VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ STUDIJŲ PROGRAMA

Giliųjų klastočių atpažinimas

Deepfake Detection

Magistro baigiamasis darbas

Atliko:	Tautvydas Skardžius	(parašas)
Darbo vadovas:	j. asist. Boleslovas Dapkūnas	(parašas)
Recenzentas:	asist. dr. Vytautas Valaitis	(parašas)

Vilnius – 2025

TURINYS

Santrauka

Šiame darbe buvo nagrinėti du giliųjų klastočių atpažinimo metodai - SBI [SY22] bei FSBI [HLK+25]. Darbo metu iš socialinių tinklų buvo surinktas nuosavas populiariausių klastočių rinkinys, kuris buvo naudojamas apmokytiems karkasams verifikuoti. Taip pat, buvo atkartoti SBI bei FSBI autorių atlikti mokymai, kurių rezultatai neprilygo autorių skelbiamiems rezultatams. Keitus karkasų architektūros komponentus buvo pasiektas tikslumo pagerėjimas visuose bandymuose tik nuosavame rinkinyje, kai CDF [LYS+20] rinkinyje geresnį tikslumą pavyko pasiekti tik viename bandyme iš trijų, o visais bandymais buvo pasiektas žymus greitaveikos paspartėjimas. Papildomų ir pakeistų duomenų augmentacijų bandymuose buvo pasiekti žymūs tikslumo priaugai CDF ir nuosavame rinkiniuose. Šiame darbe tiksliausiai klastotes atpažino FSBI25E+EfficientNetV2_S karkasas, kuris pasiekė 92.00% tikslumą CDF rinkinyje ir 74.6% nuosavame rinkinyje. Visi darbe mokyti karkasai daug prasčiau sugebėjo aptikti klastotes nuosavame rinkinyje, negu CDF rinkinyje.

Raktiniai žodžiai: su savimi sulietos nuotraukos, SBI, dažniniais požymiais patobulintos su savimi sulietos nuotraukos, FSBI, giliųjų klastočių atpažinimas

Summary

In this thesis, two deepfake detection models - SBI [SY22] and FSBI [HLK+25] were researched. A custom dataset of the most popular deepfakes was collected from social networks and used to validate the trained models. The original SBI and FSBI models trained as specified by the authors, but the resulting accuracies fell short of those reported by authors. When individual architectural components were replaced, accuracy improved across all experiments on the custom dataset, whereas on the CDF [LYS+20] dataset accuracy improved in only one of the three experiments; in every case, however, inference speed increased substantially. Experiments with additional or modified data augmentations produced substantial accuracy gains on both CDF and the custom dataset. The best overall performer was the FSBI25E+EfficientNetV2_S model, which reached 92.00% accuracy on CDF and 74.6% on the custom dataset. Nonetheless, every model trained in this work detected deepfakes considerably less reliably on the custom dataset than on CDF.

Keywords: self-blended images, SBI, frequency enhanced self-blended images, FSBI, deepfake detection

Įvadas

Dirbtiniai neuroniniai tinklai yra plačiai naudojami įvairiose srityse - medicininių nuotraukų analizei, finansinių nusikaltimų aptikimui, autonominiams automobiliams ir t.t. Dažniausiai tokios sistemos yra paremtos tokiais neuroniniais tinklais, kurie geba atlikti segmentavimo, klasifikavimo bei objektų aptikimo uždavinius. Vis dažniau yra tyrinėjami bei naudojami kūrybiniai tinklai - gebantys generuoti nuotraukas, vaizdo bei garso įrašus ir t.t. Tokių tinklų architektūrų pavyzdžiai yra variaciniai autoenkoderiai (AE, angl. *autoencoders*), generatyviniai besivaržantys neuroniniai tinklai (GANs, angl. *generative adversarial networks*) bei difuzijos modeliai (DM, angl. *diffusion models*). Šie du tinklų tipai yra labai dažnai naudojami kurti giliųjų klastočių nuotraukas ir vaizdo įrašus [Ver20]. Giliosios klastotės yra tokios, ypač realistiškai atrodančios, nuotraukos ar vaizdo įrašai, kuriuose žmonių veidai ir/ar burnos judesiai yra pakeičiami taip, kad šie žmonės būtų pavaizduoti darantys ar sakantys dalykus, kurie nebuvo iš tikrųjų įvykę. [Wes19] (1 pav.). Šis neuroninių tinklų panaudojimas yra labai kontraversiškas bei kelia daug diskusijų žiniasklaidoje, socialiniuose tinkluose bei tarp įstatymų leidėjų.

1 pav. Giliosios klastotės pavyzdys. Aktorės Amy Adams veidas (kairėje) buvo sukeistas su aktoriaus Nicolas Cage veidu (dešinėje)

Giliąsias klastotes galima išskirti į dvi kategorijas pagal jų sukūrimo ir platinimo intencijas - piktavališkas bei ne piktavališkas. Piktavališkos klastotės gali būti kuriamos siekiant skleisti politinę propagandą, paveikti visuomenės nuomonę, atlikti nusikaltimus, šantažuoti asmenis, manipuliuoti akcijų rinkomis ir t.t. [JAA+20; Ver20; Wes19]. Kaip pavyzdį galima paimti 2022 metais kovo mėnesį išplatintą giliąją klastotę, kurioje Ukrainos prezidentas V.Zelenskis tariamai praneša apie pasidavimą Rusijos vykdomai invazijai į Ukrainą [Sim22], arba uždedant įprastų žmonių ar įžymybių veidus ant pornografinių vaizdo įrašų, taip siekiant pakenkti jų reputacijai ir/arba juos šantažuoti [Mah23]. Tuo tarpu, ne piktavališkos klastotės gali būti naudojamos filmų kūrimui, fotografijoje, vaizdo žaidimuose, virtualioje realybėje, skaitmeninėje komunikacijoje, mados industrijoje, elektroninėje komercijoje ar interneto juokelių kūryboje [Ver20; Wes19]. To pavyzdys būtų, kad filme „Rogue One: A Star Wars Story", 15 sekundžių scenoje, giliųjų klastočių technologija buvo panaudota sugeneruoti devyniolikmetės Princesės Lėjos personažės veidą, pasitelkiant akto-

rės Carrie Fisher jaunystės nuotraukas [Win18]. Taip pat, kitame „Žvaigždžių karų" filme „Solo: A Star Wars Story", aktoriaus Harrison Ford nuotraukos iš 1977 metų buvo panaudotos scenoje, kurioje buvo vaizduojamas Hanas Solo, šio aktoriaus personažas, jaunystėje [Rad18].

Taip pat, svarbu paminėti, kad egzistuoja daug, ganėtinai kokybiškai giliąsias klastotes kuriančių mobiliųjų programėlių bei interneto programų, kurios nereikalauja jokių specifinių žinių apie giliųjų klastočių kūrimą ir dėl to jomis gali sėkmingai klastotes kurti net ir nedaug patirties, dirbant su giliaisiais neuroniniais tinklais, turintys vartotojai. Tokių aplikacijų pavyzdžiai būtų: Reface[1], FaceApp[2], DeepSwap[3].

Kadangi įtikinamų giliųjų klastočių kūrimas tapo lengvai prieinamas beveik ne kiekvienam ir žinant, kad klastotės žaibiškai paplinta tokiuose socialiniuose tinkluose kaip Twitter, YouTube, Facebook ar Reddit, yra svarbu turėti automatizuotų metodų atpažinti ar vaizdinis turinys yra autentiškas ar tai yra klastotė, kad būtų užkardytas kelias potencialiai piktaviliškų giliųjų klastočių plitimui ir jų daromai neigiamai įtakai.

Šiame darbe yra nagrinėjami neuroniniai tinklai gebantys atpažinti giliąsias klastotes. Daugelis nagrinėtų metodų gali atpažinti ir nuotraukų, ir vaizdo įrašų giliąsias klastotes. Visi metodai geba tiksliai atpažinti klastotes iš tokių veidų kriminalistikos duomenų rinkinių kaip FaceForensics++[RCV+19], DFDC[DBP+20], Celeb-DF[LYS+20] ir ForgeryNet[HGC+21].

Darbo tikslas - pagerinti klastočių atpažinimo metodų tikslumą, keičiant jų architektūros komponentus bei pridedant papildomas duomenų augmentacijas. Šiam tikslui pasiekti, buvo iškelti šie uždaviniai:

1. Surinkti internete labiausiai paplitusių giliųjų klastočių duomenų rinkinį ir jį darbe papildomai naudoti mokytų karkasų verifikavimui.

2. Atlikti pasirinktų karkasų architektūros komponentų pakeitimus ir įvertinti pakeitimų įtaką atpažinimo tikslumui ir greitaveikai.

3. Į pasirinktų karkasų architektūrą įtraukti papildomas duomenų augmentacijas ir įvertinti šių augmentacijų įtaką atpažinimo tikslumui.

[1] https://hey.reface.ai

[2] https://www.faceapp.com

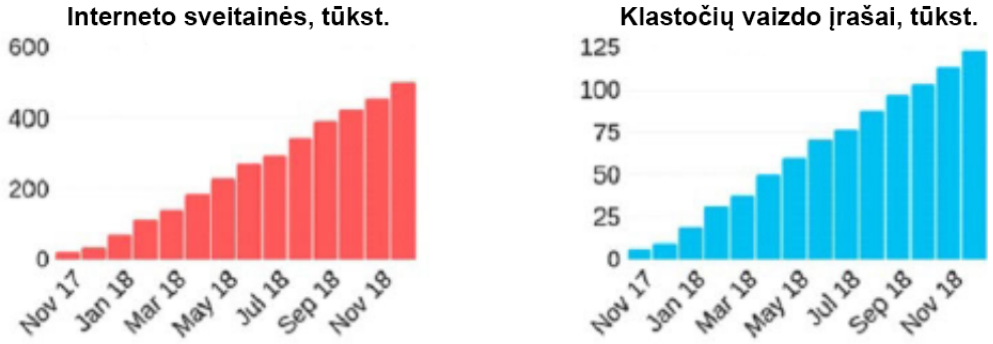
[3] https://www.deepswap.ai

1. Giliosios klastotės

1.1. Kas yra giliosios klastotės

Giliosios klastotės (angl. *deepfakes*), dar kitaip žinomos kaip vaizdinės klastotės, yra dviejų žodžių junginys - gilusis mokymas (angl. *deep learning*) bei klastotės (angl. *fakes*). Giliosios klastotės yra tokie suklastoti vaizdo įrašai, kuriuose žmogaus veidas buvo sukeistas su kito žmogaus veidu ir tam padaryti buvo pasitelktas gilusis mokymas.

Giliųjų klastočių sąvoka pirmą kart buvo panaudota 2017 metais, kai vienas *Reddit* socialinio tinklo vartotojas pradėjo naudoti mašininį mokymąsi sukeisti pornografiniuose vaizdo įrašuose esančių žmonių veidus su įžymybių veidais ir vėliau šiuos suklastotus vaizdo įrašus platino tame pačiame socialiniame tinkle [YXF+21]. Netrukus buvo išleistos *FakeApp*, *FaceSwap*[4] ir kitos aplikacijos, leidžiančios nesunkiai kurti vaizdines klastotes. Laikui einant visuomenės susidomėjimas giliosiomis klastotomis tik augo ir buvo generuojama vis daugiau klastočių (2 pav.) [RNM+22].



2 pav. *Google* paieškos rezultatų kiekis „DeepFake“ frazei: interneto svetainės (kairėje) ir klastočių vaizdo įrašai (dešinėje) [RNM+22]

1.2. Giliųjų klastočių taikymas

Giliųjų klastočių technologija lengvai leidžia kurti kokybišką suklastotą vaizdinį turinį, tam turint didelius kiekius duomenų [Ver20]. Iki šiol, didžiausia dalis internete viešai platinamų giliųjų klastočių vaizdavo įžymybes ar politikus. Tokie vaizdo įrašai yra dažnai naudojami gadinti įžymybių reputaciją ir formuoti viešąją nuomonę, taip keliant pavojų viešajai rimčiai [YXF+21]. Pati giliųjų klastočių technologija neturi gėrio ar blogio atributų, tačiau yra plačiai naudojama kenkėjiškiems tikslams. Kad geriau suprasti šią technologiją ir išvengti jos potencialios žalos, giliosios klastotės yra vis labiau tyrinėjamos akademinėje visuomenėje (3 pav.) [RNM+22].

[4]https://github.com/deepfakes/faceswap

3 pav. Giliąsias klastotes tyrinėjančių mokslinių straipsnių kiekio augimas [RNM+22]

1.2.1. Giliųjų klastočių taikymas piktaviškiems tikslams

Panaudojus specifinio žmogaus veido atvaizdą ir balsą ir pasitelkus giliųjų klastočių technologiją galima sugeneruoti tokį humoristinio, pornografinio ar politinio pobūdžio vaizdo įrašą, kuriame žmogus, be jo žinios ar sutikimo, sako būtent tai, ką pasirenka klastotės autorius. Esminis giliųjų klastočių bruožas yra tai, kad beveik kiekvienas, panaudodamas internete viešai prieinamą programinę įrangą, gali kurti kokybiškas klastotes, praktiškai neatskiriamas nuo originalaus vaizdinio turinio [Ver20; Wes19].

Giliosios klastotės kelia didelį pavojų visuomenei, politinei sistemai ir verslams. Klastotės yra platinamos siekiant skleisti propagandą ir kištis į rinkimus - taip keliant grėsmę nacionaliniam saugumui, sumenkinti visuomenės pasitikėjimą valdžios organais, šantažuoti ir gadinti privačių ar juridinių asmenų reputaciją, kurti pornografinį turinį, kurti netikrus įkalčius teisminiams procesams, manipuliuoti finansinėmis rinkomis, atlikti finansinius nusikaltimus, skleisti melagienas (angl. *fake news*) ir pan. [JAA+20; TVF+20; Ver20; Wes19]. Piktaviškos giliųjų klastočių panaudojimo galimybės yra apribotos tik žmonių fantazijos.

1.2.2. Giliųjų klastočių taikymas doriems tikslams

Nors giliosios klastotės iš pradžių ir buvo sukurtos, ir dažniausiai yra naudojamos piktaviškiems tikslams, šią technologiją galima pasitelkti ir doriems tikslams. Jos yra pasitelkiamos net keliose industrijose: filmų kūrimui, fotografijoje, edukacinėje žiniasklaidoje, skaitmeninėje komunikacijoje, vaizdo žaidimų industrijoje, virtualios realybės industrijoje, socialiniuose tinkluose, sveikatos apsaugoje, mados industrijoje bei elektroninėje komercijoje [JAA+20; TVF+20; Ver20; Wes19].

1.3. Veidų manipuliacijos metodai

Veidų manipuliacijos metodai yra skirstomi į 4 pagrindines grupes: tapatybės apkeitimo (angl. *identity swap*), veido išraiškų apkeitimo (angl. *expression swap*), išvaizdos bruožų manipuliacijos (angl. *attribute manipulation*) bei pilno veido sintezės (angl. *entire face synthesis*) [TVF+20].

1.3.1. Tapatybės apkeitimas

Šio tipo manipuliacijos sukeičia vieno žmogaus veidą su kito žmogaus veidu vaizdo įrašuose, išsaugant originalaus veido judesius bei mimikas (5 pav). Šio tipo manipuliacijos dažniausiai yra įgyvendinamos pasinaudojant trijų tipų metodais:

- Kompiuterine grafika grįsti metodai kaip FaceSwap[5]. Šis tapatybės apkeitimo metodas iš pradžių aptinka veido orientyrus (angl. *landmarks*) ir perkelia pasirinktą tikslinį veidą ant originalaus vaizdo įrašo žmogaus veido. Pasiekti kokybiškesniam galutiniam rezultatui - kad sutaptų veido kontūrų dydis, lygiavimas, veido pokrypis - kadrų suliejimui yra naudojamas „Gauss Newton“ algoritmas.

- Variaciniais autoenkoderiais grįsti metodai kaip Deepfakes[6]. Šiuo tapatybės apkeitimo metodu du variacinius autoenkoderius (angl. *autoencoders*), kurie yra naudoja tą patį enkoderį (angl. *encoder*), yra mokomi atkurti šaltinio (angl. *source*) bei tikslo (angl. *target*) veidus. Be to, yra naudojamas veidų detektorius iškirpti bei sulygiuoti veidus. Kad sugeneruoti klastotę, apmokytam autoendkoderiui, kaip įvestis, yra perduodama veido nuotrauka ir to rezultatas yra suliejamas panaudojant „Poisson“ suliejimą.

- Difuzijos modeliais grįsti metodai kaip LDFaceNet [MMN25], DreamId [YHZ+25] ar ReFace [BLL+25]. Šie tapatybės apkeitimo metodai iš pradžių tikslo nuotraukai prideda triukšmą (angl. *noise*) ir tada vykdo triukšmo mažinimo procesą, tuo pat metu įterpdami šaltinio nuotraukos informaciją tinklui. Kai triukšmas yra mažinamas, modelis išsaugo tikslo asmens pozą, apšvietimą bei foną, bet veido bruožus pritaiko prie šaltinio asmens, tad galutinis kadras atrodo kaip šaltinio asmuo, tačiau dera prie tikslo asmens charakteristikų.

[5]https://github.com/MarekKowalski/FaceSwap
[6]https://github.com/deepfakes/faceswap

4 pav. Tapatybės apkeitimo pavyzdžiai iš Celeb-DF [LYS+20] suklastotų veidų rinkinio [TVF+20]

1.3.2. Veido išraiškų apkeitimas

Veido išraiškų apkeitimas, dar žinomas kaip veido judesių atkartojimas (angl. *face reenactment*), yra metodika, kai vaizdo įraše yra keičiamos vieno žmogaus veido išraiškos kito žmogaus veido išraiškomis. Veidų išraiškų apkeitimui galima pasitelkti daug skirtingų GAN modelių, tačiau populiariausi yra Face2Face [TZS+16] ir NeuralTextures [TZN19].

5 pav. Veido išraiškų apkeitimo pavyzdžiai iš FaceForesincs++ [RCV+19] suklastotų veidų rinkinio [TVF+20]

1.3.3. Išvaizdos bruožų manipuliacija

Šis metodas, dar žinomas kaip veido redagavimas (angl. *face editing*) ar veido retušavimas (angl. *face retouching*), koreguoja tokius žmogaus išvaizdos bruožus kaip: plaukų spalva, odos spalva, lytis, amžius, ar uždeda akinius ir pan. (6 pav.). Šiam metodui dažniausiai yra pasitelkiami GAN architektūros modeliai, tokie kaip StarGAN v2 [CUY+20].

6 pav. Išvaizdos bruožų manipuliacijos pavyzdžiai. Šaltinio veido nuotraukoms yra pritaikomi tikslo veido nuotraukų bruožai [CUY+20]

1.3.4. Pilno veido sintezė

Šis manipuliacijos metodas sugeneruoja neegzistuojančio žmogaus veidą, dažniausiai pasitelkiant difuzijos modelius tokius kaip Stable Diffusion 3 [EKB+24] ar DALL-E 3 [BGJ+23]. Pasitelkus šiuos metodus galima sugeneruoti aukštos kokybės ir realistiškai atrodančias veidų nuotraukas (7 pav.).

7 pav. Sugeneruotų veidų pavyzdžiai iš https://thispersondoesnotexist.com ir https://www.whichfaceisreal.com [TVF+20]

1.4. Giliųjų klastočių kūrimo programinė įranga

Atviro kodo įrankiai:

- FaceFusion - https://github.com/facefusion/facefusion;
- faceswap - https://github.com/deepfakes/faceswap;
- DeepFaceLab - https://github.com/iperov/DeepFaceLab;

iOS ir Android aplikacijos:

- Reface - https://hey.reface.ai;
- FaceMagic - https://www.facemagic.net/download;
- FaceApp - https://www.faceapp.com;
- ZAO - https://zaodownload.com;
- Wombo Deepfake App - https://www.wombo.ai;

Interneto programos:

- DeepSwap - https://www.deepswap.ai;
- Deepfakes Web - https://deepfakesweb.com;
- Synthesia - https://www.synthesia.io;
- SwapFace - https://swapface.org;
- Stable Diffusion - https://stability.ai/stable-assistant;
- ChatGPT - https://chatgpt.com;

2. Giliųjų klastočių atpažinimas

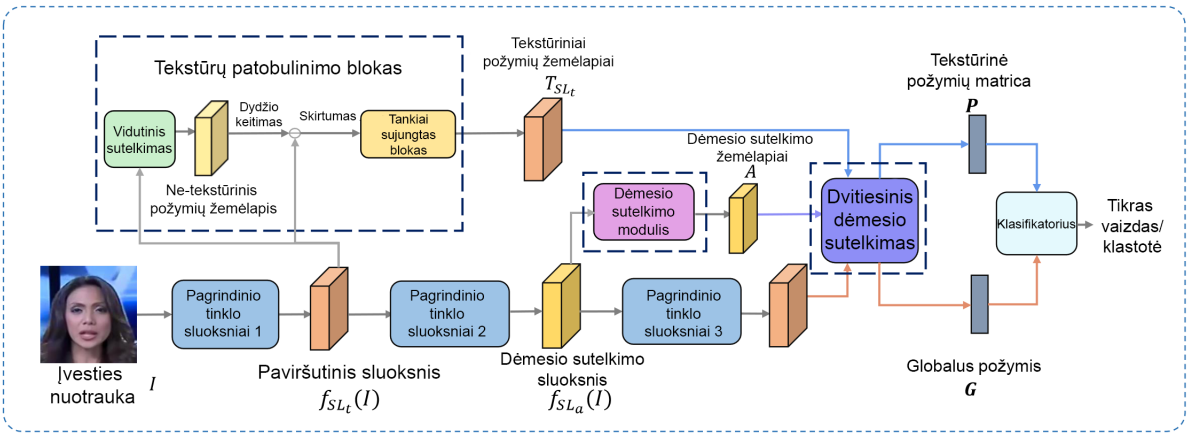
Dauguma dabartinių giliųjų klastočių atpažinimo tinklų geba labai dideliu tikslumu nustatyti giliąsias klastotes, kurios buvo sukurtos jau matytomis metodikomis, tačiau jų tikslumas labai ženkliai smunka bandant atpažinti nematytomis metodikomis sukurtas klastotes [HVP+21]. Taip pat, nesenai pasiūlyti tinklai išsiskyrė geresniu generalizuotu klastočių atpažinimu, t.y. kai nėra ieškoma tik vieno ar kelių specifinių artefaktų klastotei nustatyti, tačiau šių tinklų tikslumas stipriai suprastėdavo vos tik pradėjus naudoti vaizdinių failų suspaudimą (angl. *compression*) [HVP+21]. Šie detektorių bruožai kelia daug nerimo, kadangi realių vartotojų naudojamose programų sistemose giliųjų klastočių detektoriai turėtų atpažinti ir mokymosi metu nematytomis metodais kurtas klastotes.

Dauguma pasiūlytų giliųjų klastočių atpažinimo metodų remiasi binarine klasifikacija - naudojamas požymių ekstraktorius ir iš jo išgauti požymių žemėlapiai yra perduodami į binarinį klasifikatorių, kuris nusako ar nuotrauka yra tikra ar suklastota. Ši metodika puikiai tiko ankstyvųjų klastočių atpažinimui, kai klastotės pasižymėjo akivaizdžiais defektais, tačiau, kadangi skirtumai tarp tikrų ir suklastotų nuotraukų tampa vis mažiau pastebimi - būna lokalūs ir subtilūs, šie, binarine klasifikacija grįsti metodai, nebesugeba tiksliai atpažinti giliųjų klastočių [CZS+22; SY22; ZZC+21].

Šiame darbe buvo pasirinkti 5 giliųjų klastočių detektoriai, kurie geba generalizuotai atpažinti klastotes: MADD, SLADD, FTCN, SBI bei LipForensics.

2.1. MADD karkasas

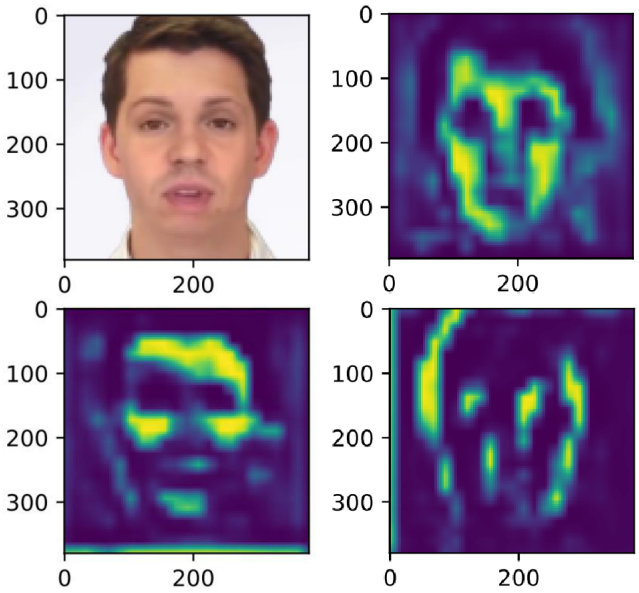
Generalizuotam giliųjų klastočių atpažinimui buvo pasiūlytas MADD (angl. *multi-attentional deepfake detection*) karkasas [ZZC+21], kuris giliųjų klastočių atpažinimui pasitelkia detalią (angl. *fine-grained*) klasifikaciją. Šiame karkase į pagrindinį tinklą (angl. *backbone network*) yra integruojami 3 esminiai komponentai: dėmesio sutelkimo modulis (angl. *attention module*), tekstūrų patobulinimo blokas (angl. *texture enhancement block*) bei dvitiesinis dėmesio sutelkimas (BAP, angl. *bilinear attention pooling*). Šio karkaso architektūra yra pavaizduota 8-ame paveikslėlyje. Karkaso tikslumo įvertinimai yra nurodyti 2-oje lentelėje, 2.6 poskyryje.



8 pav. MADD karkaso architektūra [ZZC+21]

2.1.1. Dėmesio sutelkimo modulis

Dėmesio sutelkimo modulis yra naudojamas sugeneruoti dėmesio sutelkimo žemėlapius (angl. *attention maps*), kuriuose dėmesys yra sutelkiamas į tam tikrus veido regionus - akis, burną ar veido kontūrą (žr. 9 pav.).



9 pav. Dėmesio sutelkimo regionų pavyzdžiai. Tiksliniai regionai yra atskirti bei sutelkia dėmesį į skirtingas veido bruožus [ZZC+21]

Modulis susideda iš 1 X 1 konvoliucijos, rinkinio normalizacijos bei ReLU sluoksnių (žr. 10 pav.). Jo įvestis yra požymių žemėlapis iš specifinio pagrindinio tinklo sluoksnio ir jo išvestis yra dėmesio sutelkimo žemėlapiai.

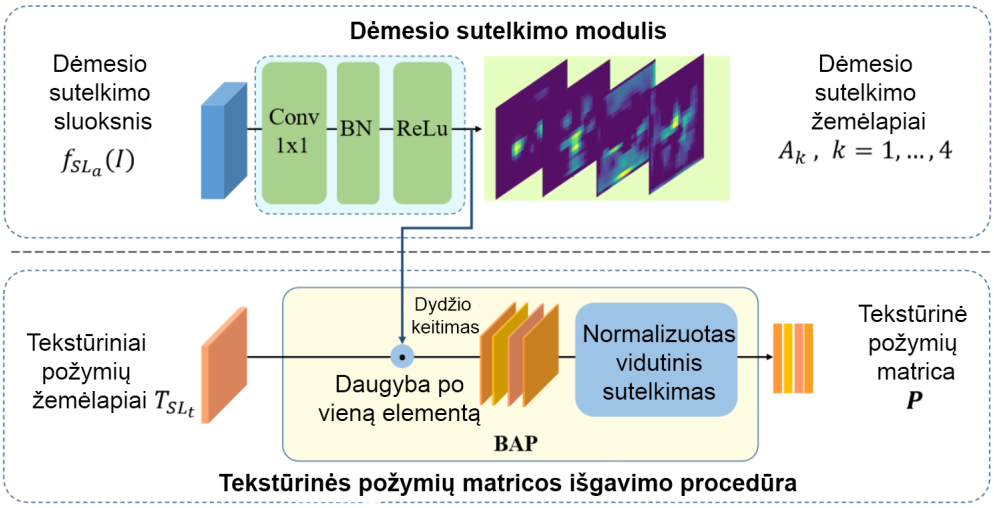
2.1.2. Tekstūrų patobulinimo blokas

Karkaso autoriai pastebi, kad dauguma kitų giliųjų klastočių atpažinimo karkasų visiškai neskiria dėmesio klastočių kūrimo metu atsiradusiems artefaktams, kurie dažniausiai yra ryškūs paviršutinių požymių žemėlapių tekstūrinėje informacijoje (angl. *textural information*). Kad neprarasti šios informacijos ir aptikti šiuos artefaktus, buvo sukurtas tekstūrų patobulinimo blokas.

Šiame bloke iš pradžių yra pasirenkamas specifinis pagrindinio tinklo sluoksnio požymių žemėlapis ir jam yra pritaikomas lokalus vidurkio sutelkimas (angl. *local average pooling*), kad išgauti sutelktą požymių žemėlapį. Tada iš to pačio pasirinkto pagrindinio tinklo sluoksnio požymių žemėlapio yra atimamas sutapatinto dydžio sutelktas požymių žemėlapis. Ir galiausiai to rezultatas yra perduodamas į iš 3 sluoksnių susidedantį tankiai sujungtą (angl. *densely connected*) konvoliucinį bloką. Tekstūrų patobulinimo bloko išvestis yra tekstūrų požymių žemėlapiai (žr. 8 pav.).

2.1.3. Dvitiesinis dėmesio sutelkimas

Jau turint dėmesio sutelkimo žemėlapius bei tekstūrų požymių žemėlapius, BAP yra naudojamas išgauti požymių žemėlapius. BAP yra dvikrypčiai naudojamas ir paviršutiniams, ir giliems požymių žemėlapiams. 10 paveiksliuke yra vaizduojamas paviršutinių požymių žemėlapių išgavimas - dėmesio sutelkimo žemėlapių dydis yra suvienodinamas su tekstūrų požymių žemėlapių dydžiu ir šie žemėlapių rinkiniai yra sudauginami tarpusavyje paelemenčiui (angl. *element-wise multiplication*). Šiai sandaugai yra pritaikomas dvitiesinis dėmesio sutelkimas ir tada autorių pasiūlytas normalizuotas vidutinis sutelkimas (angl. *normalized average pooling*). To rezultatas yra tekstūrinė požymių matrica (angl. *textural feature matrix*).



10 pav. Dėmesio sutelkimo modulio struktūra bei tekstūrinės požymių matricos išgavimas [ZZC+21]

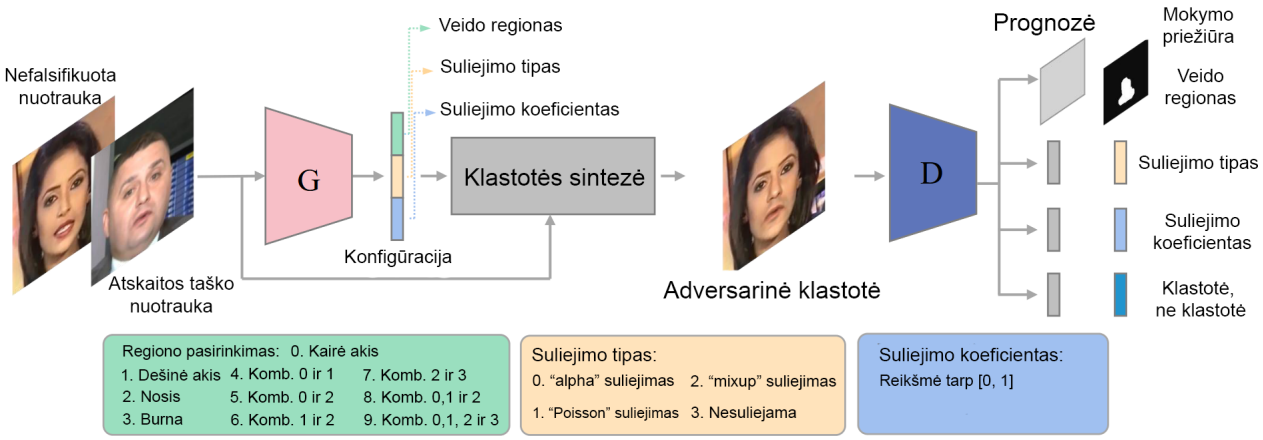
2.1.4. MADD karkaso apribojimai

MADD karkasas kaip įvesties nepriima vaizdo įrašų, o tik jų pavienius kadrus, dėl to jis negeba aptikti giliųjų klastočių laikinio neatitikimo artefaktų.

2.2. SLADD karkasas

SLADD karkasas [CZS+22] pasižymi tuo, kad pats tinklas generuoja vaizdiniais artefaktais pasižyminčias vaizdines klastotes, kad išvengti persimokymo (angl. *overfitting*) problemos, kai atpažinimo modelis labai gerai geba atpažinti klastotes iš tų pačių klastočių duomenų rinkinių iš kurių ir mokėsi, nes mokymosi, testavimo ir validacijos poaibių klastotės būna sugeneruotos naudojant jau modeliui matytas metodas su jau pažįstamais artefaktais. Karkaso tikslumo įvertinimai yra nurodyti 2-oje lentelėje, 2.6 poskyryje.

Karkaso autoriai siūlo pagerinti klastočių detektorių generalizaciją panaudojant adversarines duomenų augmentacijas, kurios padeda sukurti naujus klastočių tipus, bei panaudojant savaiminio mokymosi (angl. *self-supervised*) užduotis, kurios užtikrina karkaso jautrumą klastojimo konfigūracijoms. 11 paveikslėlyje yra pavaizduota karkaso architektūra.



11 pav. SLADD karkaso architektūra. Konfigūracijos sintezatorius išvestys yra naudojamos klastotės generavimui. Mokymo metu yra naudojama adversarinio mokymo strategija, kurioje klastočių sintezatorius yra laikomas kaip generatorius, o klastočių detektorius - kaip diskriminatorius. [CZS+22]

Mokymo metu yra atsitiktinai pasirenkama atskaitos taško (angl. *reference*) nuotrauka, jeigu įvesties nuotrauka nėra suklastota. Tada abi šios nuotraukos yra perduodamos į sintezatoriaus tinklą, kuris sugeneruoja klastotės kūrimo konfigūracijas: veido regioną, kuris bus klastojamas, suliejimo tipą bei suliejimo koeficientą (angl. *mixup blending ratio*). Tada klastotė yra sintezuojama panaudojant abi įvesties nuotraukas su gauta konfigūracija. Galiausiai klastotė yra perduodama į klastočių detektorių, kuris nuspėja ar klastotė yra tikra ar suklastota, bei bando nuspėti panaudotą konfigūraciją. Svarbu paminėti, kad jeigu įvesties nuotrauka jau iš karto yra suklastota, klastotės generavimo žingsnis bus praleistas.

Mokymo metu yra naudojama adversarinio mokymo strategija, kurioje klastočių sintezatorius yra laikomas kaip generatorius, o klastočių detektorius - kaip diskriminatorius. Mokymosi metu sintezatorius yra mokomas kurti tokias naujas klastotes, kurias detektoriui yra sunku atpažinti. Mokymuisi pasibaigus, klastočių atpažinimui yra naudojamas tik detektorius.

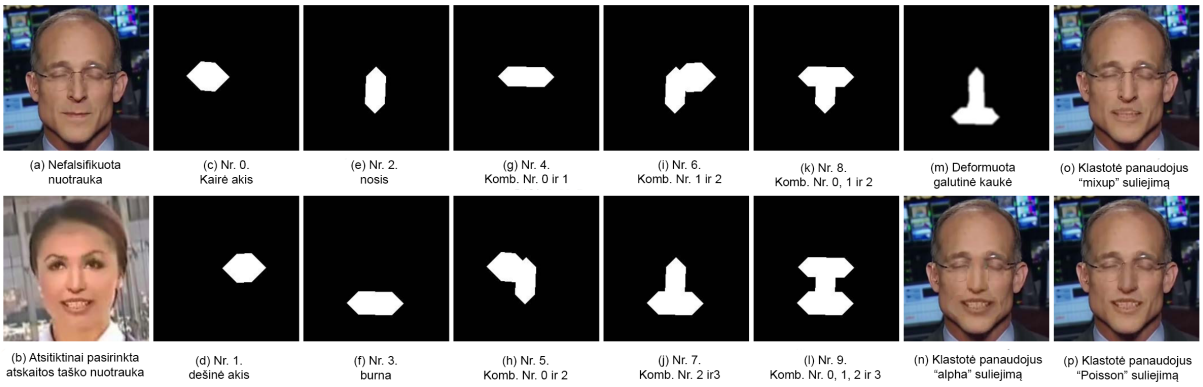
2.2.1. Konfigūracijų bei klastočių sintezatoriai

Konfigūracijos sintezatoriaus tinklas sugeneruoja trijų tipų parinktis: veido regioną, suliejimo tipą bei suliejimo koeficientą. Jie yra parenkami iš žemiau išvardintų galimų reikšmių:

- Veido regionai - kairė akis, dešinė akis, nosis, burna ar bet kuri šių regionų kombinacija - iš viso 10 galimų parinkčių.
- Suliejimo tipas - „alpha“, „Poisson“, „mixup“ suliejimai, arba jokio suliejimo - iš viso 4 galimos parinktys.
- Suliejimo koeficientas - efektyvus tik tada, kai buvo pasirinktas „mixup“ suliejimas. Koeficiento galimos reikšmės yra tarp 0 ir 1 (imtinai).

Prieš perduodant duomenis į klastočių sintezatorių, pasirinktai veido regiono konfigūracijai yra pritaikomas „Gauso“ suliejimas (angl. *Gaussian blur*) su atsitiktinio dydžio branduoliu (angl.

kernel). Galutinės kaukės pavyzdį galima matyti 12 paveiksliuko (m) dalyje. Tada pasirinktas veido regionas yra iškerpamas iš atskaitos taško nuotraukos ir yra suliejamas su nefalsifikuota nuotrauka pagal suliejimo konfigūracija. Kaip ir kituose klastočių kūrimo metoduose, iškirpti veido regionai yra suliejami pritaikant spalvų perdavimą (angl. *color transfer*), bei prieš „alpha“ ir „Poisson“ suliejimus veidai yra sulygiuojami - taip yra išvengiama akivaizdžių artefaktų. Klastočių pavyzdžiai yra vaizduojami (n) - (p) dalyse 12-ame paveikslėlyje.



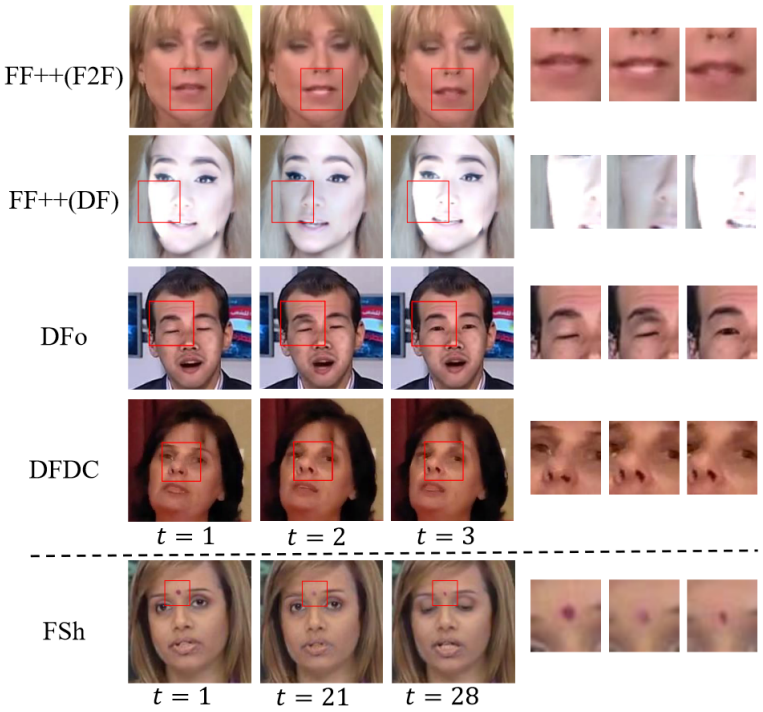
12 pav. Nefalsifikuotos ir atsitiktinai pasirinktos atskaitos taško nuotraukos pavyzdžiai ((a) - (b)), veido regionai ((c) - (m)), kur (m) yra deformuotas veido regionas, sugeneruotos klastotės ((n) - (p)), kurių generavimo metu buvo pasirinktas (j) veido regionas [CZS+22]

2.2.2. SLADD karkaso apribojimai

SLADD karkasas kaip įvesties nepriima vaizdo įrašų, o tik jų pavienius kadrus, dėl to jis negeba aptikti giliųjų klastočių laikinio neatitikimo artefaktų.

2.3. FTCN tinklu grįstas karkasas

FTCN tinklu grįstas karkasas (angl. *framework*) atpažįsta veido klastotes pasinaudodamas jų laikino nenuoseklumo (angl. *temporal coherence*) savybe. Tinklas susideda iš 2 pagrindinių dalių: pilnai laikino konvoliucijų tinklo (FTCN, angl. *fully temporal convolution network*) bei iš laikinio transformerio tinklo (angl. *temporal transformer network*). Karkaso tikslumo įvertinimai yra nurodyti 2-oje lentelėje, 2.6 poskyryje.



13 pav. Laikino nenuoseklumo pavyzdžiai suklastotų duomenų rinkiniuose: FaceForensic++ [RCV+19], DFo [JLW+20], DFDC [DBP+20] bei FSh [LBY+20]. Pirmuose keturiose eilutėse matosi laikino nenuoseklumo pavyzdžiai gretimuose kadruose, o paskutinėje eilutėje yra pavyzdys tarp toli vienas nuo kito esančiuose kadruose [ZBC+21]

Straipsnio autoriai pastebi, kad dauguma giliųjų klastočių vaizdo įrašų kadrai yra generuojami po vieną skirtingą kadrą atskirai, kas lemia tai, kad galutinis klastotės vaizdo įrašas turės akivaizdžių mirgėjimo (angl. *flickering*) bei veido nenuoseklumų trukumų. Taip pat, buvo pastebėta erdvės ir laiko ar konvoliuciniai tinklai ne taip gerai geba išmokti atpažinti laikinus nenuoseklumus.

Buvo pastebėta, kad suklastoti veidai dažniausiai pasižymi dviejų tipų artefaktais: erdviniais, tokiais kaip veido ribų suliejimo (angl. *blending boundary*), „šaškių lentos" efekto (angl. *checkboard*) ar miglos (angl. *blur*), bei laikinu nenuoseklumu. Be specifinės architektūros, dažniausiai yra ieškoma tik erdvinių artefaktų.

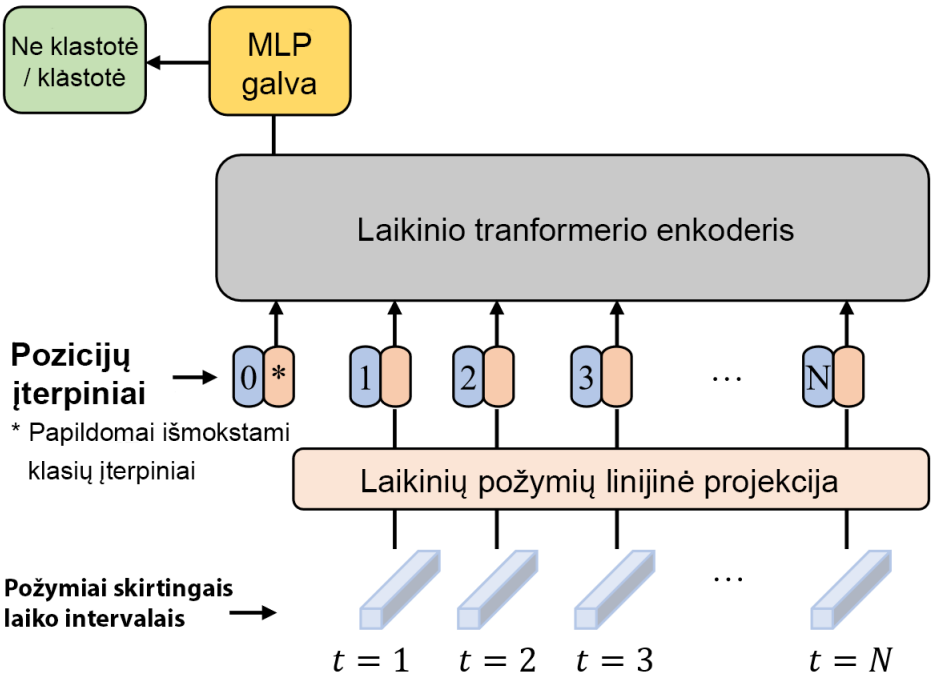
2.3.1. FTCN architektūra

Buvo paimtas erdvės ir laiko (angl. *spatio temporal*) konvoliucijų tinklas ir apribota jo galimybė aptikti erdvinius artefaktus (žr. 1 lent.). Tai buvo padaryta sumažinus konvoliucijos filtro erdvines dimensijas iki 1 ir palikus įprastas dimensijas 3D konvoliucijos operatoriaus konvoliucijos filtro laiko dimensijoje. Dėl šios priežasties tinklas išmoksta aptikti laikinius artefaktus ir beveik nepritaiko erdvinių artefaktų aptikti veidines klastotes.

1 lentelė. Autorių pasiūlytas 3D R50-FTCN modelis kuris atpažįsta giliąsias klastotes. Lyginant su originaliu 3D ResNet-50 modeliu [CZ17], konvoliucijos filtro erdvinis dydis buvo pakeistas į 1. Skliaustuose yra nurodomi liekanų blokai [ZBC+21]

Sluoksnis		Išvesties dydis
$conv_1$	$\mathbf{5 \times 1 \times 1}$, 64, žingsnis 1, 1, 1	$64 \times 32 \times 224 \times 224$
$pool_1$	$1 \times 5 \times 5$ max, žingsnis 1, 4, 4	$256 \times 32 \times 56 \times 56$
res_2	$\begin{bmatrix} 1 \times 1 \times 1, 64 \\ 3 \times 1 \times 1, 64 \\ 1 \times 1 \times 1, 256 \end{bmatrix} \times 3$	$256 \times 32 \times 56 \times 56$
$pool_2$	$2 \times 1 \times 1$ max, žingsnis 2, 1, 1	$256 \times 16 \times 56 \times 56$
res_3	$\begin{bmatrix} 1 \times 1 \times 1, 128 \\ \mathbf{3 \times 1 \times 1}, 128 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 4$	$512 \times 16 \times 28 \times 28$
res_4	$\begin{bmatrix} 1 \times 1 \times 1, 256 \\ \mathbf{3 \times 1 \times 1}, 256 \\ 1 \times 1 \times 1, 1024 \end{bmatrix} \times 6$	$1024 \times 16 \times 14 \times 14$
res_5	$\begin{bmatrix} 1 \times 1 \times 1, 512 \\ \mathbf{3 \times 1 \times 1}, 512 \\ 1 \times 1 \times 1, 2048 \end{bmatrix} \times 3$	$2048 \times 16 \times 7 \times 7$
Erdvinis sutelkimas vidurkinant		$2048 \times 16 \times 1 \times 1$

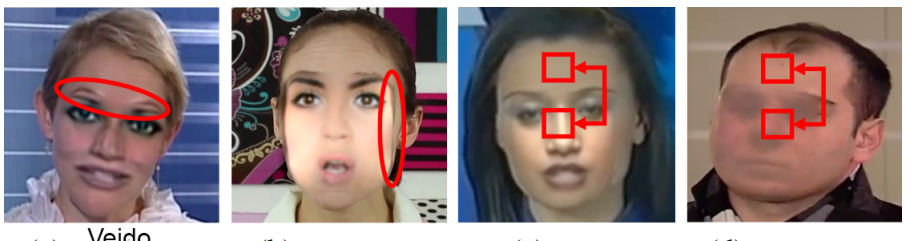
Kadangi nenuoseklumai gali būti tarp toliau nutolusių kadrų, pavyzdžiui, raukšlės ar apgamai iš lėto atsirasti ar tai dingti, buvo nuspręsta pasitelkti transformerius. Prie FTCN galo buvo pridėtas laikinis transformerio tinklas (žr. 14 pav.). Iš abiejų šių dalių susidedantį karkasą galima mokyti vienu metu - t.y. nereikia jų mokyti atskirai.



14 pav. Vaizdinių klastočių atpažinimui naudojamo laikinio transformerio architektūra. FTCN išvestys - požymiai yra padalijami pagal laiko dimensiją ir jų sekos, kaip įvestys, yra perduodamos į standartinio transformerio enkoderį. Į sekas, taip pat, yra pridedami klasifikacijos žetonai, kad būtų galima išmokti klastotės kūrimo manipuliaciją. MLP galva nusako ar vaizdo įrašas yra klastotė ar ne [ZBC+21]

2.4. SBI karkasas

Dauguma giliųjų klastočių aptikimo tinklų labai tiksliai aptinka mokymosi metu išmoktus klastočių bruožus, tačiau jų tikslumas labai ženkliai smunka bandant aptikti klastotes iš nematytų suklastotų veidų duomenų rinkinių, kuriose klastotės buvo generuojamos su nematytais metodais. Vienas iš labiausiai efektyvių šios problemos sprendimo būdų yra mokyti detektorius su sintetiniais duomenimis, siekiant kad modeliai išmoktų bendrines giliųjų klastočių charakteristikas. Remiantis būtent tokiu principu buvo sukurtas su savimi sulietų nuotraukų (SBI, angl. *self-blended images*) karkasas. Karkaso tikslumo įvertinimai yra nurodyti 2-oje lentelėje, 2.6 poskyryje.

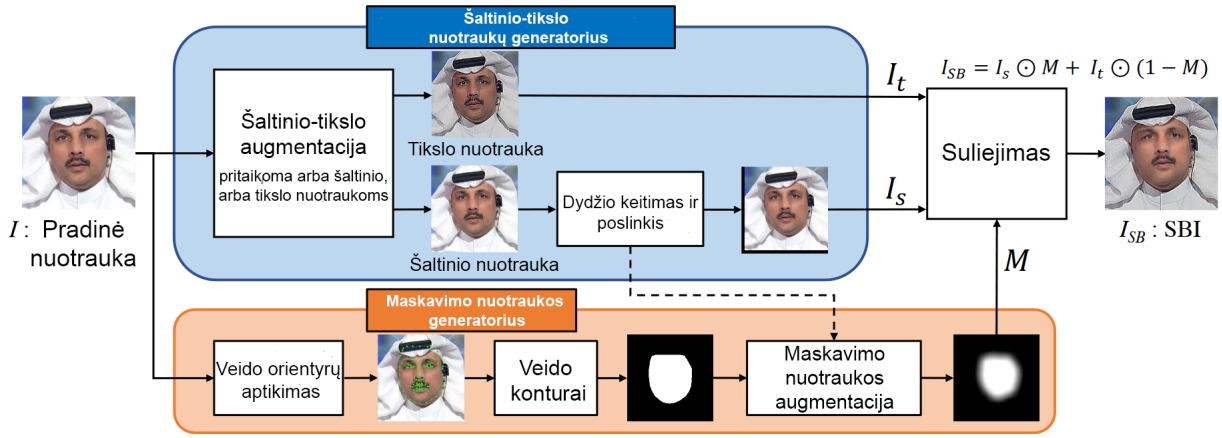


15 pav. Tipiniai giliųjų klastočių veidų artefaktai: (a) veido orientyrų neatitikimas, (b) veido ribų neatitikimas, (c) spalvos neatitikimas ir (d) dažnio neatitikimas [SY22]

SBI karkaso tikslas yra aptikti statistinius neatitikimus tarp koreguotų veidų ir jų fonų giliosiose klastotėse. Kad apmokyti bendrai panaudojamus ir aukštos kokybės detektorius, yra gene-

ruojamos sintetinės klastotės su dažnai sutinkamais bei sunkiai atpažįstamais klastočių pėdsakais. Autoriai pastebi, kad giliųjų besivaržančių neurorinių tinklų sukurtos klastotės taps vis tikroviškesnės, dėl to sukūrė sintetinių klastočių kūrimo karkasą, kuriame klastotės yra kuriamos suliejant iš vienos tos pačios nuotraukos išskirtas ir apdorotas tikslinę nuotrauką bei pseudo šaltinį (angl. *pseudo source*), siekiant mokymo metu detektoriams duoti kuo labiau sudėtingesnę bei bendrinę klastočių atpažinimo užduotį.

SBI karkasą sudaro dvi dalys - šaltinio-tikslo nuotraukų generatorius bei maskavimo nuotraukos generatorius.



16 pav. SBI generavimo architektūra. Pradinė nuotrauka yra perduodama ir į šaltinio-tikslo nuotraukų generatorių, ir į maskavimo nuotraukos generatorių. Šaltinio-tikslo nuotraukų generatorius iš pradinės nuotraukos, panaudodamas nuotraukų transformacijas, sugeneruoja tikslo ir šaltinio nuotraukas. Maskavimo nuotraukos generatorius sugeneruoja maskavimo nuotrauką ir pritaiko jai deformacijas, kad būtų išgauta didesnė veido artefaktų įvairovė. Galiausiai, tikslo, šaltinio ir maskavimo nuotraukos yra suliejamos į vieną [SY22]

2.4.1. Šaltinio-tikslo nuotraukų generatorius

Šaltinio-tikslo nuotraukų generatorius sugeneruoja šaltinio ir tikslo nuotraukas, kurios vėliau bus sulietos tarpusavyje. Taip pat, šios nuotraukos yra atsitiktinai koreguojamos: keičiamas jų raudonos, žalios, mėlynos spalvų kanalų reikšmės, pritaikomos atspalvio, spalvų prisotinimo, šviesumo bei kontrasto transformacijos siekiant sukurti spalvinius bei dažninius nuotraukų tarpusavio neatitikimus. Šaltinio nuotrauka yra paslenkama (angl. *translated*) bei pakeičiamas jos dydis siekiant sukurti netinkamo veido suliejimo (angl. *blending boundaries*) bei veido orientyrų (angl. *landmarks*) defektus.

2.4.2. Maskavimo nuotraukos generatorius

Maskavimo nuotraukos generatorius sukuria pilkos spalvos tonų maskavimo nuotrauką su papildomais defektais. Iš pradžių generatorius panaudoja veido orientyrų detektorių aptikti veido kontūrus ir sukuria pradinę maskavimo nuotrauką apskaičiuodamas veido kontūrus. Tada maskavimo nuotrauka yra deformuojama.

2.4.3. Suliejimas

Galiausiai, šaltinio, tikslo bei maskavimo nuotraukos yra suliejamos į vieną ir taip yra sukuriamos su savimi sulietos nuotraukos.

2.4.4. SBI karkaso architektūra

Tinklo autoriai kartu su šaltinio-tikslo nuotraukų generatoriumi naudojo EfficientNet-b4 [TL19] konvoliucinį neuroninį tinklą, iš anksto apmokytą su ImageNet [RDS+15] nuotraukų rinkiniu. Mokymo metu šis karkasas ima po vieną kadrą iš mokymo poaibio ir padaro jo kopiją. Vienai kopijai yra atsitiktinai pritaikomos duomenų augmentacijos ir ta nuotrauka yra laikoma kaip originalus žmogaus veidas, o iš kitos nuotraukos yra generuojama su savimi sulieta nuotrauka, kuriai vėliau irgi yra atsitiktinai pritaikomos duomenų augmentacijos.

2.4.5. SBI karkaso naudojamos duomenų augmentacijos

SBI karkasas atsitiktinai naudoja tokias augmentacijas iš *Albumentations* Python bibliotekos [BIK+20]:

- RGB poslinkio (angl. *RGB shift*);
- Atspalvio, sodrumo ir spalvos reguliavimo (angl. *hue saturation*);
- Šviesumo ir kontrasto keitimo (angl. *brightness contrast*);
- Nuotraukos suspaudimo (angl. *image compression*);
- Ryškinimo (angl. *sharpen*);

2.4.6. SBI karkaso apribojimai

SBI karkasas kaip įvesties nepriima vaizdo įrašų, o tik jų pavienius kadrus, dėl to jis negeba aptikti giliųjų klastočių laikinio neatitikimo artefaktų.

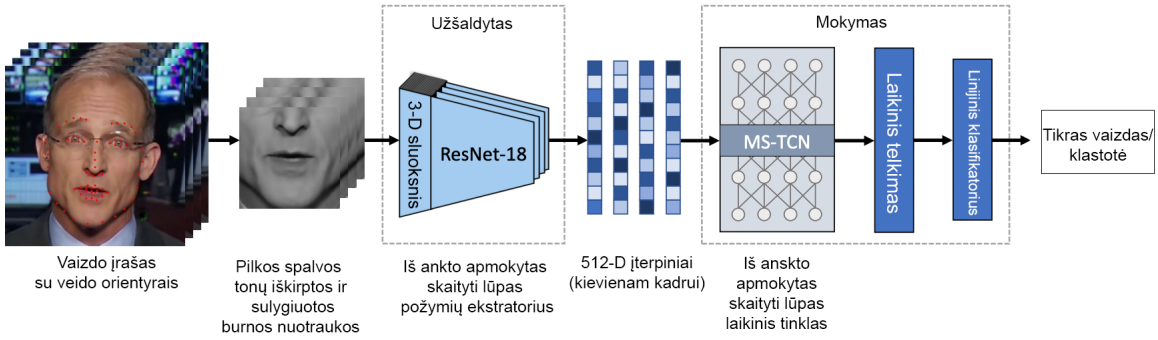
2.5. LipForensics karkasas

Generalizuotam giliųjų klastočių atpažinimui buvo pasiūlytas LipForensics [HVP+21] karkasas, kurio veikimo principas yra paremtas netaisyklingų burnos judesių klasifikavimu, kurie atsiranda dėl vaizdinių klastočių kūrimo metodų defektų. Tinklo autoriai kelia hipotezę, kad nenatūralūs burnos judesiai klastotėse egzistuoja nepriklausomai nuo to, koks klastotės kūrimo metodas buvo naudotas. Taip pat, yra pastebima, kad dauguma klastočių keičia burnos judesius taip, kad sutaptų su asmens tapatybe, kalba ir išraiška. Dėl sunkumo atkartoti natūralius burnos judesius, dabartiniai klastočių kūrimo metodai dažnai sugeneruoja tokius judesius, kurie net ir neįgudusiai akiai neatrodo itin realistiški. Pavyzdžiui, suklastotos burnos dažnai pilnai neužsidaro, kai yra tariamos specifinės fonemos, burnos judesiai vyksta nenatūraliu greičiu bei atsiranda staigūs burnos

išvaizdos ar dantų pakitimai, kad ir tuo atveju, kai vaizdo įrašo subjektas net nekalba [HVP+21]. Karkaso tikslumo įvertinimai yra nurodyti 2-oje lentelėje, 2.6 poskyryje.

2.5.1. LipForensics architektūra

Šio tinklo architektūra (17 pav.) susideda iš 2 dalių - konvoliuciniu neuroniniu tinklu grįstu požymių ekstraktoriaus (angl. *feature extractor*) bei kelių mastelių laiko konvoliuciniu tinklu (angl. *multi-scale temporal convolutional network*). Požymių ekstraktoriui buvo pasirinktas ResNet-18 [HZR+16] tinklas su pradiniu 3D konvoliuciniu sluoksniu. Ekstraktoriaus išvestis yra 512 dimensijų vektorius kiekvienam įvesties kadrui. Kelių mastelių laikinas konvoliucinis tinklas yra MS-TCN [MMP+20] architektūros. Jis apjungia trumpalaikę ir ilgalaikę laikiną informaciją kiekviename sluoksnyje, apjungiant išvestis iš kelių šakų, kuriose egzistuoja skirtingos laikinosios matymo sritys (angl. *receptive fields*). Šio tinklo išvestys yra perduodamos į laiko globalų sutelkimo sluoksnį (angl. *temporal global average pooling layer*) ir galiausiai patenka į tiesinį klasifikatorių, kuris apskaičiuoja klastotės tikimybes.



17 pav. LipForensics tinklo architektūra mokymo metu. [HVP+21]. Tinklo įvestį sudaro 25 pilkos spalvos atspalvių iškirptos burnos nuotraukos, tada jos yra kaip įvestis perduodamos į užšaldytą, ResNet-18 tinklu paremtą, su pradiniu 3-D konvoliucijos sluoksniu, požymių ekstraktorių, kuris iš pradžių buvo apmokytas skaityti lūpas ir kurio išvestis yra burnos judesiams jautrūs įterpiniai, kurie yra perduodami į MS-TCN tinklą. MS-TCN tinklas, taip pat, yra iš pradžių apmokytas skaityti lūpas. Mokymo metu MS-TCN yra permokomas atpažinti giliąsias klastotes pagal nereguliarius burnos judesius [HVP+21]

2.5.2. LipForensics karkaso apribojimai

Šis tinklas veikia tik su vaizdo įrašais, todėl jo negalima naudoti su pavieniais kadrais. Be to, šis karkasas nesugebėtų atpažinti klastočių vaizdo įrašų, jeigu juose žmogaus burna būtų dėl kažkokios priežasties nematoma (kažkas ją dengtų - pvz. veido kaukė) arba burna nebūtų manipuliuota. Taip pat, autoriai pastebėjo, kad dauguma blogų karkaso prognozių įvyko dėl staigių galvos judesių, mokymo rinkinyje retai matytų pozų arba labai ribotų burnos judesių.

2.6. Giliųjų klastočių atpažinimo karkasų palyginimas

2-oje lentelėje yra pavaizduotas aprašytų giliųjų klastočių atpažinimo karkasų palyginimas. Svarbu pastebėti, kad MADD ir SLADD karkasuose buvo pateiktas tikslumo įvertinimas, iš pra-

džių apmokius karksą su FaceForesics++, tik su kitiems karkasams bendru CDF suklastotų veidų rinkiniu, todėl negalima plačiau palyginti šių dviejų karkasų tikslumo su FTCN, LipForensics bei SBI karkasais. Be to, FTCN bei LipForensics modeliai kaip įvestį priima vaizdo įrašus, o likę modeliai - tik pavienius kadrus iš vaizdo įrašų, todėl šie du tinklai geba papildomai atpažinti klastotes pagal jų laikinio neatitikimo artefaktus.

Karkasų tikslumą įvertinus pagal vienintelį bendrą verifikavimo rinkinį - CDF - tiksliausias yra SBI karkasas, tačiau šis karkasas nusileidžia FTCN bei LipForensics karkasams juos visus vertinus pagal DFDC rinkinį.

2 lentelė. Karkasų palyginimas. Šioje lentelėje visi karkasai buvo apmokyti su FaceForensincs++ suklastotų veidų rinkiniu ir buvo testuojamas jų tikslumas su CDF [LYS+20], DFDC [DBP+20], FSh [LBY+20], DFo [JLW+20] rinkiniais. Nurodyti skaičiai yra AUC reikšmės (žr. 3.9 poskyrį) procentais

Karkasas	Įvesties tipas	CDF	DFDC	FSh	DFo
FTCN	Vaizdo įrašas	86.9	74.0	98.8	98.8
LipForensics	Vaizdo įrašas	82.4	73.5	97.1	97.6
SBI	Pavienis kadras	92.87	72.42	-	-
MADD	Pavienis kadras	67.44	-	-	-
SLADD	Pavienis kadras	79.7	-	-	-

3. Suklastotų veidų duomenų rinkiniai

Giliųjų neuroninių tinklų mokymui yra reikalingi labai dideli kiekiai duomenų, kad pasiekti optimalų siekiamą rezultatą [SM19]. Taip pat, sprendžiant tą pačia problemą su giliaisiais neuroniniais tinklais, reikia būdo objektyviai įvertinti skirtingų neuroninių tinklų veikimo korektiškumą. Dėl šios priežasties mokslininkai renka didelius ir įvairius duomenų rinkinius ir vėliau juos platina. Šiame skyriuje yra apžvelgiami FaceForensics++, DFDC, CDF, ForgeryNet, FSh, DFo, DF40 suklastotų veidų duomenų rinkiniai bei klastočių atpažinimo tikslumo vertinimo metrikos.

3.1. FaceForensics++ suklastotų veidų rinkinys

FaceForensics++ (FF++) [RCV+19] yra suklastotų veidų rinkinys (angl. *forensics dataset*) susidedantis iš 1000 originalių vaizdo įrašų, kurie buvo apdoroti 4 automatiniais veidų manipuliacijos metodais Deepfakes, Face2Face, FaceSwap ir NeuralTextures.

Šį suklastotų veidų duomenų rinkinį sudaro 5 000 vaizdo įrašai (4 000 suklastoti ir 1 000 nesuklastoti). Suklastotų vaizdo įrašų kūrimui buvo pasitelkti 5 skirtingi metodai, panaudota 2 unikalios vaizdinės perturbacijos.

3.2. DFDC suklastotų veidų rinkinys

DFDC [DBP+20] suklastotų veidų duomenų rinkinį sudaro 128 064 vaizdo įrašai (104 500 suklastoti ir 23 564 nesuklastoti). Suklastotų vaizdo įrašų kūrimui buvo pasitelkta 8 skirtingos metodikos, panaudota 19 unikalių vaizdinių perturbacijų. Vaizdo medžiagoje iš viso buvo nufilmuoti 960 skirtingi asmenys.

3.3. CDF suklastotų veidų rinkinys

Celeb-DFv2 (CDF) [LYS+20] suklastotų veidų duomenų rinkinį sudaro 6 229 vaizdo įrašai (5 639 suklastoti ir 590 nesuklastoti). Suklastotų vaizdo įrašų kūrimui buvo pasitelktas tik 1 metodas ir nebuvo panaudota jokių vaizdinių perturbacijų. Vaizdo medžiagoje iš viso buvo nufilmuoti 59 skirti asmenys.

3.4. ForgeryNet suklastotų veidų rinkinys

ForgeryNet [HGC+21] yra didžiausias viešai prieinamas suklastotų veidų duomenų rinkinys, kurį sudaro 221 247 vaizdo įrašai (121 617 suklastoti ir 99 630 nesuklastoti) bei 2,9 milijonai nuotraukų (1 457 861 suklastotos ir 1 438 201 nesuklastotos). Suklastotų vaizdo įrašų kūrimui buvo pasitelkta 15 skirtingų metodikų, panaudotos 36 unikalios vaizdinės perturbacijos, kurios buvo taip pat ir kombinuojamos tarpusavyje klastočių generavimo metu. Vaizdo medžiagoje iš viso buvo nufilmuoti ar nufotografuoti 5 400 skirti asmenys.

Šis duomenų rinkinys tinka nuotraukų klastojimo klasifikavimui, nuotraukų erdvinei klastojimo lokalizacijai, vaizdo įrašų klastojimo klasifikavimui bei laikinų suklastotų vaizdo įrašų kadrų

sekų lokalizacijai:

- Nuotraukų klastojimo klasifikavimo uždaviniams spręsti yra sukurtos anotacijos ar nuotrauka yra suklastota ar ne, bei anotacijos ar nuotrauka yra tikra, suklastota su pakeista tapatybe, ar suklastota su ta pačia tapatybe.
- Nuotraukų erdviniams klastojimo lokalizacijos uždaviniams spręsti yra sukurtos anotacijos, kuriose yra nurodomi nuotraukos suklastoti segmentai.
- Vaizdo įrašų klastojimo klasifikavimui uždaviniams spęsti sukurtos anotacijos ar vaizdo įrašas yra suklastotas ar ne, bei anotacijos ar vaizdo įrašas yra tikras, suklastotas su pakeista tapatybe, ar suklastotas su ta pačia tapatybe.
- Laikinų suklastotų vaizdo įrašų kadrų sekų lokalizacijos uždaviniams spęsti sukurtos anotacijos nusako kurie vaizdo įrašų segmentai buvo suklastoti.

3.5. FSh suklastotų veidų rinkinys

FaceShifter (FSh) yra tas pats FaceForensics++ rinkinys, tik papildomai panaudotas FaceShifter [LBY+20] giliųjų klastočių kūrimo metodas ir dėl to suklastotų vaizdo įrašų poaibis pasipildė 1 000 klastočių.

3.6. DFo suklastotų veidų rinkinys

DeeperForensics-1.0 (DFo) [JLW+20] suklastotų veidų duomenų rinkinį sudaro 60 000 vaizdo įrašų (10 000 suklastoti ir 50 000 nesuklastoti). Suklastotų vaizdo įrašų kūrimui buvo pasitelkta 1 metodika, panaudotos 35 unikalios vaizdinės perturbacijos. Vaizdo medžiagoje iš viso buvo nufilmuoti 100 skirtingų asmenų.

3.7. DF40 suklastotų veidų rinkinys

DF40 [YYC+24] suklastotų veidų rinkinį sudaro daugiau kaip 100 000 suklastotų vaizdo įrašų, 1890 tikrų vaizdo įrašų ir per 1 000 000 suklastotų nuotraukų. Suklastotų įrašų generavimui buvo pasitelktos 40 unikalių metodikų. Vaizdo medžiagoje iš viso buvo nufilmuoti 100 skirtingų asmenų.

3.8. Rinkinių palyginimas

Aprašytų suklastotų veidų rinkinių skirtingos charakteristikos yra vaizduojamos 3-ioje lentelėje. Galima pastebėti, kad kuo veidų rinkinys yra naujesnis - tuo jis turi daugiau vaizdo įrašų, vaizdo įrašuose vaizduojamų žmonių bei joms sugeneruoti yra panaudojama vis daugiau vaizdinių perturbacijų - taip siekiant išmokyti tikslesnius giliųjų klastočių detektorius. Taip pat, ankstyvesni rinkiniai, tokie kaip FaceForensics++ ir CDF, komponuoja vaizdo įrašus surastus internete, o jau naujesni rinkiniai, kaip ForgeryNet ar DFo, komponuoja tokius vaizdo įrašus, kuriems nufilmuoti buvo pasamdyti aktoriai.

3 lentelė. Suklastotų veidų rinkinių charakteristikos: tikrų ir suklastotų vaizdo įrašų, panaudotu klastočių kūrimo metodų, skirtingų subjektų ir panaudotų unikalių vaizdinių perturbacijų kiekiai

Rinkinys	Tikrų vaizdo įrašų #	Suklastotų vaizdo įrašų #	Klastočių kūrimo metodų #	Subjektų #	Perturbacijų #
FaceForensics++	1 000	4 000	4	-	2
DFDC	23 564	104 500	8	960	19
CDF	590	95 639	1	59	-
ForgeryNet	99 630	121 617	15	5 400	36
FSh	1 000	5 000	5	-	2
DFo	50 000	10 000	1	100	35
DF40	1 890	100 000	40	100	

3.9. Giliųjų klastočių atpažinimo tikslumo vertinimo metrikos

- Plotas po (ROC) kreive (AUC, angl. *area under curve*) - tai yra metrika nusakanti klasifikatoriaus tikslumą. Sprendimus priimančiojo ypatybių kreivė (ROC, angl. *receiver operating characteristic*) yra kreivė, nurodanti klasifikatoriaus jautrumo (angl. *sensitivity*) bei specifiškumo (angl. *specificity*) sąryšį. AUC reikšmės patenka į $[0, 1]$ rėžius. Klasifikatorius veikia nepriekaištingai, kai jo AUC metrikos reikšmė yra 1.

- Tikslumas (angl. *accuracy*) - yra metrika nusakanti, kiek procentų vaizdo įrašų buvo teisingai atpažinti kaip klastotės.

4. Eksperimento projektavimas

4.1. Giliųjų klastočių atpažinimo karkasų vertinimo kriterijai

Tolimesniems tyrimams - karkaso architektūros keitimo, papildomų duomenų augmentacijų naudojimo bei modelių verifikavimo su nuosavu giliųjų klastočių rinkiniu, reikėjo išsirinkti vieną karkasą. Tam buvo sudarytas vertinimo kriterijų rinkinys:

- Ar karkaso verifikavimo kodas yra viešai prieinamas, kad jį būtų galima verifikuoti identiškai, kaip tą darė karkaso autoriai.
- Ar karkaso mokymo kodas yra viešai prieinamas, kad jį būtų galima apmokyti identiškai, kaip tą darė karkaso autoriai.
- Karkaso veikimo tikslumas AUC (%), iš pradžių karkasą apmokius su FF++(HQ) rinkiniu ir tada verifikavus su Celeb-DFv2 (CDF) duomenų rinkiniu.
- Karkaso veikimo tikslumas AUC (%), iš pradžių karkasą apmokius su FF++(HQ) rinkiniu ir tada verifikavus su Deepfake Detection Challenge Dataset (DFDC) duomenų rinkiniu.

4.2. Giliųjų klastočių atpažinimo karkaso pasirinkimas tolimesniems tyrimams

Sudarius vertinimo kriterijų sąrašą ir pagal juos įvertinus karkasus aprašytus 2 skyriuje buvo gauta 4 lentelė.

4 lentelė. Išnagrinėtų giliųjų klastočių atpažinimo karkasų atitikimas vertinimo kriterijams

Karkasas	Prieinamas atviras verifikavimo kodas	Prieinamas atviras mokymo kodas	CDF AUC (%)	DFDC AUC (%)
MADD	Taip	Taip	67.44	-
SLADD	Taip	Taip	79.7	-
FTCN	Taip	Ne	86.9	74.0
SBI	Taip	Taip	92.87	72.42
LipForensics	Taip	Ne	82.4	73.5

Svarbu atkreipti dėmesį į tai, kad MADD ir SLADD karkasai neturi AUC reikšmių verifikuojant su DFDC rinkiniu. Buvo ieškota šių reikšmių moksliniuose straipsniuose, susijusiuose su giliųjų klastočių atpažinimu, tačiau šių reikšmių surasti nepavyko. To pasekoje, pats autorius norėjo verifikuoti šiuos karkasus su DFDC rinkiniu, tačiau darbo rašymo metu nebuvo galimybės atsisiųsti DFDC rinkinio.

Analizuojant 4 lentelę yra matyti, kad FTCN ir LipForensics karkasų autoriai nėra paskelbę mokymo atvirojo kodo, kas iš esmės priverčia atmesti šiuos modelius nuo jų pasirinkimo tolimesniems tyrimams. Iš likusių 3 karkasų didžiausią AUC reikšmę su CDF rinkiniu, 92.97%, pasiekė SBI karkasas. Taip pat, atsižvelgiant į tai, kaip yra susijusios FTCN ir LipForensics karkasų AUC reikšmės juos verifikuojant su CDF ir DFDC rinkiniais, galima daryti prielaidą, kad panašiai bus

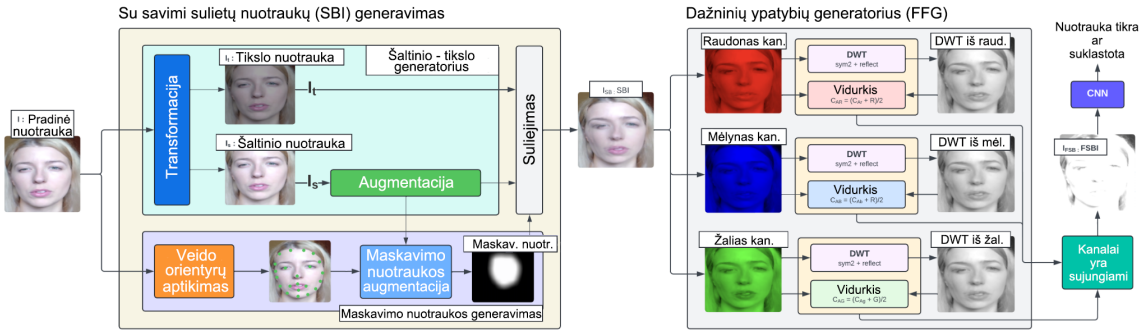
susijusios MADD ir SLADD karkasų AUC reikšmės verifikuojant šiuos karkasus su CDF ir DFDC rinkiniais - t.y. MADD ir SLADD karkasų AUC reikšmės turėtų būti mažesnės DFDC rinkinyje, negu likusių 3 karkasų: SBI, FTCN ir LipForensics. Dėl šių priežasčių tolimesniems tyrimams buvo pasirinktas SBI karkasas.

4.3. FSBI karkasas

Dažniniais požymiais patobulintų su savimi sulietų nuotraukų (FSBI, angl. *frequency enhanced self-blended images*) giliųjų klastočių aptikimo karkasas yra SBI karkaso patobulinta versija [HLK+25]. Šiame darbe taip pat bus tiriamas ir šis karkasas su jame pasiūlytais patobulinimais.

4.3.1. Dažninių požymių generatorius

Mokymo metu, su savimi sulietos nuotraukos yra perduodamos į dažninių požymių generatorių (FFG, angl. *frequency features generator*), kuris su savimi sulietą nuotrauką išskaido į raudonos, mėlynas bei žalios spalvų kanalus. Tada kiekvienam kanalui pritaikoma keletas diskrečių vilnelių transformacijų (DWT, angl. *discrete wavelet transforms*), kurios išryškiną dažninę informaciją su savimi sulietose nuotraukose. Tada šie transformuoti kanalai yra suliejami tarpusavyje ir perduodami konvoliuciniam neorininiam tinklui, kad klasifikuoti ar nuotrauka yra tikra ar tai yra klastotė (18 pav.). Geriausią rezultatą pavyko išgauti naudojant *Symlet* transformacijų šeima diskrečių vilnelių transformacijoms, kuri pasižymi tuo, kad subalansuoja glotnumą ir osciliacijas, todėl tinka signalams, kuriuose yra ir tolygių, ir staigių pokyčių. Be to, siekiant sumažinti iš signalų ekstrapoliacijos kylančius artefaktus, yra naudojami išplėsties rėžimai (angl. *extension*) modes, iš kurių geriausias rezultatas buvo pasiektas su atspindies (angl. *reflective*) rėžimu.



18 pav. FSBI karkaso veikimo principas mokymo metu [HLK+25]

4.3.2. FSBI architektūra

Karkaso autoriai naudojo EfficientNetB5 bei EfficientNetB4 konvoliucinius neuroninius tinklus. Pasak straipsnio autorių, FSBI+EfficientNetB5 karkasas pasiekė 95.49% AUC iš anksto jį apmokius su FF++(HQ) rinkiniu ir jį verifikuojant su CDF rinkiniu, o FSBI+EfficientNetB4 karkasas pasiekė 94.29%. Straipsnyje, pagrindinėje karkasų palyginimo lentelėje, autorių pasi-

rinkimas nurodyti FSBI+EfficientNetB5 AUC reikšmę ir lyginti ją su SBI+EfficientNetB4 karkaso AUC reikšme, išreikštinai nenurodant naudotų konvoliucinių tinklų, pasirodė kaip abejotinas sprendimas. Tokiu atveju autoriams reikėjo nurodyti FSBI karkaso pasiektą AUC reikšmę su EfficientNetB4 konvoliuciniu neuroniniu tinklu.

4.3.3. FSBI karkaso naudojamos duomenų augmentacijos

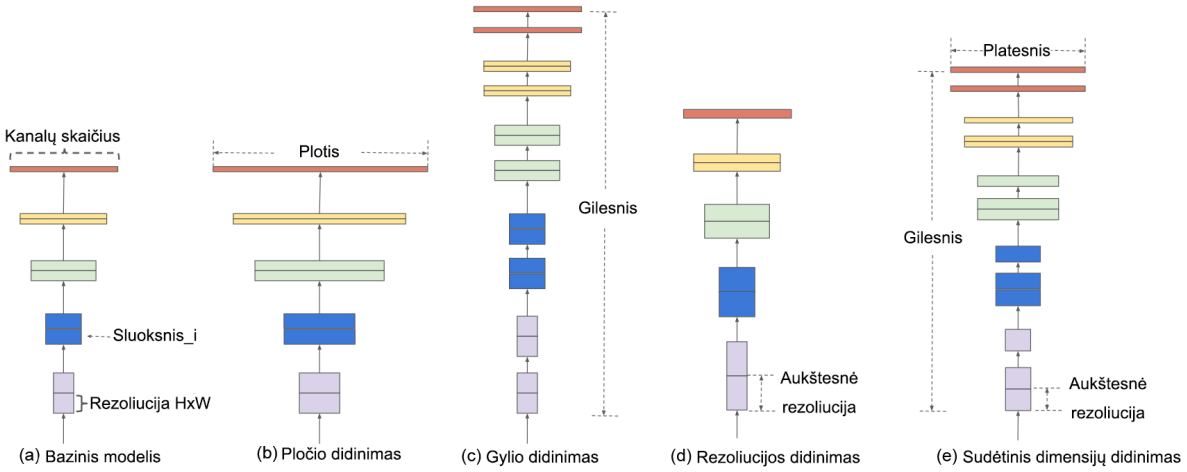
FSBI karkasas naudoja tas pačias augmentacijas kaip ir SBI karkasas, apart dažninių požymių generatoriaus.

Pats dažninių požymių generatorius yra visados naudojamas augmentuoti su savimi sulietas nuotraukas, kas šio darbo autoriaus nuomone nėra efektyviausias sprendimas, nes tai turėtų naudojama tik kaip papildoma duomenų augmentacija, nes tikrai ne visos klastotės pasižymi dažniniais defektais. Taip pat, FSBI karkaso autoriai tikras (ne su savimi sulietas) nuotraukas visados augmentuoja su dažninių požymių generatoriumi, kas šio darbo autoriui irgi pasirodė kaip netinkamas sprendimas, kuris turėtų sumažinti šio karkaso tikslumą, nes karkasas visados susidurs su kadrais, kuriose yra artefaktai ir prasčiau atpažins tikras, nesuklastotas nuotraukas.

4.4. EfficientNet konvoliucinių neuroninių tinklų šeimos

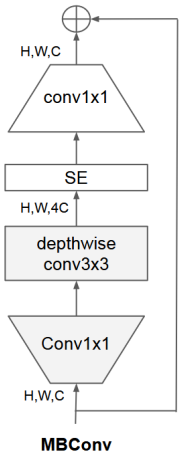
4.4.1. EfficientNet tinklų šeima

EfficientNets yra konvoliucinių neuroninių tinklų šeima pasižyminti aukštu tikslumu turint sąlyginai mažą parametrų skaičių [TL19]. Šie tinklai pasižymi tokia pačia tikslia bei našia bazine architektūra (EfficientNet-B0), kuriai sukurti buvo pasitelkta neuroninė architektūros paieška (angl. *neural architecture search*). Sukurti skirtingo dydžio šios šeimos tinklus buvo panaudotas sudėtinis dimensijų didinimas (angl. *compound scaling*), kuris tolygiai didina tinklo gylį, plotį ir rezoliuciją - taip padarant tinklą labiau tikslų (19 pav.). Kitaip nei ankstesniuose dimensijų didinimo metoduose, kur būdavo didinamos tik viena arba dvi dimensijos, pavyzdžiui, gylis ir plotis, EfficientNets tinklai buvo padidinti naudojant specifinius koeficientus, kurie užtikrina optimalų santykį tarp modelių tikslumo ir našumo.



19 pav. Modelio dimensijų didinimas [TL19]. (a) yra bazinio tinklo pavyzdys; (b) - (d) yra tipinių dimensijų didinimo metodų pavyzdžiai, kur yra didininama tik pločio, gylio ar rezoliucijos dimensijos; (e) yra sudėtinis dimensijų didinimas, su kuriuo dimensijos yra didinamos lygiavertiškai pagal iš anksto nustatytus koeficientus

Bazinė architektūra yra sudaryta pagrinde iš mobilių invertuotų sutraukimo konvoliucijos blokų (MBConv, angl. *mobile inverted bottleneck*), kurių sudėtyje yra pridėta suspaudimo ir sužadinimo optimizacija (SE, angl. *squeeze-and-excitation optimization*) (20 pav.).

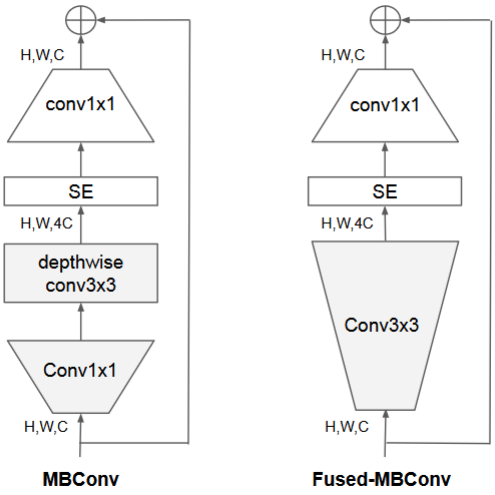


20 pav. MBConv bloko struktūra [TL21]

4.4.2. EfficientNetV2 tinklų šeima

EfficientNetV2 šeimos tinklai yra tikslesni, našesni bei trumpiau besimokantys tinklai, lyginant juos su įprastais EfficientNet šeimos tinklais, kurie ir taip jau turėjo gerą tikslumo bei našumo santykį [TL21].

Vienas iš pagrindinių architektūrinių pakeitimų yra pradėjimas naudoti Fused-MBConv blokus, kurie yra tie patys MBConv blokai, tik juose konvoliuciją į gylį 3x3 yra (angl. *depthwise conv3x3*) ir išplėtimo konvoliucija 1x1 (angl. *expansion conv1x1*) yra pakeista į įprastą konvoliuciją 3x3 (angl. *conv3x3*) (21 pav.). Fused-MBConv blokai yra naudojami tinklo pradžioje, o pabaigoje yra naudojami įprasti MBConv blokai.



21 pav. Skirtumai tarp MBConv ir Fused-MBConv blokų [TL21]

Be to, buvo EfficientNetV2 pradėjo naudoti mažesnius išplesties santykius (angl. *expansion ratios*) MBConv blokuose, mažesnius 3x3 branduolius, bet pridėjo daugiau sluoksnių, kad kompensuoti dėl to sumažintas matymo sritis, bei dar kelis architektūrinius bei mokymo pakeitimus, kurie leido EfficientNetV2 tinklams pranokti įprastus EfficientNet tinklus.

4.4.3. EfficientNet tinklų palyginimas

Bazinis SBI karkasas komponuoja EfficientNetB4 konvoliucinį neuroninį tinklą, o FSBI - turi dvi variacijas - EfficientNetB4 bei EfficientNetB5.

Naujesnės architektūros EfficientNetV2_S konvoliucinis neuroninis tinklas turėtų būti nežymiai tikslesnis ir šiek tiek greitesnis, negu EfficientNetB4 bei EfficientNetB5 tinklai (žr. 5 lentelę). Šiame darbe abu karkasai yra mokomi su EfficientNetV2_S konvoliuciniu neuroniniu tinklu, siekiant patikrinti šią hipotezę.

5 lentelė. EfficientNetB4, EfficientNetB5 bei EfficientNetV2_S tinklų palyginimas [TL21]

Tinklas	ImageNet Top-1 tiksl.	Parametrų skaičius	FLOPs	Greitis(ms)
EfficientNetB4	82.9%	19M	4.2B	30
EfficientNetB5	83.7%	30M	10B	60
EfficientNetV2_S	83.9%	22M	8.8B	24

4.5. Mokymo aplinka

Karkasų mokymas bei verifikavimas buvo atliekami VU Informacinių technologijų atviros prieigos centre (ITAPC), naudojant NVIDIA Tesla V100 32GB vaizdo plokštes. Mokymas bei verifikavimas buvo atliekami pasitelkiant Jupyter Notebook aplinką ir naudojant Torch karkasą. Mokymui ir verifikavimui buvo naudojamas SBI[7] ir FSBI[8] autorių viešai skelbiamas kodas, kuris

[7]https://github.com/mapooon/SelfBlendedImages
[8]https://github.com/gufranSabri/FSBI

vėliau buvo pritaikytas ir optimizuotas veikimui ITAPC'e.

4.6. Mokymo bei verifikavimo aprašas

Visi konvoliuciniai neuroniniai tinklai iš pradžių buvo apmokyti su ImageNet [RDS+15] duomenų rinkiniu. Įvesties kadro dydis yra 380x380 pikselių. Naudotas paketo dydis (angl. *batch size*) yra 16 mokymo metu, o verifikavimo metu 32 - taip daryta prisitaikant prie SBI bei FSBI karkasų aprašų.

Kadangi baziniai SBI bei FSBI karkasai buvo mokomi iki 100 epochų, todėl šiame darbe mokymas taip pat buvo vykdomas iki 100 epochų. Verifikavimui buvo imami epochos svoriai pasiekę aukščiausią AUC reikšmę mokymosi metu su FaceForensics++(HQ) rinkiniu.

4.7. Nuosavas giliųjų klastočių rinkinys

Siekiant ištirti kaip karkasai geba atpažinti populiariausias internete paplitusias giliąsias klastotes, buvo nuspręsta surinkti nuosavą giliųjų klastočių rinkinį ir jį naudoti tolimesnių bandymu metu mokytų karkasų verifikavimui.

Iš tokių socialinių tinklų kaip YouTube, Reddit bei Instagram buvo surinktos 50 giliųjų klastočių bei 50 tikrų vaizdo įrašų - iš viso 100 vaizdo įrašų. Surinktos giliosios klastotės yra sukurtos naudojantis įvairiais metodais ir yra įvairaus kokybės lygio. Klastotės dažniausiai yra juokų formos, kuriose įvairiems personažams yra uždedami kitų aktorių veidai, tačiau yra ir kelios klastotės, kuriose yra vaizduojami politikai siekiant pakenkti jų reputacijai. Surinkti tikri vaizdo įrašai pagrinde yra iškarpos iš įvairių interviu, spaudos konferencijų ar vaizdo įrašų tinklaraščių.

5. Eksperimentas

5.1. Karkasų architektūros keitimo tyrimas

5.1.1. Bazinio SBI karkaso mokymas

Siekiant užtikrinti kuo korektiškesnį tolimesnių tyrimų rezultatų palyginimą su baziniu SBI karkasu ir turėti svarų atskaitos tašką, buvo nuspręsta bazinį SBI karkasą apmokyti nuo pradžių.

Bazinio karkaso architektūra yra SBI+EfficientNetB4. EfficientNetB4 konvoliucinis neuroninis tinklas buvo jau iš pradžių apmokytas su ImageNet [RDS+15] nuotraukų rinkiniu.

SBI+EfficientNetB4 karkasas, apmokytas su FaceForensics++(HQ) rinkiniu iki 100 epochų, geriausią rezultatą pasiekė ties 44 epocha - mokymo AUC 99.74%. Verifikavus su CDF rinkiniu, buvo gauta AUC reikšmė 91.15%, o su nuosavu rinkiniu - 68.64%. Vieno paketo verifikavimas vidutiniškai truko 0.81 ms.

Lyginant su SBI straipsnyje autorių gautais tokio pačio karkaso rezultatais, šis apmokytas karkasas yra 1.72% mažiau tikslus CDF rinkinyje.

5.1.2. SBI+EfficientNetV2_S karkaso mokymas

Tikintis pasiekti tikslesnį ir greitesnį klastočių atpažinimą su SBI karkasu, buvo nuspręstą apmokyti šį karkasą su EfficientNetV2_S konvoliuciniu neuroniniu tinklu.

SBI+EfficientNetV2_S karkasas, apmokytas su FaceForensics++(HQ) rinkiniu iki 100 epochų, geriausią rezultatą pasiekė ties 78 epocha - mokymo AUC 99.98%. Verifikavus su CDF rinkiniu, buvo gauta AUC reikšmė 91.02%, o su nuosavu - 69.80%. Vieno vaizdo verifikavimas vidutiniškai truko 0.69 ms.

Lyginant su baziniu SBI karkasu, šis karkasas yra 0.13% mažiau tikslus CDF rinkinyje ir 1.16% labiau tikslus nuosavame rinkinyje. Greitaveika paspartėjo ~17.4%.

5.1.3. Bazinio FSBI+EfficientNetB5 karkaso mokymas

Siekiant užtikrinti kuo korektiškesnį tolimesnių tyrimų rezultatų palyginimą su šia bazinio FSBI karkaso variacija ir turėti svarų atskaitos tašką, buvo nuspręsta bazinį FSBI+EfficientNetB5 karkasą apmokyti nuo pradžių.

EfficientNetB5 konvoliucinis neuroninis tinklas buvo jau iš pradžių apmokytas su ImageNet [RDS+15] nuotraukų rinkiniu.

FSBI+EfficientNetB5 karkasas, apmokytas su FaceForensics++(HQ) rinkiniu iki 100 epochų, geriausią rezultatą pasiekė ties 34 epocha - mokymo AUC 99.74%. Verifikavus su CDF rinkiniu, buvo gauta AUC reikšmė 89.53%, o su nuosavu - 68.52%. Vieno vaizdo verifikavimas vidutiniškai truko 1.02 ms.

Lyginant su FSBI straipsnyje autorių gautais tokio pačio karkaso rezultatais, šis apmokytas karkasas yra net 5.96% mažiau tikslus CDF rinkinyje.

5.1.4. Bazinio FSBI+EfficientNetB4 karkaso mokymas

Siekiant užtikrinti kuo korektiškesnį tolimesnių tyrimų rezultatų palyginimą su šia bazinio FSBI karkaso variacija ir turėti svarų atskaitos tašką, buvo nuspręsta bazinį FSBI+EfficientNetB4 karkasą apmokyti nuo pradžių.

EfficientNetB4 konvoliucinis neuroninis tinklas buvo jau iš pradžių apmokytas su ImageNet [RDS+15] nuotraukų rinkiniu.

FSBI+EfficientNetB4 karkasas, apmokytas su FaceForensics++(HQ) rinkiniu iki 100 epochų, geriausią rezultatą pasiekė ties 57 epocha - mokymo AUC 99.72%. Verifikavus su CDF rinkiniu, buvo gauta AUC reikšmė 87.77%, o su nuosavu - 66.24%. Vieno vaizdo verifikavimas vidutiniškai truko 0.82 ms.

Lyginant su FSBI straipsnyje autorių gautais tokio pačio karkaso rezultatais, šis apmokytas karkasas yra net 6.52% mažiau tikslus CDF rinkinyje.

Lyginant su pačio autoriaus apmokytu FSBI+EfficientNetB5 karkasu, šis yra 1,76% mažiau tikslus CDF rinkinyje ir 2.28% mažiau tikslus nuosavame rinkinyje. Greitaveika paspartėjo 24%.

5.1.5. FSBI+EfficientNetV2_S karkaso mokymas

Tikintis pasiekti tikslesnį ir greitesnį klastočių atpažinimą su FSBI karkasu, buvo nuspręstą apmokyti šį karkasą su EfficientNetV2_S konvoliuciniu neuroniniu tinklu.

FSBI+EfficientNetV2_S karkasas, apmokytas su FaceForensics++(HQ) rinkiniu iki 100 epochų, geriausią rezultatą pasiekė ties 73 epocha - mokymo AUC 99.95%. Verifikavus su CDF rinkiniu, buvo gauta AUC reikšmė 89.48%, o su nuosavu - 69.46%. Vieno paketo verifikavimas vidutiniškai truko 0.74 ms.

Lyginant su baziniu FSBI+EfficientNetB5 karkasu, šis karkasas yra 0.05% mažiau tikslus CDF rinkinyje ir 0.94% labiau tikslus nuosavame rinkinyje. Greitaveika paspartėjo ~37.8%.

Lyginant su baziniu FSBI+EfficientNetB4 karkasu, šis karkasas yra 1.71% tikslesnis CDF rinkinyje ir 3.22% labiau tikslus nuosavame rinkinyje. Greitaveika paspartėjo ~10.8%.

5.1.6. Visų tirtų karkasų rezultatų apibendrinimas

Nei vienam apmokytam baziniam karkasui nepavyko pasiekti autorių skelbiamų AUC reikšmių CDF rinkinyje nors mokymas buvo vykdomas naudojant jų skelbiamą kodą ir pagal jų straipsniuose aprašytą mokymo konfigūraciją.

Karkasų greitaveika pasikeitė kaip ir buvo tikimasi pakeitus bazinius konvoliucinius neuroninius tinklus į EfficientNetV2_S. Taip pat, buvo pasiekti geresni tikslumo rodikliai nuosavame rinkinyje visais bandymais. Tačiau, to pat nebūtų galima pasakyti apie tikslumą CDF rinkinyje, nes geresnį tikslumą pavyko pasiekti tik lyginant su baziniu FSBI+EfficientNetB4 karkasu, o kitais 2 atvejais tikslumas smuko, kad ir labai nežymiai.

FaceForensics++ rinkinys, su kuriuo buvo mokomi visi karkasai, yra panašesnis į CDF rinkinį komponuojamais vaizdo įrašais, negu į nuosavą rinkinį ir būtent nuosavame rinkinyje pavyko pasiekti tikslumo pagerėjimą pakeitus konvoliucinį tinklą - galbūt EfficientNetV2_S sugebėjo išmokti

atpažinti tam tikrus klastočių artefaktus, kurių nesugebėjo atpažinti EfficientNetB4 ar EfficientNetB5 tinklai.

6 lentelė. Tirtų karkasų tikslumas CDF rinkinyje, nuosavame rinkinyje bei greitaveika. * pažymėtose eilutėse yra pateikiamos reikšmės iš SBI bei FSBI straipsnių

Karkasas	CDF rinkinio AUC (%)	Nuosavo rinkinio AUC (%)	Vaizdo verifikavimo trukmė (ms)
SBI+EfficientNetB4	91.15	68.64	0.81
SBI+EfficientNetB4*	92.87	-	-
SBI+EfficientNetV2_S	91.02	69.80	0.69
FSBI+EfficientNetB5	89.53	68.52	1.02
FSBI+EfficientNetB5*	95.49	-	-
FSBI+EfficientNetB4	87.77	66.24	0.82
FSBI+EfficientNetB4*	94.29	-	-
FSBI+EfficientNetV2_S	89.48	69.46	0.74

5.2. Papildomų duomenų augmentacijų tyrimas

5.2.1. FSBI50+EfficientNetV2_S karkaso mokymas

Pastebėjus tai, kad FSBI karkaso autoriai dažninių požymių generatorių naudoja visoms su savimi sulietoms nuotraukoms ir netgi visoms tikroms nuotraukoms, buvo nuspręsta atlikti FSBI karkaso mokymą su pakeitimais:

- Dažninį požymių generatorių naudoti su savimi sulietoms nuotraukoms tik su 50% tikimybe;
- Tikroms nuotraukoms iš viso nebenaudoti dažninių požymių generatoriaus;

Kaip konvoliucinį neuroninį tinklą šiam karkasui buvo nuspręsta naudoti EfficientNetV2_S. Šis karkasas bus vadinamas FSBI50+EfficientNetV2_S.

FSBI50+EfficientNetV2_S karkasas, apmokytas su FaceForensics++(HQ) rinkiniu iki 100 epochų, geriausią rezultatą pasiekė ties 93 epocha - mokymo AUC 99.91%. Verifikavus su CDF rinkiniu, buvo gauta AUC reikšmė 90.57%, o su nuosavu - 71.12%.

Šių modifikacijų dėka, lyginant su FSBI+EfficientNetV2_S karkasu, buvo pasiektas 1.09% geresnis tikslumas CDF rinkinyje, o 1.66% geresnis tikslumas nuosavame rinkinyje.

5.2.2. FSBI25+EfficientNetV2_S karkaso mokymas

Pastebėjus FSBI50+EfficientNetV2_S karkaso tikslumo pagerėjimą, buvo nuspręsta dar labiau sumažinti tikimybę naudoti dažninių požymių generatorių iki 25% - šis karkasas bus vadinamas FSBI25+EfficientNetV2_S.

FSBI25+EfficientNetV2_S karkasas, apmokytas su FaceForensics++(HQ) rinkiniu iki 100 epochų, geriausią rezultatą pasiekė ties 100 epocha - mokymo AUC 99.95%. Verifikavus su CDF rinkiniu, buvo gauta AUC reikšmė 91.80%, o su nuosavu - 72.26%.

Šios modifikacijos dėka, lyginant su FSBI+EfficientNetV2_S karkasu, buvo pasiektas 2.32% geresnis tikslumas CDF rinkinyje, o nuosavame 2.8% geresnis tikslumas. Lyginant su FSBI50+EfficientNetV2_S - 1.23% geresnis tikslumas CDF rinkinyje ir 1.14% geresnis tikslumas nuosavame rinkinyje.

5.2.3. FSBI25E+EfficientNetV2_S karkaso mokymas

Ties dažninių požymių generatoriaus panaudojimo tikimybės keitimu buvo nuspręsta neapsistoti - į karkasą FSBI25+EfficientNetV2_S buvo įtrauktos papildomos kadrų augmentacijos iš *Albumenations* python bibliotekos, kurios dar nebuvo naudojamos nei SBI, nei FSBI karkasuose. Šis karkasas bus vadinamas FSBI25E+EfficientNetV2_S.

Įtrauktos kadrų augmentacijos:

- Gauso suliejimas (angl. *Gaussian blur*) - mažina triukšmą ir detales, sukurdamas glotninimo efektą.
- Kontrasto ribotas adaptyvus histogramos išlyginimas (CLAHE, angl. *contrast limited adaptive histogram equalization*) - lokaliais regionais pakeičia nuotraukos kontrastą;
- Šviesumo ir kontrasto keitimo (angl. *random gamma*) - reguliuoja nuotraukos ryškumą, išlaikant tamsių ir šviesių sričių santykį, taip imituojant skirtingus apšvietimo lygius;

FSBI25+EfficientNetV2_S karkasasas, apmokytas su FaceForensics++(HQ) rinkiniu iki 100 epochų, geriausią rezultatą pasiekė ties 84 epocha - mokymo AUC 99.95%. Verifikavus su CDF rinkiniu, buvo gauta AUC reikšmė 92.00%, o su nuosavu - 74.60%.

Šių modifikacijų dėka, lyginant su FSBI+EfficientNetV2_S karkasu, buvo pasiektas 2.52% geresnis tikslumas CDF rinkinyje ir 5.14% geresnis tikslumas nuosavame. Lyginant su FSBI50+EfficientNetV2_S - 1.43% geresnis tikslumas CDF rinkinyje ir 3.48% geresnis tikslumas nuosavame. O lyginant su FSBI25+EfficientNetV2_S - 0.2% geresnis tikslumas CDF rinkinyje ir 2.34% gerenis tikslumas nuosavame.

5.2.4. Visų tirtų karkasų rezultatų apibendrinimas

Kiekvienas mokymas su pakeistomis ar papildomomis duomenų augmentacijomis rodė vis tikslesnius klastočių atpažinimo rezultatus CDF ir nuosavame rinkiniuose (7 lent.). Geriausi rezultatai, 92% AUC CDF rinkinyje bei 74.6% AUC nuosavame rinkinyje, buvo pasiekti FSBI25E+EfficientNetV2_S karkaso, kuriame dažninių požymių generatorius buvo taikomas tik su savimi sulietomis nuotraukomis ir tik su 25% tikimybe, bei su papildomomis augmentacijomis iš *Albumenations* bibliotekos.

7 lentelė. Karkasų su papildomomis ir pakeistomis duomenų augmentacijomis tikslumas CDF bei nuosavoje aibėse

Karkasas	CDF rinkinio AUC (%)	Nuosavo rinkinio AUC (%)
FSBI+EfficientNetV2_S	89.48	69.46
FSBI50+EfficientNetV2_S	90.57	71.12
FSBI25+EfficientNetV2_S	91.80	72.26
FSBI25E+EfficientNetV2_S	92.00	74.60

Rezultatai

1. Buvo atkartoti SBI bei FSBI karkasų autorių atlikti mokymai, tačiau gauti rezultatai buvo nuo 1.72% iki 6.52% prastesni tikslumo atžvilgiu verifikuojant su CDF rinkiniu, lyginant su tais, kurie yra skelbiami autorių straipsniuose.

2. Pakeitus SBI karkaso konvoliucinį neuroninį tinklą į EfficientNetV2_S, tikslumas CDF rinkinyje sumažėjo 0.13%, tačiau nuosavame rinkinyje padidėjo 1.15%, o greitaveika paspartėjo ~17.4%.

3. Pakeitus FSBI karkaso tinklą į EfficientNetV2_S: vietoj EfficientNetB5 - tikslumas CDF rinkinyje sumažėjo 0.05%, nuosavame rinkinyje pagerėjo 0.94%, o sparta išaugo apie 38%; tuo tarpu vietoj EfficientNetB4 - tikslumas padidėjo 1.71% CDF rinkinyje ir 3.22% nuosavame, o našumas padidėjo ~11%.

4. FSBI karkase apribojus dažninio požymių generatoriaus taikymą tik su savimi sulietoms nuotraukoms iki 25%, jo nenaudojant tikroms nuotraukoms bei pridėjus 3 papildomas duomenų augmentacijas buvo pasiektas 2.52% tikslumo pagerėjimas CDF rinkinyje ir 5.14% nuosavame.

Išvados

1. Publikuotų SBI ir FSBI karkasų tikslumo reikšmių nepavyko gauti atkartojus bandymus, todėl straipsniuose pateikti rezultatai kelia abejonių dėl jų reprodukuojamumo.

2. SBI bei FSBI karkasuose tinklo pakeitimas į EfficientNetV2_S tikslumą nežymiai padidina tik rinkiniuose, kuriuose gausu klastočių kurtų nematytomis metodikomis, tačiau visais atvejais spartina greitaveiką.

3. FSBI karkase įvesti papildomi duomenų augmentacijų pakeitimai ženklai padidino karkaso tikslumą CDF bei nuosavame rinkiniuose. Didžiausi tikslumo pagerėjimai buvo pasiekti nustojus naudoti dažninį požymių generatorių realioms nuotraukoms, jį naudojant 25% tikimybe su savimi sulietomis nuotraukomis bei pradėjus atsitiktinai taikyti 3 papildomas kadrų augmentacijas.

4. Visi apmokyti karkasai prasčiau aptiko giliąsias klastotes nuosavame giliųjų klastočių rinkinyje negu CDF rinkinyje. Tikėtina dėl to, kad nuosavo rinkinio klastotės buvo kokybiškesnės, sukurtos iki šiol karkasams nematytais metodais ir turėjo mažiau defektų.

Literatūra

[BGJ+23] James Betker, Gabriel Goh, Li Jing, Tim Brooks ir k.t. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[BIK+20] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin ir Alexandr A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2):125, 2020-02. ISSN: 2078-2489. DOI: 10.3390/info11020125. URL: http://dx.doi.org/10.3390/info11020125.

[BLL+25] Sanoojan Baliah, Qinliang Lin, Shengcai Liao, Xiaodan Liang ir Muhammad Haris Khan. Realistic and efficient face swapping: A unified approach with diffusion models. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, p. 1062–1071. IEEE, 2025.

[CUY+20] Yunjey Choi, Youngjung Uh, Jaejun Yoo ir Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, p. 8188–8197, 2020.

[CZ17] Joao Carreira ir Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 6299–6308, 2017.

[CZS+22] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu ir Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, p. 18710–18719, 2022.

[DBP+20] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang ir Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset. https://arxiv.org/abs/2006.07397, 2020. arXiv: 2006.07397 [cs.CV].

[EKB+24] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari ir k.t. Scaling rectified flow transformers for high-resolution image synthesis. *Forty-first international conference on machine learning*, 2024.

[HGC+21] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao ir Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, p. 4360–4369, 2021.

[HLK+25] Ahmed Abul Hasanaath, Hamzah Luqman, Raed Katib ir Saeed Anwar. FSBI: Deepfake detection with frequency enhanced self-blended images. *Image and Vision Computing*:105418, 2025.

[HVP+21] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis ir Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, p. 5039–5049, 2021.

[HZR+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren ir Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 770–778, 2016.

[YHZ+25] Fulong Ye, Miao Hua, Pengze Zhang, Xinghui Li, Qichao Sun, Songtao Zhao, Qian He ir Xinglong Wu. DreamID: High-Fidelity and Fast diffusion-based Face Swapping via Triplet ID Group Learning. *arXiv preprint arXiv:2504.14509*, 2025.

[YYC+24] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao ir k.t. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024.

[YXF+21] P. Yu, Z. Xia, J. Fei ir Y. Lu. A Survey on Deepfake Video Detection. *IET Biom*:607–624, 2021.

[JAA+20] Mousa Tayseer Jafar, Mohammad Ababneh, Mohammad Al-Zoube ir Ammar Elhassan. Forensics and analysis of deepfake videos. *2020 11th international conference on information and communication systems (ICICS)*, p. 053–058. IEEE, 2020.

[JLW+20] Liming Jiang, Ren Li, Wayne Wu, Chen Qian ir Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, p. 2889–2898, 2020.

[LBY+20] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen ir Fang Wen. Advancing high fidelity identity swapping for forgery detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, p. 5074–5083, 2020.

[LYS+20] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi ir Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, p. 3207–3216, 2020.

[Mah23] Arwa Mahdawi. Nonconsensual deepfake porn is an emergency that is ruining lives, 2023. URL: `https://www.theguardian.com/commentisfree/2023/apr/01/ai-deepfake-porn-fake-images`.

[MMN25] Dwij Mehta, Aditya Mehta ir Pratik Narang. LDFaceNet: Latent Diffusion-Based Network for High-Fidelity Deepfake Generation. *International Conference on Pattern Recognition*, p. 386–400. Springer, 2025.

[MMP+20] Brais Martinez, Pingchuan Ma, Stavros Petridis ir Maja Pantic. Lipreading using temporal convolutional networks. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6319–6323. IEEE, 2020.

[Rad18] Petrana Radulovic. Harrison Ford is the star of Solo: A Star Wars Story thanks to deepfake technology, 2018. URL: https://www.polygon.com/2018/10/17/17989214/harrison-ford-solo-movie-deepfake-technology.

[RCV+19] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies ir Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE/CVF international conference on computer vision*, p. 1–11, 2019.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause ir k.t. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[RNM+22] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali ir Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513, 2022.

[SY22] Kaede Shiohara ir Toshihiko Yamasaki. Detecting deepfakes with self-blended images. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, p. 18720–18729, 2022.

[Sim22] Tom Simonite. A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be, 2022. URL: https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook.

[SM19] Ajay Shrestha ir Ausif Mahmood. Review of deep learning algorithms and architectures. *IEEE access*, 7:53040–53065, 2019.

[TL19] Mingxing Tan ir Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *International conference on machine learning*, p. 6105–6114. PMLR, 2019.

[TL21] Mingxing Tan ir Quoc Le. Efficientnetv2: Smaller models and faster training. *International conference on machine learning*, p. 10096–10106. PMLR, 2021.

[TVF+20] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales ir Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.

[TZN19] Justus Thies, Michael Zollhöfer ir Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[TZS+16] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt ir Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 2387–2395, 2016.

[Ver20] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE journal of selected topics in signal processing*, 14(5):910–932, 2020.

[Wes19] Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019.

[Win18] Erin Winick. How acting as Carrie Fisher's puppet made a career for Rogue One's Princess Leia, 2018. URL: https://www.technologyreview.com/2018/10/16/139739/how-acting-as-carrie-fishers-puppet-made-a-career-for-rogue-ones-princess-leia.

[ZBC+21] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng ir Fang Wen. Exploring temporal coherence for more general video face forgery detection. *Proceedings of the IEEE/CVF international conference on computer vision*, p. 15044–15054, 2021.

[ZZC+21] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang ir Nenghai Yu. Multi-attentional deepfake detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, p. 2185–2194, 2021.