# Generatiivisella tekoälyllä toimivien kuvageneraattoreiden ongelmat kuvien generoinnissa

Turun yliopisto
Tietotekniikan laitos
TkK-tutkielma
Tieto- ja viestintätekniikka
Kesäkuu 2024
Roosa Naula

TURUN YLIOPISTO
Tietotekniikan laitos

Roosa Naula: Generatiivisella tekoälyllä toimivien kuvageneraattoreiden ongelmat kuvien generoinnissa

TkK-tutkielma, 27 s.
Tieto- ja viestintätekniikka
Kesäkuu 2024

---

Kuvageneraattorit ovat kasvattaneet suosiotaan huimasti viime vuosien aikana. Kuvageneraattorit ovat hämmästyttävän hyviä luomaan aidon näköisiä kuvia. Monesti generoituja kuvia on vaikea erottaa aidoista valokuvista. Kuvageneraattorit ovat kuitenkin alttiita tietyille virheille, joista generoidut kuvat voi tunnistaa. Tässä kandidaatintutkielmassa tutustutaan generatiivisella tekoälyllä toimivien kuvageneraattoreiden toimintaan. Aluksi käydään läpi keskeisiä käsitteitä liittyen tekoälyyn. Tämän jälkeen tutustutaan siihen, miten kuvageneraattorit konkreettisesti toimivat. Kuvageneraattorin rakenne jaetaan eri osa-alueisiin. Näitä osa-alueita tarkastellaan yksityiskohtaisemmin Stable Diffusionin rakennetta tarkastelemalla. Tutkielmassa esitellään erilaisia virheitä, jotka esiintyvät generoiduissa kuvissa. Tutkielmassa pohditaan millaisia ongelmia ja puutteita piilee näiden virheiden takana. Kuvageneraattorit tekevät eniten virheitä ihmisten kuvaamisessa, joten tutkielmassa keskitytään ihmisten kuvantamisen ongelmaan. Tutkielmassa tarkastellaan myös kehotteiden merkitystä generoimisprosessin ohjaajina. Kehotteiden hallinta on tärkeää, jos haluaa generoida laadukkaita kuvia. Tutkielmassa nostetaan esiin arkijärjen puute kuvageneraattoreissa. Arkijärjen mallinnus ratkaisisi luultavasti monet kuvien generoimiseen liittyvät ongelmat. Lopuksi pohditaan mikä on se kaikista kriittisin asia, josta monet virheet johtuvat, ja onko kattavan arkijärjen mallintaminen mahdollista.

Asiasanat: generatiivinen tekoäly, diffuusiomalli, kuvageneraattori, taide, virheet

# Sisällys

# 1 Johdanto

Tässä tutkielmassa tutkitaan generatiiviseen tekoälyyn perustuvien kuvageneraattoreiden toimintaa. Tutkielmassa pohditaan, miksi kuvageneraattorit tuottavat virheellisiä kuvia ja mitkä ovat yleisimmät virheet, joita ne tekevät. Tutkielmassa keskitytään pohtimaan, miksi tekoälyllä on vaikeuksia ihmisten kuvaamisessa. Tutkielmassa pohditaan myös arkijärjen mahdollista hyödyntämistä kuvageneraattoreissa. Suosituimpia kuvageneraattoreita ovat Stable Diffusion, OpenAI:n DALL-E sekä Midjourney. Tutkimuksen pääkohteeksi valikoitui AI-kuvageneraattori Stable Diffusion, sillä sen toiminta on läpinäkyvämpää kuin muiden kuvageneraattorien. Stable Diffusion on oikeastaan ainoa, jota on pystytty tutkimaan, sillä se on avoimen lähdekoodin kuvageneraattori.

Todella moni käyttää kuvageneraattoreita joko huvikseen tai kaupalliseen tarkoitukseen. Kovinkaan moni ei tiedä, miten kuvageneraattorit toimivat tai että kuvageneraattoreita käyttäessä saattaa tukea tekijänoikeuksia loukkaavaa toimintaa. Tekoälyn kehitykseen liittyy paljon eettisiä kysymyksiä, joihin olisi hyvä saada vastauksia samalla kun tekoäly tulee yhä isommaksi osaksi yhteiskuntaa. Generatiivisella tekoälyllä luodut kuvat ovat usein joitain ihmisryhmiä syrjiviä opetusdatan yksipuolisuuden takia. Valheellisten generoitujen kuvien avulla voidaan levittää tahallista disinformaatiota. Ihminen oppii parhaiten virheistään, ja tekoälyn tekemistä virheistä voidaan oppia lisää sen toiminnasta. Tekoälyn toimintaa voidaan parantaa tekoälyn tekemien virheiden tarkastelun kautta. Virheettömän tekoälyn kehitys

on tärkeää myös tulevaisuuden kannalta. Tulevaisuudessa tullaan tekemään entistä tehokkaampia ja kykenevämpiä tekoälyn sovelluksia, jolloin virheisiin ei ole varaa. Tekoälyn tutkimusta ei voi tehdä liikaa, ja siksi generatiivisella tekoälyllä toimivat kuvageneraattorit valikoituivat tämän tutkielman pääaiheeksi.

Tässä tutkielmassa etsittiin vastauksia seuraaviin kysymyksiin:

**TK1** Millaisia ongelmia kuvageneraattoreilla on kuvien generoinnissa?

**TK2** Mistä johtuvat kuvageneraattoreiden tuottamat virheelliset kuvat?

Tiedonhaku alkoi suosituimpien tekniikan alan tietokantojen (Web of Science, IEEE sekä ACM) hakupalveluihin tutustumisella. Etsin tietoa seuraavanlaisella hakulausekkeella sekä sen variaatioilla: Generative AI AND (errors OR mistakes) AND (pictures OR images). Aika nopeasti kävi selväksi, ettei hakukannoista löytynyt kunnolla tutkimuksia, joissa olisi tutkittu kuvageneraattoreissa ilmeneviä ongelmia sekä niistä johtuvia virheitä.

Tutustuin aluksi alan keskeisiin papereihin ja löysin keskeisiä tutkimuksia, joissa esiteltyjä tekniikoita käytetään kuvageneraattoreissa. Itse virheitä tarkastelevat tutkimuspaperit löysin Google Scholarista. Hakulauseella "mistakes in generated images"löysin muutamia artikkeleita. Näiden artikkeleiden lähteitä seuraamalla löysin lisää artikkeleita.

Generatiivisella tekoälyllä toimivat kuvageneraattorit ovat vielä erittäin uusi tutkimusala. Kuvageneraattoreiden suosio on kasvanut räjähdysmäisesti parin viime vuoden aikana. Generoiduissa kuvissa olevista virheistä on käyty paljon keskustelua netissä. Generoiduissa kuvissa ilmenevistä virheistä ei kuitenkaan olla vielä tehty kovin kattavaa tutkimusta. Tässä tutkielmassa on koottu yhteen erilaisia menetelmiä, jotka osaltaan vaikuttavat generoiduissa kuvissa ilmeneviin virheisiin. Tämä aihe on kuitenkin niin monipuolinen ja laaja, ettei aiheesta pysty tekemään kaiken kattavaa kirjallisuuskatsausta.

Tämä tutkielma koostuu johdannon lisäksi viidestä muusta luvusta. Toisessa luvussa esitellään aihetta ja selitetään lyhyesti keskeisiä tekoälyyn liittyviä käsitteitä. Kolmannessa luvussa tutustutaan kuvageneraattorien perusrakenteeseen Stable Diffusionin toimintaa tarkastelemalla. Neljännessä luvussa tarkastellaan kuvageneraattoreiden kuvantuoton ongelmia sekä niistä johtuvia virheitä generoiduissa kuvissa. Viidennessä luvussa on neljännen luvun yhteenvetoa. Kuudes luku sisältää yleisen yhteenvedon sekä tulevaisuuden näkymien pohdintaa.

# 2 Taustaa ja termistöä

Tekoäly on yksi isoimmista puheenaiheista tällä hetkellä. Se tuli pari vuotta sitten näkyväksi osaksi ihmisten jokapäiväistä elämää. Tekoälystä on ollut aiemminkin puhetta, mutta nykyään se on ajankohtaisempi aihe kuin koskaan ennen. Tekoäly on koko ajan ottamassa yhä isompaa asemaa yhteiskunnassa. Tekoälystä tuli kaikilla aloilla tunnettu termi viimeistään vuoden 2022 lopulla kun OpenAI julkaisi chattibotti ChatGPT:n. Tämän jälkeen monet yritykset ja tahot alkoivat kehittämään ja julkaisemaan uusia generatiivisen tekoälyn sovellutuksia. Varsinkin erilaiset kuvageneraattorit ovat kasvattaneet huimasti suosiotaan muutaman viime vuoden aikana. [1]

Kuvageneraattoreiden toimintaan liittyy paljon monimutkaisia käsitteitä ja menetelmiä, joita käydään tässä tutkielmassa läpi. Seuraavaksi käsitellään muutamia tekniikan alan peruskäsitteitä liittyen tekoälyyn. Näiden omaksuminen on tärkeää, jotta haastavampien menetelmien ymmärtäminen on mahdollista, kun siirrytään syvemmälle asiaan.

## 2.1 Generatiivinen vs perinteinen tekoäly

Generatiivinen tekoäly tarkoittaa nimensä mukaisesti tekoälyä, joka luo jotain uutta dataa. Tekstigeneraattorit generoivat uutta tekstiä ja kuvageneraattorit generoivat uusia kuvia. Perinteinen tekoäly ei pysty tuottamaan uutta dataa. Perinteisellä tekoälyllä on yleensä tietty tehtävä, johon se on suunniteltu. Perinteinen tekoäly

pystyy tekemään johtopäätöksiä ja yhteenvetoja suurista datamassoista. Perinteistä tekoälyä käytetään muun muassa konenäössä, kuvien segmentoinnissa sekä asioiden ennustamisessa.

Generatiivinen tekoäly oppii koulutusdatan kuvioita ja rakennetta tarkastelemalla luomaan uutta dataa, joka muistuttaa sille syötettyä koulutusdataa. Generatiivisen tekoälyn generoimaa dataa voidaan käyttää myös muiden algoritmien koulutukseen. Generatiivinen tekoäly nopeuttaa monia laskuprosesseja ja säästää resursseja. Nyt monilla aloilla puhutaan nimenomaan generatiivisen tekoälyn kehityksestä. Jotkut uskovat generatiivisen tekoälyn tuovan ratkaisuja ihmiskunnan isoihin ongelmiin kuten ilmastomuutokseen ja köyhyyteen. Voidaankin sanoa, että generatiivisen tekoälyn aikakausi on alkanut. [2]

## 2.2 Koneoppiminen

Koneoppiminen on tekoälyn osa-alue, joka vastaa tekoälyn oppimisesta. Koneoppimisen algoritmi oppii parantamaan itsenäisesti omaa suoritustaan. Se pystyy tähän löytämällä toistuvia kaavoja sekä ennustamalla. Koneoppimista käytetään monimutkaisissa tehtävissä, joita ihminen ei pysty ratkaisemaan. Koneoppiminen voidaan jakaa kahteen luokkaan: ohjattuun ja ei-ohjattuun. Ohjatussa oppimisessa algoritmille syötetään syötedata ja vastaus. Se pystyy näiden data-vastaus-parien riippuvuuksia tarkastelemalla muodostamaan funktion. Ei-ohjatussa oppimisessa algoritmille annetaan vain syötedata. Sille ei anneta mitään ennalta määrättyjä sääntöjä. Se oppii tekemään omia johtopäätöksiä datasta luokittelemalla annetun datan samankaltaisuuksia tai eroja. Koneoppimisen algoritmit kykenevät suuren datamäärän ja sille määrätyn tehtävän onnistumisarvon avulla muuttamaan omaa toimintaansa. [3] [4]

## 2.3 Neuroverkot

Neuroverkot ovat yksi koneoppimisen algoritmeista. Neuroverkot on suunniteltu jäljittelemään ihmisen aivojen toimintaa. Perseptroni on matemaattinen mallinnus biologisesta neuronista. Perseptroni laskee saatujen syötearvojen painotetun summan. Aktivointifunktio määrittää sen, mikä arvo lähtee tulosteena eteenpäin. Perseptronit ovat yhteydessä toisiinsa synapseilla. Nämä yhteydet joko heikentyvät tai vahvistuvat, mikä mahdollistaa neuroverkon kyvyn oppia. Neuroverkon yksi taso koostuu monista kerroksista perseptroneja. Neuroverkko voi koostua useista eri tasoista. Mitä useammasta tasosta neuroverkko koostuu, sen tehokkaampi neuroverkko on. Useimmat neuroverkot ovat syväoppivia neuroverkkoja. Syväoppivat neuroverkot pystyvät tekemään monimutkaisia asioita sillä ne koostuvat monista erilaisia tehtäviä laskevista neuroverkkotasoista. [3]

## 2.4 Yhteenveto

Tekoälyn algoritmit ja tilastolliset mallit kuuluvat koneoppimisen alueelle. Koneoppimista käytetään joskus synonyyminä tekoälylle, vaikka oikeasti koneoppiminen on tekoälyn osa-alue. Neuroverkot ovat pääasiallinen algoritmi mitä käytetään kuvageneraattoreiden eri vaiheissa. Neuroverkoista erityisesti konvoluutioneuroverkkoja käytetään paljon kuvageneraattoreissa.

# 3 Kuvageneraattorin toiminta

Tässä luvussa tarkastellaan kuvageneraattoreiden toimintaa. Suosituimmat kuvageneraattorit koostuvat pääasiassa samoista rakenneosista, mutta joitain pieniä eroja saattaa olla. Tämä luku perustuu Stable Diffusion-kuvageneraattorin rakenteeseen. Stable Diffusion on yksi suosituimmista kuvageneraattoreista. Stable Diffusion on koulutettu LAION-5B tietoaineistolla [5]. LAION-5B on laajin julkisesti saatavilla oleva tietoaineisto sisältäen yli viisi miljardia kuva-sana-paria [5]. Muiden kuvageneraattoreiden opetusaineistoa ei ole kokonaan saatavilla. Stable Diffusion on avoimen lähdekoodin kuvageneraattori. Stable Diffusionia on luonnollisesti tutkittu eniten, sillä se on oikeastaan ainoa kuvageneraattori, jota on pystytty kunnolla tutkimaan. Stable Diffusionia koskevan lähdemateriaalin löytäminen on helpompaa kuin muiden vastaavien kuvageneraattoreiden.

Kuvageneraattorit koulutetaan generoimaan kuvia valtavalla määrällä opetusdataa. Opetusdata koostuu kuva-sana-pareista. Tämä koulutusaineiston laatu osaltaan määrittää generoitujen kuvien laadun. Kuvageneraattorit toimivat siten, että ne tuottavat niille annettujen tekstimuotoisten syötteiden mukaisen kuvan. Kuvageneraattoreiden toiminta koostuu useista eri osa-alueista. Stable Diffusionin toiminta voidaan jakaa kolmeen eri osa-alueeseen: VAE:seen, U-net:iin ja teksti-enkooderiin.

## 3.1 Variationaalinen autoenkooderi (VAE)

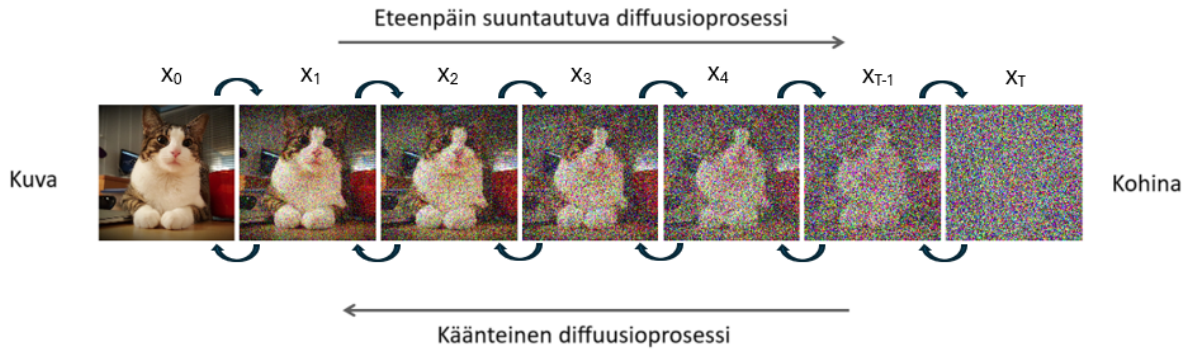
Kuvissa on paljon pikseleitä ja kolme eri väriulottuvuutta. Kuvat vievät paljon tilaa pikselimuodossa, minkä takia kuvien käsittely vie paljon muistitilaa sekä aikaa. Variationaalinen autoenkooderi eli VAE on neuroverkkotekniikka, joka koostuu enkooderista ja dekooderista. VAE muuttaa kuvan matalaulotteiseksi latenttiesitykseksi, näitä kuvien esityksiä kutsutaan latenteiksi. VAE on häviöllinen pakkausmenetelmä, eli joitain tietoja saattaa kadota kun kuva muutetaan pienempään latenttimuotoon. Kaikki pääpiirteet, jotka ovat oleellisia myöhemmissä vaiheissa, kuitenkin säilyy. Kuvien tarkempi käsittely myöhemmissä vaiheissa tapahtuu latenttiavaruudessa. [6] [7]

Enkooderi muuttaa kuvan latenttiesitykseksi. Esimerkiksi 512X512X3 – kokoinen kuva muutetaan 64X64X4 – kokoiseksi latenttiesitykseksi. Kuva vie 48 kertaa vähemmän muistia ja kuvaa on huomattavasti nopeampaa käsitellä. [6] [7]

Dekooderi muuttaa kuvan latenttiesityksen takaisin kuvaksi. Se siis tekee päinvastaisen prosessin kuin enkooderi. Varsinaisen generaattorin käytön aikana tarvitaan vain dekooderia. [6] [7]

## 3.2 Diffuusiomalli

Diffuusiomalli on generatiivinen syväoppiva neuroverkko. Diffuusiomallia käytetään kuvageneraattorin koulutuksessa. Diffuusiomallin tavoitteena on onnistua muodostamaan kuva kohinasta. Diffuusiomalli koostuu kahdesta eri vaiheesta: eteenpäin suuntautuvasta diffuusioprosessista sekä käänteisestä diffuusioprosessista. Diffuusiomalli lisää koulutusdatana annettuihin kuviin kohinaa (Gaussin kohinaa). Diffuusiomallille annetaan myös askeleet, kuinka monta kertaa kohinaa on lisätty kuviin. Tällä tavoin se oppii lisäämään kohinaa tietyn verran. Käänteisessä prosessissa diffuusio-




Kuva 3.1: Havainnekuva diffuusiomallista. Kuvan tekstit on muokattu, alkuperäisen kuvan lähde: [2]

malli oppii poistamaan kohinaa kuvasta. Lopulta diffuusiomalli pystyy muodostamaan kuvan kohinasta. [8] [9] [10] [11]

### Diffuusiomalli vs GAN-verkot

Ennen diffuusiomallin kehitystä generatiiviset kilpailevat verkostot eli GAN-verkot olivat paras tapa generoida kuvia. Nykyään diffuusiomallilla toimivat kuvageneraattorit ovat yleisesti parempia kuvien generoinnissa [12]. GAN-verkoissa on kaksi eri konvoluutioneuroverkkoa, jotka kilpailevat keskenään. Ensimmäisen verkon nimi on generaattori, jonka tehtävä on luoda epäaitoja kuvia. Toisen verkon nimi diskriminaattori, jonka tehtävä on arvioida, onko kuva aito vai tuotettu. [13] [14]

Aluksi molemmat verkot koulutetaan aidoilla sekä epäaidoilla kuvilla. Koulutuksen jälkeen generaattori alkaa luoda epäaitoja kuvia. Koulutuksen jälkeen neuroverkot alkavat kilpailla keskenään. Diskriminaattori arvioi ovatko sille syötetyt kuvat aitoja kuvia vai generaattorin luomia kuvia. Kilpailun voittaja saa pysyä muuttumattomana ja häviäjä joutuu kehittämään itseään. Jos generaattori onnistuu huijaamaan diskriminaattoria, joutuu diskriminaattori kehittämään omaa algoritmiaan paremmaksi. Jos taas diskriminaattori havaitsee generoidun kuvan olevan epäaito, joutuu generaattori kehittämään omaa algoritmiaan. Tätä jatkuu niin kauan, kun-

nes generaattorin tuottamat kuvat menevät poikkeuksetta diskriminaattorin tutkan läpi. GAN-verkko on nyt kykenevä tuottamaan generoimaan kuvia generaattoriverkon avulla. [13] [14]

GAN-verkkojen koulutus on tarkkaa sillä pienikin virhesäätö voi aiheuttaa toimintaongelmia. GAN-verkot eivät myöskään sovellu monimutkaisten kuvien generoimiseen. Niiden oppimisdata on homogeenistä, minkä seurauksena generaattori alkaa helposti tuottamaan toisiaan muistuttavia kuvia [11]. Diffuusiomalleja on helpompi kouluttaa ja ohjata, toisaalta niiden koulutus vie enemmän aikaa. Diffuusiomallit ovat parempia luomaan erilaisia yksityiskohtaisia kuviteltuja skenaarioita. Molemmilla on käyttötarkoituksensa, mutta diffuusiomallit ovat syrjäyttämässä GAN-verkot kuvageneraattoreissa. Diffuusiomallit soveltuvat paremmin kuvageneraattoreihin niiden monipuolisuuden vuoksi. [10] [11] [12]

## 3.3 U-Net

U-Net on konvoluutioneuroverkoista koostuva arkkitehtuuri, joka muistuttaa muodoltaan U-kirjainta. U-Net:in esitti alun perin Ronneberge et al. tieteellisessä julkaisussa "U-Net: Convolutional Networks for Biomedical Image Segmentation"(2015) [15]. U-net keksittiin alun perin biolääketieteellisten kuvien segmentointiin. Stable Diffusionissa U-Net muodostaa kuvan ikään kuin käänteisen diffuusiomallin tapaan. U-Nettiin syötetään kohinaa sisältäviä latentteja sekä haluttu tekstisyöte latenttimuodossa. Ulostulona on kuvien sisältämä kohina. Kun saatu ulostulo vähennetään syötteenä olleesta latentista, saadaan alkuperäinen latentti ilman kohinaa. [15] [16]

## 3.4 Tekstienkooderi

Tekstienkooderin tehtävä on muuttaa sanallinen syöte muotoon, jota U-Net osaa lukea. Aluksi tekstienkooderi muuttaa tekstimuodossa olevan syötteen vektorimuo-

toon (eng. embedding). Lopuksi tekstienkooderi muuttaa vektorimuodossa olevan sanan latenttimuotoon. Tämä menee eteenpäin U-Net:iin, joka generoi halutun kuvan. [11]

### Vektorointi

Sanallinen kuvaus halutusta kuvasta syötetään kehotteina (eng. prompt). Kehotteet koostuvat erillisistä sanoista ja merkeistä. Jokainen yksittäinen sana on siis yksi osa kehotetta. Kehotteet ovat tarkkaan valittuja lauseita, joita käytetään tietyn tyylisten generoitujen kuvien saavuttamiseksi. [17]

Tietokoneet eivät ymmärrä ihmisten kieltä. Kirjoitetulla kielellä annetut kehotteet pitää kääntää koneille ymmärrettävään muotoon. Vektorointi (eng. embedding) tarkoittaa sitä, että sanalliset kehotteet tallennetaan vektorimuodossa. Jokainen kehotteen sana tallennetaan yksittäisenä vektorina. Vektori on pitkä lista lukuja. Lukuja voi olla vaikka tuhansia, jos niin halutaan. Jokainen luku merkitsee jotain arvoa tai piirrettä. Mitä enemmän sanoilla on yhteistä, sen lähempänä toisiaan ne sijaitsevat vektoriavaruudessa. Vektoriavaruus on moniulotteinen tila, jossa kukin ulottuvuus kuvastaa yhtä sanan piirrettä. Esimerkiksi sanat Helsinki ja Tukholma sijaitsisivat suhteellisen lähellä toisiaan, sillä ne ovat kahden pohjoismaan pääkaupunkeja. Sen sijaan Turku ja Tokio sijaitsisivat kaukana toisistaan vektoriavaruudessa, sillä ne sijaitsevat kaukana toisistaan ja niiden väkiluvuissa on suuri ero. Vektoriavaruus antaa tekoälylle mahdollisuuden tarkastella sanojen välisiä suhteita ja poimia sitä kautta lisää tietoa sanojen merkityksistä. [18] [19] [20]

### CLIP

Stable Diffusion käyttää tekstienkooderina OpenAi:n julkaisemaa CLIP-mallia. CLIP on koulutettu netistä löydetyillä, yli 400 miljoonalla kuva-sana-parilla. CLIP

sisältää vektoriavaruudet sekä kuville että kehotteille. CLIP pystyy täten tarkastelemaan kuvien ja kehotteiden välisiä suhteita tehokkaasti. [21]

## 3.5 Yhteenveto

Kuvageneraattorit koulutetaan isolla määrällä kuva-sana-pareja, joiden pohjalta kuvageneraattorit oppivat miltä asiat näyttävät. Kaikki kuvien käsittely, koulutus sekä uusien kuvien generointi tapahtuu latenttiavaruudessa. Nykyään lähes kaikki kuvageneraattorit generoivat kuvansa diffuusiomallia käyttäen.

Kuvageneraattoreiden toimintaan vaikuttaa suuresti koulutuksessa käytetyn datan laatu. Jos data on virheellistä, niin myös lopputuote on virheellinen. Myös koulutusdatan määrä vaikuttaa lopputulokseen. Kuvageneraattorit generoivat parhaiten asioita, joista ne ovat saaneet eniten opetusdataa.

Kuvageneraattoreiden generointiprosessia ohjataan kehotteiden avulla. Taitava kehotteiden osaaja saa todennäköisemmin generoitua halutunlaisen kuvan, kuin sellainen käyttäjä, jolla ei ole kokemusta kuvageneraattoreiden käytöstä.

# 4 Generoiduissa kuvissa ilmenevät virheet

Kuvageneraattorit ovat kehittyneet valtavasti muutamien viime vuosien aikana. Nykyään ne tuottavat erittäin totuudenmukaisia kuvia, joita on vaikea erottaa aidoista reaalimaailman kuvista. Kuvien tarkempi tarkastelu voi kuitenkin kertoa totuuden siitä, onko kuva tehty kuvageneraattoria hyödyntämällä. Generoiduissa kuvissa toistuvat tietyt asiat, joita kuvageneraattorit eivät aina onnistu kuvaamaan.

Kuvageneraattorit tulevat kehittymään entisestään tulevaisuudessa jatkuvasti kehittyvien algoritmien ja datan kasvun myötä. On todennäköistä, että kuvien erottaminen epäaidoista kuvista tulee olemaan tulevaisuudessa lähes mahdotonta.

Seuraavaksi käsitellään joitain yleisimpiä ongelmia, joita kuvageneraattoreilla on kuvien generoimisessa. Nämä ongelmat usein johtavat virheellisiin kuviin. Tässä tutkielmassa ei käsitellä generoituja kuvia, jotka esittävät tietyt ihmisryhmät epäedustavassa valossa. Kuvageneraattorit saattavat generoida sisältöä, joka on rasistista tai seksististä [22]. Tässä tutkielmassa ei käsitellä tätä ongelmaa, sillä se on laaja ja vaikea ongelma.

Taulukkoon 4 on kerätty tässä tutkielmassa esiintyviä tutkimuksia. Tutkimuksissa esitetään ongelma, johon lähdetään etsimään ratkaisua. Ratkaisu oli monessa tapauksessa tutkijoiden itse kehittelemä malli, joka nojautui jo olemassa oleviin tekniikoihin. Monissa tutkimuksissa tarkasteltiin erilaisia ongelmia liittyen ihmisten

| Tutkimuksen tekijä | Julkaisu-vuosi | Ongelma | Mitä esitetään ratkaisuksi/ mitä tutkimuksessa tehtiin |
|---|---|---|---|
| Lu ja muut | 2023 | Käsien generointi | HandRefiner hyödyntäen ControlNet:iä |
| Narasimhaswamy ja muut | 2024 | Käsien generointi | Diffuusiomalliin perustuva HandDiffuser |
| Cao ja muut | 2021 | Ihmisten kehonosien havainnointi | OpenPose, konvoluutio-neuroverkkomalli |
| Qian ja muut | 2017 | Ihmisten asento muuttuu eri kuvakulmissa | Malli perustuen GAN-verkkoon |
| Chen ja muut | 2016 | Yhteisten piirteiden löytäminen valitusta kuvajoukosta | InfoGAN |
| Ma ja muut | 2018 | Ihmisen asennon kuvaaminen oikein | U-Nettiä ja GAN-verkkoja muistuttava arkkitehtuuri |
| Ma ja muut | 2018 | Ihmisen kuvaaminen oikein | Monia eri metodeja yhdistelevä arkkitehtuuri |
| Siarohin ja muut | 2018 | Ihmisen asennon kuvaaminen oikein | Malli GAN-verkkoon |
| Bitton-Guetta ja muut | 2023 | Arkijärjen mallinnus | Datasetti ja sen arviointi kielimallien avulla |
| Oppenlaender ja muut | 2023 | Kehotteiden käyttö | Ihmisten kuvageneraattorien käytön arviointia |

Taulukko 4.1: Tutkimuksia, jotka liittyvät kuvageneraattoreiden ongelmiin

kuvaamiseen. Ihmisten kuvantaminen tuottaa jatkuvasti vaikeuksia kuvageneraattoreille. Kasvojen generointi onnistuu yleensä paremmin kuin vartaloiden ja raajojen generointi. Ihmisten virheellinen kuvantaminen on yksi yleisimmistä virheistä, jonka voi havaita generoiduista kuvista.

Generaattorit oppivat kaksiulotteisten kuvien kautta generoimaan kuvia, joten generaattoreilta puuttuu kyky hahmottaa asioita kolmiulotteisina. Tämän seurauksena generaattorit saattavat generoida oudon muotoisia asioita.

Kuvageneraattorien koulutusmenetelmän takia niiltä puuttuu ihmisenkaltainen ymmärrys maailmasta. Generoiduissa kuvissa ilmeneekin välillä järjenvastaisia asioita.

Generaattoreiden tuotokset riippuvat hyvin vahvasti siitä, millaisen kehotteen käyttäjä antaa generaattorille. Kehotteiden käytön osaaminen on tärkeää käyttäjille, jotta he pystyisivät generoimaan halutunlaisia kuvia.

## 4.1 Ihmiskasvot, raajat ja laskettavat asiat

Kuvageneraattoreilla on monenlaisia ongelmia ihmisten generoimisessa. Ensisilmäyksellä generoidussa kuvassa esiintyvä ihminen saattaa vaikuttaa aidolta. Lähempi kuvan tarkastelu voi paljastaa totuuden. Kuvassa 4.1 on esitelty erilaisia virheitä.

Generaattorit osaavat kuvata ihmisten kasvot usein melko hyvin. GAN-verkoilla oli vaikeuksia kuvata ihmisten pupillit oikein, mutta nykyään generaattorit osaavat kuvata silmät pääsääntöisesti hyvin, sillä ne ovat yksi naamojen pääpiirteistä. Yksityiskohtaisemmat asiat kuten korvat voivat paljastaa generoidun kasvon. Korvan sisälehden kuvio voi olla omituisen muotoinen. Ihmisten hiukset ovat myös erittäin vaikeita kuvata luonnollisen näköisinä. Hiusten rakenne on yleensä epäluonnollisen sileä ja hiustuppoja saattaa kasvaa oudoista paikoista. [23]

Ihmiskasvon symmetrian ymmärtäminen ei ole generaattoreille helppoa. Generoiduissa naamoissa korvat saattavat sijaita eri tasolla tai ne voivat olla erikokoiset.

Toisesta korvasta voi puuttua korvakoru tai korut ovat eri paria. Asusteista erityisesti silmälasien kuvaaminen on hankalaa. Ne näyttävät hieman erilaisilta eri kuvakulmien mukaan. Silmälasit saattavat olla epäsymmetriset. Silmälasit voivat sulautua ihoon tai osa sangoista puuttua kokonaan. Silmälasit eivät välttämättä osu oikeaan kohtaan kasvoja, jolloin silmät näyttävät omituisilta lasien takaa. [23]

Yksi helpoimmista keinoista tunnistaa generoidut kasvot on niiden tekstuurin puute. Generoitujen ihmiskasvojen iho näyttää liian sileältä. Ihohuokoset saattavat puuttua kokonaan. Kasvoryppyjä ei ole ollenkaan tai niitä on liian vähän. [23]

Ihmisten kasvojen kuvaaminen on generaattoreille verrattain helppoa verrattuna ihmisraajojen kuvaamiseen. Generoitujen ihmiskehojen raajat taipuvat epänormaalisti. Raajat saattavat alkaa tyhjästä ja roikkua kehon ulkopuolella. Käsiä tai jalkoja saattaa olla kolme tai useampi kahden sijasta. Raajat saattavat olla epäsuhdassa muuhun kehoon nähden ja ne voivat olla vääntyneenä epäluonnollisiin asentoihin. [23]

Raajoista kädet ovat vaikeinta kuvata. Kuvageneraattorit epäonnistuvat sormien kuvaamisessa systemaattisesti. Sormia on liikaa tai ne ovat sulautuneet yhteen oudon näköisesti. Sormet taipuvat vääristä paikoista ja vääntyvät erikoisiin asentoihin. Kun kädet pitelevät asioita, käsien ja sormien kuvaaminen vaikeutuu entisestään. Sormet sojottavat eri suuntiin kuin mihin niiden pitäisi. Sormet saattavat sulautua pideltävään esineeseen. Myös pideltävä esine saattaa muuttaa muotoaan tai se on epäloogisessa asennossa. Kädet ovat erittäin yksityiskohtaiset. Sormissa on monta pientä niveltä, jotka taipuvat moneen eri kulmaan. Nämä asiat tekevät sormista kuvageneraattoreille erittäin vaikeat hahmottaa. Sormien kuvantamisen parantamisesta on tehty muutamia tutkimuksia. [24] [25]

Ihmisten kehojen tunnistaminen on ollut koneille aina hankalaa. Cao et al. käsittelivät julkaisussaan "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields"(2018) [26] konenäön soveltamista ihmisten eri kehonosien tun-

Kuva 4.1: Virheitä ihmisten kuvantamisessa: silmät, tekstuurin puute, lasit, sormet, hevosen ja ihmisen keho, perspektiivi. Kuvien lähde: [23]

nistamiseen. Tutkimuksessa konvoluutioneuroverkoista koostuva arkkitehtuuri tunnisti kuvista ihmiskehojen pääpiirteitä ja piirsi vektorikarttoja niiden hahmottamiseksi. OpenPose tunnisti hyvin ihmiskehon, jalan, käsien ja kasvojen avainkohtia. Se myös teki virheitä, joissa se tunnisti vain osan kehonosista tai tunnisti patsaan ihmiseksi. Tutkimus kuitenkin osoitti, että ihmisten tunnistaminen on kriittistä koneiden toiminnan kehityksen kannalta. [26]

## 4.2 Kolmiulotteisuuden ymmärtäminen

Kuvageneraattorit oppivat mallintamaan objektit vain kuvien kautta. Kaikki objektit näyttävät erilaisilta eri kuvakulmista katsottuna. Tarkasteltavien asioiden koko, väri ja muoto muuttuvat kuvakulman muuttuessa. Ihmisten raajojen väliset suhteet näyttävät erilaisilta eri kuvakulmissa. Ihmisten ulkomuotoon vaikuttaa myös ihmisen asento. Sormia näkyy eri määrä käden eri asennoissa, mikä vaikeuttaa kuvageneraattoria oppimasta sormien oikeaa lukumäärää. [27] [28] [24]

Asioiden ymmärtäminen kolmiulotteisina objekteina on tärkeää, jotta onnistuneiden kaksiulotteisten kuvien tuottaminen onnistuu. Asioiden oppiminen eri kuvakulmista auttaa kuvageneraattoreita ymmärtämään, että asiat näyttävät erilaisilta eri kuvakulmista. Tutkimusta on tehty syöttämällä GAN-verkoille kuvia samoista ihmisistä eri asennoissa. Tutkimusta on tehty myös syöttämällä GAN-verkoille tikku-ukkomaisia havainnekuvia halutuista asennoista. Tutkimusta on myös tehty yhdistelemällä eri asentoja kuvaavia kuvia sekä pääpiirteitä havainnollistavia malleja asennoista. Näillä tutkimuksilla on yritetty löytää keinoja saada ihmiset näyttämään oikeanlaisilta tietyissä asennoissa ja kuvakulmissa. Tutkimuksia voi myös soveltaa muihinkin objekteihin kuin ihmisiin. [27] [28] [29] [30].

Tutkimus GAN-verkon sovelluksesta nimeltään "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets", Chen et al. (2016) [31] tarjoaa varteenotettavan tavan lisätä tietoa generaattoreihin.

InfoGAN maksimoi yhteisen tiedon pienen latenttiesitysten osajoukon sekä havainnon välillä. InfoGAN pystyy etsimään näitä yhtymäkohtia ei-ohjatulla menetelmällä. Tutkimuksessa InfoGAN erotti onnistuneesti eri kirjoitustyylit annetusta numerojoukosta. InfoGAN pystyi tunnistamaan samanlaiset asennot, kuvakulmat, valotukset sekä pään leveyden 3D-mallinnutetuista ihmisten päistä. [31]

InfoGAN pystyi tarkastelemaan ja yhdistämään erilaisia visuaalisia piirteitä CelebA-kasvotietojoukosta. Näihin piirteisiin kuului hiustyylit, silmälasit sekä tunteet. Nämä tutkimukset osoittavat, että InfoGAN pystyy oppimaan ja etsimään tiettyjä piirteitä kuvista. InfoGAN:ia voisi mahdollisesti hyödyntää kuvageneraattoreiden ongelmiin liittyen oikeanlaisten asentojen kuvaamiseen. InfoGAN:ia voisi myös soveltaa haluttujen piirteiden oppimiseen. [31]

## 4.3 Arkijärki

Arkijärjen omaavan tekoälyn mallintaminen on nykypäivänä yksi tärkeimmistä osa-alueista tekoälyn saralla [32]. Arkijärki on ihmisille luontainen osa elämää. Se perustuu ihmisten kokemuksiin ja tietoon maailmasta. Se on tietoa ja toimintaa, jota ei tarvitse sen kummemmin miettiä. Arkijärjen konkreettista mallinnusta kuvageneraattoreille ei olla vielä tutkittu. Tutkimusta on teorian tasolla tehty hieman. Tutkimuksessa havaittiin, että kuvageneraattoreilta puuttuu arkijärjen ymmärrys kuvia luodessa.

Joskus voidaan haluta generoida kuvia, joissa tapahtuu asioita, joita ei voi oikeasti tapahtua. Toisaalta, jos halutaan luoda realistisia kuvia, voi tällaiset kuvat silloin olla virheellisiä. Kuvageneraattoreiden olisi hyvä osata kertoa milloin ne generoivat järjettömiä kuvia. Arkijärjen vastaiset kuvat eivät sinänsä ole ongelma, mutta arkijärjen puute on.

Ryhmä tutkijoita loivat arkijärjen vastaisen datasetin tutkimuksessa "WHOOPS! Breaking Common Sense", Bitton-Guetta et al. (2023) [33]. Tutkijat generoivat datasettiin kuvia, jotka esittivät skenaarioita, joita ei voi tapahtua oikeassa elämässä 4.2. Yksi kuva oli Albert Einstein pitelemässä matkapuhelinta, mikä on järjenvastaista, sillä matkapuhelimia ei ollut Einsteinin aikana. Kuva luotiin niin, että aluksi luotiin kuva, jossa oli kaksi elementtiä, jotka voivat esiintyä yhdessä, Einstein ja vihko. Tämän jälkeen tehtiin uusi kuva, jossa vihko korvattiin matkapuhelimella. Kaikki kuvat luotiin tällä tavalla. Kuvissa on minimaalinen muutos, jotta epätavallisuuden huomaaminen olisi mahdollisimman vaikeaa. [33]

Tutkijat tekivät erilaisia kokeita eri tekoälyn kielimalleilla. Aluksi WHOOPS!-datasetti annettiin ihmisten arvioitavaksi. Heidän piti keksiä kuvaavia lauseita mitä kuvassa tapahtuu, sekä arvioida onko kuva outo vai ei. Näitä ihmisten tekemiä päätelmiä käytettiin metodien koulutuksessa. Yhdessä kokeessa datasetin kuvia syötettiin erilaisille kielimalleille. Kielimallien piti arvioida esiintyykö kuvassa jotain taval-

Kuva 4.2: Tutkimuksessa [33] generoituja arkijärjenvastaisia kuvia: Einstein pitelemässä puhelinta ja kynttilä palaa suljetussa tilassa.

lisesta poikkeavaa, vai onko kuva normaali. Kielimallien piti myös kertoa, mikä teki kuvasta oudon. Parhaimmat mallit pystyivät lajittelemaan oudot ja normaalit kuvat 78 prosentin tarkkuudella. Kielimalleilla oli pääsy ihmisten tekemiin selitteisiin ja paras malli sai outojen kuvien selitteet oikein vain 68-prosenttisesti. Ihmisten onnistumisprosentti oli 95%. Kielimallit jäivät siis huomattavasti jälkeen ihmisarvioijista. [33]

Toisessa kokeessa kielimalleille annettiin kolmen kuvan setti arvioitavaksi. Setti sisälsi aidon kuvan, normaalin generoidun kuvan, sekä oudon generoidun kuvan. Setin kuvat kuvasivat samaa tapahtumaa, mutta oudossa kuvassa oli yksi epänormaali asia. Kielimallien piti tuottaa selitteet kuville. Mallit tuottivat oikeat selitteet normaaleille kuville, riippumatta siitä olivatko ne generoituja vaiko aitoja, 89% tarkkuudella. Outojen kuvien selitteet menivät oikein vain 49% tapauksissa. Malleilla oli siis huomattavia vaikeuksia tunnistaa oudot kuvat epänormaaleiksi. Mielenkiintoista oli myös, että mallit eivät osanneet kuvata aikakausien välisiä ristiriitoja, mutta epänormaalien ympäristöjen kuvailu onnistui hyvin. [33]

WHOOPS!-datasettiin liittyvä tutkimus osoitti, että tekoälyn arkijärjen mallinnus on vielä kaukana toimivasta. Tutkimus kuitenkin antoi työkaluja parempien tutkimuksien tekemiseen ja arkijärjen mallinnukseen tulevaisuudessa. Tutkimus osoit-

ti, että ratkaisu arkijärjen mallintamiseen kuvageneraattoreissa voisi olla kielimalli, joka osaa tunnistaa outoudet. [33]

Arkijärjen tutkimusta on tehty enemmän tekstin generoinnin saralla. Arkijärkeä on koitettu soveltaa antamalla kielimalleille erilaisia tehtäviä tekstin generointiin liittyen. [34] [35] Arkijärkeä tekstin generoinnin parantamiseen on myös tutkittu kuvien kautta. Tutkimuksessa "Visual Grounding for Enhancing Commonsense in Text Generation Models", Feng et al. (2022) [36] kielimalleille annettiin kuvia, joiden pohjalta niiden tuli generoida tekstiä. Tutkijat uskoivat, että arkijärjen perusta voisi piillä visuaalisessa kyvyssä. [36]

## 4.4 Kehotteiden tehokas käyttö

Generoitujen kuvien laatu riippuu myös annetuista kehotteista. Kehotteet ohjaavat generointiprosessia ja kehotteita joudutaan monesti muotoilemaan uudestaan. Oikeiden kehotteiden keksiminen voi olla vaikeaa ja se onkin nykyään taito, jonka oppiminen vie aikaa. Aina ei ole edes varmaa millaisista kehotteista saadaan haluttu kuva. Vääränlaisten kehotteiden syöttäminen voi helposti johtaa virheisiin kuvissa. Kehotesuunnittelu (eng. prompt engineering) on nykyään yksi isoista tutkimussuuntauksista liittyen tekstigeneraattoreihin. Kuvageneraattorit ovat vielä verrattain uusia eikä tutkimusta olla vielä ehditty tekemään yhtä paljoa. Kehotteiden ymmärtäminen ja tutkiminen on tärkeää, jotta tiedetään miten kuvageneraattorit reagoivat mihinkin kehotteisiin. Tämän tiedon pohjalta voidaan kehittää työkaluja ja ohjeita, jotka auttavat käyttäjiä generoimaan parempia kuvia. [37] [38] [39]

Kehote on sanallinen kuvaus siitä, millaisen kuvan kuvageneraattorin käyttäjä haluaa luoda. Laadukkaimmat kuvat kuitenkin saa sellaisilla kehotteilla, joita ihmiset eivät normaalisti käyttäisi kuvien kuvailuun, esimerkkinä kuva 4.3. Kehotteet sisältävät usein kuvan kuvailun lisäksi myös erilaisia avainsanoja, joita lisätään kehotteeseen kuvan muokkaamiseksi. Avainsanat sisältävät muun muassa tyylin ku-

Kuva 4.3: Esimerkki kuvasta jonka generoimiseen on käytetty seuraavaa kehotetta "A beautiful painting of a singular lighthouse, shining its light across a tumultuous sea of blood by greg rutkowski and thomas kinkade, Trending on artstation."Lähde: [39]

vauksia, laadun parantajia, sisällön tarkempia kuvauksia tai taikatermejä. Avainsanoja kutsutaan kehotteen muokkaajiksi. Tyylin muokkaajia ovat esimerkiksi sanat öljyvärimaalaus, surrealismi tai jonkun taiteilijan tietty tyyli (eng. by [artist]). Laadun parantajiin kuuluvat sanat eeppinen, suosittu artstationissa tai mestariteos. Tarkennusta kuvan sisältöön voi saada toisteisella sanoilla esimerkiksi kehotteella "erittäin erittäin erittäin erittäin kaunis maisema"saa hienomman kuvan kuin kehotteella "erittäin kaunis maisema". Taikatermit ovat satunnaisia termejä, joilla ei ole suoraan mitään tekemistä muun kuvan kanssa. Taikatermit tuovat satunnaisuutta generoitaviin kuviin, minkä tavoitteena on monipuolistaa generoitujen kuvien joukkoa. Taikatermit kuvaavat jotain ei-nähtävää ominaisuutta, esimerkiksi "ruoki sielua". [17] Tehokkaiden kehotteiden muokkaajien löytäminen on pitkä prosessi, joka tapahtuu yrityksen ja erehdyksen kautta. Kehotesuunnittelijat etsivät tehokkaita kehotteita, joita kuvageneraattoreiden arkikäyttäjät voivat käyttää. Netti sisältää monenlaisia opuksia kehotteiden käyttöön, sekä tietokantoja, jotka sisältävät valmiita kehotteita käytettäväksi. [17] [37] [38] [39]

Kehotteiden käytön ongelmana on vain, että normikäyttäjät eivät ymmärrä mitä tietyt kehotteet tekevät. Monet eivät edes tiedä kehotteiden muokkaajien olemassaolosta. Tutkimuksessa "Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering", Oppenlaender et al. (2023) [38] tutkittiin ihmisten luontaisia kehotesuunnittelutaitoja. Yhdessä tutkimuksessa 125 kokelasta pyydettiin luomaan kehotteita kuvageneraattorille. Heidän tehtävänään oli luoda mahdollisimman hieno ja laadukas taidekuva annetusta aiheesta. Vähän alle puolella kokelaista oli koulutus taidealalta ja 30%:lla oli jotain kokemusta kuvageneraattoreiden käytöstä. Suurin osa kokelaista osasi kuvailla haluttua kuvaa monimuotoisella ja rikkaalla kielellä. Vain yksi kokelas osasi käyttää tarkoituksellisesti suosittuja kehotteen muokkaajia. Tämän jälkeen kokelaita pyydettiin parantamaan heidän kehotteittaan. Selvä vähemmistö onnistui vain hieman parantamaan kuviensa laatua. Tutkimus osoittaa, että kuvageneraattoreiden tehokas käyttö vaatii siis asiantuntevuutta. Kehotesuunnittelu on taito, jota ei muutamassa päivässä saavuteta. Kehotesuunnittelun ongelmakohdat herättävät myös laajempia kysymyksiä ihmisen ja tekoälyn välisestä kommunikaatiosta, ja miten sen saisi sujuvammaksi. [38] [39]

# 5 Päätelmät

Monet tässä tutkielmassa esitellyistä tutkimuksista keskittyvät ihmisten asentojen oikeanlaiseen kuvantamiseen. Kuvageneraattoreilla on vaikeuksia kuvata ihmisiä halutuissa asennoissa. Tämä kertoo siitä, että ihmisen keho on kokonaisuutena monimutkainen. Kuvageneraattorit eivät ymmärrä, että asiat näyttävät erilaisilta eri kuvakulmista tarkasteltuna. Jotta kuvageneraattorit osaisivat kuvata ihmisiä paremmin, tulee opetusdatan ja kuvageneraattoreiden toiminnan tukea kolmiulotteisuuden hahmottamista.

Kuvageneraattoreilla on ongelmia ihmisten kuvantamisessa. Ihmisten kasvoja on käytetty opetusdatana moninkertaisesti enemmän kuin esimerkiksi ihmisten sormia. Siksi kuvageneraattorit eivät osaa generoida sormia yhtä hyvin kuin kasvoja. Opetusdatan laadun, sekä määrän lisääminen on yksinkertaisin keino parantaa kuvageneraattoreiden toimintaa.

Arkijärjen lisääminen tavalla, tai toisella kuvageneraattoreihin voisi tuoda ymmärrystä siitä, miten ihminen toimii ja miten ihmisen raajat taipuvat ja liikkuvat. Tällöin ihmisten kuvantaminen luonnollisesti olisi paljon helpompaa. Kuvageneraattoreiden tulisi ymmärtää paremmin ihmistä kokonaisuutena, joka on parhaiten mahdollista arkijärjen kautta.

Huonojen kehotteiden käyttäminen altistaa virheellisten kuvien generoimiseen, kun kuva ei vastaakaan syötettyä kehotetta. Kehotesuunnittelu tutkii parempien

kehotteiden etsimistä. Tutkimuksien mukaan normaalit käyttäjät eivät osaa välttämättä hyödyntää kehotteita tehokkaasti.

Huomattavaa on, että iso osa tutkimuksista on tehty GAN-verkkoja tutkimalla. Tutkimuksista vain neljä on julkaistu sen jälkeen kun diffuusiomallilla toimivat kuvageneraattorit julkaistiin. Tämä kertoo, siitä että diffuusiomallilla toimivat kuvageneraattorit ovat niin uusia, että kattavia tutkimuksia ei olla ehditty vielä tekemään. Käydessäni tutkimuksia läpi huomasin, että GAN-verkoilla on ollut samanlaisia ongelmia kuin mitä on havaittu myös diffuusiomalleilla, esimerkkinä ihmisten kuvaaminen. Periaatteessa GAN-verkoilla tehtyjä tutkimuksia voidaan käyttää hyvänä pohjana, kun pohditaan diffuusiomallilla toimivien kuvageneraattoreiden ongelmia.

Monessa tutkimuksessa myös todettiin, että esiteltyä tutkimusta voidaan käyttää pohjana lisätutkimuksien tekemiselle. Toisin sanoen näissä tutkimuksissa löydetyt ratkaisut ongelmiin ovat vielä keskeneräisiä. Eli vielä ei ole olemassa mitään yhtä hyvää ratkaisua, joka poistaisi jonkun tietyn ongelman kokonaan.

# 6 Yhteenveto

Kuvageneraattoreiden generoimien kuvien virheitä tarkastelevia tutkimuksia oli vaikea löytää. Kuvageneraattorit ovat aiheena vielä uusi asia. Kuvageneraattoreiden suosio on kasvanut räjähdysmäisesti parin viime vuoden aikana. Tieteellisiä artikkeleita ei täten löydy vielä kovin paljoa. Keskustelua ja uutisia kuvageneraattoreiden tuottamista kuvista löytyy paljon, mutta tieteellisiä tutkimuksia ollaan vasta tekemässä. Löydetyt tieteelliset tutkimukset olivat vielä teorian tasolla. Tutkimuksissa esitetyt menetelmät toimivat hyvin tutkimuksissa, mutta lisää kokeita on tehtävä, jotta menetelmiä voisi soveltaa kuvageneraattoreihin. Joissain tutkimuksissa todettiin, että menetelmät toimivat puutteellisesti tehden virheitä. Muutaman vuoden päästä tieteellisiä artikkeleita aiheesta on paljon enemmän saatavilla. Kuvien generoinnin ongelmakohtia tarkastelevan tutkielman tekeminen on varmasti helpompaa muutaman vuoden päästä.

Yleisesi ottaen opetusdatan laadun ja määrän lisääminen parantaa generoitujen kuvien laatua. Toisaalta kuvageneraattoreille on jo syötetty valtavat määrät dataa. Kuvageneraattorit lähestyvät pikkuhiljaa kriittistä pistettä, missä opetusdatan parantaminen ei enää takaa parempia tuloksia.

Ihmisten kuvantamisessa ilmenevät virheet johtuvat siitä, etteivät kuvageneraattorit ymmärrä miten ihmiset toimivat. Kuvageneraattorit oppivat kaiken ympäröivästä maailmasta vain kuvia tarkastelemalla. Kuvageneraattorit eivät hahmota maailmaa samalla tavalla kuin ihmiset. Ne eivät tiedä mitään maailmassa vallitsevista

säännöistä. Kuvageneraattoreiden oppimista voi verrata ihmiseen, joka asuu koko elämänsä taidemuseossa ja oppii ympäröivästä maailmasta vain taideteoksia tarkastelemalla. Kaikki mitä tämä ihminen tietää maailmasta perustuu vain siihen, mitä hän näkee taidemuseon seinillä. Pelkkien maalauksien tarkastelu ei anna kattavaa kuvaa maailman toiminnasta.

Arkijärjen soveltaminen tekoälyn kaikkiin sovelluksiin tulee tulvaisuudessa olemaan iso tutkimusala. Arkijärjen hyödyntäminen kuvageneraattoreissa antaisi paljon lisää sellaista tietoa, joka hyödyttäisi kuvien generoimista huomattavasti. Uskon, että arkijärjen soveltaminen voisi olla ratkaisu moniin tekoälyn puutteisiin. Kattavan arkijärjen soveltaminen tekoälyyn ja kuvageneraattoreihin on tosin erittäin vaikeaa, jollei mahdotonta.

# Lähdeluettelo


[1] Y. Cao, S. Li, Y. Liu et al., *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT*, 2023. arXiv: 2303.04226 [cs.AI].

[2] C. Meijer ja L. Y. Chen, *The Rise of Diffusion Models in Time-Series Forecasting*, 2024. arXiv: 2401.03006 [cs.LG].

[3] I. Goodfellow, Y. Bengio ja A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[4] Y. LeCun, Y. Bengio ja G. Hinton, "Deep learning", *nature*, vol. 521, nro 7553, s. 436–444, 2015.

[5] C. Schuhmann, R. Beaumont, R. Vencu et al., "LAION-5B: An open large-scale dataset for training next generation image-text models", teoksessa *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho ja A. Oh, toim., vol. 35, Curran Associates, Inc., 2022, s. 25 278–25 294. url: https://proceedings.neurips.cc/paper_files/paper/2022/file/a1859debfb3b59d094f3504d5ebb6c25-Paper-Datasets_and_Benchmarks.pdf.

[6] D. P. Kingma ja M. Welling, *Auto-Encoding Variational Bayes*, 2022. arXiv: 1312.6114 [stat.ML].

[7] A. van den Oord, O. Vinyals ja K. Kavukcuoglu, *Neural Discrete Representation Learning*, 2018. arXiv: 1711.00937 [cs.LG].

[8] F.-A. Croitoru, V. Hondru, R. T. Ionescu ja M. Shah, "Diffusion Models in Vision: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, nro 9, s. 10 850–10 869, 2023. DOI: 10.1109/TPAMI.2023.3261988.

[9] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan ja S. Ganguli, *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*, 2015. arXiv: 1503.03585 [cs.LG].

[10] C. Saharia, W. Chan, S. Saxena et al., *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*, 2022. arXiv: 2205.11487 [cs.CV].

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser ja B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models", teoksessa *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, s. 10 674–10 685. DOI: 10.1109/CVPR52688.2022.01042.

[12] P. Dhariwal ja A. Nichol, "Diffusion Models Beat GANs on Image Synthesis", teoksessa *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang ja J. W. Vaughan, toim., vol. 34, Curran Associates, Inc., 2021, s. 8780–8794. url: https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial networks", *Commun. ACM*, vol. 63, nro 11, s. 139–144, lokakuu 2020, ISSN: 0001-0782. DOI: 10.1145/3422622. url: https://doi.org/10.1145/3422622.

[14] A. Brock, J. Donahue ja K. Simonyan, *Large Scale GAN Training for High Fidelity Natural Image Synthesis*, 2019. arXiv: 1809.11096 [cs.LG].

[15] O. Ronneberger, P. Fischer ja T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, 2015. arXiv: 1505.04597 [cs.CV].

[16] P. Esser, E. Sutter ja B. Ommer, "A Variational U-Net for Conditional Appearance and Shape Generation", teoksessa *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, kesäkuu 2018.

[17] J. Oppenlaender, "A taxonomy of prompt modifiers for text-to-image generation", *Behaviour & Information Technology*, s. 1–14, marraskuu 2023, ISSN: 1362-3001. DOI: 10.1080/0144929x.2023.2286532. url: http://dx.doi.org/10.1080/0144929X.2023.2286532.

[18] A. Frome, G. S. Corrado, J. Shlens et al., "DeViSE: A Deep Visual-Semantic Embedding Model", teoksessa *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani ja K. Weinberger, toim., vol. 26, Curran Associates, Inc., 2013. url: https://proceedings.neurips.cc/paper_files/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

[19] P. D. Turney ja P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics", *Journal of Artificial Intelligence Research*, vol. 37, s. 141–188, helmikuu 2010, ISSN: 1076-9757. DOI: 10.1613/jair.2934. url: http://dx.doi.org/10.1613/jair.2934.

[20] E. H. Huang, R. Socher, C. D. Manning ja A. Y. Ng, "Improving word representations via global context and multiple word prototypes", teoksessa *Proceedings of the 50th annual meeting of the association for computational linguistics (Volume 1: Long papers)*, 2012, s. 873–882.

[21] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu ja M. Chen, "Hierarchical text-conditional image generation with clip latents", *arXiv preprint arXiv:2204.06125*, vol. 1, nro 2, s. 3, 2022.

[22] Ananya, *AI image generators often give racist and sexist results: can they be fixed?*, 2024.

[23] A. Borji, *Qualitative Failures of Image Generation Models and Their Application in Detecting Deepfakes*, 2024. arXiv: 2304.06470 [cs.CV].

[24] W. Lu, Y. Xu, J. Zhang, C. Wang ja D. Tao, "HandRefiner: Refining Malformed Hands in Generated Images by Diffusion-based Conditional Inpainting", *ArXiv*, vol. abs/2311.17957, 2023. url: https://api.semanticscholar.org/CorpusID:265506151.

[25] S. Narasimhaswamy, U. Bhattacharya, X. Chen, I. Dasgupta, S. Mitra ja M. Hoai, "HanDiffuser: Text-to-Image Generation With Realistic Hand Appearances", *ArXiv*, vol. abs/2403.01693, 2024. url: https://api.semanticscholar.org/CorpusID:268247775.

[26] Z. Cao, G. Hidalgo, T. Simon, S. Wei ja Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 43, nro 01, s. 172–186, tammikuu 2021, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2019.2929257.

[27] X. Qian, Y. Fu, T. Xiang et al., *Pose-Normalized Image Generation for Person Re-identification*, 2018. arXiv: 1712.02225 [cs.CV].

[28] A. Siarohin, E. Sangineto, S. Lathuiliere ja N. Sebe, *Deformable GANs for Pose-based Human Image Generation*, 2018. arXiv: 1801.00055 [cs.CV].

[29] L. Ma, Q. Sun, S. Georgoulis, L. V. Gool, B. Schiele ja M. Fritz, *Disentangled Person Image Generation*, 2018. arXiv: 1712.02621 [cs.CV].

[30] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars ja L. V. Gool, *Pose Guided Person Image Generation*, 2018. arXiv: 1705.09368 [cs.CV].

[31] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever ja P. Abbeel, *InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets*, 2016. arXiv: 1606.03657 [cs.LG].

[32] E. Davis ja G. Marcus, "Commonsense reasoning and commonsense knowledge in artificial intelligence", *Commun. ACM*, vol. 58, nro 9, s. 92–103, elokuu 2015, ISSN: 0001-0782. DOI: 10.1145/2701413. url: https://doi.org/10.1145/2701413.

[33] N. Bitton-Guetta, Y. Bitton, J. Hessel et al., *Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images*, 2023. arXiv: 2303.07274 [cs.CV].

[34] B. Y. Lin, W. Zhou, M. Shen et al., *CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning*, 2020. arXiv: 1911.03705 [cs.CL].

[35] A. Talmor, J. Herzig, N. Lourie ja J. Berant, "CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge", teoksessa *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran ja T. Solorio, toim., Minneapolis, Minnesota: Association for Computational Linguistics, kesäkuu 2019, s. 4149–4158. DOI: 10.18653/v1/N19-1421. url: https://aclanthology.org/N19-1421.

[36] S. Y. Feng, K. Lu, Z. Tao et al., *Retrieve, Caption, Generate: Visual Grounding for Enhancing Commonsense in Text Generation Models*, 2022. arXiv: 2109.03892 [cs.CL].

[37] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover ja D. H. Chau, *DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models*, 2023. arXiv: 2210.14896 [cs.CV].

[38] J. Oppenlaender, R. Linder ja J. Silvennoinen, *Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering*, 2023. arXiv: 2303.13534 [cs.HC].

[39] J. Oppenlaender, "A taxonomy of prompt modifiers for text-to-image generation", *Behaviour & Information Technology*, vol. 0, nro 0, s. 1–14, 2023. DOI: 10.1080/0144929X.2023.2286532.