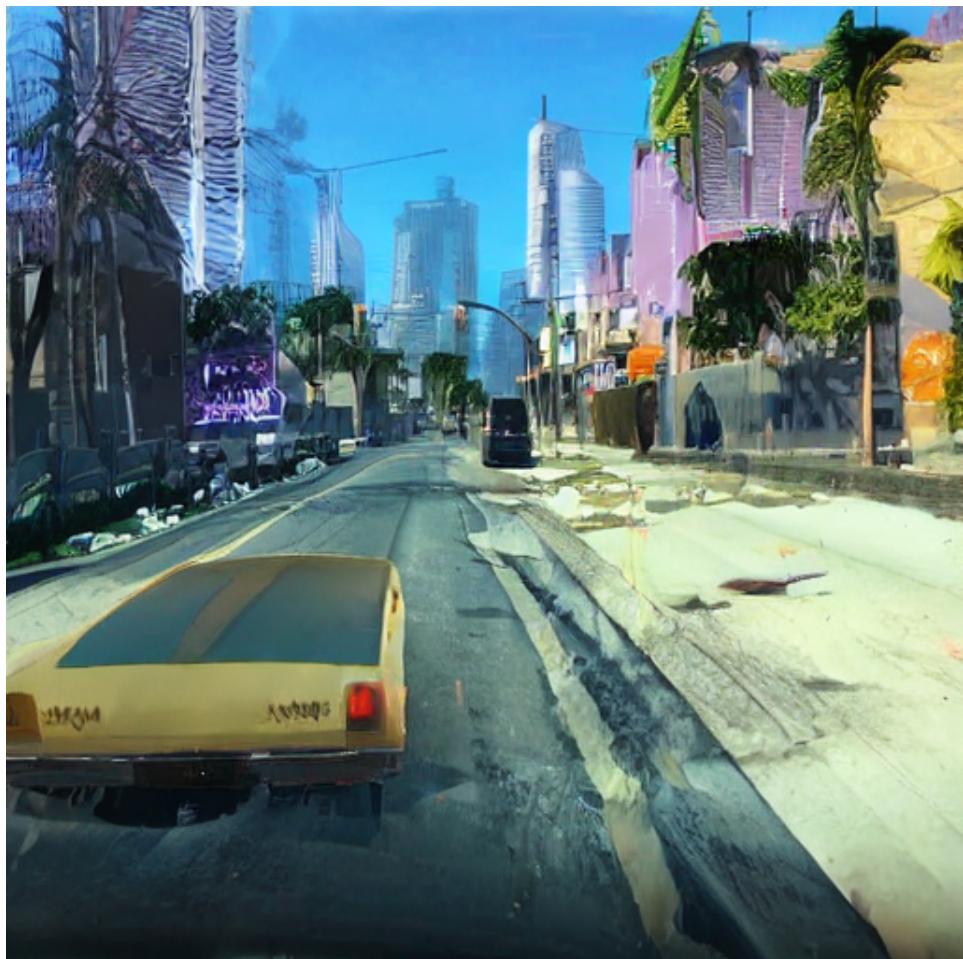


14 pav. InstructPix2Pix modelio pavyzdinis rezultatas.

2.5.3. CycleDiffusion modelis

Šis modelis taip pat buvo atsisiųstas ir panaudotas naudojant HuggingFace biblioteką bei galima būtų pritaikyti panašų pavyzdinį kodą iš InstructPix2Pix skyriaus, du esminiai skirtumai: modelio pavadinimas šiuo atveju būtų „CompVis/stable-diffusion-v1-4“ ir reikėtų pateikti ne tik rezultato tekstinę reikšmę, bet ir įvesties nuotraukos tekstinį apibūdinimą.

Šis modelis yra mažiau stabilus, daugely nuotraukų yra triukšmo, panaikinamos detalės, įterpiami artefaktai arba paprasčiausiai netransformuoojamos detalės (žr. 15 pav.)



15 pav. CycleDiffusion modelio pavyzdinis rezultatas.

2.6. Modelių vertinimo metrikos

Modelių nuotraukų transformavimo gebėjimams vertinti buvo pasirinktos dvi metrikos: FID ir SSIM, šie trumpiniai atitinkamai reiškia Fréchet pradžios atstumas (angl. *Fréchet inception distance*) ir struktūrinis panašumas (angl. *structural similarity index*).

2.6.1. FID

Pirmoji pasirinkta metrika modelių nuotraukų turinio transformavimo kokybei vertinti buvo FID, dažniausiai naudojama vertinti modelių generuojamų nuotraukų kokybei. Šia metrika apskaičiuojamas atstumas tarp realių ir sugeneruotų nuotraukų požymių vektorių.

Šis įvertis parodo, kaip nuotraukos iš dviejų grupių yra struktūriškai panašios. Žemesnis įvertis reiškia kokybiškesnes nuotraukas, o didelis – prastesnes. FID metrika pirmą kartą pristatyta ir panaudota 2017 metais Linco Johaneso Keplerio universiteto tyrimo straipsnyje pavadinimu „GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium“ [HRU⁺17]. Šis įvertis pristatytas kaip geresnė metrika vietoj pradžios įverčio (angl. *inception score*, trumpinama IS), kadangi IS vertina tik sugeneruotų nuotraukų kokybę, o FID metrika vertina sugeneruotų nuotraukų kokybę lyginant su šaltinio domeno duomenų rinkiniu.

2.6.2. SSIM

SSIM [WBS⁺04] yra antroji pasirinkta metrika modelių vertinimui. Tai yra metrika skirta įvertinti panašumą tarp dviejų nuotraukų. Šiuo atveju, tai buvo vidurkis tarp visų transformuotų nuotraukų ir jų originalų (angliškai dar vadinama *ground truth*). Ši metrika apskaičiuoja įvertį pagal tris pagrindinius punktus: struktūra, kontrastas ir apšvietimas.

Šios metrikos įverčių ribos yra nuo 0 iki 1, kas reiškia panašumą tarp dviejų nuotraukų (ar vidurkį tarp dviejų duomenų rinkinių). 0 reiškia, kad nuotraukos visiškai nepanašios, o 1 – labai panašios arba vienodos.

Šio uždavinio atveju, idealu būtų didesnis SSIM įvertis, nes tai reikštų, jog pakankamai gerai išlaikoma turinio struktūra, tačiau, kad nebūtų arti 1, nes tai reikštų, kad yra pakeista per mažai turinio, nėra atvaizduojama pakankamai vaizdo žaidimo vaizdinio stilus ir yra paliekama per daug realaus eismo vaizdo aspektų.

2.7. Gauti įverčiai

Šiame skyriuje pristatomi rezultatai gauti apmokius CycleGAN, CUT, MSPC modelius su pirmąja ir antrąja duomenų rinkinio iteracija, apmokius ir sukūrus MPSC Car metodą ir panaudojus apmokytus CycleDiffusion, ControlNet-Seg ir InstructPix2Pix difuzijos modelius.

2.7.1. FID įverčiai

Apmokius modelius buvo apskaičiuotos FID metrikos indikuojančios, kaip kokybiškai modeliai sugeba transformuoti nuotraukas lyginant su tikslu domenu.

Remiantis 4 lentelės duomenimis, matoma, jog geriausias FID nebūtinai reiškia geriausius rezultatus. Antroji apmokyta CycleGAN versija patobulino savo FID, tačiau, nors ir minimaliai, nukrito atpažistamų automobilių skaičius (žr. 5 lentelę), o CUT modeliui yra atvirkštinė situacija – antroji modelio versija padidino aptiktų automobilių skaičių, tačiau dėl to sumažėjo FID įvertis. Vienintelis modelis visiškai patobulėjės nuo pagerinto GTA:VC duomenų rinkinio yra MSPC, jo atpažintų automobilių skaičius padidėjo ir FID įvertis sumažėjo. Pagal šiuos rezultatus matoma, jog iš trijų bandytų modelių, MSPC teikia geriausius rezultatus.

4 lentelė. Apmokyti modelių gauti FID įverčiai.

Modelio pavadinimas	FID
MSPC V2	145
CycleGAN V2	161
MSPC V1	167
CycleGAN V1	170
MSPC Car	173
Cut V1	181
Cut V2	186
CycleDiffusion	190
ControlNet-Seg	194
InstructPix2Pix	206

MSPC Car modelis gerokai suprastino savo FID įverti, tačiau padvigubino savo aptiktų automobilių skaičių. Taip pat difuzijos modelių FID įverčiai teigtų, jog nuotraukos yra prasčiau transformuojamos, tačiau pažvelgus į pavyzdžius, matome, jog turinys geresnis iš detalių išlaikymo perspektyvos nors ir nuotraukų stilius iškraipomas.

2.7.2. Atpažintų automobilių kiekiai

5 lentelėje ryškiai matomas difuzijos modelių pranašumas. Jie daug geriau sugeba atpažinti išlaikyti automobilius. Naujo MSPC Car modelio atpažintų automobilių kiekis padidėjo daugiau nei dvigubai lyginant su paprastu pavieniu MSPC V2 modeliu.

5 lentelė. Apmokyti modelių atpažintų automobilių kiekiai ir dalis visų atpažintų automobilių realaus domeno testavimo duomenų rinkinje.

Modelio pavadinimas	Atpažintų automobilių kiekis, vnt.	Atpažintų automobilių kiekis, %
ControlNet-Seg	231	54,2 %
CycleDiffusion	171	40,1 %
InstructPix2Pix	121	28,4 %
MSPC Car	44	10,3 %
MSPC V2	17	3,9 %
CycleGAN V1	15	3,5 %
CycleGAN V2	14	3,2 %
MSPC V1	10	2,3 %
Cut V2	10	2,3 %
Cut V1	7	1,6 %

2.7.3. SSIM įverčiai

Iš statistikos, taip pat buvo apskaičiuota SSIM metrika, kuri skirta apskaičiuoti struktūrinį panašumą tarp nuotraukų. Šioje metrikoje šiame uždavinyste geras rezultatas neturėtų būti 1, nes tai reikštų, kad nuotrauka buvo mažai pakeista, o kai yra toks skirtumas tarp domenų stilių, tas yra reikalinga. Taip pat 0 būtų prastas rezultatas, kadangi tai reikštų, jog pakeista buvo per daug ir visiškai neišlaikyta struktūra. Dėl įvardintų priežasčių, pavienė SSIM metrika mažai ką sako,

kadangi tiek gerai išlaikytas, tiek stipriai pakeistas turinys nėra vienareikšmiškai nei blogas, nei geras rezultatas. Šio eksperimento atveju, tai turėtų leisti pamatyti akivaizdžiai blogus rezultatus ir kartu vertinant su FID metrika leistų atpažinti, kurie modeliai tiek gerai perima tikslą domeno stilių, tiek išlaiko šaltinio domeno struktūrą.

Iš panaudotų modelių, geriausią SSIM metriką pasiekė CycleDiffusion modelis, tačiau iš pavyzdinių nuotraukų (žr. 4 priedą), akivaizdu, jog tai pasiekta dėl gan mažo nuotraukų pakeitimo. Be CycleDiffusion modelio, geriausią SSIM ir FID metriką pasiekė MSPC V2 modelis, kuris taip pat turi didžiausią atpažintų automobilių kiekį neskaitant difuzijos ir MPSC Car modelio.

6 lentelė. Apmokyti modelių gauti SSIM įverčiai.

Modelio pavadinimas	SSIM įvertis, nuo 0 iki 1
CycleDiffusion	0,7116
MSPC V2	0,6316
CycleGAN V2	0,6226
MSPC Car	0,5996
MSPC V1	0,5927
InstructPix2Pix	0,5879
CycleGAN V1	0,5772
Cut V2	0,5091
CUT V1	0,4726
ControlNet-Seg	0,2449

Rezultatai ir išvados

Gausesni transformuotų nuotraukų pavyzdžiai yra prisegti priedų skiltyje.

Modelis pasiekės žemiausią FID (145) yra MSPC V2.

Modelis, kurio transformuotuose vaizduose aptinkama daugiausia automobilių (54,2 %) yra ControlNet.

Modelis pasiekės aukščiausią SSIM įvertį (0,7116) yra CycleDiffusion.

Vaizdų transformavimas buvo įgyvendintas iki tokio lygio, kad galima nesunkiai atpažinti transformuoto vaizdo domeną, tačiau GAN modeliai kenčia nuo architektūrinių problemų neleidžiančių jiems išlaikyti tam tikro turinio tokiam plačios apimties uždavinyje, o difuzijos modeliams neužtenka bendro mokymo spręsti tokiam uždavinui. Modeliai dažnai panaikina automobilius iš transformuojamų nuotraukų ir gerai sugeba transformuoti kelius, dangų ir gamtinius objektus arba gerai išlaiko transformuojamo turinio detales, bet nuotrauka būna transformuojama per stipriai arba per mažai. Tyrimo išvados:

1. Kokybiškiausius rezultatus duodantys modeliai yra MSPC ir ControlNet.
2. Duomenų rinkinių struktūriniai ir objektų dažnumo skirtumai stipriai įtakoja nuotraukų transformavimo kokybę.
3. Žemesnis FID įvertis ir didesnis SSIM įvertis nebūtinai reiškia kokybiškesnį nuotraukos turinio išlaikymą.

Santrumpos

1. **GTA:VC** – Žaidimas „Grand Theft Auto: Vice City“.
2. **GTA V** – Žaidimas „Grand Theft Auto 5“.
3. **FID** – angl. *Fréchet inception distance* – Fréchet pradžios atstumas, yra metrika naudojama ivertinti generatyvinių modelių kokybę.
4. **IS** – angl. *Inception score* – pradžios ivertis, yra metrika naudojama ivertinti generatyvinių modelių kokybę.
5. **SSIM** – angl. *Structural similarity index* – struktūrinis panašumas, metrika naudojama ivertinti generatyvių modelių kokybę.

Šaltiniai

- [BHE22] Tim Brooks, Aleksander Holynski, Alexei A Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. *arXiv preprint arXiv:2211.09800*. 2022.
- [COR⁺16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, Bernt Schiele. *The Cityscapes Dataset for Semantic Urban Scene Understanding*. arXiv, 2016. Pasiekiamas per DOI: 10.48550/ARXIV.1604.01685.
- [DBX⁺22] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, Qi Tian. *CenterNet++ for Object Detection*. arXiv, 2022. Pasiekiamas per DOI: 10.48550/ARXIV.2204.08394.
- [GPM⁺14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. *Generative Adversarial Networks*. arXiv, 2014. Pasiekiamas per DOI: 10.48550/ARXIV.1406.2661.
- [HJA20] Jonathan Ho, Ajay Jain, Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. Pasiekiamas per arXiv: 2006.11239 [cs.LG].
- [HRU⁺17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. 2017. Pasiekiamas per DOI: 10.48550/ARXIV.1706.08500.
- [YCW⁺18] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, Trevor Darrell. *BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning*. arXiv, 2018. Pasiekiamas per DOI: 10.48550/ARXIV.1805.04687.
- [JW22] Taesung Park Jun-Yan Zhu, Tongzhou Wang. *CycleGAN and pix2pix in PyTorch*. 2022. **urlalso:** <https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix>.
- [LMB⁺14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev ir kt. Microsoft COCO: Common Objects in Context. *CoRR*. 2014, t. abs/1405.0312. Pasiekiamas per arXiv: 1405.0312.
- [PLQ⁺21] Yingxue Pang, Jianxin Lin, Tao Qin, Zhibo Chen. *Image-to-Image Translation: Methods and Applications*. arXiv, 2021. Pasiekiamas per DOI: 10.48550/ARXIV.2101.08629.
- [RAK21] Stephan R. Richter, Hassan Abu Alhaija, Vladlen Koltun. Enhancing Photorealism Enhancement. *arXiv:2105.04619*. 2021.
- [RBL⁺21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. Pasiekiamas per arXiv: 2112.10752 [cs.CV].

- [SWM⁺15] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, Surya Ganguli. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. 2015. Pasiekiamas per arXiv: 1503.03585 [cs.LG].
- [WBS⁺04] Zhou Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*. 2004, t. 13, Nr. 4, p. 600–612. Pasiekiamas per DOI: 10.1109/TIP.2003.819861.
- [XXW⁺22] Yanwu Xu, Shaoan Xie, Wenhao Wu, Kun Zhang, Mingming Gong, 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)*. 2022-06, p. 18311–18320.
- [ZA23a] Lvmin Zhang, Maneesh Agrawala. *Adding Conditional Control to Text-to-Image Diffusion Models*. arXiv, 2023. Pasiekiamas per DOI: 10.48550/ARXIV.2302.05543.
- [ZA23b] Lvmin Zhang, Maneesh Agrawala. *Adding Conditional Control to Text-to-Image Diffusion Models*. 2023. Pasiekiamas per arXiv: 2302.05543 [cs.CV].
- [ZPI⁺17] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *Computer Vision (ICCV), 2017 IEEE International Conference on*. 2017.

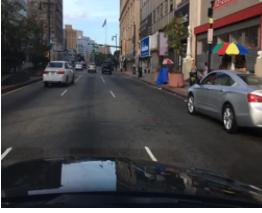
Priedas nr. 1

Darbo saugykla

Darbo saugykla su duomenų rinkiniu, pavyzdinėmis nuotraukomis ir gautais modeliais yra pasiekiamta nuo-roda: <https://github.com/0ksas/bachelors>.

Priedas nr. 2
**Pirminiu modeliu (CycleGAN, CUT, MSPC) nuotraukų pavyzdžiai
(pirmosios duomenų rinkinio iteracijos)**

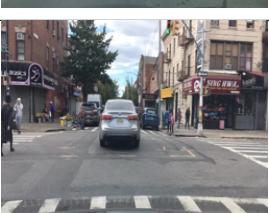
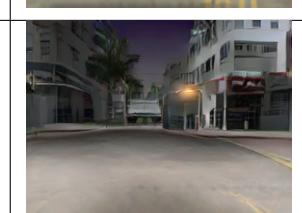
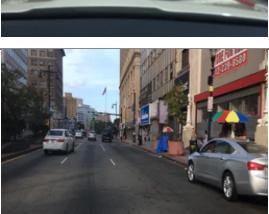
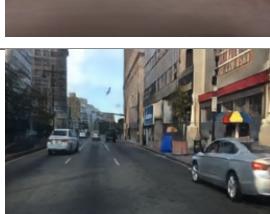
7 lentelė. Transformuojamų nuotraukų ir jų rezultatų pavyzdžiai

Originali nuotrauka	CycleGAN	CUT	MSPC
			
			
			
			
			
			

Priedas nr. 3

**Pirminiu modeliu (CycleGAN, CUT, MSPC) nuotraukų pavyzdžiai
(antrosios duomenų rinkinio iteracijos)**

8 lentelė. Transformuojamų nuotraukų ir jų rezultatų pavyzdžiai

Originali nuotrauka	CycleGAN V2	CUT V2	MSPC V2
			
			
			
			
			
			

Priedas nr. 4

Difuzijos modelių nuotraukų pavyzdžiai

9 lentelė. Transformuojamų nuotraukų ir difuzijos modelių rezultatų pavyzdžiai

Originali nuotrauka	ControlNet-Set	InstructPix2Pix	CycleDiffusion

Priedas nr. 5

MSPC Car modelio nuotraukų pavyzdžiai

10 lentelė. Transformuojamų nuotraukų ir difuzijos modelių rezultatų pavyzdžiai

Originali nuotrauka	MSPC V2	MSPC Car
		
		
		
		
		
		