

Generatyvinių modelių kolapso prevencija

Atliko: Greta Virpšaitė

Darbo vadovas: j. asist. Boleslovas Dapkūnas

Vilniaus universitetas

Matematikos ir informatikos fakultetas

Programų sistemų bakalauro studijų programa

2025

Generatyvinių modelių kolapsas

Generatyvinių modelių kolapsas - degeneratyvus procesas, kuris atsiranda, kai generatyviniai modeliai mokomi naudojant jų pačių sugeneruotus duomenis praranda gebėjimą tiksliai atspindėti pradinę duomenų informaciją.

Generatyvinių modelių kolapso tyrimo svarba

Didžiulės apmokymo sąnaudos:

- Elektros energijos, laiko ir techninės įrangos sąnaudos.
- Generatyvinių modelių treniravimo sąnaudos nuo 2016m. kasmet auga vidutiniškai **2,4 karto** (2024m. duomenys), o net senesni modeliai, tokie kaip **GPT-4**, kainavo apie **40 milijonų JAV dolerių** (2024m. duomenys).

Kodėl atsiranda modelių kolapsas

Generatyvių modelių kolapsas įvyksta, kai jų naudotojai su sugeneruotais duomenimis elgiasi taip:

- talpina viešai internete, kur jie tampa prieinami būsimiems modelių apmokymo procesams.
- siunčia generatyviniam modeliui atgal jų naujai sugeneruotus duomenis.

- Ištirti ar žmogaus įsitraukimas į duomenų filtravimo procesą bei edukacija apie generatyvinių modelių kolapsą gali prisidėti prie šio reiškinio prevencijos GenAI modeliuose.

- Išanalizuoti literatūrą, susijusią su DI modelių kolapsu, apimant pagrindinius mechanizmus, rizikas ir galimus sprendimus. Aprašyti DI modelių kolapso atpažinimo technikas, remiantis šiuolaikiniais tyrimais.
- Apibrėžti pagrindines DI modelių kolapso prevencijos strategijas, įtraukiant Europos Sajungos DI aktą ir jo įtaką DI modelių stabilumui.

- Atliekti eksperimentą su StyleGAN2 modeliu, siekiant ištirti generatyvinių modelių kolapso reiškinį bei įvertinti naudotojų įsitraukimo į duomenų filtravimą galimybes kaip prevencinę priemonę šiam reiškinui.
- Įvertinti vartotojų motyvaciją ir gebėjimą prisidėti prie duomenų kokybės vertinimo bei jų požiūrį į sąmoningą sprendimų priėmimą, keliamų duomenų kontekste.

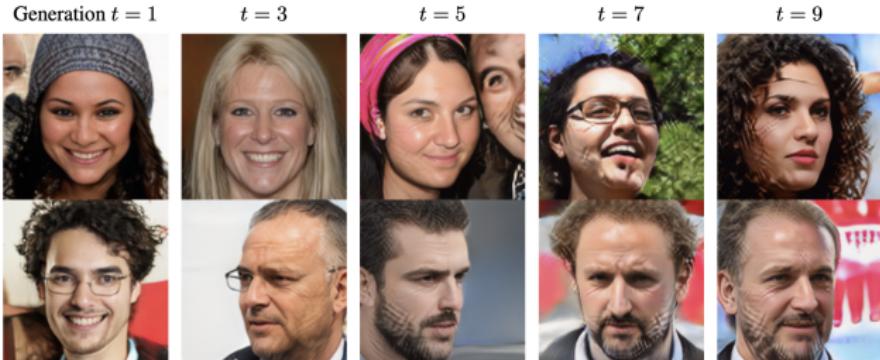
- Pateikti rekomendacijas, pagristas atlikto eksperimento ir literatūros apžvalgos rezultatais, įvertinant pasiūlyto metodo taikymo galimybes, privalumus ir trūkumus tiek moksliniame, tiek praktiniame kontekste.

Generatyvinių modelių kolapso atpažinimo bruožai



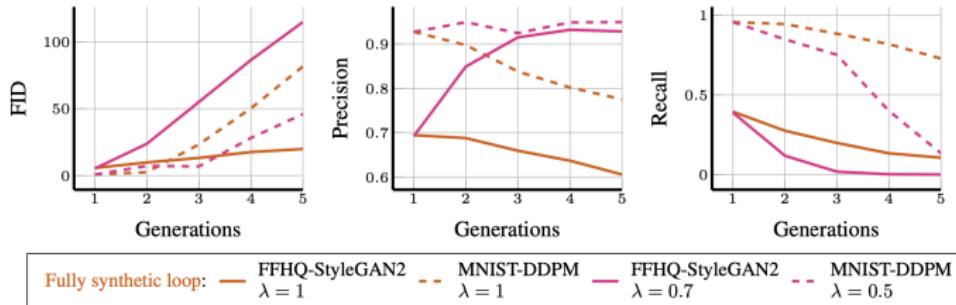
pav. 1: **FFHQ-StyleGAN2 kolapso pavyzdys.** Didėjant generacijų (angl. *Generation*) skaičiui t, matomas išvesčių suvienodėjimas.

Generatyvinių modelių kolapso atpažinimo bruožai



pav. 2: FFHQ-StyleGAN2 kolapso pavyzdys. Didėjant generacijų (angl. *Generation*) skaičiui t, išvestyse matomi labiau ryškėjantys kryželiniai artefaktai.

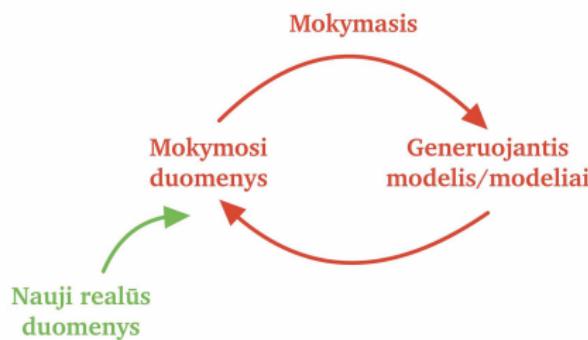
Generatyvinių modelių kolapso atpažinimo bruožai



pav. 3: Grafike vaizduojami pokyčiai, atsirandantys mokant modelį pilnai sintetiniu duomenų ciklu. Pateiktos trys metrikos: **Frécheto įsivaizdavimo atstumas** (angl. *Fréchet Inception Distance*, trump. *FID*), **vaizdų kokybė** (angl. *precision*) ir **įvairovė** (angl. *recall*). Skirtingos spalvos žymi skirtingas λ reikšmes, o linijų tipai nurodo skirtingus modelius. Matoma, kad kolapso metu *FID* reikšmės didėja, *recall* mažėja, o *precision* pokyčiai priklauso nuo λ : kai $\lambda \neq 1$, kokybė didėja, o esant $\lambda = 1$ - mažėja.

Pagrindinės generatyvinių modelių kolapso prevencijos strategijos

- **Realių ir sintetinių duomenų derinimas** – į mokymo rinkinius reguliariai įtraukiami švieži realūs duomenys, kad būtų išlaikyta jvairovė ir sumažintas duomenų užterštumas.



pav. 4: **Šviežių duomenų ciklas:** save valgančių ciklų tipas.

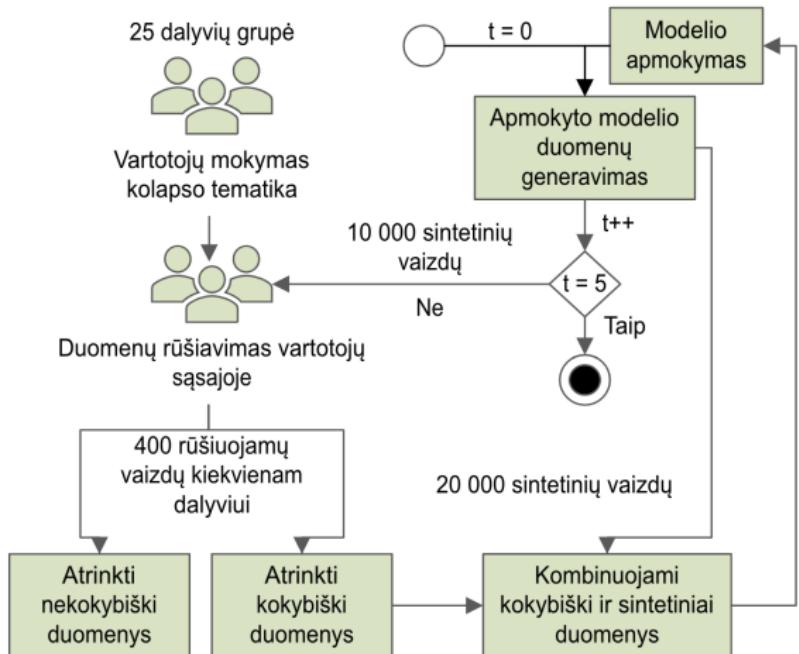
Pagrindinės generatyvinių modelių kolapso prevencijos strategijos

- **Duomenų žymėjimas** – sintetiniai duomenys žymimi (pvz., vandensženkliais), kad būtų išvengta jų atsitiktinio įtraukimo į mokymo rinkinius.

Eksperimento tikslas

- Išsiaiškinti, ar apmokyti vartotojų elgesys gali sumažinti generatyvinių modelių kolapso riziką StyleGAN2-FFHQ modelyje.

Eksperimento eiga



pav. 5: Eksperimento eigos pagalbinė diagramma (nėra darbe)

Rezultatai - įvairovė



pav. 6: FFHQ–StyleGAN2 kolapso pavyzdys. Didėjant generacijų skaičiui (t), matomas išvesčių vienodėjimas.

Rezultatai – artefaktai

$t = 1$



$t = 2$



$t = 3$



$t = 4$

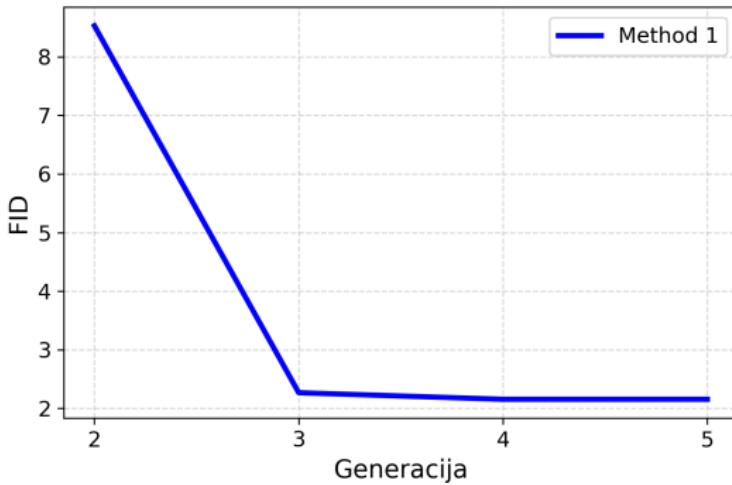


$t = 5$



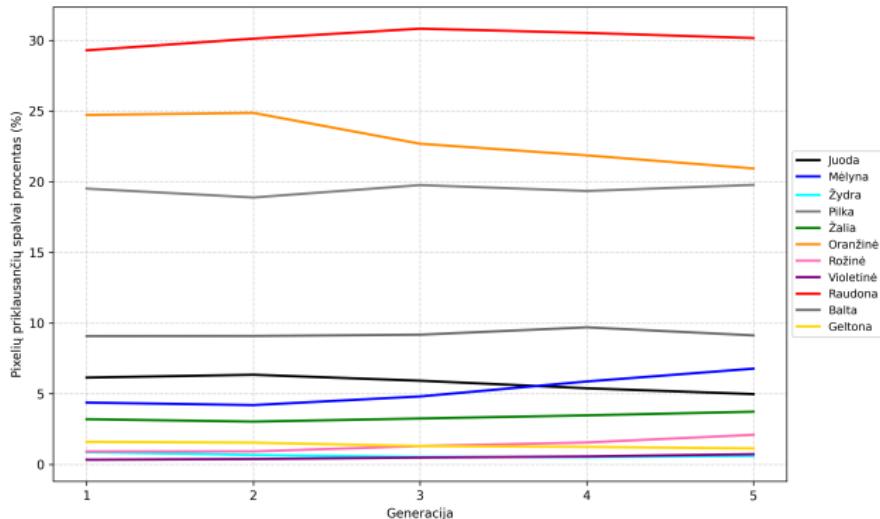
pav. 7: Kryžiavimo artefaktų (angl. cross-hatching) raida skirtinėse generacijose.

Rezultatai – FID metrika



pav. 8: **FID reikšmių pokytis per penkias generacijas.**

Rezultatai - spalvų pasiskirstymas



pav. 9: Sugeneruotų paveikslų spalvų pasiskirstymo pokytis tarp generacijų (%).

Rezultatai – naudotojų apklausa

Ientelė 1: Naudotojų anketos rezultatai. Atsakymai pateikti procentais.
Paryškintos dažniausiai pasirinktos reikšmės.

Klausimai	Atsakymai (%)			
	Visiškai sutinku	Sutinku	Nesutinku	Visiškai nesutinku
Ar ateityje būsite mažiau linkę viešai talpinti dirbtinio intelekto sugeneruotus duomenis internete?	28	44	24	4
Ar gerai supratote kokius paveikslus reikia išmesti pirmoje atrankoje ($t = 1$)?	32	44	16	8
Ar gerai supratote kokius paveikslus reikia išmesti antroje atrankoje ($t = 2$)?	12	72	16	0
Ar gerai supratote kokius paveikslus reikia išmesti trečioje atrankoje ($t = 3$)?	32	68	0	0
Ar gerai supratote kokius paveikslus reikia išmesti ketvirtijoje atrankoje ($t = 4$)?	32	68	0	0
Ar atrinkinėjote vaizdus pilnai susikaupę?	24	76	0	0
Ar manote, kad jūsų darbas atrinkinėjant vaizdus buvo kokybiškas?	28	68	4	0

- **Naudotojų gebėjimas filtruoti duomenis:** Eksperimentas parodė, kad edukuoti naudotojai geba atpažinti nekokybiskus generatyvius vaizdus, ypač tuos, kuriuose pasireiškia modelio kolapso požymiai, tokie kaip „kryželiai“ (angl. *cross-hatching*) artefaktai, pasikartojantys veido ir aplinkos bruožai. Rezultatai rodo, kad vaizdų įvairovė reikšmingai nesumažėjo, tačiau artefaktai ryškėjo su kiekviena karta, todėl naudotojų įsitraukimas gali padėti sumažinti įvairovės praradimo požymių skliaudą.

- **Motyvacija ir įsitraukimo iššūkiai:** Nors dauguma dalyvių suprato užduotį ir vertino savo atliktą darbą kaip kokybišką, apklausa atskleidė, kad ne visi naudotojai būtų linkę keisti savo elgesį internte ar prisidėti prie generatyvinių modelių sugeneruoto turinio filtravimo be papildomos motyvacijos. Tai rodo, kad vien edukacijos gali nepakakti, todėl svarbu ieškoti būdų, kaip įtraukti naudotojus - pavyzdžiui, per vartotojo sasajas ar žymėjimo sistemas.

- **Vartotojų filtravimas kaip prevencijos priemonės dalis:**
Naudotojų filtravimas galėtų tapti papildomu sugeneruotų duomenų reguliavimo sluoksniu greta kitų generacinių modelių kolapso prevencijos būdu.
- **Žmogaus vaidmuo generatyvinių modelių kolapso prevencijoje:** rankinis rūšiavimas yra prasmingas ir praktiskai pritaikomas. Tačiau norint šį metodą taikyti, būtina derinti technologinius sprendimus su edukacinėmis ir motyvacinėmis priemonėmis.

Mokslininkams:

- Toliau tyrinėti naudotojų įtraukimą į skirtingu modelių duomenų filtravimo ir priežiūros procesus.
- Plėsti tyrimus už tekstinių ir vaizdų generatyvinių modelių ribų – įtraukti garso generatyvinius modelius.
- Analizuoti, kaip šiuose modeliuose pasireiškia kolapso požymiai ir ilgalaikės pasekmės.

Praktikams:

- Įtraukti naudotojus į modelio generuotų vaizdų kokybės vertinimą.

Bendruomenei:

- Skatinti naudotojų sąmoningumą dėl duomenų kokybės ir DI modelių poveikio.
- Suteikti galimybę žymėti generuotą turinį kaip DI sukurtą.
- Įtraukti naudotojus į generuotų duomenų kokybės vertinimą – tai prisdėtų prie skaidrumo ir modelių patikimumo.

Recenzijos diskusija - pastabos

Pastaba	Atsakymas
Eksperimente naudotas StyleGAN2-FFHQ modelis turėjo būti aprašytas detaliau . Taip pat, vertėjo detaliau aprašyti sukurtą programėlę ir jos integraciją.	Šiame tyime svarbiausias buvo ne pats modelis, o vartotojų elgsena ir jo kolapso požymiai. Modelis buvo pasirinktas remiantis literatūra, kad būtų galima palyginti rezultatus su ankstesniu tyrimu ir jvertinti, ar naudotojų edukacija gali sumažinti kolapso efektą.
2.4 poskyryje pateikti rezultatai apima ir kiekybiinius įvertinimus , tačiau 2.5 poskyryje pateiktos tik apklausos išvados .	Tokia struktūra pasirinkta vadovaujantis darbo vadovo rekomendacijomis. Kadangi pagrindinis tyrimo tikslas buvo jvertinti siūlomos prevencinės priemonės efektyvumą, kiekybiniai ir kokybiniai vertinimai buvo pateikti atskirai, norit geriau atspindėti ar tokia prevencijos priemonė būtų geriau pritaikoma realiame pasauliye.

Recenzijos diskusija - klausimai

Klausimai	Atsakymai
Darbe teigiamas, kad klaidos dėl variacijos sumažėjimo kyla dėl modelių mokymosi ir optimizavimo ypatybių. Ką reiškia optimizavimo ypatybės? Ir kuo jos skiriasi nuo mokymosi?	Optimizavimo ypatybės susijusios su tuo, kaip keičiasi modelio parametrai, siekiant pagerinti generuojamų vaizdų kokybę specialiai keičiant skirtinges parametrus (kaip naudojimas skirtingu funkcijų). Tuo tarpu mokymasis yra platesnis procesas, apimantis modelio gebėjimą generalizuoti duomenis ir gerėti su kiekviena mokymosi kata.
Kaip sudaryta 25 dalyvių grupė, kuri dalyvavo apklausoje?	Dalyviai buvo atrinkti iš artimos aplinkos, siekiant įvairovės pagal amžių, profesiją ir patirtį. Grupėje buvo tiek studentų, programuotojų, menininkų, mokytojų ir net pensininkų. Svarbiausias atrankos kriterijus buvo noras dalyvauti tyime, atsižvelgiant į ribotus resursus.

Papildoma informacija – vartotojų gebėjimas atrinkti tinkamus duomenis

Generacija t = 4



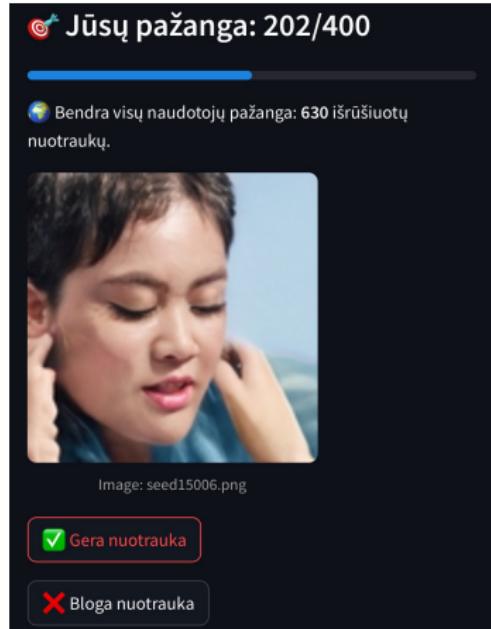
pav. 10: Naudotojų pažymėtos **geros** nuotraukos ketvirtijoje generacijoje.

Generacija t = 4



pav. 11: Naudotojų pažymėtos **blogos** nuotraukos ketvirtijoje generacijoje.

Papildoma informacija - vartotojų sasaja



pav. 12: Naudotojų sasaja. Sąsaja pritaikyta ir silpnaregystę turintiems naudotojams naudojant emoji. Matomas tiek bendras dalyvių progresas, tiek individuali pažanga iš 400 paveikslėlių.

- **Programavimo aplinka:** Visual Studio Code;
- **Naudoti resursai:** GPU – NVIDIA RTX 4080 Super; diegimas atliktas naudojant *Hetzner* debesijos serverius su *Ubuntu* operacine sistema;
- **Programavimo kalba:** Python;
- **Naudoti karkasai:** naudotojo sąsajos kūrimui naudotas *Streamlit*; modelių apmokymui – *PyTorch*;
- **Pagrindinės bibliotekos:** *PyTorch*, *NumPy* ir kitos, reikalingos *StyleGAN2* veikimui;
- **Modelio architektūra:** generatyvinis priešiškas tinklas, paremtas Wang et al. pasiūlyta *StyleGAN2* architektūra (angl. *generative adversarial network*, trumpinama GAN);
- **Duomenų rinkinys:** Flickr-Faces-HQ (FFHQ).

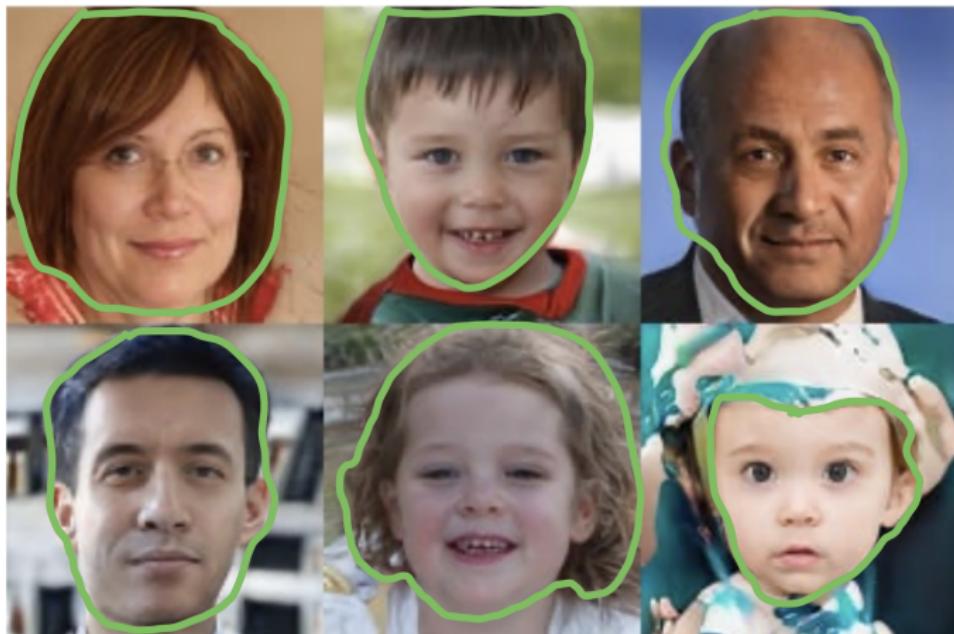
- **Teisinis reguliavimas** – ES ir JAV teisės aktai, skatinantys duomenų atsekiškumą ir generuoto turinio skaidrumą.
- **Kokybės ir įvairovės stebėsena** – nuolatinis sugeneruotų duomenų tikrinimas naudojant FID, precision ir recall metrikas.
- **Tikslinis duomenų kuravimas** – papildomas retų, įvairių pavyzdžių įtraukimas siekiant apsaugoti skirtinio „uodegas“.
- **Šališkumo parametrų derinimas modelių apmokyme** – derinami šališkumo mažinimo parametrai (λ) tam, kad būtų reguliuojamas modelio dėmesys duomenų įvairovei.

Papildoma informacija - tyrimui vykdyti naudoti edukaciniai paveikslai



pav. 13: Kokybiškų sugeneruotų vaizdų pavyzdys iš [?] su išskirtais unikaliais bruožais

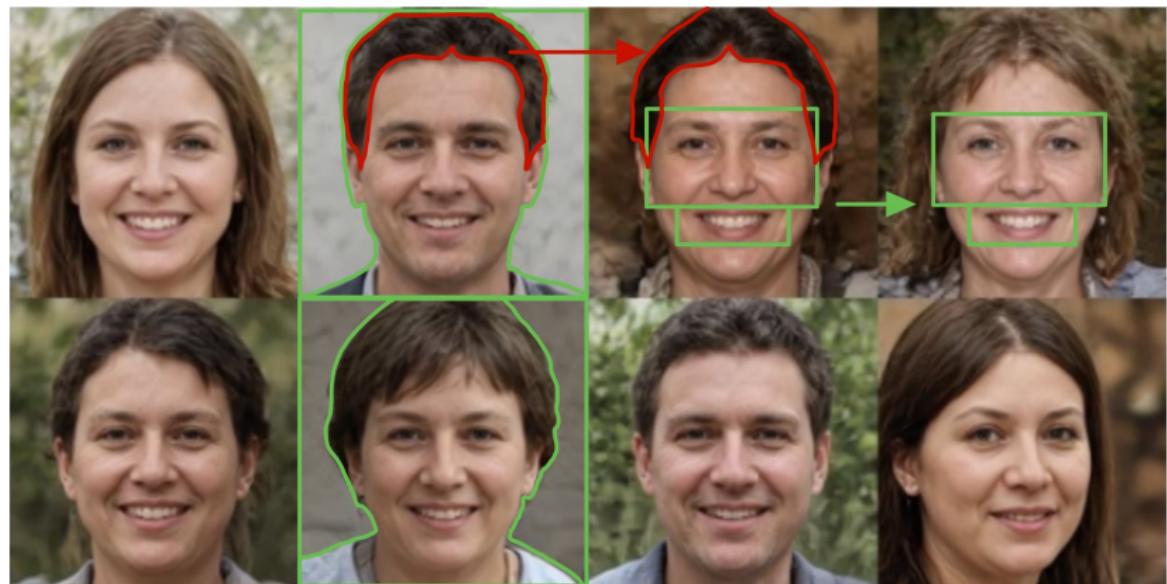
Papildoma informacija - tyrimui vykdyti naudoti edukaciniai paveikslai



Veidai be defektų ir
artefaktų

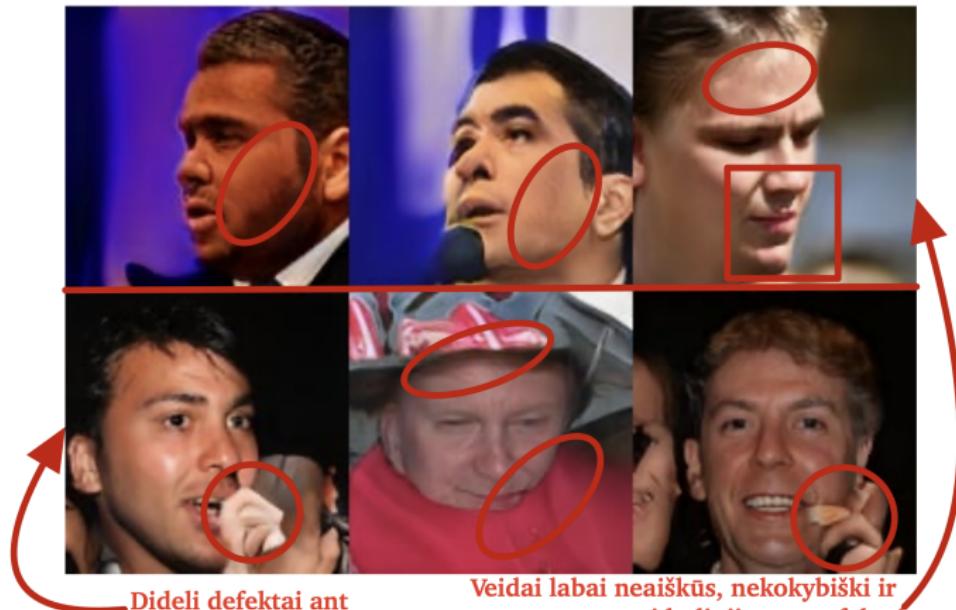
pav. 14: Kokybiškų veidų pavyzdžiai

Papildoma informacija - tyrimui vykdyti naudoti edukaciniai paveikslai



pav. 15: Įvairovės nykimo pavyzdžiai su išskirtais suvienodėjimo
bruožais

Papildoma informacija - tyrimui vykdyti naudoti edukaciniai paveikslai

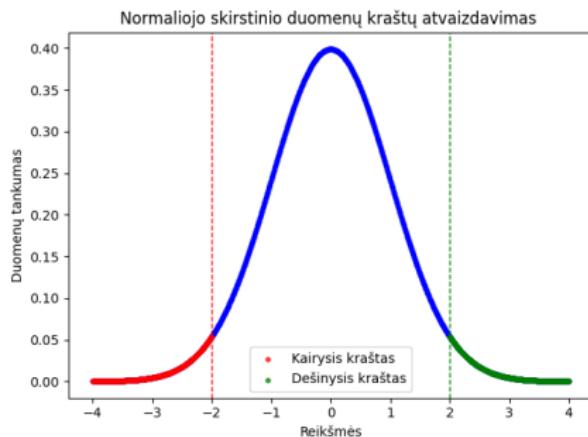


Dideli defektai ant
veidų

Veidai labai neaiškūs, nekokybiški ir
matomos ant veidų linijos - artefaktai

pav. 16: Brūkšnių arba pilnai nesusiformavusių kryželių (angl. *cross-hatching*) artefaktų pavyzdžiai

Papildoma informacija - teorinis normaliojo skirstinio duomenų kraštų atvaizdavimas



pav. 17: Teorinis normaliojo skirstinio duomenų kraštų atvaizdavimas pagal knygoje pateikiamą informaciją. Tai tokis skirstinys, kai duomenys yra simetriškai pasiskirstę aplink vidurį, o retesni duomenys yra kraštuose. **Kairysis kraštas (raudona)** vaizduoja mažo tankio, retai pasitaikančius duomenis vienoje skirstinio pusėje, o **dešinysis kraštas (žalia)** rodo panašaus pobūdžio retus duomenis kitoje pusėje.