TURUN YLIOPISTO

Kuvien suojaaminen tekoälykuvageneraattoreilta harhauttavilla menetelmillä

Tieto- ja viestintätekniikan tutkinto-ohjelma
Tietotekniikan laitos, Teknillinen tiedekunta
TkK-tutkielma

Laatija:
Tuula Rantala

Toukokuu 2024

TkK-tutkielma
Tietotekniikan laitos, Teknillinen tiedekunta
Turun yliopisto

Tutkinto-ohjelma: Tieto- ja viestintätekniikka
Tekijä: Tuula Rantala
Otsikko: Kuvien suojaaminen tekoälykuvageneraattoreilta harhauttavilla menetelmillä
Sivumäärä: 40 sivua, 5 liitesivua
Päivämäärä: Toukokuu 2024

Tekoälykuvageneraattorien tuottamien kuvien laatu on kehittynyt viime vuosina nopeammin kuin monet osasivat odottaa. Kuvien laatu on moniin käyttötarkoituksiin jo riittävän hyvä kilpaillakseen suoraan ihmistaiteilijoiden kanssa. Teknologiaan liittyvät tekijänoikeusongelmat ovat nousseet kriittiseksi kysymykseksi, sillä suosittujen diffuusiomallipohjaisten kuvageneraattorien kouluttamiseen on käytetty verkosta kerättyä kuvadataa, usein ilman tekijänoikeudenhaltijoiden suostumusta.

Koska nykyinen lainsäädäntö ei riittävästi suojele taitelijoiden oikeuksia, tarvitaan vaihtoehtoisia tapoja suojata verkkoon esille laitettavia töitä. Huomiota ovat saaneet erityisesti harhauttaviin hyökkäyksiin (engl. adversarial attack) perustuvat suojauskeinot. Harhauttavan suojausmenetelmän perusidea on lisätä kuvaan ihmissilmälle huomaamattomia häiriöitä, jotka estävät tekoälymallia oppimasta kuvan piirteitä oikein.

Tässä kirjallisuuskatsauksessa perehdytään diffuusiomalleihin kohdistuviin harhauttaviin suojausmenetelmiin. Tavoite on selvittää, mihin suojausvaikutus perustuu, millaisia eri keinoja häiriöiden optimointiin on kehitetty ja arvioida, ovatko menetelmät todella riittävä keino kuvien suojaamiseen. Tyypillisiä eroja eri menetelmien välillä ovat esimerkiksi se, kohdistuuko hyökkäys pelkästään kuvan latenttiesityksen muodostamiseen vai koko diffuusioprosessiin, ja millaista kohdekuvaa häiriöiden laskemisen apuna käytetään.

Suojausmenetelmien perimmäinen heikkous on, että kerran julkaistun kuvan suojausta ei voi jälkikäteen päivittää. Menetelmät ovat myös jossain määrin hauraita erilaisia kuvankäsittelytekniikoita vastaan, koska suojattu kuva ei saisi visuaalisesti muuttua paljon. Vaikka suojausmenetelmät ovat puutteellisia, on niillä kuitenkin moraalinen ja sosiaalinen merkitys aikana, jona generatiivisen tekoälyn pelisääntöjä vielä laaditaan.

Asiasanat: diffuusiomallit, kuvan generointi, adversarial attack

Sisällysluettelo

1 Johdanto

Syväoppiviin neuroverkkoihin perustuvien kuvageneraattorien huima kehitys viime vuosina on yllättänyt monet. Parhaita generoituja kuvia on ensisilmäyksellä jo vaikea erottaa oikeiden taiteilijoiden teoksista tai valokuvista. Generoidut kuvat ovat herättäneet ihastusta ja ihmetystä, mutta myös huomattavasti kritiikkiä. Huolta ovat aiheuttaneet esimerkiksi generoiduissa kuvissa toistuvat haitalliset stereotyypit ja muut vääristymät sekä generaattorien käyttö loukkaavan, harhaanjohtavan tai muuten haitallisen materiaalin tuottamiseen (Bird, Ungless & Kasirzadeh. 2023). Vastarintaa ovat synnyttäneet myös mahdollinen opetusdatan plagiointi sekä generaattorien vaikutus taiteilijoiden työmarkkinoihin ja kulttuuriin tuotantoon (Jiang et al. 2023). Vaikka osa taiteilijoista suhtautuukin tekoälyyn kiinnostavana uutena työkaluna, ovat toiset olleet pöyristyneitä siitä, että heidän teoksiaan on päätynyt mallien opetusdatakokoelmiin ilman heidän tietämystään, saati suostumustaan. Erityisen loukkaaviksi on koettu mallit, jotka on jatkokoulutettu tuottamaan nimenomaan tietyn taiteilijan tyyliä mukailevia kuvia. (Ali & Breazeal 2023.)

Tekijänoikeussuojatun materiaalin käyttämiseen tekoälymallien opetuksessa liittyy lainsäädännöllisesti vielä paljon avoimia kysymyksiä. Oikeuskanteita kuvageneraattorien kehittäjiä vastaan ovat nostaneet ainakin arkistokuvapalvelu Getty Images sekä sarjakuvapiirtäjä Sarah Andersenista ja yhdeksästä muusta taiteilijasta koostuva ryhmä (Brittain 2023, Saveri & Butterick 2023), mutta tapaukset ovat tutkielman kirjoitushetkellä vielä ratkeamatta. Mikäli selkeitä vaatimuksia opetusdatan lisensoinnille ei määrätä, on todennäköistä, että tekijänoikeussuojatun materiaalin käyttö generaattorien koulutuksessa tulee jatkumaan. Vaikka osa tekoälymallien kehittäjistä on jo luvannut olla jatkossa käyttämättä materiaalia vastoin oikeudenomistajien tahtoa, on pois jättäytymistä vaadittava erikseen, mistä mallien kehittäjät voivat tehdä juuri niin hankalan prosessin kuin tahtovat (Weatherbed 2024). Lupausten noudattamista on myös vaikea todistaa (Liang et al. 2023).

Välttääkseen teostensa käytön tekoälymallien opettamisessa, osa taiteilijoista on valmis jopa vähentämään töidensä jakamista verkossa. Riittävän näkyvyyden saaminen on itsenäisille taiteilijoille valmiiksi hankalaa, ja verkkoläsnäolon tarkoituksellinen rajoittaminen heikentää uramahdollisuuksia entisestään. (Shan et al. 2023a, Jiang et al. 2023.) Kysyntää siis olisi keinoille, joilla taiteilijat voisivat suojella julkaisemiaan töitä lainsäädännön uudistuksia odotellessa.

Huomattavan määrän mediahuomiota ovat saaneet Chicagon yliopistossa kehitetyt Glaze- ja Nightshade-ohjelmat (Sand Lab 2023a, 2024a). Työkalujen tavoitteena on muokata kuvia ihmissilmälle huomaamattomasti, mutta niin, että tekoälymalli oppii niistä jotain muuta kuin mitä kuvassa ihmiselle näyttäytyy (Shan et al. 2023a). Kyse on siis harhauttavista hyökkäyksistä (engl. adversarial attacks). Harhauttavia esimerkkejä kuvantunnistuksessa on aiemmin tutkittu erityisesti yksityisyydensuojan kannalta, mutta kuvageneraattorien kehittyessä on tekijänoikeusnäkökulma noussut yhä merkittävämmäksi (Shan et al. 2023a).

Tässä tutkielmassa perehdytään harhauttaviin hyökkäyksiin ja esitellään, miten erityisesti taideteosten suojaamiseen moderneilta diffuusiomalleilta tähtäävät tekniikat toimivat. Tarkastelu avaa näkökulmia neuroverkkojen yleiseen toimintaan ja siihen, miten niiden oppimisprosessi eroaa ihmisestä. Lisäksi pyritään arvioimaan, mikä näiden tekniikoiden käytännön merkitys on taiteilijoiden kamppailussa tekoälyä vastaan. Onko kuvan suojeleminen oikeasti mahdollista, vai onko kyseessä ennemminkin vain protesti?

Tutkielman varsinaiset tutkimuskysymykset ovat:

1. Mikä on harhauttava hyökkäys?
2. Mitä ovat kuvateosten suojaamiseen tarkoitettujen harhauttavien hyökkäysten tyypilliset ominaisuudet?
3. Mitkä tekijät rajoittavat näiden hyökkäysten käytännön hyödyllisyyttä?

Tekijänoikeuksien suojelemiseen on esitetty myös vaihtoehtoisia tutkimussuuntauksia. Yksi näistä ovat näkymättömät vesileimat, jotka harhauttavien hyökkäysten tapaan perustuvat kuvaan tehtäviin lähes näkymättömiin muutoksiin. Toisin kuin harhauttavien hyökkäysten, vesileimojen tarkoitus ei ole kuitenkaan häiritä kuvien generointia, vaan opettaa tekoälymallille mahdollisimman tehokkaasti vesileimapiirteet, joiden avulla datan luvaton käyttö mallin koulutuksessa voidaan tunnistaa generoiduista kuvista. (Ma et al. 2023, Cui et al. 2023.) Tämänhetkinen lainsäädäntö ei kuitenkaan anna takeita siitä, että tunnistetun vesileiman perusteella olisi mahdollista hakea korvauksia tai ryhtyä muihinkaan toimenpiteisiin. Myös vesileimoihin liittyvät tekniset haasteet ovat osittain erilaisia suojausmenetelmiin verrattuna. Tässä tutkielmassa keskitytään siksi vain menetelmiin, joiden pyrkimys on estää kuvien hyödyntäminen tekoälykäytössä kokonaan.

1.1 Menetelmät

Tutkielma on toteutettu kirjallisuuskatsauksena. Aineistohaku aloitettiin hakemalla laajemmin kuvageneraattorien tekijänoikeusongelmiin liittyviä artikkeleja hakulausekkeella: *("image synthesis" OR "diffusion model" OR "diffusion models" OR "text-to-image" OR "ai art" OR "generative art" OR "image generation") AND (infring* OR replic* OR forgery OR copy* OR mimic* OR protect*)*

Haku kohdistettiin ensisijaisesti arXiv-, IEEE- ja ACM-tietokantoihin. Näistä tärkeimmäksi osoittautui artikkelien ennakkojulkaisuihin keskittyvä arXiv, jonka käyttö on perusteltua aiheen ajankohtaisuuden vuoksi.

Hakutuloksista nousivat merkittäväksi omaksi kategoriakseen kuvien suojaamiseen tarkoitetut harhauttavat hyökkäykset. Tutkielma rajattiin näihin, koska ne ovat ainoita konkreettisia keinoja, joita taiteilijoille on tarjottu kuviensa väärinkäytön ehkäisemiseen. Menetelmät ovat saaneet huomattavasti mediahuomiota (Sand Lab 2023a, 2024a), mutta toisaalta niiden tehokkuus on myös kyseenalaistettu (Sandoval-Segura, Geiping & Goldstein 2023, Cao et al. 2023, Zhao et al. 2023b, Qin et al. 2023).

Tutkielman rajauksen tarkennuttua täydennettiin aineistoa vielä artikkelien lähdeviitteitä sekä Google Scholarin käänteisiä lähdeviitteitä hyödyntäen. Lopullisen pääaineiston muodostavat 17 harhauttavia suojausmenetelmiä esittelevää tai jatkokehittävää artikkelia, sekä 6 harhauttavien menetelmien arviointiin tai puhdistamiseen keskittyvää artikkelia. Aineisto on lajiteltu taulukossa 1, ja käsitellään tarkemmin luvussa 3.

Taulukko 1. Tutkielman pääaineisto.

Harhauttavat suojausmenetelmät		**Harhautusmenetelmien puhdistus/arviointi**
Ahn et al. 2024	Tan et al. 2024	Cao et al. 2023
Chen et al. 2023	Van Le et al. 2023	Qin et al. 2023
Liang et al. 2023	Wang et al. 2023	Sandoval-Segura, Geiping & Goldstein 2023
Liang & Wu 2023	Wu et al. 2023	Shan et al. 2023c
Liu et al. 2023	Xue et al. 2024	Zhao et al. 2023b
Rhodes et al. 2023	Ye et al. 2023	Zhang et al. 2023
Salman et al. 2023	Zhao et al. 2023a	
Shan et al. 2023a	Zheng et al. 2023	
Shan et al. 2023b		

Suojausmenetelmien tehokkuuden arvioimiseen käytännön kokeilla ei tämän tutkielman puitteissa ole resursseja. Osaa menetelmistä on kuitenkin testattu kirjoittajan omiin kuviin sen verran, että suojauksesta aiheutuvien artefaktien häiritsevyyttä on voitu subjektiivisesti arvioida. Joitain näistä kuvista on esillä liitteessä 1.

1.2 Tutkielman rakenne

Luvussa 2 esitellään aineiston käsittelyn kannalta olennaisia konsepteja, kuten diffuusiomallien ja niiden kouluttamisen perusperiaatteet, sekä tiettyjä aineistossa esiintyviä diffuusiomallien hienosäätömenetelmiä. Samalla pyritään tuomaan esille, miksi harhauttavien suojausmenetelmien tutkiminen ja kehittäminen on ylipäätään tarpeellista.

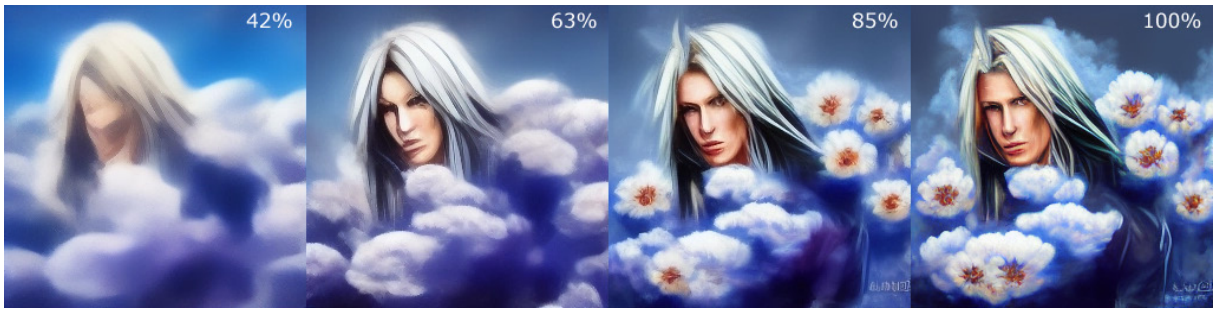
Luvussa 3 käsitellään tutkielman varsinainen aineisto. Ensin esitellään harhauttavien suojausmenetelmien yleiset piirteet, jonka jälkeen arvioidaan menetelmien käyttökelpoisuutta ja sitä rajoittavia tekijöitä.

Luvussa 4 muodostetaan yhteenveto tuloksista ja vastataan tutkimuskysymyksiin. Lopuksi pohditaan harhauttaviin suojausmenetelmiin keskittyvän tutkimuksen merkitystä taiteilijanäkökulmasta.

2 Tausta

2.1 Diffuusiomallit

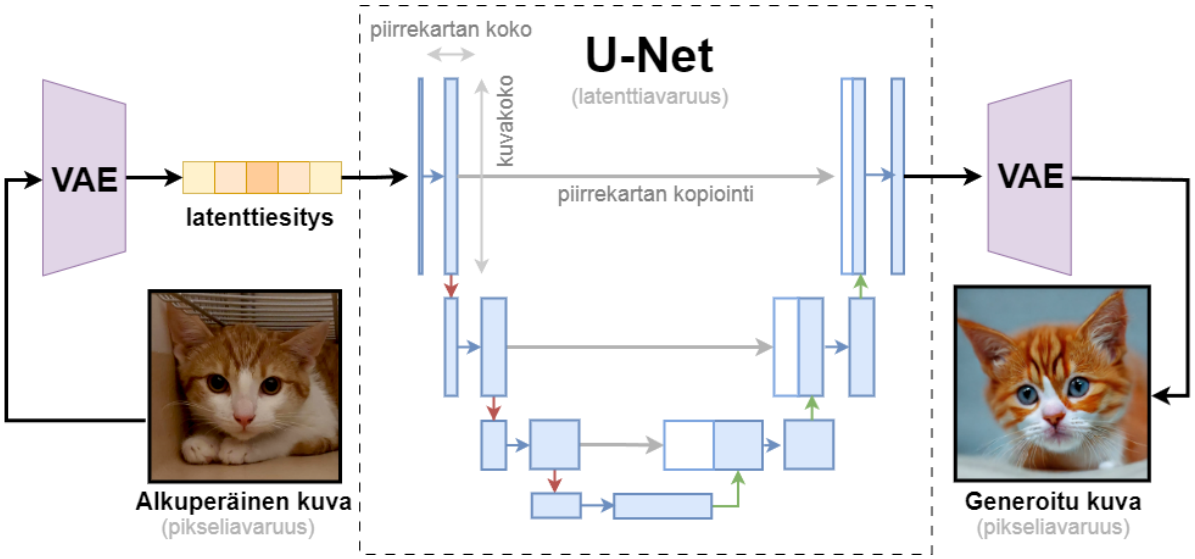
Generatiivinen tekoäly on kehittynyt viime vuosina nopeasti monilla eri osa-alueilla. Kuvien generoinnissa johtavaksi teknologiaksi ovat nousseet diffuusiomallit, jotka päihittävät monilla mittareilla aiemmin parhaita tuloksia tuottaneet GAN-verkot (generative adversarial networks) (Dhariwal & Nichol 2021). Diffuusiomalleihin perustuvat monet suositut kaupalliset sovellukset, kuten Midjourney (Salkowitz 2022, kuva 1), DALL-E 3 (Betker et al. 2023), Adobe Firefly (Adobe 2024a), Imagen 2 (Google DeepMind 2023), sekä myös avoimena lähdekoodina saatavilla oleva Stable Diffusion (Rombach, Esser & Ha 2022).



Kuva 1. Generoinnin eteneminen vanhalla vuoden 2022 Midjourney-versiolla. Käytetty syöte: ”sephiroth final fantasy flowers clouds mucha portrait”

Diffuusion koneoppimisen menetelmänä esittelivät ensimmäisenä Sohl-Dickstein et al. vuonna 2015. Mallin kehitystyöhön saatiin inspiraatio epätasapainoisen termodynamiikan ilmiöstä. Menetelmän perusidea on ensin tuhota datajakauman rakenne lisäämällä siihen asteittain kohinaa (eteenpäindiffuusio), ja sitten oppia kääntämään prosessi (taaksepäindiffuusio). Kyse on Markovin ketjusta, jonka jokaisen askeleen todennäköisyys on analyyttisesti arvioitavissa. Diffuusiomallien soveltuvuuden laadukkaiden kuvien generointiin osoittivat Ho, Jain ja Abbeel vuonna 2020. (Sohl-Dickstein et al. 2015, Ho, Jain & Abbeel 2020.)

Varhaiset diffuusiomallit operoivat suoraan kuvien pikselitasolla, mikä on laskennallisesti hyvin raskasta. Latentit diffuusiomallit, kuten Stable Diffusion, ratkaisevat ongelman kouluttamalla ensin autoenkooderin, joka muuttaa datan matalaulotteisempaan latenttiesitykseen. Latenttiesitys abstrahoi pois pienet, näkymättömät yksityiskohdat, jolloin generoiva malli voi keskittyä datan semanttisesti tärkeisiin elementteihin. Koulutus matalaulotteisemmassa avaruudessa on selvästi tehokkaampaa. (Rombach et al. 2022.)



Kuva 2. Yksinkertaistettu latentti diffuusioarkkitehtuuri

Kuvan latenttiesityksen muodostamiseen käytetään tyypillisesti varioivaa autoenkooderia (engl. variational autoencoder, VAE). Varioiva autoenkooderi kuvaa annetun syötteen yksittäisen datapisteen sijaan jakaumaksi, mikä tekee siitä soveltuvan generoivaan käyttöön (Rocca 2019). Monet harhauttavista hyökkäyksistä kohdistuvat juuri diffuusiomallin VAE-komponenttiin ja pyrkivät sotkemaan kuvan latenttiesityksen (Zheng et al. 2023).

Monimutkaisemmat hyökkäykset kohdistuvat itse diffuusioprosessiin, josta tyypillisesti vastaa U-Net-komponentti (Zheng et al. 2023). U-Net kehitettiin alun perin biolääketieteelliseen kuvien segmentointiin (Ronneberger, Fischer & Brox 2015), mutta se on nykyään yleisesti käytössä myös diffuusiomalleissa, kuten Stable Diffusion ja Imagen (Rombach, Esser & Ha 2022, Saharia et al. 2022). U-Net-arkkitehtuuri koostuu supistavasta ja laajentavasta polusta, jotka ovat keskenään suunnilleen symmetrisiä (kuva 2). Supistava polku on konvoluutioverkko, joka pienentää syötekuvaa askel kerrallaan ja samalla muodostaa siitä piirrekartat. Laajentava polku suurentaa piirrekartan takaisin alkuperäisresoluutioon. Supistusvaiheessa laskettujen piirrekarttojen hyödyntäminen mahdollistaa suurentamisen paremmalla tarkkuudella. (Ronneberger, Fischer & Brox 2015.)

Joissain uusimmissa diffuusiomalleisssa, kuten Stable Diffusion 3 ja OpenAI:n SORA, U-Net-arkkitehtuuri on korvattu muuntajilla, joiden skaalautuvuus on todettu erinomaiseksi monilla tekoälyn osa-alueilla. (Stability AI 2024, Brooks et al. 2024, Peebles & Xie 2023.) Tämän tutkielman aineistossa esitellyt harhauttavat menetelmät on kuitenkin kehitetty ja testattu ensisijaisesti U-Net-arkkitehtuuria käyttävillä malleilla.

2.2 Tekstienkooderit

Jotta kuvan generointia voisi ohjata monimutkaisilla tekstisyötteillä, tarvitaan tehokkaita semanttisia tekstienkoodereita (Saharia et al. 2022).

Eräs suosittu tekniikka on CLIP (Contrastive Language-Image Pre-training), jota hyödyntävät esimerkiksi DALL-E 2 (Ramesh et al. 2022) ja Stable Diffusion (Rombach, Esser & Ha 2022). Perinteiset konenäkömenetelmät kykenevät tunnistamaan vain ennalta määrättyjä kategorioita, mutta CLIP-malli oppii luokittelemaan kuvia luonnollisen kielen opastamana, hyödyntämällä satoja miljoonia internetistä kerättyjä kuva–teksti-pareja. Mallin koulutustehtävänä on ennustaa, mitkä kuvat ja kuvatekstit kuuluvat yhteen. Yhteenkuuluvien parien vektoriupotusten kosinietäisyys minimoidaan, kun taas väärien parien etäisyys maksimoidaan. Kouluttamalla kuvaenkooderia ja tekstienkooderia yhdessä, CLIP oppii näin multimodaalisen upotusavaruuden. (Radford et al. 2021.) CLIP siis esittää kuvan ja tekstin samassa latenttiavaruudessa, ja pystyy kuvaamaan kattavasti sekä kuvien semanttista sisältöä että tyyliä (Ramesh et al. 2022).

Saharia et al. (2022) esittävät, että myös pelkällä tekstillä esikoulutetut suuret kielimallit ovat huomattavan hyviä tekstin enkoodauksessa kuvangenerointia varten. Suurissa kielimalleissa on tyypillisesti enemmän parametreja ja ne on koulutettu suuremmalla datamäärällä kuin kuva–teksti-pareihin perustuvat mallit. Tekstienkooderin skaalaus voi parantaa kuvien laatua ja varsinkin syötteeseen täsmäämistä jopa enemmän kuin diffuusiomallin U-Netin skaalaus. (Saharia et al. 2022.)

Käytetyllä tekstienkooderilla on mielenkiintoinen vaikutus esimerkiksi tyylin imitointiin liittyvissä kysymyksissä. Yksi suosituimmista Stable Diffusion -tekstisyötteisiin liitettävistä nimistä on ollut puolalainen fantasiakuvittaja Greg Rutkowski (Heikkilä 2022). Koulutukseen käytetyn LAION-datakokoelman osittainen tarkastelu kuitenkin viittaa siihen, että opetusdatan joukossa olisi verrattain vähän Rutkowskin töitä (Baio 2022). Stability AI:n perustaja Emad Mostaque (2022) onkin esittänyt nimen tehon tietyissä Stable Diffusion -versioissa johtuvan käytetystä CLIP-mallista. Kyseisen CLIP-mallin koulutusdata ei ole julkista (Mostaque 2022), joten epäselvää on, ovatko Rutkowskin työt siinä huomattavan vahvasti edustettuina, vai onko nimi muusta syystä assosioitunut erityisen haluttuihin tyylipiirteisiin.

2.3 Mallin koulutus

Neuroverkkojen avulla voidaan approksimoida monimutkaisia matemaattisia funktioita yhdistelemällä laskennallisia neuroneja monessa kerroksessa. Yhden kerroksen ulostulo toimii syötteenä seuraavan kerroksen neuroneille, kerrottuna jollain painokertoimella. Verkon koulutuksessa pyritään saamaan viimeisen kerroksen ulostulo vastaamaan odotusarvoa säätämällä painokertoimia koulutuskierrosten välissä. Tavoitteena on siis minimoida tappiofunktio. Koulutuksen lopputuotteena ovat halutunlaista dataa tuottavat painokertoimet.

Käytännössä esimerkiksi Stable Diffusion v2:n koulutus alkaa kuvien ja tekstisyötteiden enkoodaamisella latenttiavaruuteen. Tekstienkooderin ulostulo syötetään diffuusiprosessista vastaavalle U-Netille ristihuomiomekanismin avulla. Tappiofunktio lasketaan latenttiesitykseen lisätyn kohinan ja U-Netin tekemän ennusteen väliltä. (Rombach, Esser & Ha 2022.)

Tekijänoikeuskysymyksien kannalta olennaista on, että koulutettu malli ei siis sisällä suoria kopioita opetusdatastaan ainakaan pikselimuodossa. Esimerkiksi esikoulutetun Stable Diffusion v2 -mallin painokertoimien koko on vain noin 5 gigatavua (Rombach, Esser & Ha 2022), vaikka koulutukseen käytetyn LAION-5B -kokoelman koko ladattuna on 80-240 teratavua, riippuen kuvien latauskoosta (Beaumont 2022). Kuvageneraattorien käyttäjät toisinaan esittävätkin, että mallit vain ”inspiroituvat” koulutusdatastaan, verraten niiden oppimisprosessia ihmisen oppimisprosessiin (Jiang et al. 2023). Diffuusiomallien, kuten Stable Diffusion ja DALL-E 2, on kuitenkin osoitettu kykenevän replikoimaan ainakin osan koulutusdatastaan lähes täysin suoraan (Somepalli et al. 2023, Nichol 2022), mikä ei tue inspiroitumisnarratiivia. Myös harhauttavien esimerkkien olemassaolo osoittaa, että tekoäly ei opi konsepteja samalla tavalla kuin ihminen (Goodfellow, Shlens & Szegedy 2015). Jiang et al. (2023) argumentoivat kuvageneraattorien inhimillistämiseen tähtäävän retoriikan olevan haitallista, sillä se aliarvostaa mallien kehityksen mahdollistaneiden taiteilijoiden työpanosta ja siirtää vastuuta pois mallien kehittäjiltä.

2.3.1 Opetusdatan kerääminen

Kuvageneraattoreihin liittyvien tekijänoikeusongelmien juurisyy on, että tekoälymallien kouluttamiseen käytettävät datakokoelmat eivät useinkaan koostu ainoastaan materiaalista,

jonka käyttöön on erikseen hankittu lupa. Esimerkiksi Stable Diffusion -diffuusiomallien kouluttamiseen on käytetty avoimen LAION-5B-datakokoelman osajoukkoja (Rombach, Esser & Ha 2022), ja myös Googlen ensimmäinen Imagen-versio koulutettiin osaksi LAION-400M:llä (Saharia et al. 2022). Monet taiteilijat ovat löytäneet teoksiaan LAION-datakokoelmista, vaikka eivät ole antaneet käyttöön suostumustaan (Jiang et al. 2023). Monet muista malleista eivät ole julkistaneet koulutusdatojensa sisältöä, mutta myös esimerkiksi Midjourneyn ja OpenAI:n DALL-E 2:n kouluttamiseen tiedetään käytetyn internetistä kaavittua dataa (Salkowitz 2022, Nichol 2022).

LAION-5B on yli viidestä miljardista kuva–teksti-parista koostuva datakokoelma, joka on laadittu verkosta aineistoa keräävän julkisen Common Crawl -verkkoarkiston pohjalta. Kokoelman rakentamiseksi arkiston metadatatiedostoista luetaan verkkosivujen HTML IMG-eli kuvatagit keskittyen kuviin, jotka on varustettu kuvan sisältöä kuvaavalla alt-text-attribuutilla. Tämä attribuutti mahdollistaa kuva–teksti-parien koostamisen. Parien kuva- ja tekstienkoodausten samankaltaisuus mitataan CLIP-mallia hyödyntäen, ja liian huonon tuloksen saavat parit poistetaan lopullisesta datakokoelmasta. (Schuhmann et al. 2022.) Lopullisen kuva–teksti-parin metadata sisältää uniikin tunnisteen, kuvan verkko-osoitteen, kuvaan liittyvän tekstin, kuvan mitat, arvion kuvan ja tekstin samankaltaisuudesta sekä arvion siitä, sisältääkö kuva K18-sisältöä tai vesileiman (Schuhmann et al. 2022). Tekijänoikeuskysymysten kannalta merkittävää on, että LAION ei itse säilö kuvadataa, vaan kokoelman käyttäjän on ladattava sisällöt itse alkuperäisosoitteista (LAION FAQ).

Osaa verkkosisällön kaapijoista voi ohjeistaa olemaan kaapimatta sivuston sisältöä robots.txt -tiedoston avulla, jonka rinnalle on ehdotettu myös spesifimpää ai.txt -tiedostoa. Näillä ei voi kuitenkaan suojella sellaisia kuvien kopioita, jotka on ladattu ulkopuolisten tahojen hallinnassa oleville verkkosivuille. (Miller 2023.) Pahantahtoinen toimija voi myös jättää verkko-osoitteeseen tai kuvan metadataan lisätyt ohjeistukset yksinkertaisesti huomioimatta.

Lisäksi massiivisilla datakokoelmilla esikoulutetut mallit ovat vasta ensimmäinen osa ongelmaa. Jos mallin on tarkoitus esimerkiksi imitoida tietyn taiteilijan tyyliä, on mallia yleensä jatkokoulutettava kyseisen taiteilijan töillä. Nykyisillä hienosäätömenetelmillä tämä on mahdollista niin pienillä datakokoelmilla, että ne on mahdollista kerätä myös käsin. Vaikka mallin alkuperäinen kehittäjä olisi lupautunut kunnioittamaan tekijänoikeudenomistajien tahtoa, mallia jatkokouluttavien yksityishenkilöiden ei voida olettaa niin tekevän. Kun Stable Diffusion v2:sta poistettiin mahdollisuus käyttää tiettyjen taiteilijoiden nimiä avainsanoina,

tähän reagoitiin käyttäjäkunnassa julkaisemalla esimerkiksi Greg Rutkowskin tyyliä jäljittelevä LoRA -malli (Lanz 2023). Rutkowski on mukana Stable Diffusionin kehittäjiä vastaan nostetussa oikeuskanteessa (Saveri & Butterick 2023).

2.3.2 Mallin hienosäätö ja kustomointi

Mallin hienosäätö (engl. fine-tuning) tarkoittaa esikoulutetun tekoälymallin jatkokoulutusta uudella datalla, jotta malli tuottaisi haluttuun tarkoitukseen sopivampaa materiaalia. Täysi hienosäätö, joka kouluttaa kaikki mallin parametrit uudelleen, on nykyisillä suurilla malleilla huomattavan kallista ja epäkäytännöllistä (Hu et al. 2021). Siksi on kehitetty tekniikoita, jotka mahdollistavat mallin kustomoinnin tehokkaammin. Aineistossa esille nousseita diffuusiomallien kustomointitekniikoita ovat erityisesti LoRA, DreamBooth ja Textual Inversion.

Alun perin kielimalleja varten kehitetty LoRA (Low-Rank Adaptation) -menetelmä tekee hienosäädöstä laskennallisesti kevyempää vähentämällä koulutettavien parametrien määrää jopa monituhatkertaisesti. Tämä saavutetaan jäädyttämällä esikoulutetun mallin painokertoimet ja kouluttamalla niiden sijaan neuroverkkoarkkitehtuurin tiiviisiin kerroksiin upotettuja matala-asteisia matriiseja. Koulutetut matriisit yhdistetään lopulta jäädytettyihin painokertoimiin. (Hu et al. 2021). LoRA on myöhemmin sovitettu myös diffuusiomalleille. Se mahdollistaa mallien hienosäätämisen huomattavasti tavallista nopeammin ja kevyemmällä laitteistolla. Lopputuloksena syntyvät painokertoimet vievät vain muutaman megatavun verran tallennustilaa ja ovat näin helposti jaettavissa muille käyttäjille. (Cuenca & Paul 2023.)

DreamBooth on hienosäätötekniikka, jolla tekstistä kuvaksi -malli opetetaan liittämään jokin valittu kohde (esim. käyttäjän kissa) uniikkiin tunnistetermiin. Termin käyttäminen tekstisyötteissä mahdollistaa kohteen toistamisen eri konteksteissa ja asennoissa kohteen yksilöivien piirteiden säilyessä tunnistettavina. Oppimista varten tarvitaan vain muutama kuva kohteesta, sillä tekniikka hyödyntää myös mallin aiempaa tietoa kohteen edustamasta yleisemmästä luokasta (esim. kissa). Tekniikkaan sisältyy lisäksi varotoimia, jotta malli ei koulutuksen aikana unohda aiemmin oppimaansa ja ala assosioida koko luokkatermiä vain opetettuun kohteeseen (Ruiz et al. 2023.) Tavallisesti DreamBooth-koulutus päivittää koko diffuusiomallin, mutta sitä voi käyttää myös yhdessä LoRA-tekniikan kanssa (Hugging Face).

Myös Textual Inversion pyrkii mahdollistamaan tietyn konseptin toiston eri konteksteissa. Kyseessä ei kuitenkaan ole varsinainen hienosäätötekniikka, sillä se ei kajoa itse generatiiviseen malliin. Textual Inversion -tekniikassa esikoulutettu malli jäädytetään, ja sen tekstinupotusulottuvuudesta etsitään uusi termi kuvaamaan haluttua konseptia. Konsepti siis muunnetaan mallin latenttiavaruuteen ja lisätään mallin tuntemaan sanastoon. Tätä varten tarvitaan 3–5 kuvaa, jotka kuvaavat esimerkiksi haluttua objektia tai tyyliä. (Gal et al. 2022.) Mallin jäädyttäminen rajoittaa tekniikan ilmaisukykyä, ja esimerkiksi objektin identiteetin säilyttämisessä Textual Inversion häviää selvästi DreamBoothille (Ruiz et al. 2023).

2.4 Harhauttavat esimerkit

Harhauttavat esimerkit (engl. adversial examples) ovat tunnettu ilmiö neuroverkkopohjaisissa luokittelujärjestelmissä. On huomattu, että koneoppimismallit saattavat luokitella väärin esimerkkejä, jotka eroavat jopa huomaamattoman vähän oikein luokitelluista esimerkeistä. Sama luokitteluvirhe voi toistua monilla eri malleilla, vaikka niiden arkkitehtuuri ja käytetty koulutusdata olisi eri. Ilmiöstä voidaan päätellä, että hyvinkään suoriutuvat luokittelumallit eivät välttämättä tunnista luokkia tosiasiassa määrittäviä perusperiaatteita. (Goodfellow, Shlens & Szegedy 2015.)

Pienet satunnaiset häiriöt kuvassa eivät tyypillisesti vääristä luokittelijan toimintaa. Harhauttavan esimerkin voi kuitenkin löytää optimoimalla syötteen ennustevirheen maksimoimiseksi. Näin löydetyn harhauttavan esimerkin ero alkuperäiseen on lähes huomaamaton. (Szegedy et al. 2013.)

Goodfellow, Shlens ja Szegedy (2015) esittävät harhauttavien esimerkkien olemassaolon johtuvan mallien lineaarisuudesta: pienet muutokset riittävän monessa syötteen ulottuvuudessa johtavat suureen muutokseen ulostulossa. Tätä suosittua hypoteesia ei kuitenkaan ole tähän mennessä täysin todistettu eikä kumottu. Selityksiä ilmiön olemassaololle on haettu myös muista mallin, koulutusprosessin ja datan ominaisuuksista. Täyttä ymmärrystä ilmiön juurisyistä ei ole vielä saavutettu. (Han et al. 2023.)

Harhauttavien hyökkäysten laatimista ja niitä vastaan puolustautumista on tästä huolimatta tutkittu laajasti, varsinkin luokittelumalleissa. Hyökkäykset kategorisoidaan tyypillisesti joko white-box- tai black-box-hyökkäyksiksi, riippuen siitä, ovatko hyökkäyksen kohteena olevan mallin arkkitehtuuri ja parametrit tunnettuja vai eivät. Black-box-hyökkäykset ovat

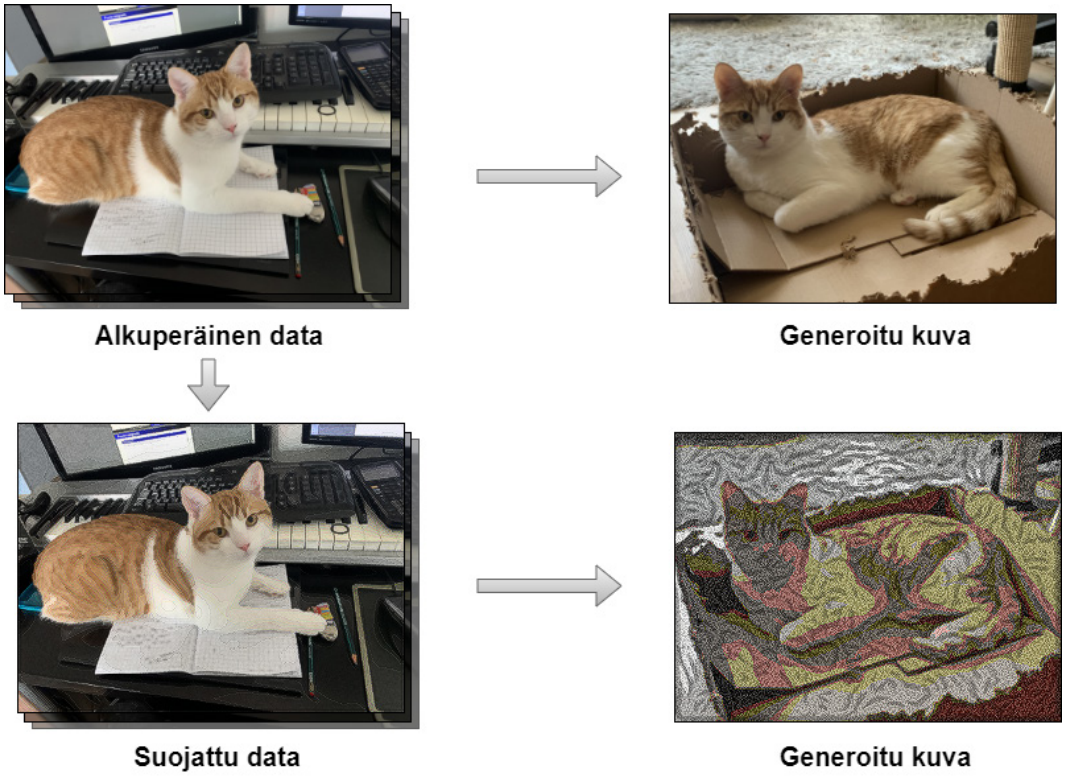
tosimaailmassa käytännöllisempiä. Tällainen hyökkäys voidaan toteuttaa esimerkiksi kohdemallin ulostuloja hyödyntäen, tai se voidaan laatia tarkoitukseen koulutetulla sijaismallilla, sillä harhauttavat esimerkit ovat ainakin osin siirrettäviä eri mallien välillä. (Han et al. 2023.)

Yleisiä puolustuskeinoja harhauttavia hyökkäyksiä vastaan ovat syötteiden käsittely harhautuksen kumoamiseksi, tai harhauttavien esimerkkien sisällyttäminen koulutusprosessiin (adversarial training). Jälkimmäinen lähestymistapa vaatii huomattavasti resursseja ja saattaa johtaa huonompaan suorituskykyyn puhdasta dataa käsiteltäessä. Myös arkkitehtuuriratkaisuilla on vaikutusta siihen, miten hyvin malli sietää harhauttavia esimerkkejä. Vaihtoehtoisesti harhauttavat esimerkit voi pyrkiä tunnistamaan ja poistamaan opetusdatan joukosta. (Han et al. 2023.)

Generoivien mallien harhauttaminen on vaativampaa kuin luokittelumallien. Luokittelijan piirreavaruuden tarvitsee sisältää pääasiassa vain kohteen identiteettiin liittyvää tietoa, kun taas generoiva malli vaatii huomattavasti enemmän tietoa kuvan sisältämien objektien sijoittelusta ja muista ominaisuuksista. Saman vaikutuksen aikaansaaminen generoivassa mallissa vaatii siis näkyvämpien häiriöiden lisäämistä kuvaan – tai parempia optimointistrategioita. (Shan et al. 2023a.)

3 Kuvien suojaaminen harhauttavilla menetelmillä

Kaikkien aineiston kuvansuojausmenetelmien perusperiaate on sama: kuvaan lisätään häiriöitä, jotka on suunniteltu harhauttamaan tekoälymalli ”näkemään” kuva väärin. Manipulointi kohdistuu siis kuvadataan itseensä, eikä esimerkiksi sen metadataan. Riippuen menetelmästä, malli joko ohjataan oppimaan kuvasta tarkoituksella vääriä piirteitä, tai sitä yritetään estää oppimasta hyödyllisiä piirteitä. Käsitellyllä datalla koulutetun mallin on tarkoitus tuottaa huonolaatuisia tai muuten odotuksista poikkeavia kuvia, kun sillä yritetään generoida suojattuun dataan perustuvia kuvia (kuva 3).



Kuva 3. Kuvitteellinen esimerkki siitä, mitä harhauttavalla suojauksella voitaisiin tavoitella.

Käsitellyt menetelmät on esitelty taulukossa 2. Joidenkin menetelmien pääasiallinen tarkoitus on nimenomaan tekijänoikeuksien ja taidetyylien suojelu, kun taas toiset keskittyvät enemmän yksityisyydensuojaan ja sosiaalisesti haitallisen käytön ehkäisemiseen. Kummassakin tapauksessa tavoite on, että suojatun datan käyttäminen mallin koulutuksessa tai hienosäädössä ei opeta mallille datan todellisia piirteitä. Lisäksi varsinkin monet yksityisyyden suojaamiseen keskittyvät menetelmät pyrkivät häiritsemään tekoälypohjaista kuvamanipulaatiota, jossa kuva toimii suoraan syötteenä diffuusiomallille.

Taulukko 2. Aineiston harhauttavat suojausmenetelmät.

	Lähde	Motivaatio	Tavoite
ADAF (Adversarial Decoupling Augmentation Framework)	Wu et al. 2023	yksityisyydensuoja	DreamBooth-menetelmän häirintä, kuva-teksti-fuusiomoduulin huomiointi
AdvDM (Adversarial Example for DMs)	Liang et al. 2023	tekijänoikeudet	Textual Inversion- ja kuvasta kuvaksi -menetelmien häirintä
Anti-DreamBooth	Van Le et al. 2023	yksityisyydensuoja	hienosäätömenetelmien häirintä (DreamBooth, Textual Inversion)
DUAW (Data-free Universal Adversarial Watermark)	Ye et al. 2023	tekijänoikeudet	uudelleenkäytettävyys, hienosäätömenetelmien häirintä (DreamBooth, LoRA)
EditShield	Chen et al. 2023	yksityisyydensuoja	ohjeperusteisen kuvamanipulaation häirintä
Glaze	Shan et al. 2023a	tekijänoikeudet	tyylin oppimisen estäminen
IMPASTO (IMperceptible Protection Against STyle imitation)	Ahn et al. 2024	tekijänoikeudet	muiden menetelmien lisäämien häiriöiden häiritsevyyden vähentäminen
ITA / ITA+ (Improved Targeted Attack)	Zheng et al. 2023	tekijänoikeudet, yksityisyydensuoja	muistivaatimusten vähentäminen, ”few-shot”-tekniikoiden häirintä (LoRA + DreamBooth, kuvasta kuvaksi)
MetaCloak	Liu et al. 2023	yksityisyydensuoja	hienosäätömenetelmien häirintä (DreamBooth, Textual Inversion), suojauksen sietokyvyn parantaminen
MAMC (My Art My Choice)	Rhodes et al. 2023	tekijänoikeudet	häiriöiden ja suojauksen tasapainon kustomoitavuus eri tarkoituksiin
Mist	Liang & Wu 2023	tekijänoikeudet	DreamBooth-, Textual Inversion- ja kuvasta kuvaksi -menetelmien häirintä
Nightshade	Shan et al. 2023b	tekijänoikeudet	mallin myrkyttäminen
PAG (Protecting Artworks from Personalizing Image Generative Models)	Tan et al. 2024	tekijänoikeudet	tyylin suojelu hienosäädöltä (LoRA) black-box-periaatteella
PhotoGuard	Salman et al. 2023	yksityisyydensuoja	kuvasta kuvaksi -manipulaation häirintä
SDS (Score Distillation Sampling)	Xue et al. 2024	tekijänoikeudet	menetelmien muistivaatimusten vähentäminen
SimAC (Simple Anti-Customization)	Wang et al. 2023	yksityisyydensuoja	menetelmien tehostaminen aika-askeleen optimoinnilla
UDP / EUDP (Enhanced Unlearnable Diffusion Perturbation)	Zhao et al. 2023a	tekijänoikeudet, yksityisyydensuoja	hienosäätömenetelmien häirintä (DreamBooth, Textual Inversion)

Aineiston menetelmistä Nightshade on ainoa, joka yksittäisten kuvien suojaamisen sijaan pyrkii tekoälymallin kokonaisvaltaisempaan häiritsemiseen, myrkyttämällä kokonaisia konsepteja. Myrkytetyt kuvat näyttävät visuaalisesti normaaleilta, mutta opettavat mallia toimimaan vastoin oletuksia – esimerkiksi generoimaan kuvan kissasta, kun mallilta pyydetään kuvaa koirasta. Kuvien suojausvaikutus perustuu tällöin pelotteeseen: käyttämällä kuvia ilman lupaa riskeeraa myrkytetyn datan päätymisen opetusdataan. (Shan et al. 2023b.)

Suurin osa menetelmistä on testattu avoimesti saatavissa olevilla Stable Diffusion v1 ja v2 -malleilla. Stable Diffusion on latentti diffuusiomalli, jonka kyseiset mallit perustuvat U-Net-arkkitehtuuriin (Rombach, Esser & Ha 2022).

3.1 Häiriöiden optimointi

Kuvaan lisättävät häiriöt on tarkoitus pitää niin hienovaraisina, etteivät ne ihmissilmään muuta kuvaa merkittävästi. Häiriöiden sallittua määrää kuvataan niin sanotulla häiriöbudjetilla (engl. perturbation budget tai noise budget) – suurempi budjetti johtaa tehokkaampaan suojaukseen, mutta heikentää kuvan laatua selvemmin (Van Le et al. 2023). Mahdollisimman tehokkaiden mutta huomaamattomien häiriöiden laatiminen on siis optimointiongelma. Menetelmien erot ovat siinä, miten kuvaan lisättävät häiriöt lasketaan.

Teknisesti hyökkäykset voivat kohdistua joko mallin VAE-autoenkooderiin tai diffuusioprosessista vastaavaan U-Net-komponenttiin. VAE-hyökkäykset, kuten Glaze, EditShield ja DUAW, pyrkivät vääristämään kuvan latenttiesityksen kasvattamalla sen etäisyyttä alkuperäiseen latenttiesitykseen (Shan et al. 2023a, Chen et al. 2023, Ye et al. 2023). U-Net-hyökkäykset pyrkivät estämään mallin kohinanennustajaa ennustamasta manipuloitujen latenttimuuttujien kohinaa oikein. Koska U-Net-hyökkäys huomioi myös enkoodausvaiheen, on se empiirisesti tehokkaampi. (Zheng et al. 2023.) Xue et al. (2024) esittävät kuitenkin, että nimenomaan autoenkooderi on merkittävin tekijä tällaisessakin hyökkäyksessä, sillä se on kohinanpoistajaa haavoittuvaisempi hyökkäyksille. Tulos on ristiriidassa Zhang et al. (2023) aiemmin esittämän tuloksen kanssa, jonka mukaan juuri kohinanpoistaja ja varsinkin sen ResNet-moduuli (residual neural network) on diffuusiomallin heikoin kohta.

Koko diffuusioprosessiin kohdistuva hyökkäys on laskennallisesti raskaampi kuin pelkkä enkooderihyökkäys. PhotoGuard käyttää optimointiongelman ratkaisemiseksi PGD-menetelmää (projected gradient descent), mutta koska koko diffuusioprosessin läpi

taaksepropagointi veisi liikaa GPU-muistia, voidaan askelia käydä läpi vain muutama. (Salman et al. 2023.) AdvDM-menetelmä hyödyntää häiriöiden laskemisessa Monte Carlo -simulaatiota, sillä tarvittavaa gradienttia ei voi suoraan laskea. Askelten määrä prosessissa on lopputuloksen kannalta olennainen. (Liang et al. 2023.) EUDP- ja SimAC-menetelmät hakevat lisätehoa myös optimoimalla aika-askeleen valinnan (Zhao et al. 2023a, Wang et al. 2023). Nämä menetelmät ovat silti käytännössä liian raskaita varsinkin yksityiskäyttäjille. Xue et al. (2024) esittävät laskennan keventämiseksi SDS-menetelmää, joka approksimoi gradientin U-Netin Jacobin matriisin avulla ja näin jopa puolittaa tarvittavan GPU-muistin AdvDM:ään verrattuna. (Xue et al. 2024). Myös ITA-menetelmä tavoittelee muistivaatimusten vähentämistä usein eri keinoin (Zheng et al. 2023).

Häiriöiden optimointiin vaikuttaa myös se, onko kuvia tarkoitus suojella kiinteää diffuusiomallia vai muuntuvia hienosäätömenetelmiä, kuten DreamBoothia, vastaan. Näissä skenaarioissa pätevät osin eri mekanismit. (Liu et al. 2023.) Ensimmäisessä tapauksessa häiriöiden laatimiseen voidaan käyttää samaa esikoulutettua diffuusiomallia kuin mihin hyökkäys kohdistuu. Hienosäätömenetelmiin kohdistuvassa hyökkäyksessä tehokkaampaa on kouluttaa yksi tai mieluiten useampi sijaismalli, ja päivittää vuorotellen häiriöitä ja sijaismalleja. Tällaista tekniikkaa käyttävät esimerkiksi Anti-Dreambooth ja MetaCloak (Van Le et al. 2023, Liu et al 2023).

Zhao et al. (2023b) huomioivat, että eri hienosäätömenetelmien sietokyky harhauttavia menetelmiä vastaan vaihtelee. Pelkän tekstienkooderin koulutukseen perustuvat menetelmät, kuten Textual Inversion, ovat kaikkein helpoimmin harhautettavia. Pelkkää U-Netiä kouluttavat menetelmät sietävät häiriöitä parhaiten. (Zhao et al. 2023b.)

Kohdistettu ja kohdistamaton hyökkäys

Harhauttavan hyökkäyksen tavoite on siis maksimoida etäisyys mallin ulostulon ja todellisen jakauman välillä. Empiiristä toimivuutta on tyypillistä tehostaa kääntämällä tavoite niin, että pyritäänkin minimoimaan etäisyys ulostulon ja valitun kohdejakauman välillä. Näin malli saadaan luulemaan harhauttavaa esimerkkiä kohteen kaltaiseksi. Kohteeksi valitaan yleensä jokin toinen kuva. (Zheng et al. 2023.)

Esimerkiksi PhotoGuard valitsee kohdekuvaksi täysharmaan kuvan ja ADAF täysmustan kuvan (Salman et al. 2023, Wu et al. 2023). Liang ja Wu (2023) esittävät kuitenkin, että tyhjä kuva ei

ole kohteena erityisen tehokas harhaanjohtamaan mallin esikoulutusprosessia. Sen sijaan kohdekuvina toimivat hyvin korkeakontrastiset ja teräväreunaiset kuvat, joissa on toistuvia, tiheitä kuvioita. Tällaisten kuvioiden vaikutus on selkeästi havaittavissa myrkytetyn datan pohjalta generoiduissa kuvissa. (Liang & Wu 2023, Zheng et al. 2023.) Toistuvat kuviot myös parantavat menetelmän sietokykyä (Liang & Wu 2023).

Kohdekuvan voi valita myös tarkemmin kriteerein, kuten esimerkiksi Glaze ja Nightshade -menetelmissä. Glaze pyrkii suojaamaan imitaattoreilta nimenomaan kuvan tyyliä, ei sen sisältöä. Häiriöbudjetin tuhlaaminen muiden kuin tyylipiirteiden vääristämiseen ei tällöin ole optimaalista. Olennaisten tyylipiirteiden tunnistaminen on kuitenkin vaikeaa, sillä sen lisäksi, että generoivien mallien käyttämät piirteet ovat vaikeasti tulkittavia, myös taidetyylin määrittäminen matemaattisesti on hankalaa. Glazen oivallus on hyödyntää tyylinsiirtoteknologiaa (engl. style transfer). Alkuperäisestä kuvasta generoidaan ensin tyylinsiirtomallin avulla riittävän erilaista tyyliä, esimerkiksi Van Goghin öljymaalausta, mukaileva versio, jonka sisältö on samankaltaista kuin alkuperäiskuvassa. Tyylisiirretyn kuvan avulla voidaan laskea, millaisia häiriöitä kuvaan on lisättävä, jotta tyyliverhon vaikutus kohdistuisi nimenomaan kuvan tyylipiirteisiin. Melko pientenkin muutosten tyylipiirteissä kerrotaan riittävän imitoinnin häiritsemiseen, sillä generoivien mallien ulostuloavaruus on jatkuva, mikä johtaa tyylien sekoittumiseen. Kuvan piirteiden ekstraktointiin ja tyylinsiirtoon voidaan käyttää samaa diffuusiomallia. (Shan et al. 2023a.)

Nightshade sen sijaan pyrkii yksittäisten kuvien suojaamisen sijasta kokonaisten konseptien myrkyttämiseen. Tämän esitetään olevan mahdollista, koska suuressakin koulutusdatakokoelmassa on vain hyvin rajallinen määrä tiettyyn konseptiin liittyviä esimerkkejä. Myrkytetyn datan vaikutus pyritään maksimoimaan tuottamalla ensin kuva–teksti-pareja, joissa kukin tekstisyöte esittää mahdollisimman selkeästi konseptia C, mutta siihen liitetty kuva täysin eri konseptia A. Teho maksimoidaan generoimalla tarkoitusta varten prototyyppisiä kuvia konseptista A. Jotta yhteensopimattomia pareja ei kyetä tunnistamaan automaattisesti eikä myöskään ihmissilmin, korvataan konseptia A esittävät generoidut kuvat luonnollisilla kuvilla konseptista C, joihin on kuitenkin lisätty näiden piirreavaruutta muuntavia häiriötä. Generoidut kuvat konseptista A toimivat ankkureina, joiden pohjalta tarvittavat häiriöt lasketaan. (Shan et al. 2023b.)

Hyökkäys voi olla myös kohdistamaton ja pyrkiä mallin koulutushäviön maksimoimiseen. Tämä johtaa kaoottiseen sisältöön generoiduissa kuvissa (Liang & Wu 2023). Rhodes et al.

(2023) pyrkivät tekemään MAMC-menetelmän kohdistamattomasta hyökkäyksestä eri tarkoituksiin kustomoitavan hyödyntämällä sen optimoimiseen useaa eri tappiofunktiota, jotka ovat säädettävissä hyperparametreilla. Xue et al. (2024) huomioivat, että itseasiassa myös koulutushäviön minimoiminen tuottaa tehokkaita häiriöitä, jotka ovat myös luonnollisemman näköisiä. Tällaiset häiriöt johtavat sumentuneisiin generoituihin kuviin. (Xue et al. 2024.) Luokittelijaa koulutettaessa minimoiva lähestymistapa tarkoittaa, että esimerkissä on vähemmän opittavaa (Zhao et al. 2023a).

Zheng et al. (2023) esittävät, että käytännössä hyvin laadittu kohdistettu hyökkäys on tehokkaampi kuin teoriassa optimaalinen kohdistamaton hyökkäys. Tämän spekuloidaan johtuvan eroavaisuuksista siinä, miten ihminen ja tietokone tulkitsevat kuvien hyvyyttä tai huonoutta. (Zheng et al. 2023.)

Häiriöiden uudelleenkäytettävyys

Tyypillisesti häiriöt on laskettava uudestaan jokaista suojattavaa kuvaa kohti. Suuria datamääriä suojatessa tämä vie huomattavasti aikaa ja resursseja.

Ye et al. (2023) pyrkivät luomaan yleiskäyttöisen harhauttavan vesileiman, joka suojelee kuvia hienosäätötekniikoita, kuten DreamBoothia tai LoRA:a vastaan. DUAW-tekniikka perustuu kahteen havaintoon: ensinnäkin tällä hetkellä suositut hienosäätömallit perustuvat VAE-autoenkooderiin, jonka parametrit on jäädytetty kustomoinnin ajaksi. Toiseksi kustomointitekniikat usein vahventavat pieniä häiriöitä opetusdatassaan. Tavoitteena on siis laatia häiriöitä, jotka sekä säilyvät hienosäätöprosessissa että häiritsevät autoenkooderia, johtaen vääristyneisiin kuviin. (Ye et al. 2023.)

Tekijänoikeussyistä vesileima laaditaan ilman suoraa pääsyä suojeltaviin kuviin, hyödyntäen synteettistä dataa. Synteettinen, monia eri tyylejä ja syötteitä edustava koulutusdata generoidaan Stable Diffusionilla ja vesileimataan. Vesileima optimoidaan minimoimalla rakenteellista yhtäläisyyttä kuvaava MS-SSIM-arvo (multiscale structural similarity index) ulostulokuvien ja alkuperäiskuvien välillä. Data käsitellään erissä, ja vesileima päivitetään koko datakokoelman läpi, jotta se kykenee suojelemaan monia eri kuvia. Lopulta optimoitu vesileima voidaan lisätä suojausta vaativaan kuvaan (Ye et al. 2023.)

Myös Chen et al. (2023) tutkivat mahdollisuutta uudelleenkäytettävien häiriöiden generoimiseen.

3.2 Suojausmenetelmien sietokyky

Jotta harhauttaviin esimerkkeihin perustuva suojausmenetelmä ei olisi täysin käyttökelvoton, on sen siedettävä vähintään yksinkertaisimpia kuvanmuokkauskeinoja, kuten kuvan koon muuttaminen, rajaaminen tai JPEG-pakkaus. Verkossa kiertävä kuva saattaa altistua näille käsittelyille, vaikka mikään taho ei erityisesti pyrkisi purkamaan suojausta. Lisäksi on tyypillistä, että kuvat skaalataan ja rajataan sopivaan kokoon ennen käyttöä mallin koulutuksessa (Podell et al. 2023).

Taulukko 3 esittää yleisimmät kuvien esiprosessointikeinot, jotka suojausmenetelmiä esittelevissä artikkeleissa on huomioitu. Kattavimmin eri käsittelyjä vastaan testaavat Liu et al. (2023), sillä yksinkertaisten kuvankäsittelykeinojen sietäminen on MetaCloak-menetelmän keskeisiä tavoitteita.

Taulukko 3. Kuvien esiprosessointimenetelmiä ja niistä raportointi suojausmenetelmien esittelyissä.

	JPEG	sumennus	kohina	skaalaus	rajaus	super-resoluutio	Adverse Cleaner
ADAF							
AdvDM	x					x	
Anti-DreamBooth	x	x					x
DUAW	x	x	x				
EditShield	x	x					
Glaze	x		x				x
ITA / ITA+	x		x	x		x	
IMPASTO	x	x	x				
MAMC	x	x					
MetaCloak	x	x		x	x	x	
Mist				x	x		
Nightshade							
PAG							
Photoguard							
SDS	x			x	x		x
SimAC							
UDP / EUDP	x	x	x				

Menetelmien tehon JPEG-pakkausta vastaan esitetään olevan pääosin hyvä, sillä vaikka pakkaus jonkin verran heikentää suojausta, se heikentää samalla myös kuvanlaatua merkittävästi (Liu et al. 2023, Shan et al. 2023a, Van Le et al. 2023, Ye et al. 2023, Zhao et al.

2023a, Zheng et al. 2023). Sandoval-Segura, Geiping ja Goldstein (2023) osoittavat kuitenkin, että ainakin PhotoGuard on huomattavan heikko JPEG-pakkausta vastaan, ja huomauttavat, että pakkauksen vaikutusta on tärkeää testata monilla eri pakkausvahvuuksilla.

Intuitiivinen tapa häiriöiden siivoamiseen on kuvan sumennus. Gaussin sumennus heikentää suojausten tehoa (Rhodes et al. 2023, Ye et al. 2023), mutta myös kuvanlaatua (Sandoval-Segura, Geiping & Goldstein 2023, Van Le et al. 2023, Zhao et al. 2023a,). Shan et al. (2023a), Van Le et al. (2023) ja Xue et al. (2023) näyttävät, että myöskään bilateraalisiin pehmennyssuodattimiin perustuva AdverseCleaner-menetelmä ei riitä suojausten purkamiseen. Häiriöiden häiritsevyyttä kaitsemaan pyrkivä IMPASTO-menetelmä kärsii sumennuksesta hieman tavallista enemmän, sillä hienovaraisemmat häiriöt ovat käsittelylle alttiimpia (Ahn et al. 2024).

Satunnaisen kohinan lisääminen ensisijaisesti vain heikentää kuvanlaatua (Shan et al. 2023a, Zhao et al. 2023a, Zheng et al. 2023).

Ainoastaan Liang & Wu (2023), Liu et al. (2023) ja Xue et al. (2024) demonstroivat menetelmiensä kestävyyttä kuvan skaalausta ja rajausta vastaan. Kuvien skaalaus ja rajaaminen sopivaan kokoon ennen mallin koulutusta on tyypillistä (Podell et al. 2023), joten tämä on yllättävä puute tutkimuksissa. Cao et al. (2023) näyttävät, että Glaze selviää kuvan skaalauksesta, mutta PhotoGuard ei. Xue et al. (2024) esittävät kuvan rajaamisen ja skaalaamisen takaisin alkuperäiseen kokoon olevan suhteellisen tehokas puhdistuskeino useita menetelmiä vastaan.

Superresoluutiotekniikoiden alkuperäinen tarkoitus on kuvanlaadun parantaminen kuvaa skaalatessa, mutta niiden on huomattu toimivan myös harhauttavien esimerkkien puhdistajana, kun kyseessä on luokittelumalli (Mustafa et al. 2020). Mielenkiintoisesti Zheng et al. (2023) raportoivat superresoluution vain parantavan ITA-suojauksen tehokkuutta, kun taas Liu et al. (2023) kertovat superresoluution selvästi palauttavan generoidun kuvan laatua MetaCloak-suojausta käytettäessä.

Suojauksenpuhdistusmenetelmät

Suojauksen voi pyrkiä mitätöimään myös monimutkaisemmin keinoin. Diffuusiomalleihin perustuvat menetelmät ovat viime aikoina saavuttaneet erityisen hyviä tuloksia luokittelijoihin kohdistuvien hyökkäysten puhdistamisessa (Nie et al 2022). Luokittelijoita varten kehitetyt

menetelmät eivät kuitenkaan ole optimaalisia generoiviin malleihin kohdistuvia hyökkäyksiä vastaan.

GrIDPure (Zhao et al. 2023b) perustuu luokittelijoita varten kehitettyyn DiffPure-menetelmään, jossa kuvaan ensin lisätään kohinaa eteenpäindiffuusiolla ja sen jälkeen poistetaan taaksepäindiffuusiolla (Nie et al. 2022). DiffPure ei toimi hyvin generoivien mallien kanssa, sillä vaikka luokan määrityksen kannalta kuvassa tapahtuvat pienet muutokset eivät ole kriittisiä, generatiivisessa käytössä nämä pienetkin yksityiskohdat voivat olla tärkeitä. Lisäksi DiffPure toimii pienemmällä resoluutiolla kuin esimerkiksi Stable Diffusion vaatii. GrIDPure kehittää menetelmää jakamalla korkearesoluutioisen kuvan ensin useaksi limittyväksi ruudukoksi, joista kukin puhdistetaan erikseen diffuusiomallilla käyttäen pieniä askelia. Lopuksi ruudukot yhdistetään takaisin alkuperäistä resoluutiota vastaavaksi kuvaksi ja sulautetaan alkuperäiseen kuvaan määrätyssä suhteessa. Prosessia iteroidaan moneen kertaan. Menetelmän tehokkuutta testataan Glazea, AdvDM:ää ja Anti-DreamBoothia vastaan hyvin tuloksin. Vahvasti teksturoiduissa taidetyyleissä, kuten öljymaalauksissa, kuvanlaadun myönnetään kuitenkin kärsivän. (Zhao et al. 2023b.)

Qin et al. (2023) ehdottavat puhdistukseen DiffCleaner-menetelmää, jonka ideana on ensin tuhota ja sitten palauttaa kuvanlaatua. Suojattu kuva käsitellään ensin perinteisillä korruptoivilla operaatioilla, kuten JPEG-pakkauksella tai Gaussin sumennuksella, harhauttavien häiriöiden kumoamiseksi. Tämän jälkeen kuvanlaatua palautetaan diffuusiopohjaisella kohinanpoistomenetelmällä. Menetelmän esitetään toimivan hyvin ainakin Anti-Dreamboothilla ja PhotoGuardilla käsiteltyjä kuvia vastaan, johtaen laadukkaisiin generoituihin kuviin. (Qin et al. 2023.)

Cao et al. (2023) esittelevät IMPRESS-menetelmän, joka perustuu diffuusiomallin luontaiseen kykyyn uudelleenrakentaa sille syötetty kuva. Jos syötteenä käyttää harhauttavasti suojattua kuvaa, eroaa luotu rekonstruktio syötekuvasta, kun taas puhtaan kuvan perusteella luotu rekonstruktio vastaa itseään. Tästä muodostuu optimointiongelma, jossa sekä puhdistettu kuva että sen perusteella luotu rekonstruktio pyritään saamaan visuaalisesti vastaamaan alkuperäistä suojattua kuvaa. Menetelmän raportoidaan saavuttavan hyviä tuloksia Glaze- ja Photoguard -suojauksia vastaan, kun kokeissa käytetään kuvia julkisista datakokoelmista. (Cat et al. 2023.) Shan et al. (2023c) kuitenkin kyseenalaistavat puhdistuksen tehokkuuden tosielämän skenaarioissa, joissa esikoulutettu malli ei valmiiksi tunne suojeltavaa tyyliä, ja kiinnittävät huomiota myös puhdistuksen aiheuttamaan laadunmenetykseen.

Menetelmien arviointi

Suojausten ja niiden puhdistuskeinojen arviointi on jossain määrin hankalaa. Automatisoidut kuvien laadun arviointimenetelmät, kuten FID (Fréchet inception distance) ja CLIP-pisteytys, ovat rajoittuneita ja saattavat erota ihmisen havaitsemasta laadusta (Saharia et al. 2022). Varsinkin tyylin suojaamisen toimivuutta on vaikea arvioida, koska taidetyylien arviointi on osin subjektiivista (Sandoval-Segura, Geiping & Goldstein 2023), ja vaatisi erityisasiantuntemusta (Shan et al. 2023c). Esimerkiksi Glaze-suojauksen toimivuuden varmistamiseen käytettiin alun perin sekä taiteilijoiden arvioita että CLIP-malliin perustuvaa taidegenrejen luokittelua (Shan et al. 2023a). Myöhemmissä versioissa CLIP-arvioinnista on kuitenkin luovuttu, sillä se ei ota huomioon kuvan laadun heikkenemistä. (Shan et al. 2023c). Paras arvioija menetelmän toimivuudesta olisi varmasti taiteilija itse.

3.3 Menetelmien rajoitukset

Tulevaisuudenkestävyys

Suojausmenetelmien perimmäinen ongelma on, että kun suojattu kuva on kerran ladattu datakokoelmaan, suojausta ei voi enää jälkikäteen päivittää. Teknologinen kehitys hyvin todennäköisesti lopulta rikkoo suojauksen toimivuuden. (Radiya-Dixit et al. 2021.) Niin kauan kuin ihminen pystyy näkemään, mitä kuva esittää, tämä todistaa kuvan semanttisen sisällön palauttamisen olevan ainakin teoriassa mahdollista (Sandoval-Segura, Geiping & Goldstein 2023).

Uuden diffuusiomallin julkaisu ei silti välttämättä automaattisesti mitätöi kaikkia aiempia suojausmenetelmiä. Zhang et al. (2023) esittävät, että esimerkiksi Stable Diffusion v2 on itseasiassa jopa heikompi Stable Diffusion v1 -malleilla generoituja harhauttavia esimerkkejä vastaan kuin toisinpäin. Tämä implikoi, että myöhemmät mallit voivat paitsi periä aiemman mallin haavoittuvuudet, myös lisätä uusia, mihin mallien kehittäjien kannattaisi kiinnittää enemmän huomiota (Zhang et al. 2023). Suojausmenetelmien siirrettävyyttä arvioivat kokeet osoittavat, että suojaus säilyy hyvin ainakin eri Stable Diffusion -mallien välillä (Liu et al. 2023, Van Le et al. 2023, Xue et al. 2024, Ye et al. 2023), mutta heikkenee, jos häiriöiden generointiin käytetty malli on huomattavan erilainen kuin lopulliseen diffuusioprosessiin käytetty (Zhao et al. 2023a). Huomioitavaa on myös, että latenttia diffuusiomallia vastaan suunniteltu

gradienttipohjainen hyökkäys ei toimi pikselitasolla toimivia diffuusiomalleja vastaan (Xue et al. 2024).

Kuvansuojaustekniikoiden toimivuutta kaikkia tulevaisuuden teknologioita vastaan on kuitenkin mahdotonta taata, etenkin, koska vastatoimia kehitetään aktiivisesti (Shan et al. 2023a). Salman et al. (2023) toivovat siksi yhteistyötä diffuusiomallien kehittäjien, käyttäjien ja potentiaalista koulutusdataa isännöivien tahojen välillä. Heidän ehdotuksensa on, että mallien kehittäjät voisivat antaa käyttäjien immunisoida kuvansa kehittäjän oman sovellusrajapinnan kautta. Tulevaisuudenkestävyys voitaisiin tällöin taata lisäämällä mallien koulutusvaiheessa takaportti olemassa oleville suojaaville häiriöille. (Salman et al. 2023.) Myös Qin et al. (2023) esittävät, että datan suojaaminen ei ole riittävää, vaan suojelumekanismit tulisi sisäänrakentaa esikoulutettuun diffuusiomalliin. Epäselväksi jää, löytyykö tällaisiin järjestelyihin tahtoa mallien kehittäjiltä.

Shan et al. (2023c) huomauttavat, että suojausmenetelmän käyttö saattaa itsessään toimia ilmaisuna siitä, että kuvan käyttö tekoälyn koulutuksessa ei ole sallittua. Esimerkiksi Yhdysvaltojen DMCA-tekijänoikeuslain osa 1201 kieltää teosten teknologisen suojauksen kiertämisen (17 U.S.C § 1201). Myös Suomen tekijänoikeuslain pykälä 50 a § kieltää teosta suojaavan tehokkaan teknisen toimenpiteen kiertämisen (L 404/1961). Nähtäväksi jää, pätevätkö lakipykälät harhauttavien suojausmenetelmien käyttöön.

Puhtaan koulutusdatan määrä

Tyyliä suojelevien menetelmien teho heikentyy, mikäli koulutukseen käytetyssä datakokoelmassa on myös saman taiteilijan suojaamatonta materiaalia. Aivan kaiken datan ei kuitenkaan tarvitse olla suojattua. Shan et al. (2023a) väittävät, että kohtuulliseen tyylin suojaukseen riittää, jos noin neljännes koulutusdataan sisällytetyistä taiteilijan töistä on Glaze-suojattu. Aktiiviset taiteilijat voivat siis edelleen pyrkiä estämään uusia generaattorimalleja imitoimasta tyyliään, vaikka vanhoja töitä olisi jo päätynyt olemassa oleviin koulutusdatoihin. (Shan et al. 2023a.) Zhao et al. (2023a) taasen raportoivat huomattavan vaikutuksen, kun puolet konseptia edustavasta opetusdatasta on UDP-suojattu. Myös Anti-Dreambooth-suojauksen esitetään toimivan, jos hienosäätö tehdään pienellä kuvamäärällä, josta vähintään puolet on suojattu (Van Le et al. 2023).

Moniin käyttötarkoituksiin opetusdataa lienee kuitenkin jo nyt riittävästi saatavilla. Esimerkiksi tyyliä imitoivan hienosäätömallin kouluttamiseen todennäköisesti riittävät alalla vähänkään pidempään toimineen taiteilijan jo julkaisemat työt. Voi olla, että myös perusmallien koulutuksessa saavutetaan lopulta piste, jossa suuremmista datakokoelmista ei ole enää merkittävää hyötyä. Saharia et al. (2022) raportoivat, että käytetyn kielimallin skaalaamisella on diffuusiomallin kokonaissuorituskykyyn jopa suurempi vaikutus, kuin diffuusioprosessista vastaavan U-Netin skaalaamisella.

Toistaiseksi avoin kysymys on myös se, miten tekoälygeneroidun kuvamateriaalin määrän räjähdysmäinen kasvu viime vuosina vaikuttaa tekoälymallien kehitykseen jatkossa. Synteettistä dataa päätyy mallien koulutusdataan yhä enemmän joko tarkoituksella tai vahingossa. Alustavien tutkimustulosten mukaan generoidun datan käyttäminen tekoälymallin koulutuksessa heikentää mallin laatua tai monimuotoisuutta, mikäli myös uutta aitoa opetusdataa ei lisätä riittävästi. (Alemohammad et al. 2023.)

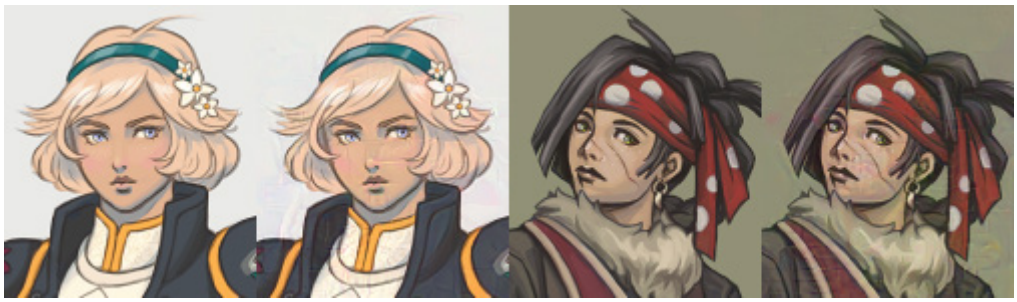
Käytettävyys

Oletetaan, että suojausmenetelmät toimivat niin kuin esitetty nykyisiä diffuusiomallipohjaisia kuvageneraattoreita vastaan. Jotta menetelmillä olisi vaikutusta tekoälymallien kehitykseen jatkossa, on niiden levittävä laajempaan käyttöön. Suojaustyökaluista esimerkiksi Glaze on ladattu yli 2.3 miljoonaa kertaa, ja Nightshade yli 300 000 kertaa (Sand Lab 2024c). Aktiivisten käyttäjien määrästä ei kuitenkaan ole tietoa saatavilla. Todennäköisesti määrä on pienempi, sillä suojauskeinojen käytettävyyteen liittyy vielä joitain haasteita.

Menetelmät ovat laskennallisesti raskaita, sillä häiriöiden laskemiseen tarvitaan samankaltaisia tekoälymalleja kuin mitä vastaan kuvia yritetään suojata. Suojaustyökaluista kevyimpiä ovat Glaze, jonka käyttö vaatii 3.6Gt GPU-muistia (Sand Lab 2023b), ja yli 4Gt vaativa Nightshade (Sand Lab 2024b). Muut menetelmät vaativat GPU-muistia selvästi enemmän (Zheng et al. 2023, Mist 2023). Ilman riittävän tehokasta näytönohjainta esimerkiksi Nightshaden käyttäminen CPU-laskennalla on todella hidasta. Vertauksena alan standardina pidetyn kuvankäsittelyohjelma Adobe Photoshopin nykyinen versio vaatii vähimmillään vain 1.5Gt GPU-muistia (Adobe 2024b). Lisäksi kaikilla nykytaiteilijoilla ei edes ole tietokoneita, vaan työt saatetaan tehdä mobiililaitteilla (Sand Lab 2023c). Ei siis voida olettaa, että kaikki kuvansuojaustyökaluista kiinnostuneet voisivat käyttää niitä omilla laitteillaan. Ongelmaa ratkaisemaan on esitetty verkon yli toimivia versioita (Sand Lab 2023c), tai suojauksen

integroimista suoraan kuvien julkaisualustoihin (Cara Project 2023). Sosiaalisen median alustojen suhtautuminen aiheeseen on kuitenkin vaihtelevaa, ja osa palveluntarjoajista on jopa tehnyt tekoälymallien kehittäjien kanssa sopimuksia, jotka antavat käyttäjien luoman sisällön näiden käyttöön (Robertson 2024).

Toinen ongelma on, että vaikka kuviin lisättävien häiriöiden väitetään olevan ihmissilmään huomaamattomia, tosiasiassa tämä ei useinkaan pidä paikkaansa (kuva 4). Varsinkin tasaisiin värialueisiin perustuvissa piirtotyyleissä artefaktit voivat olla hyvinkin häiritseviä, kun taas vahvemmin teksturoituun kuvaan kohinan piilottaminen voi onnistua kohtuullisen hyvin (Ahn et al. 2024, liite 1). Huomionarvoista kuitenkin on, että monet verkossa toimivat taiteilijat ovat jo tottuneet käyttämään kuvien väärinkäytön ehkäisemiseen niiden ulkoasua heikentäviä keinoja, kuten näkyviä vesileimoja (Shan et al. 2023a). Kuvien laadun heikentyminen ei siis välttämättä ole aina niin suuri ongelma kuin voisi luulla. Suojaustekniikoita on myös mahdollista vielä kehittää kuvanlaadun saralla. Esimerkiksi IMPASTO-menetelmä mahdollistaa häiriöiden intensiteetin säätämisen kuvan eri alueilla riippuen siitä, miten helppoa häiriöiden huomaaminen alueella on (Ahn et al. 2024). PAG-menetelmä taas antaa taitelijalle suoran mahdollisuuden määrittää kuvan osia, joihin häiriöitä ei lisätä (Tan et al. 2024). Tämän vaikutusta suojauksen tehoon ei tosin täsmennetä.



Kuva 4. Glaze voi aiheuttaa selkeitä häiriöitä kuvan tärkeisiin osiin, kuten hahmojen kasvoihin.

Yksittäisen taiteilijan näkökulmasta suojausmenetelmä toimii, jos sen seurauksena kuva ei päädy tekoälymallin opetusdataan, tai vähintään vääristää generoinnin lopputuloksen. Todennäköisimmin tyyli-imitaation kohteeksi joutunevat valmiiksi tunnetut taiteilijat, joiden töitä on jo paljon verkossa. Aloittelevan tai muuten tuntemattoman taiteilijan kohdalla riski on pienempi. Tällaisessa tapauksessa kuvan suojauksesta saatava koettu hyöty ei välttämättä korvaa laadunmenetyksestä aiheutuvaa haittaa, vaikka taiteilija vastustaisikin kuvien luvatonta hyödyntämistä.

4 Yhteenveto

Tässä tutkielmassa perehdyttiin harhauttaviin hyökkäyksiin, joiden tavoite on estää kuvien hyödyntäminen diffuusiomalleihin perustuvien kuvageneraattorien kouluttamisessa. Harhauttaviin hyökkäyksiin turvaudutaan tekijänoikeuksien tai yksityisyyden suojaamiseksi, koska koulutusdatan kerääminen verkosta on yleinen käytäntö, eikä tämänhetkinen lainsäädäntö aseta datan lisensoinnille selkeitä vaatimuksia. Lisäksi nykyiset hienosäätömenetelmät mahdollistavat esimerkiksi tietyn taiteilijan tyyliä jäljittelevän mallin kouluttamisen hyvin vähäisellä opetusdatan määrällä ja resursseilla.

Harhauttava esimerkki on tekoälymallille annettava syöte, joka vaikuttaa näennäisesti normaalilta, mutta on tarkoituksella suunniteltu niin, että malli tulkitsee sen väärin. Harhauttavien esimerkkien olemassaolo on syväoppivien neuroverkkojen yleinen piirre. Perinteisen harhauttavan hyökkäyksen tavoite voi olla esimerkiksi saada kuvanluokittelija luokittelemaan kuva väärin.

Kuvan suojaamiseen diffuusiomalleilta tähtäävä harhauttava hyökkäys perustuu kuvaan lisättäviin häiriöihin, jotka opettavat mallille vääriä piirteitä, tai muuten häiritsevät piirteiden oppimista. Hyökkäyksen tavoite voi olla vain kuvan latenttiesityksen sotkeminen, tai se voi ottaa huomioon koko diffuusioprosessin, mikä on kuitenkin laskennallisesti vaativampaa. Olennaista on pitää häiriöt visuaalisesti huomaamattomina, mikä johtaa optimointiongelmaan. Monet tekniikoista käyttävät häiriöiden laskemisen apuna suojeltavasta kuvasta selvästi eroavaa kohdekuvaa, jonka piirre-esitystä lähemmäs suojeltava kuva viedään. Koulutushäviön maksimoimiseen tai minimoimiseen voi pyrkiä myös ilman kohdekuvaa. Häiriöiden laskemiseen voidaan käyttää joko hyökkäyksen kohteena olevaa diffuusiomallia tai erikseen koulutettua sijaismallia, mikä on hyödyllistä etenkin hienosäätötekniikoihin kohdistuvissa hyökkäyksissä.

Suojausmenetelmien perimmäinen heikkous on, että suojausta ei voi jälkikäteen päivittää, kun kuva on kerran julkaistu ja mahdollisesti sisällytetty koulutusdatakokoelmaan. Datan luvaton käyttäjä voi tämän jälkeen vapaasti pyrkiä suojauksen purkamiseen erilaisilla kuvanprosessointitekniikoilla. Yksinkertaiset kuvankäsittelymenetelmät, kuten JPEG-pakkaus tai sumennus, heikentävät suojausten tehoa, mutta johtavat yleensä myös kuvanlaadun liialliseen huononemiseen. Hienostuneempia vastatekniikoita kehitetään kuitenkin aktiivisesti,

eikä suojauksen toimivuutta myöskään uusia kuvageneraattorimalleja vastaan ole mahdollista taata.

Menetelmien käyttökelpoisuutta rajoittavat myös niistä aiheutuva kuvanlaadun menetys, suojauksen lisäämiseen vaadittava laskentateho, sekä jo olemassa olevan puhtaan opetusdatan määrä.

Pohdinta

Arviot harhauttavien suojausmenetelmien toimivuudesta jäävät pitkälti menetelmien kehittäjien oman sanan varaan, sillä kattavaa systemaattista tutkimusta aiheesta on tehty vähän. Osassa julkaisuista on myös selviä puutteita suojauksen sietokyvyn arvioinnissa. Lisänäyttö toimivuudesta olisi tarpeen, sillä vaakakupissa painaa suojauksen aiheuttama kuvanlaadunmenetys, sekä häiriöiden lisäämiseen käytetty laskenta-aika. Asiaa hankaloittaa se, että suojausten toimivuuden arviointi on osin subjektiivista, eivätkä tekoälytutkijat välttämättä ole taiteen asiantuntijoita. Koska suojausmenetelmien pääasiallinen kohderyhmä ei todennäköisesti itse käytä tekoälykuvageneraattoreita, ei omien kokeiden tekeminen ole useimmille mahdollista.

Vaikka menetelmien tehokkuudesta ei ole täyttä varmuutta, voi niiden käytöllä olla toinen merkitys. Suojausmenetelmän käytön voi tulkita olevan tekijänoikeudenhaltijan ilmaisu siitä, ettei teosta ole sallittua käyttää tekoälymallien koulutuksessa. Epäselvää on, onko tällaisella ilmaisulla mitään virkaa lain edessä. Suojattujen kuvien luvaton käyttäjä asettuu kuitenkin vähintään eettisesti arveluttavaan valoon.

Suojausmenetelmien saama mediahuomio toimii lisäksi keskustelunavauksena, joka tuo taiteilijoiden ahdingon laajempaan tietoisuuteen. Tämä voi osaltaan johtaa yleisen mielipiteen kääntymiseen tekoälykielteisemmäksi, mikä muodostaa tekoälymallien kehittäjille paineita toimia reilummin eri osapuolia kohtaan. Vaikka suojausmenetelmistä saatava hyöty jäisikin teknisesti lyhytkestoiseksi, on tärkeää, että tekoälytutkimusta tehdään myös taiteilijoiden näkökulma huomioiden. Harhauttaviin suojausmenetelmiin keskittyvällä tutkimuksella on varmasti roolinsa keskustelussa tekoälyn pelisäännöistä.

Lähteet

Artikkelit

Ahn, N., Ahn, W., Yoo, K., Kim, D. ja Nam, S.H., 2024. *Imperceptible Protection against Style Imitation from Diffusion Models.* arXiv:2403.19254 [cs.CV]. DOI: 10.48550/arXiv.2403.19254

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A.I., Babaei, H., LeJeune, D., Siahkoohi, A. ja Baraniuk, R.G., 2023. *Self-Consuming Generative Models Go MAD.* arXiv:2307.01850 [cs.LG]. DOI: 10.48550/arXiv.2307.01850

Ali, S. ja Breazeal, C., 2023. *Studying Artist Sentiments around AI-generated Artwork.* arXiv:2311.13725 [cs.HC]. DOI: 10.48550/arXiv.2311.13725

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y. ja Ramesh, A., 2023. *Improving Image Generation with Better Captions*. URL: https://cdn.openai.com/papers/dall-e-3.pdf

Bird, C., Ungless, E. ja Kasirzadeh, A., 2023. Typology of Risks of Generative Text-to-Image Models. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23). Association for Computing Machinery, New York, NY, USA*, s. 396–410. DOI: 10.1145/3600211.3604722

Cao, B., Li, C., Wang, T., Jia, J., Li, B. ja Chen, J., 2023. IMPRESS: Evaluating the Resilience of Imperceptible Perturbations Against Unauthorized Data Usage in Diffusion-Based Generative AI. *Advances in Neural Information Processing Systems*, *36*, s. 10657-10677. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/222dda29587fbc2979ca99fd5ed00735-Paper-Conference.pdf

Chen, R., Jin, H., Chen, J. ja Sun, L., 2023. *EditShield: Protecting Unauthorized Image Editing by Instruction-guided Diffusion Models.* arXiv:2311.12066 [cs.CR]. DOI: 10.48550/arXiv.2311.12066

Cui, Y., Ren, J., Lin, Y., Xu, H., He, P., Xing, Y., Fan, W., Liu, H. ja Tang, J., 2023. *FT-Shield: A Watermark Against Unauthorized Fine-tuning in Text-to-Image Diffusion Models.* arXiv:2310.02401 [cs.CV]. DOI: 10.48550/arXiv.2310.02401

Dhariwal, P. ja Nichol, A., 2021. Diffusion Models Beat GANs on Image Synthesis. *Advances in Neural Information Processing Systems*, *34*, s. 8780-8794.

URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G. ja Cohen-Or, D., 2022. *An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion.* arXiv:2208.01618 [cs.CV]. DOI: 10.48550/arXiv.2208.01618

Goodfellow, I.J., Shlens, J. ja Szegedy, C., 2015. Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations (ICLR 2015).* DOI: 10.48550/arXiv.1412.6572

Han, S., Lin, C., Shen, C., Wang, Q. ja Guan, X., 2023. 2023. Interpreting Adversarial Examples in Deep Learning: A Review. *ACM Computing Surveys, 55(14s),* s.1-38. DOI: 10.1145/3594869

Ho, J., Jain, A. ja Abbeel, P., 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, *33*, s.6840-6851. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. ja Chen, W., 2021. *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv:2106.09685 [cs.CL]. DOI: 10.48550/arXiv.2106.09685

Jiang, H.H., Brown, L., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J. ja Gebru, T., 2023. AI Art and its Impact on Artists. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23). Association for Computing Machinery, New York, NY, USA,* s. 363-374. DOI: 10.1145/3600211.3604681

Liang, C. ja Wu, X., 2023. *Mist: Towards Improved Adversarial Examples for Diffusion Models.* arXiv:2305.12683 [cs.CV]. DOI: 10.48550/arXiv.2305.12683

Liang, C., Wu, X., Hua, Y., Zhang, J., Xue, Y., Song, T., Xue, Z., Ma, R. ja Guan, H., 2023. *Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples*. arXiv:2302.04578 [cs.CV]. DOI: 10.48550/arXiv.2302.04578

Liu, Y., Fan, C., Dai, Y., Chen, X., Zhou, P. ja Sun, L., 2023. *Toward Robust Imperceptible Perturbation against Unauthorized Text-to-image Diffusion-based Synthesis.* arXiv:2311.13127 [cs.CV]. DOI: 10.48550/arXiv.2311.13127

Ma, Y., Zhao, Z., He, X., Li, Z., Backes, M. ja Zhang, Y., 2023. *Generative Watermarking Against Unauthorized Subject-Driven Image Synthesis.* arXiv:2306.07754 [cs.CV]. DOI: 10.48550/arXiv.2306.07754

Mustafa, A., Khan, S.H., Hayat, M., Shen, J. ja Shao, L., 2020 Image Super-Resolution as a Defense Against Adversarial Attacks. *IEEE Transactions on Image Processing, 29, s. 1711-1724.* DOI: 10.1109/TIP.2019.2940533

Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A. ja Anandkumar, A., 2022. Diffusion Models for Adversarial Purification. *Proceedings of the 39th International Conference on Machine Learning, PMLR 162,* s. 16805-16827. URL: https://proceedings.mlr.press/v162/nie22a/nie22a.pdf

Peebles, W. ja Xie, S., 2023. Scalable Diffusion Models with Transformers. *IEEE/CVF International Conference on Computer Vision (ICCV), Pariisi, Ranska, 2023,* s. 4172-4182. DOI: 10.1109/ICCV51070.2023.00387

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J. ja Rombach, R., 2023. *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis.* arXiv:2307.01952 [cs.CV]. DOI: 10.48550/arXiv.2307.01952

Qin, T., Gao, X., Zhao, J. ja Ye, K., 2023. Destruction-Restoration Suppresses Data Protection Perturbations against Diffusion Models. *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI), Atlanta, GA, USA*, s. 586-594. DOI: 10.1109/ICTAI59109.2023.00093

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. ja Sutskever, I., 2021. Learning Transferable Visual Models from Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning, PMLR 139,* s. 8748-8763. URL: http://proceedings.mlr.press/v139/radford21a/radford21a.pdf

Radiya-Dixit, E., Hong, S., Carlini, N. ja Tramèr, F., 2021. *Data Poisoning Won't Save You From Facial Recognition.* arXiv:2106.14851 [cs.LG]. DOI: 10.48550/arXiv.2106.14851

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. ja Chen, M., 2022. *Hierarchical Text-Conditional Image Generation with CLIP Latents.* arXiv:2204.06125 [cs.CV]. DOI: 10.48550/arXiv.2204.06125

Rhodes, A., Bhagat, R., Ciftci, U.A. ja Demir, I., 2023. *My Art My Choice: Adversarial Protection Against Unruly AI.* arXiv:2309.03198 [cs.CV]. DOI: 10.48550/arXiv.2309.03198

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. ja Ommer, B., 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA,* s. 10674-10685. DOI: 10.1109/CVPR52688.2022.01042

Ronneberger, O., Fischer, P. ja Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science(), vol 9351. Springer, Cham.* DOI: 10.1007/978-3-319-24574-4_28

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M. ja Aberman, K., 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Kanada,* s. 22500-22510. DOI: 10.1109/CVPR52729.2023.02155

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D.J. ja Norouzi, M., 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Advances in Neural Information Processing Systems*, *35*, s. 36479-36494. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf

Salman, H., Alaa, K., Leclerc, G., Ilyas, A. ja Madry, A., 2023. Raising the Cost of Malicious AI-Powered Image Editing. *Proceedings of the 40th International Conference on Machine Learning, PMLR 202,* s. 29894-29918. URL: https://proceedings.mlr.press/v202/salman23a.html

Sandoval-Segura, P., Geiping, J. ja Goldstein, T., 2023. *JPEG Compressed Images Can Bypass Protections Against AI Editing.* arXiv:2304.02234 [cs.LG]. DOI: 10.48550/arXiv.2304.02234

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R. ja Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems, 35,* s. 25278-25294. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/a1859debfb3b59d094f3504d5ebb6c25-Paper-Datasets_and_Benchmarks.pdf

Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R. ja Zhao, B.Y., 2023a. Glaze: Protecting Artists from Atyle Mimicry by Text-to-Image Models. *32nd USENIX Security Symposium (USENIX Security 23),* s. 2187-2204. URL: https://www.usenix.org/conference/usenixsecurity23/presentation/shan

Shan, S., Ding, W., Passananti, J., Zheng, H. ja Zhao, B.Y., 2023b. *Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models.* arXiv:2310.13828 [cs.CR]. DOI: 10.48550/arXiv.2310.13828

Shan, S., Wu, S., Zheng, H. ja Zhao, B.Y., 2023c. *A Response to Glaze Purification via IMPRESS.* arXiv:2312.07731 [cs.CR]. DOI: 10.48550/arXiv.2312.07731

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. ja Ganguli, S., 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *Proceedings of the 32nd International Conference on Machine Learning, PMLR 37,* s. 2256-2265. URL: http://proceedings.mlr.press/v37/sohl-dickstein15.pdf

Somepalli, G., Singla, V., Goldblum, M., Geiping, J. ja Goldstein, T., 2023. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Kanada,* s. 6048-6058, DOI: 10.1109/CVPR52729.2023.00586

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. ja Fergus, R., 2013. *Intriguing properties of neural networks.* arXiv:1312.6199 [cs.CV]. DOI: 10.48550/arXiv.1312.6199

Tan, Z., Wang, S., Yang, X. ja Huang, K., 2024. PAG: Protecting Artworks from Personalizing Image Generative Models. *Neural Information Processing. ICONIP 2023. Lecture Notes in Computer Science, vol 14450. Springer, Singapore.* DOI: 10.1007/978-981-99-8070-3_33

Van Le, T., Phung, H., Nguyen, T.H., Dao, Q., Tran, N.N. ja Tran, A., 2023. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. *2023 IEEE/CVF International Conference on Computer Vision (ICCV), Pariisi, Ranska,* s. 2116-2127, DOI: 10.1109/ICCV51070.2023.00202

Wang, F., Tan, Z., Wei, T., Wu, Y. ja Huang, Q., 2023. *SimAC: A Simple Anti-Customization Method against Text-to-Image Synthesis of Diffusion Models.* arXiv:2312.07865 [cs.CV]. DOI: 10.48550/arXiv.2312.07865

Wu, R., Wang, Y., Shi, H., Yu, Z., Wu, Y. ja Liang, D., 2023. *Towards Prompt-robust Face Privacy Protection via Adversarial Decoupling Augmentation Framework.* arXiv:2305.03980 [cs.CV]. DOI: 10.48550/arXiv.2305.03980

Xue, H., Liang, C., Wu, X. ja Chen, Y., 2024. Toward effective protection against diffusion-based mimicry through score distillation. *The Twelfth International Conference on Learning Representations.* URL: https://openreview.net/forum?id=NzxCMe88HX

Ye, X., Huang, H., An, J. ja Wang, Y., 2023. *DUAW: Data-free Universal Adversarial Watermark against Stable Diffusion Customization.* arXiv:2308.09889 [cs.CV]. DOI: 10.48550/arXiv.2308.09889

Zhang, J., Xu, Z., Cui, S., Meng, C., Wu, W. ja Lyu, M.R., 2023. *On the Robustness of Latent Diffusion Models.* arXiv:2306.08257 [cs.CV]. DOI: 10.48550/arXiv.2306.08257

Zhao, Z., Duan, J., Hu, X., Xu, K., Wang, C., Zhang, R., Du, Z., Guo, Q. ja Chen, Y., 2023a. *Unlearnable Examples for Diffusion Models: Protect Data from Unauthorized Exploitation.* arXiv:2306.01902 [cs.CV]. DOI: 10.48550/arXiv.2306.01902

Zhao, Z., Duan, J., Xu, K., Wang, C., Guo, R.Z.Z.D.Q. ja Hu, X., 2023b. *Can Protective Perturbation Safeguard Personal Data from Being Exploited by Stable Diffusion?.* arXiv:2312.00084 [cs.CV]. DOI: 10.48550/arXiv.2312.00084

Zheng, B., Liang, C., Wu, X. ja Liu, Y., 2023. *Understanding and Improving Adversarial Attacks on Latent Diffusion Model.* arXiv:2310.04687v2 [cs.CV]. DOI: 10.48550/arXiv.2310.04687

Verkkosivut

17 USC § 1201: Circumvention of copyright protection systems. URL: https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title17-section1201&num=0&edition=prelim. Viitattu 7.5.2024.

Adobe, 2024a. *Adobe vs. Stable Diffusion: tuo työnkulkuihisi enemmän ideoita nopeasti Fireflyn avulla*. Adobe 4.1.2024. URL: https://www.adobe.com/fi/products/firefly/discover/firefly-vs-stable-diffusion.html. Viitattu 10.3.2024.

Adobe, 2024b. *Photohop system requirements*. Adobe 23.4.2024. URL: https://helpx.adobe.com/photoshop/system-requirements.html. Viitattu 11.3.2024.

Baio, A., 2022. *Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator*. Waxy.org 30.8.2022. URL: https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/. Viitattu 10.3.2024.

Beaumont, R., 2022. *LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS*. LAION blog 31.3.2022. URL: https://laion.ai/blog/laion-5b/. Viitattu: 17.3.2024.

Brittain, B., 2023. *Getty Images lawsuit says Stability AI misused photos to train AI.* Reuters 2.6.2023. https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/. Viitattu 6.3.2024.

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R. ja Ramesh, A., 2024. *Video generation models as world simulators.* OpenAI 15.2.2024. URL: https://openai.com/research/video-generation-models-as-world-simulators. Viitattu 18.4.2024.

Cara Project, 2023. *Introducing: Cara Glaze*. Cara blog 1.12.2023. URL: https://blog.cara.app/blog/cara-glaze-about. Viitattu 11.3.2024.

Cuenca, P. ja Paul, S., 2023. *Using LoRA for Efficient Stable Diffusion Fine-Tuning.* Hugging Face 26.1.2023. URL: https://huggingface.co/blog/lora. Viitattu 2.3.2024.

Google DeepMind, 2023. *Imagen 2 - our most advanced text-to-image technology*. Google DeepMind Blog. URL: https://deepmind.google/technologies/imagen-2/. Viitattu 10.3.2024.

Heikkilä, M., 2022. *This artist is dominating AI-generated art. And he's not happy about it.* MIT Technology Review 16.9.2022. URL: https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/. Viitattu 10.3.2024.

Hugging Face. *DreamBooth.* Hugging Face documentations. URL: https://huggingface.co/docs/diffusers/en/training/dreambooth. Viitattu 5.5.2024.

L 404/1961. Tekijänoikeuslaki 8.7.1961/404. URL: https://finlex.fi/fi/laki/ajantasa/1961/19610404. Viitattu 7.5.2024.

LAION FAQ. LAION. URL: https://laion.ai/faq/. Viitattu 10.3.2024.

Lanz, J., 2023. *Greg Rutkowski Was Removed From Stable Diffusion, But AI Artists Brought Him Back*. Decrypt 30.7.2023. URL: https://decrypt.co/150575/greg-rutkowski-removed-from-stable-diffusion-but-brought-back-by-ai-artists. Viitattu 6.3.2024.

Miller, C., 2023. *ai.txt: A new way for websites to set permissions for AI.* Spawning Blog 30.5.2023. URL: https://spawning.substack.com/p/aitxt-a-new-way-for-websites-to-set. Viitattu 10.3.2024.

Mist, 2023. *Device Requirements*. Mist Documentation. URL: https://mist-documentation.readthedocs.io/en/latest/content/device.html. Viitattu 7.4.2024

Mostaque, E. [@EMostaque], 2022. "*The reason Greg Rutkowski and many other artists "work" in #StableDiffusion is not actually due to the LAION dataset but the @OpenAI CLIP model (L14) used to teach it language We don't know dataset it's based on (closed) so even if an artist removed from LAION still "knows" style."* [Twiitti] Twitter 19.9.2022. URL: https://twitter.com/EMostaque/status/1571634871084236801. Viitattu 10.3.2024.

Nichol, A., 2022. *DALL·E 2 pre-training mitigations*. OpenAI 28.6.2022. URL: https://openai.com/research/dall-e-2-pre-training-mitigations. Viitattu 6.3.2024.

Salkowitz, R., 2022. *Midjourney Founder David Holz On The Impact Of AI On Art, Imagination And The Creative Economy*. Forbes 16.9.2022. URL: https://www.forbes.com/sites/robsalkowitz/2022/09/16/midjourney-founder-david-holz-on-the-impact-of-ai-on-art-imagination-and-the-creative-economy/?sh=1f96c87a2d2b. Viitattu 6.3.2024.

Sand Lab, University of Chicago, 2023a. *Glaze Publications & Media Coverage.* Sand Lab 25.6.2023. URL: https://glaze.cs.uchicago.edu/media.html Viitattu 6.3.2024.

Sand Lab, University of Chicago, 2023b. *Glaze User's Guide.* Sand Lab 26.6.2023. URL: https://glaze.cs.uchicago.edu/userguide.html. Viitattu 11.3.2024.

Sand Lab, University of Chicago, 2023c. *Web Glaze*. Sand Lab. URL: https://glaze.cs.uchicago.edu/webglaze.html. Viitattu 11.3.2024.

Sand Lab, University of Chicago, 2024a. *Publications & Media Coverage.* Sand Lab 10.2.2024. URL: https://nightshade.cs.uchicago.edu/media.html. Viitattu 6.3.2024.

Sand Lab, University of Chicago, 2024b. *Nightshade User's Guide*. Sand Lab 18.1.2024. URL: https://nightshade.cs.uchicago.edu/userguide.html. Viitattu 11.3.2024.

Sand Lab, University of Chicago, 2024c. *About The Glaze Project.* Sand Lab. URL: https://nightshade.cs.uchicago.edu/aboutus.html. Viitattu 7.4.2024.

Saveri, J. ja Butterick, M., 2023. *We've filed a lawsuit challenging AI image generators for using artists' work without consent, credit, or compensation*. URL: https://stablediffusionlitigation.com/. Viitattu 6.3.2024.

Stability AI, 2024. *Stable Diffusion 3*. Stability AI 22.2.2024. URL: https://stability.ai/news/stable-diffusion-3. Viitattu 18.4.2024.

Robertson, A., 2024. *Tumblr's owner is striking deals with OpenAI and Midjourney for training data, says report.* The Verge 27.2.2024. URL:

https://www.theverge.com/2024/2/27/24084884/tumblr-midjourney-openai-training-data-deal-report. Viitattu 7.4.2024.

Rocca, J., 2019. *Understanding Variational Autoencoders (VAEs).* Medium 24.9.2019. URL: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73. Viitattu 4.5.2024.

Rombach,R., Esser, P. ja Ha, D., 2022. *Stable Diffusion v2-base Model Card*. Hugging Face. URL: https://huggingface.co/stabilityai/stable-diffusion-2-base. Viitattu 6.3.2024.

Weatherbed, J., 2024. *How to keep your art out of AI generators*. The Verge 7.2.2024. URL: https://www.theverge.com/24063327/ai-art-protect-images-copyright-generators. Viitattu 6.3.2024.

Liitteet

Liite 1. Glaze- ja Nightshade-käsiteltyjä kuvia

Esimerkki 1a. Alkuperäinen kuva ylhäällä, Glaze 1.1.1 -käsitelty kuva alhaalla (default intensity / medium render quality).

Glaze-suojatussa kuvassa on havaittavissa lievä tekstuuri, minkä lisäksi selkeitä värillisiä artefakteja on esimerkiksi vasemmassa yläkulmassa ja hahmon etummaisen jalan alapuolella.

Esimerkki 1b. Nightshade 1.0 -käsitelty kuva ylhäällä (default intensity / medium render quality), Nightshade + Glaze -käsitelty kuva alhaalla.

Glaze-käsittely ei enää tuota värikkäitä artefakteja, kun kuva on käsitelty ensin Nightshadella, mutta yleinen kuvanlaatu heikkenee. Ohjelmien kehittäjät suosittelevat nimenomaan tätä järjestystä (Sand Lab 2024b).

Esimerkki 2. Ylh. vas. alkuperäinen, ylh. oik. Glaze (default intensity / medium render quality), alh. vas. Nightshade (default intensity / medium render quality), alh. oik. Nightshade + Glaze.

Häiriöt erottuvat erityisesti tasaisilta värialueilta.

Esimerkki 3. Nightshade. Ylh. vas. alkuperäinen, ylh. oik. low intensity / medium render quality, alh. vas. default intensity / medium render quality, alh. oik. high intensity / medium render quality.

Muilla kuin korkealla intensiteetillä häiriöt sekoittuvat kuvaan melko huomaamattomasti, ja ovat parhaiten havaittavissa vaaleilla pinnoilla.

Esimerkki 4. Nightshade. Ylh. vas. alkuperäinen, ylh. oik. low intensity / medium render quality, alh. vas. default intensity / medium render quality, alh. oik. high intensity / medium render quality.

Häiriöt ovat parhaiten havaittavissa pilvimuodostelmien ympärillä.