Tampereen yliopisto

Teemu Haug

GENERATIIVISEN TEKOÄLYN VISUAALISET KÄYTTÖKOHTEET

Kandidaattitutkielma
Informaatioteknologian ja viestinnän tiedekunta
Tarkastaja: Tapio Elomaa
Joulukuu 2023

TIIVISTELMÄ

Teemu Haug: Generatiivisen tekoälyn visuaaliset käyttökohteet
Kandidaattitutkielma
Tampereen yliopisto
Informaatioteknologian ja viestinnän tiedekunta
Tietojenkäsittelytieteiden tutkinto-ohjelma
Joulukuu 2023

Generoivat tekoälymallit ovat saaneet paljon huomiota viime vuosien aikana niiden laajan soveltuvuuden ja alan nopean kehityksen ansiosta, mutta mediassa harvoin kerrotaan tarkemmin käytetyistä tekoälymalleista. Työ on toteutettu kirjallisuuskatsauksena, ja sen tavoitteena on selvittää, mitä erilaisia visuaaliseen generointiin kykeneviä tekoälymalleja on käytetty ja miten ne toimivat, sekä perehtyä mallien mahdollisiin käyttötapoihin ja sovelluksiin. Oleellisimpia mallityyppejä, joihin työssä keskitytään, on kolme: generatiiviset kilpailevat verkostot, autoenkooderit ja diffuusiomallit. Etenkin näiden mallityyppien tärkeimpiin alatyyppeihin, kuten variaatioautoenkoodereihin sekä näitä hyödyntäviin latentteihin diffuusiomalleihin, perehdytään tarkemmin.

Työ on jaettu kahteen osaan. Ensimmäisessä osassa esitellään keskeisten mallien arkkitehtuuria ja toimintaa, ja toisessa tarkastellaan esiteltyjen mallien visuaalisia käyttötapoja ja -tarkoituksia niin yksityishenkilöiden kuin eri toimialojenkin näkökulmasta. Visuaalisen generoinnin johtavat mallit pohjautuvat yleensä diffuusioon, mutta myös erikoistuneita GAN-malleja käytetään edelleen paikoittain. Etenkin latentit diffuusiomallit ovat saaneet suurta suosiota generointien laadun, monipuolisuuden, helppokäyttöisyyden sekä tehokkuuden ansiosta.

Työssä käsitellään erilaisia kuvien ja videoiden tuottamis- ja muokkausmenetelmiä. Tuottamisella tarkoitetaan uuden materiaalin generointia erilaisten syötteiden avulla, kun taas muokkaamisella tarkoitetaan jo olemassa olevan median muuntamista joko syötettä tai pelkkää mallia hyödyntäen. Visuaaliseen generointiin kykeneviä malleja voidaan hyödyntää monella alalla. Esimerkiksi lääketieteessä CT- ja MR-kuvattuja kuvia voidaan tarkentaa superresoluutiolla, elokuvatuotannossa hahmojen luonti voidaan toteuttaa kuvasynteesillä, ja yksityishenkilöt voivat helposti värittää isovanhempiensa mustavalkokuvia väritykseen erikoistuvilla malleilla.

Alan nopea kehitys ja uusien käyttötarkoitusten suuri määrä viittaavat tekoälyn käytännöllisen tärkeyden kasvamiseen, ja tekoäly tulee varmasti mullistamaan monia aloja tulevaisuudessa tehostamalla ja automatisoimalla manuaalista työtä vaativia prosesseja, kuten se on jo tehnyt eri toimialoilla viime vuosina. Täydellisiä, ihmisiä jatkuvasti huijaavia generointeja tuottavia malleja ei olla vielä kehitetty, mutta ollaan jo todella lähellä 50 %:n huijausastetta kuvien generoinnissa, ja uusia malleja kehitetään lähes päivittäin. Videoiden tuotto ja muokkaus tuovat uusia haasteita, joista oleellisimpana voitaisiin mainita aikakoherenssi (engl. Temporal coherence).

Avainsanat: generatiivinen tekoäly, visuaalinen media, GAN, diffuusio, autoenkooderit, visuaalinen generointi

SISÄLLYSLUETTELO

1. JOHDANTO

Viime vuosina yksi nopeimmin kehittyvimmistä käytännöllisistä tieteenaloista on ollut tekoäly. Tekoäly on mahdollistanut tavallisesti jonkin tasoista ihmisälyä vaativien tehtävien suorittamisen koneellisesti. (Nantheera & Bull, 2022) Etenkin generatiivinen tekoäly on saanut paljon huomiota myös yleisessä mediassa.

Generatiivisen tekoälyn saama suosio on viime aikoina ollut noussussa mallien laajan soveltuvuuden ansiosta. Uudet generatiivisia malleja käyttävät työkalut ovat mahdollistaneet esimerkiksi kuvien tekstipohjaisen generoinnin helppokäyttöisyyden, mikä on tuonut tekoälymallit myös ilman aiempaa osaamista tai tietämystä omaavien kuluttajien käyttöön. Työkaluja on sekä maksullisia että ilmaisia ja niin helppokäyttöisempiä kuin tarkempaa tietämystä vaativiakin, ja ne pohjautuvat usein erilaisiin syviin koneoppimismalleihin tai -algoritmeihin, kuten generatiivisiin kilpaileviin verkostoihin (engl. Generative adversarial networks, GAN), diffuusiomalleihin (engl. Diffusion models, DM) sekä autoenkoodereihin (engl. Autoencoders). Alan kehitys on luonut uusia tapoja helpottaa ja tehostaa työtä, joka aiemmin vei tunteja, päiviä tai jopa viikkoja. Esimerkiksi joidenkin erikoistehosteiden lisäys tai mustavalkoelokuvien värittäminen on mahdollista aiempaa vaivattomammin.

Tekoälyn nopean kehityksen ja uutuuden myötä voi olla hankala pysyä ajan tasalla aiheeseen liittyen. Mediassakin usein puhutaan vain tekoälystä yleisesti mainitsematta tarkemmin, mitä malleja on käytetty ja mistä on kyse. Kiinnostuin tämän takia aiheesta itse ja halusin selvittää, mitä varsinaisesti tarkoitetaan, kun puhutaan tekoälystä ja koneoppimisesta sekä mitä niiden avulla voidaan jo nykyään saavuttaa.

Työ on toteutettu kirjallisuuskatsauksena. Tutkielman tavoitteena on selvittää generatiivisten tekoälymallien toimintaa sekä perehtyä enemmän niiden visuaalisen generoinnin käyttötapoihin. Aihe on valittu generatiivisten mallien uutuuden ja nopean kehityksen sekä työn tekijän oman mielenkiinnon perusteella. Aihe on rajattu edelleen visuaaliseen generointiin, jotta voitaisiin keskittyä kirjoittajalle mielenkiintoisimpaan osaan.

Luvussa 2 kerron tarkemmin tutkimusmenetelmästä. Luvussa 3 avataan tutkielmassa käsiteltyjen keskeisten generatiivisten mallien arkkitehtuureja ja toimintaperiaatteita. Luvussa 4 keskitytään erilaisiin visuaalisen median generointi-

ja muokkaustekniikoihin näiden mallien avulla. Lopuksi pohditaan mallien ja työkalujen käyttötapoja tulevaisuudessa, ja esitellään tutkielman yhteenveto.

2. TUTKIMUSMENETELMÄ

Työ on toteutettu kirjallisuuskatsauksena. Relevantteja artikkeleita on haettu Google scholarista, proQuestistä sekä IEEExplorerista seuraavien hakusanojen eri yhdistelmillä: Generative AI, GAN, General Adversarial Networks, Diffusion, Latent Diffusion Model, autoencoder, VAE, usage, usecases, applications, movie, super-resolution, visual media. Hakuja on myös tarkennettu tiettyyn käyttökohteeseen keskityttäessä, kuten haussa (img2img OR image-to-image OR image synthesis) AND (generative adversarial models OR GAN). Myös Googlen, Redditin ja Youtuben avulla on haettu tieteellisten lähteiden tueksi alan tutkijoiden haastatteluja, käyttöesimerkkejä sekä generatiivisia malleja hyödyntävien työkalujen omia verkkosivuja.

Haut on kohdistettu 2021 tammikuun ja 2023 lokakuun välisellä aikavälillä julkaistuihin vertaisarvioituihin tieteellisiin artikkeleihin ja konferenssijulkaisuihin. Lähteet on lähtökohtaisesti valittu otsikon lisäksi relevanttiuden kriittisen arvioinnin, viittausten määrän, vertaisarvioinnin tilan sekä uutuuden perusteella. Valitut lähteet on uudelleenrajattu abstraktien ja laajemman silmäilyn mukaan ennen tarkempaa käsittelyä. Myös valituista lähteistä on otettu mukaan niissä usein esiintyviä tai niiden aiheita selventäviä artikkeleita.

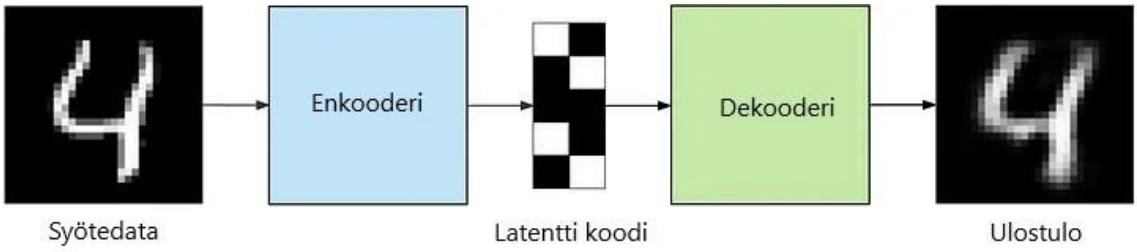
Alkuperäinen aihe oli ”Tekoäly ja koneoppiminen taiteessa”, mutta sitä on uudelleenrajattu useita kertoja tutkimuksen edetessä aiheeseen liittyvien artikkeleiden suuren määrän sekä aiheen laajuuden seurauksena. Aihetta rajattiin ensin tekoälystä generatiiviseen tekoälyyn ja edelleen visuaalista mediaa generoivaan tekoälyyn. Näistä generoivista tekoälyistä on valittu 3 todennäköisesti keskeisintä mallia, joihin on keskitytty syvemmin.

3. KESKEISTEN MALLIEN ARKKITEHTUURIT

Tässä luvussa tarkastellaan keskeisten tekoälymallien ja -algoritmien arkkitehtuureja. Luku siis pyrkii vastaamaan siihen, miten mallit toimivat ja miten ne eroavat toisistaan.

3.1 Autoenkooderit

Autoenkooderi on yksinkertaisimmillaan kahdesta neuroverkosta koostuva malli, joka on koulutettu jälleenrakentamaan saamansa syöte. Enkooderi ensin tulkitsee ja tiivistää datan niin sanottuun latenttiin tilaan, tavoitteenaan säilyttää dataa parhaiten kuvailevat piirteet alempidimensionaalisessa muodossa (Bank ja muut, 2020). Tämä kompressoitu data pyritään sitten muuttamaan takaisin lähes alkuperäiseen muotoonsa dekooderin avulla (Bank ja muut, 2020). Kuva 1 havainnollistaa tätä enkooderi-dekooderi rakennetta.



Kuva 1 - Autoenkooderin arkkitehtuuri. Kuva haettu mukaillen ositteesta towardsdatascience.com 23.10.2023.

Tavallisilla autoenkoodereilla ei kuitenkaan pystytä generoimaan uutta dataa (Bandi ja muut, 2023). Tämän ongelman ratkaisemiseksi kehitettiin esimerkiksi variaatioautoenkooderi (engl. Variational autoencoder, VAE), joka onkin ehkä nykyään yleisimmin käytetty autoenkooderityyppi (Bank ja muut, 2020). Siinä autoenkooderin arkkitehtuuriin yhdistetään variaatiopäättelyä (engl. Variational inference), mikä mahdollistaa myös uusien tietopisteiden (engl. data point) luonnin (Bandi ja muut, 2023). Toisin sanoen data kompressoidaan distribuutioiksi vektorien sijaan mahdollistaen syötteessä ennestään olemattomien kokonaisuuksien, kuten kuvien, generoimisen.

Vaikka VAE on hyvinkin tehokas laadukkaiden kuvien synteesissä, se kuitenkin häviää näytteiden laadussa generatiivisille kilpaileville verkostoille (Rombach

ja muut, 2022). Sitä on silti hyödynnetty muissa korkeaan laatuun kykenevien mallien arkkitehtuureissa.

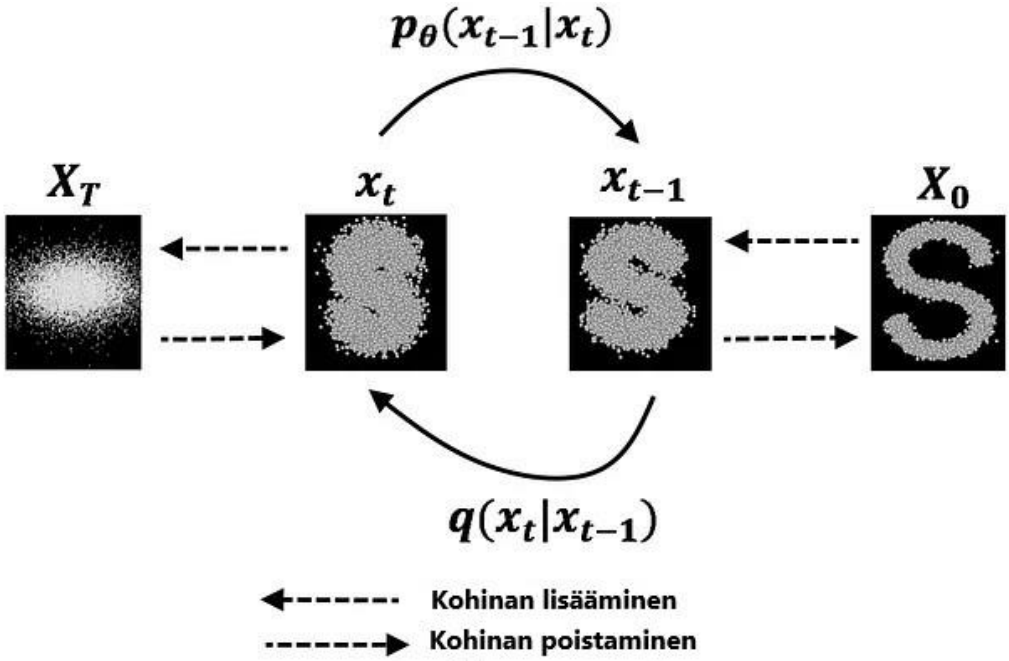
3.2 Generatiiviset kilpailevat verkostot

Generatiiviset kilpailevat verkostot ovat kahdesta neuroverkosta koostuva koneoppimisalgoritmi (Gui ja muut, 2020). Siinä yksi neuroverkko toimii uutta dataa tuottavana generaattorina ja toinen luokittelijana, joka pyrkii erottamaan generoidun ja aidon datan toisistaan (Goodfellow ja muut, 2014). Ne kilpailevat toisiaan vastaan, muodostaen eräänlaisen kaksinpelin; generaattori pyrkii luomaan mahdollisimman aidonoloista dataa harhauttaakseen luokittelijaa, kun taas tämä luokittelija pyrkii minimoimaan aidon ja generoidun datan luokittelussa tapahtuvan virheen (Goodfellow ja muut, 2014). Tapahtuma on iteratiivinen, ja jokaisen läpikäynnin jälkeen neuroverkot kehittyvät paremmiksi tehtävissään. Usein koulutuksen jälkeen luokittelija jätetään lopullisesta mallista kokonaan pois, koska ollaan kiinnostuneita vain generaattorin toiminnasta.

GAN pohjaisia malleja on satoja (The Gan Zoo github, n.d.). Niitä on kehitetty lukuisiin eri tehtäviin, ja ne ovat mullistaneet monia tekoälyn ja koneoppimisen toiminta-aloja (Bandi ja muut, 2023). Aiempiin malleihin verrattuna GAN pohjaiset generaattorit tuottavat parempia kuvia ja niiden arkkitehtuurilla on vähemmän rajoitteita (Gui ja muut, 2020). Uudemmat huippuluokan diffuusiomallit kuitenkin usein pärjäävät tehtävissään paremmin kuin GAN-mallit.

3.3 Diffuusiomallit

Diffuusiomallien toiminnan idea koostuu kahdesta vaiheesta: Ensimmäinen on rakenteen iteratiivista ”tuhoamista” datan jakaumassa diffuusion avulla. Tämä on käytännössä normaalidistribuutiota noudattavan kohinan (engl. noise) lisäämistä dataan, kunnes se on pelkkää kohinaa. Toinen vaihe on käänteisen prosessin kouluttaminen neuroverkolle, jossa iteratiivisesti kohinaa vähentämällä voidaan päästä alkuperäistä dataa muistuttavaan lopputulokseen. Tässä siis aloitetaan pelkästä kohinasta, ja jokaisessa vaiheessa kohinaa ennustetaan ja poistetaan vaiheittain. Ensimmäinen vaihe voidaan toteuttaa myös yhdellä askeleella, jolloin prosessin iteratiivisuus ohitetaan kokonaan. Toisessa vaiheessa tämä ei ole suositeltavaa tulosten huomattavan huonontumisen takia. (Sohl-Dickstein ja muut, 2015)



Kuva 2 – Diffuusiomalli. Kuva haettu mukaillen sivustolta towardsdatascience.com 23.10.2023

Yleistä prosessia on havainnollistettu kuvassa 2. Diffuusiomallit kuitenkin eroavat toisistaan hieman ja ne voidaan jakaa kolmeen alatyyppiin niiden prosessin vaiheiden toteutuksen perusteella: DDPM (Denoising Diffusion Probabilistic Models), SGM (Score-based Generative Models) sekä SDE (Stochastic Differential Equations) (Bandi ja muut, 2023).

Nämä mallit saavuttivat johtavan aseman tiheysfunktion estimoinnissa ja generointien laadussa. Ne kuitenkin kärsivät toiminnan hitaudesta sekä koulutuksen vaatimien resurssien suuresta määrästä. Latentit diffuusiomallit (engl. Latent diffusion models, LDM) ovat kehittyneempiä diffuusiomalleja, ja ne ovat edeltäjiinsä verrattuna skaalautuvampia, helpommin koulutettavia sekä tehokkaampia korkeissakin resoluutioissa. (Rombach ja muut, 2022) Niiden toimintamalli eroaa tavallisista diffuusiomalleista siten, että ne koulutetaan kompressoimaan pikselidataa pienempiulotteisempaan tilaan ja tuottamaan alkuperäinen kuva kompressoidusta tilasta autoenkooderien avulla, säästäen muistia ja nostaen laskentatehoa (Blattmann ja muut, 2023). Tarkoituksena on saada kuvaileva vektori, jolloin voidaan ohittaa kuvan mitättömät yksityiskohdat ja keskittyä datan tärkeisiin semanttisiin piirteisiin (Rombach ja muut, 2022).

Latenttia diffuusiomallia on käytetty etenkin korkealaatuisten ja vaihtelevien näytteiden, kuten kuvien tai videoiden, generoimiseen (Bandi ja muut, 2023). Viimeisimmät kuvien generointityökalut, kuten DALL-E 2 ja Stable Diffusion, käyttävät nimenomaan latentteja diffuusiomalleja. LDM:illä saavutetaan muihin malleihin verrattuna parempilaatuisia ja monipuolisempia kuvia (Rombach ja muut,

2022). Rombachin ja muiden (2022) mukaan latentti diffuusiomalli on myös multimodaalinen, eli datan ei välttämättä tarvitse olla tietynlaista, kuten kuvadataa. Näin latentteja diffuusiomalleja voitaisiin siis käyttää monellakin alalla riippumatta käytetyn datan muodosta.

4. VISUAALISET KÄYTTÖTAVAT JA -KOHTEET

Tässä luvussa käydään läpi erilaisia GANien, diffuusiomallien sekä autoenkoodereiden visuaalisia käyttötapoja. Visuaalisiin käyttötapoihin kuuluvat kuvien ja videoiden generointi sekä niiden eri muokkaustehtävät, kuten niiden väritys ja tyylinsiirto (engl. Style transfer).

4.1 Kuvien generointi

Yksi suosituimmista tavoista käyttää generatiivisia malleja on kuvasynteesi, toisin sanoen kuvien tuotto. Tämä on saanut paljon huomiota myös yleisessä mediassa. Generointi onnistuu monien eri arkkitehtuurien, algoritmien ja mallien, kuten GAN:ien, DM:ien ja VAE:iden, avulla (Bandi ja muut, 2023).

Yksi tapa on käyttää GAN-malleja. Esimerkiksi DRAGAN-mallia voi hyödyntää anime -tyylisten kuvien generoimiseen (Jin ja muut, 2017). Koulutettu malli on vapaasti käytettävissä verkossa ja tekijät mainitsevat, että sitä voidaan hyödyntää esimerkiksi pelien hahmojen suunnittelussa ja piirtämisessä. Tämä voisi heidän mukaansa vähentää viihdemedian, kuten Doujin RPG -pelien, luomiseen kulutettuja resursseja.

Vaikka ensimmäiset korkealaatuisia kuvia tuottavat mallit olivat variaatioita generatiivisista kilpailevista verkostoista ja VAE:ista, viime vuosina suosioon ovat nousseet diffuusiomallit (Borji, 2023). Ne ovat ohittaneet GAN pohjaiset mallit generoitujen kuvien laadussa ja monipuolisuudessa ja ovat lisäksi helpommin koulutettavia (Dhariwal & Nichol, 2021). Esimerkiksi tunnetut kuvangenerointityökalut Stable Diffusion (Rombach ja muut, 2022), DALL-E 2 (Ramesh ja muut, 2022) ja Imagen (Saharia ja muut, 2022b) hyödyntävät diffuusiomalleja. Ilmaisella latenttia diffuusiomallia käyttävällä Stable Diffusionilla voi kuka tahansa luoda haluamanlaisiaan kuvia luonnollisen kielen avulla. Borjin (2023) tekemän vertailevan tutkimuksen mukaan Stable Diffusion on johtavista diffuusiomallin sisältävistä työkaluista selvästi paras, ainakin kasvojen tarkkuuden suhteen.

Diffuusiomallien monipuolisen luonteen takia melkein mikä vain tyyli, kuten realistinen tai Disney-piirretyn tapainen, on mahdollista generoida yhdellä yleisellä mallilla (Rombach ja muut, 2022), mihin GAN-mallit ja variaatioautoenkooderit eivät kykene. Työkalulla on suhteellisen helppo saada niin korkealaatuista taidetta, että sitä voitaisiin käyttää esimerkiksi kirjojen kansikuvien (Reddit, 2023a) tai kokonaisten visuaalisten romaanien luomiseen (Reddit, 2023b). Myös

tekoälymalleja käyttävien taiteilijoiden määrä on lisääntynyt, ja heidän yleisimmin käyttämänsä mallit ovat diffuusiomalleja.

Erityisesti lisäämällä työkaluihin erilaisia laajennuksia, kuten ControlNet (Zhang ja muut, 2023), voidaan saavuttaa hyvinkin suuri kontrolli lopputulokseen. Esimerkiksi viime aikoina paljon mielenkiintoa herättänyt taiteellisten ja toimivien QR-koodien generointi (Stable diffusion art, 2023) viittaa nykyisen kontrollin vahvuuteen.

Myös suuret yritykset ovat alkaneet hyödyntää generoivaa tekoälyä. Esimerkiksi South Parkin tuottaja Fable Studio (2023) kertoo tutkivansa räätälöidyn diffuusiomallin käyttöä uusien hahmojen luomiseen. NVIDIA Research osoitti myös pelien tasojen automaattisen generoinnin olevan mahdollista (Anantrasirichai ja Bull, 2022).

4.2 Kuvien muokkaaminen

Uusien kuvien tuoton lisäksi jo olemassa olevien kuvien muokkaaminen on mahdollista. Esimerkiksi mustavalkokuvien väritys, ns. tyylinsiirto, sisämaalaus (engl. inpainting), vahinkojen korjaus sekä resoluution kohottaminen ovat tällaisia käyttötapoja (Anantrasirichai ja Bull, 2022). Muokkaamalla korjataan ja nykyaikaistetaan vanhoja jo olemassa olevia kuvia, muutetaan kuvien tyyliä tai luodaan erikoistehosteita.

4.2.1 Tyylinsiirto

Tyylinsiirrossa uutta kuvaa generoidessa säilytetään syötekuvan semanttiset piirteet, mutta saavutettu lopputulos on erityylinen. Kuva voidaan muuttaa piirretyn näköiseksi tai jonkin taiteilijan tekemäksi, google kartasta voidaan luoda ilmakuva ja mustavalkoisesta tai kellertävästä vanhasta kuvasta voidaan tehdä värikäs, nykyaikaistettu kuva. (Anantrasirichai ja Bull, 2022)

NVIDIAn StyleGAN on ollut yksi historiallisesti merkittävimmistä GAN-malleista tyylinsiirrossa. Se tuottaa hyvinkin realistisia ja monipuolisia kuvia ja antaa käyttäjälle hyvän hallinnan vaikuttaa kuvan ulkonäköön. (Bandi ja muut, 2023) Tyylinsiirto on mahdollista myös latenteilla diffuusiomalleilla, ja ne toimivat tehokkaina kuvasta kuvaan tekniikan (engl. image-to-image) yleismalleina (Rombach ja muut, 2022).

Tämän metodin avulla voi olla helpompi saavuttaa halutut tulokset uusien kuvien generointiin verrattuna. Esimerkiksi pelejä tehdessä voidaan käyttää oikean ihmisen kuvia, joiden tyyliä vain muutetaan, jolloin esimerkiksi ilmeet, kuvakulmat sekä lähtökohtainen vaatetus voidaan ottaa suoraan syötteestä. Näin hahmot ja

ympäristöt on myös helpompi pitää yhtenäisinä ilman, että pitäisi kouluttaa erikseen tekoälymalli jokaiselle hahmolle.

4.2.2 Kuvien väritys

Vanhojen mustavalkoisten tai kellertävien kuvien väritys voi olla monelle mielenkiintoinen käyttökohde. Voidaan esimerkiksi värittää kuvia historiallisista tapahtumista, kuten sodista tai kunkin omista isovanhemmista ja nähdä, miltä tilanteet oikeasti näyttivät. Värit tuovat yllättävän paljon eloa kuviin ja pelkistä mustavalkokuvista voi olla vaikea saada käsitystä, miltä esimerkiksi isovanhemmat oikeasti näyttivät.

Tavallisesti väritys vaatii paljon aikaa vievää manuaalista työtä, joten hyvälle tekoälymallille, joka automatisoisi työn, on tarvetta. On kuitenkin havaittu, että väritys on edelleen haastava ongelma ja vaatii lisää tutkimusta. (Anantrasirichai ja Bull, 2022) Saharia ja muut (2022a) kehittivät diffuusiomalleja sisältävän kehyksen, joka toimi aiempia paremmin tässä tehtävässä. Malli saavutti 47.8 %:n huijausasteen testissä, jossa ihmisiä pyydettiin valitsemaan alkuperäinen värikuva mallin värittämän kuvan sijaan. Tulos lähenee tavoiteltua 50 %:ia. Havaittiin myös, että kuvat olivat pelkän realistisuuden lisäksi hyvin lähellä alkuperäisiä, mikä viittaa erinoimaiseen ja yleisesti käyttökelpoiseen malliin.

4.2.3 Sisämaalaus

Anantrasirichain ja Bullin (2022) mukaan sisämaalaus on kuvan tai videon puutteellisten tai vahingoittuneiden osien, kuten halkeamien tai naarmujen, täyttävää estimointia. Rombach ja muut (2022) taas kuvailevat sitä maalattujen kohtien täyttämisenä. GAN-mallit, kuten ExGAN, CA-GAN ja PGGAN, sekä diffuusiomallit ovat hyviä tässä tehtävässä (Bandi ja muut, 2023). Erityisesti Saharian ja muiden (2022) malli sekä latentit diffuusiomallit (Rombach ja muut, 2022) vaikuttaisivat olevan hyvinkin tehokkaita sisämaalauksessa.

Tätä tekniikkaa voi käyttää yhdessä värityksen ja muiden ehostavien prosessien kanssa. Esimerkiksi vanhoja vahingoittuneita maalauksia voitaisiin digitaalisesti korjata lähes alkuperäiseen muotoonsa. Toisaalta ihmiset voivat myös hyödyntää sisämaalausta omien kuviensa muokkaamiseen; taustalta voidaan poistaa ihmisiä, kohteen hiusten väriä voidaan muuttaa napin painalluksella, rakennuksia voidaan muuttaa puiksi tai ihmisen silmät voidaan avata kuvassa. Myös mainoksia tai tekoälytaidetta tehdessä saadaan mahdollisuus muokata vain pientä osaa kerrallaan, jolloin teoksia voidaan luoda ja parannella vaiheittain.

Jopa Adobe Photoshop, yksi tunnetuimmista ja käytetyimmistä digitaalisista kuvanmuokkausohjelmistoista, sisältää sisämaalaukseen kykenevän työkalun

(Anantrasirichai ja Bull, 2022). Työkalua voidaan käyttää sekä manuaalisen että automaattisesti segmentoivan maskin avulla.

4.2.4 Superresoluutio

Superresoluutio (engl. Super-resolution) on kuvan (tai videon) resoluution kohottamista. Tehtävä on hankala, sillä kuvan pääpiirteet ja semanttinen merkitys on pidettävä samana, kun kokoa kasvatetaan generoimalla uusia pikseleitä (Bandi ja muut, 2023). Generoitujen kuvien on usein oltava laadukkaita, eikä niissä saa yleensä olla lisättyjä piirteitä, sillä tekniikkaa käytetään esimerkiksi lääketieteessä CT- ja MR-kuvien tarkentamiseen (Zhao ja muut, 2022) sekä rikostutkinnassa (AlTakrouri ja muut, 2023). Nykyaikaisia malleja ovat esimerkiksi AlTakrourin ja muiden (2023) ISRGAN sekä diffuusiomallit, jotka ovat Saharian ja muiden (2022) mukaan ohittaneet GANit tehtävässä.

Kuvien koon kohottamista ja tarkentamista voitaisiin hyödyntää myös puhelimissa käyttäjien ottamien kuvien parantamisessa etenkin digitaalista zoomausta käytettäessä. Myös lehtijulkaisuissa usein näkee huonolaatuisia, esimerkiksi valvontakameroiden ottamia kuvia, ja tällä tekniikalla voitaisiin edistää tilannetta.

4.3 Videoiden tuotto

Generatiivisia malleja voidaan myös käyttää videoiden tuottamiseen. Yksinkertaisimmillaan videot ovat vain kuvasarjoja, ja laadukkaiden kuvien tuottamisessa on jo päästy pitkälle, mutta videoformaatti kuitenkin tuo uusia haasteita, joista oleellisimmat ovat aika- ja tilakoherenssin (engl. Temporal and spatial coherence) säilyttäminen. Pohjana käytetään kuitenkin samoja luvun 4.2 läpi käymiä metodeja.

Niin sanottuja kuvasta videoon (engl. image-to-video) -malleja, joissa aloituskuvasta luodaan kuvasarja, on useita. Ni ja muut (2023) kehittivät yhden näistä malleista, joka on tarkemmin tarkoitettu konditionaaliseen kuvasta videoon -prosessiin, jossa se pystyy kuvan ja tekstiohjeen avulla tuottamaan ohjetta vastaavan kuvasarjan. Heidän saavuttamansa aikakoherenssi, kohteen ulkonäön säilyminen sekä liikkeen jatkuvuuden varmistaminen olivat parempia kuin artikkelissa vertailun kohteena olevien aiempien mallien tuotokset. Mallin arkkitehtuuri koostuu niin sanotusta latentin vuon autoenkooderista (engl. Latent flow autoencoder, LFAE), joka koulutetaan valvomattomasti ennustamaan annetun kuvan jälkeisen kuvan sen latentin optisen vuon (engl. Latent optical flow) perusteella, sekä tätä ennustetta hyödyntävästä generoivasta diffuusiomallista. Lähestymistapa on yksinkertainen ja toimii matalaulotteisesti helpottaen generointia ja vähentäen tarvittavia koneellisia resursseja. (Ni ja muut, 2023)

Tämä käyttötapa kuitenkin vaatii vielä selvää kehitystä. Nin ja muiden (2023) mallikin pystyy tuottamaan vain yhden kohteen sisältäviä videoita ja sen ohjaus toimii luokkamerkinnöillä luonnollisen kielen sijaan. Toinen tunnettu toimintamalli on tekstistä videoon (engl. Text-to-video) -käyttötapa, jossa pelkän tekstiohjeen perusteella pyritään generoimaan koherentti kuvasarja. Tämä on kuitenkin haastavampi toteuttaa.

4.4 Videoiden tyylinsiirto

Videoiden tyylinsiirto eroaa edellisestä metodista siten, että siinä syötteenä annetaan kokonainen video yhden lähtökuvan tai -tekstin sijaan. Se onkin täten helpompaa, kuin kokonaan uuden videon tuottaminen. Ongelmat ovat kuitenkin lähtökohtaisesti samat, mutta liikkeen säilyminen on huomattavasti paremmalla tasolla pelkän ohjeen avulla generoimiseen verrattuna.

Esimerkiksi Wu ja muut (2022) kehittivät toimivan latenttia diffuusiota hyödyntävän mallin, joka pystyy kokonaisen videon ja tekstiohjeen avulla generoimaan halutunlaisen videon. Se vaikuttaa ymmärtävän tekstisyötteen semanttisen merkityksen hyvin ja pystyvän luomaan kohtalaisen hyviä muokattuja videoita.

Myös Stable Diffusionilla voidaan etenkin lisäosien avulla saavuttaa tyylinsiirto videoihin. Videosta voidaan eristää kuvat, generoida jokainen kuvasta kuvaan -tyylillä ja lopulta liittää kuvat toisiinsa. Näin saavutetaan kohtalainen liikkeen säilyttävä lopputulos säädetyillä parametreilla, mutta erilliset kuvat eivät vielä tiedä toisistaan mitään. Täten saatu video sisältää usein paljon ”välkkymistä” (engl. Flickering) vaihtelevien värien ja taustojen takia. Siksi työkalun kanssa on hyvä käyttää temporaalista yhtenäisyyttä parantavia lisäosia, kuten TemporalKittiä (CiaraRowles, 2023a), TemporalNettiä (CiaraRowles, 2023b) tai aiemmin mainittua ControlNettiä (Zhang ja muut, 2023), joka auttaa pitämään lopullisen videon tasaisempana.

Elokuvien erikoistehosteissakin voidaan käyttää tätä tekniikkaa hyödyksi. Todorovic ja muut (2023) kertovat ”Wonder Studios” -nimisen työkalunsa pystyvän automatisoimaan suuren osan CGI työstä yli 25:ttä koneoppimismallia käyttämällä. Itse työkalun käyttö toimii muutamalla napin painalluksella. Hahmojen luonnin lisäksi se huolehtii esimerkiksi valaistuksesta, liikkeentunnistuksesta sekä varsinaisen näyttelijän poistamisesta kuvasta.

Tämäkin on siis monipuolinen tekniikka, jolle on monta eri mahdollista käyttötarkoitusta ja tulee luultavasti mullistamaan monet eri visuaalisen median tuottamisprosessit. Etenkin video- ja elokuvatuotannossa tullaan varmasti havaitsemaan tuottamisen vaiheiden automatisointia ja tehostamista.

5. YHTEENVETO JA POHDINTA

Tutkielmassa tutkittiin autoenkooderien, generatiivisten kilpailevien verkkojen sekä diffuusiomallien toimintaa ja niiden käyttöä visuaalisen median generoinnissa. Autoenkoodereista nykyään käytetyin ja tunnetuin tyyppi on variaatioautoenkooderi, kun taas diffuusiomalleista nykyaikaisin on tätä hyödyntävä diffuusiomalli, latentti diffuusiomalli. GAN-malleja taas on lukuisia, ja ne ovat erilaisiin käyttötarkoituksiin erikoistuneita.

Kuvien ja videoiden generoimiseen on monta eri tapaa. Työssä tarkasteltiin tekstistä kuvaan, tekstistä videoon, kuvasta videoon, kuvasta kuvaan ja videosta videoon -toimintamalleja. Vastaavasti jo olemassa olevan median muokkaamiseen on monia keinoja, joista oleellisimmat ovat työssä käsitellyt tyylinsiirto, väritys, sisämaalaus sekä superresoluutio, joilla yleensä pyritään parantamaan median laatua, lisäämään haluttuja piirteitä tai korjaamaan epäkohtia. Myös tavalliset kuluttajat pystyvät hyödyntämään malleja erilaisten työkalujen avulla.

Kuluttajille helppokäyttöisimmät ja parhaita lopputuloksia tuottavat työkalut käyttävät latentteja diffuusiomalleja, mutta myös generatiivisia kilpailevia verkostoja tutkitaan ja parannellaan edelleen paljon. Julkaistuja malleja ja työkaluja voidaan hyödyntää monella eri alalla, kuten rikostutkinnassa, lääketieteessä, elokuvatuotannossa sekä sosiaalisessa mediassa. Myös kuluttajat voivat käyttää malleja omiin tarkoituksiinsa haluamansa mukaan, kuten isovanhempiensa mustavalkokuvien automaattiseen värittämiseen.

Mahdollisia käyttötarkoituksia onkin siis paljon ja niitä tullaan varmasti keksimään lisää. Onkin mielenkiintoista seurata, miten paljon hyötyä malleista on ja mitä uusia ongelmia tullaan kohtaamaan tekoälyn alan kehittyessä. Esimerkiksi elokuvatuotannon paljon aikaa vieviä prosesseja voidaan luultavasti automatisoida ja nopeuttaa huomattavasti lähitulevaisuudessa, mutta se saattaa aiheuttaa sen, että alan ammattilaisia ei tarvita enää lähellekään yhtä paljoa johtaen työttömyyden kasvuun.

Nykyään videoiden generoinnilla ja muokkaamisella ei vielä pystytä tuottamaan täydellisiä lopputuloksia, mutta uskon vahvasti, ettei tulevaisuuden työkalujen ja mallien tuotoksia tarvitse jälkeenpäin korjailla. Ala on kuitenkin kehittynyt hyvin nopeasti jo useamman vuoden ajan eikä tahti tunnu hidastuvan lähivuosina. Tekoälyn saama huomio yleisessä mediassakin varmasti tuo paljon mielenkiintoa ja mahdollisesti rahoitusta alan jatkuvalle kehitykselle. Ehdotan, että etenkin videoita generoivien ja muokkaavien mallien parantamiseen tulisi keskittyä lähitulevaisuudessa. Myös superresoluutiota tulisi tutkia lisää, sillä sitä voidaan

hyödyntää lääketieteessä ja muilla tarkkuutta vaativilla aloilla, ja pienilläkin virheillä voi olla suuria seurauksia.

LÄHTEET

AlTakrouri, S., Norliza, M. N., Ahmad, N., Justinia, T., & Usman, S. (2023). Image Super-Resolution using Generative Adversarial Networks with EfficientNetV2. International Journal of Advanced Computer Science and Applications, 14(2) https://doi.org/10.14569/IJACSA.2023.01402100

Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: A review. The Artificial Intelligence Review, 55(1), 589-656. https://doi.org/10.1007/s10462-021-10039-7

Bandi, A., Adapa, P.V.S.R. & Kuchi, Y. E. V. P. K. (2023). The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. Future Internet, 15(8), 260. https://doi.org/10.3390/fi15080260

Bank, D., Koenigstein, N., Giryes, R. (2020). Autoencoders. Arxiv.org. arXiv:2003.05991

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., & Kreis, K. (2023). Align your latents: High-resolution video synthesis with latent diffusion models. Piscataway: The Institute of Electrical and Electronics Engineers, Inc. (IEEE). https://doi.org/10.1109/CVPR52729.2023.02161

Borji, A. (2023). Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2. Cornell University Library, arXiv.org. arXiv:2210.00586

CiaraRowles (2023a). TemporalKit. Github.com. https://github.com/CiaraStrawberry/TemporalKit

CiaraRowles (2023b). TemporalNet2. Huggingface.co. https://huggingface.co/CiaraRowles/TemporalNet2

Dhariwal, P., Nichol, A. (2021) Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233.

Fable studio. (2023). To Infinity and Beyond: SHOW-1 and Showrunner Agents in Multi-Agent Simulations. Fable Studio's GitHub page https://fablestudio.github.io/showrunner-agents/

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative Adversarial Networks. arXiv:1406.2661

Jin, Y., Zhang, J., Li, M., Tian, Y., Zhu, H., & Fang, Z. (2017). Towards the automatic anime characters creation with generative adversarial networks. Ithaca: Cornell University Library, arXiv.org. https://doi.org/10.48550/arXiv.1708.05509

Ni, H., Shi, C., Li, K., Huang, S. X., & Min, M. R. (2023). Conditional image-to-video generation with latent flow diffusion models. Piscataway: The Institute of Electrical and Electronics Engineers, Inc. (IEEE). https://doi.org/10.1109/CVPR52729.2023.01769

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. https://doi.org/10.48550/arXiv.2204.06125

Reddit. (2023a). Fantasy Book Cover Art (Workflow): Stable Diffusion as an alternative to stock photography?. https://www.reddit.com/r/StableDiffusion/comments/101pchf/fantasy_book_cover_art_wo rkflow_stable_diffusion/

Reddit. (2023b). Visual Novel chapter w/ SD. https://www.reddit.com/r/StableDiffusion/comments/10l7ejf/visual_novel_chapter_w_sd/

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Piscataway: The Institute of Electrical and Electronics Engineers, Inc. (IEEE). https://doi.org/10.1109/CVPR52688.2022.01042.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., & Norouzi, M. (2022a). Palette: Image-to-Image Diffusion Models. In ACM SIGGRAPH 2022 Conference Proceedings (SIGGRAPH '22). Association for Computing Machinery, New York, NY, USA, Article 15, 1–10. https://doi.org/10.1145/3528233.3530757

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Seyed Ghasemipour, S. K., Karagol Ayan, B., Mahdavi, S. S., Gontijo Lopes, R., Salimans, T., Ho, J., Fleet, D. J., Norouzi, M. (2022b) Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Arxiv.org. https://doi.org/10.48550/arXiv.2205.11487

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. https://doi.org/10.48550/arXiv.1503.03585

Stable diffusion art. (2023). How to generate a QR code with Stable Diffusion. Stablediffusion-art.com. https://stable-diffusion-art.com/qr-code

The Gan Zoo. (n.d.). https://github.com/hindupuravinash/the-gan-zoo

Todorovic, N., Sheridan, T., Kanazawa, A., Torralba, A. (2023), Understanding the Role of AI in Reshaping the Film & TV Industry featuring Tye Sheridan, SXSW 2023. https://youtu.be/NbpRl3YTL2Q?si=l786Ck1QpaRcCkZZ

Wu, J. Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Hsu, W., Shan, Y., Qie, X. & Shou, M.Z. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. arXiv 2022, https://doi.org/10.48550/arXiv.2212.11565

Zhang, L., Rao, A., Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Model. Arxiv.org. https://doi.org/10.48550/arXiv.2302.05543

Zhao, J., Hou, X., Pan, M., & Zhang, H. (2022). Attention-based generative adversarial network in medical imaging: A narrative review. Computers in Biology and Medicine, 149. https://doi.org/10.1016/j.compbiomed.2022.105948