Informe de Progrés I TFG

Ús de models de difusió per alterar la realitat física de les escenes

Tutor: Ramon Baldrich Caselles
Alumna: Abril Piñol Chacon - 1604159

1 INTRODUCCIÓ

Aquest informe s'escriu amb la finalitat de presentar i definir el projecte de final de grau, proporcionant una visió detallada dels objectius establerts, l'estat de l'art i la metodologia prevista, així com la planificació del treball, organitzat setmanalment i per objectius. A més, s'hi exposen els diferents reptes que s'han plantejat fins al moment i s'analitzen tant l'evolució de la feina com els resultats aconseguits fins a la data, de la mateixa manera que el perquè de cada decisió presa. A diferència de l'informe inicial en aquest, no només es parla del progrés sinó que també s'hi desenvolupa el funcionament teòric del model utilitzat, entre algún altre concepte rellevant.

El projecte es basa en un model d'intel·ligència artificial (IA) que ha experimentat un notable increment de popularitat i presència, especialment en les xarxes socials els darrers temps. Aquesta tendència s'ha fet visible en diverses ocasions, quan s'ha viralitzat contingut basat en imatges generades per IA. Els models en qüestió es coneixen com a models de generació d'imatges, i aquest projecte se centra específicament en els coneguts com a *Stable Diffusion* (SD).

2 OBJECTIUS

En aquest apartat, es plantegen els objectius que guiaran el desenvolupament i l'execució del projecte. Aquests objectius s'han establert amb la finalitat de proporcionar una estructura clara i definida a l'hora d'afrontar les tasques necessàries per assolir l'objectiu final. A través d'aquests objectius específics, es busca comprendre a fons el funcionament dels models *Stable Diffusion* i aspectes directament relacionats amb aquests, com ho serien la *ControlNet* i diferents arquitectures específiques de tasques d'image restoration, entre altres. Es vol aplicar aquest coneixement per aconseguir resultats efectius i innovadors en la modificació d'imatges. Més endavant es detallen els objectius concrets de què es parla en una taula *(figura 15, annex)*. Aquests objectius que es tracten de manera setmanal es van actualitzant en funció del camí més òptim en cada moment de cara a assolir la meta.

L'objectiu últim és replicar la tasca presentada a l'article *"Relighting from a Single Image: Datasets and Deep Intrinsic-based Architecture"*[1], substituint l'aproximació basada en les *Generative Adversarial Networks* (GANs) per models basats en Stable Diffusion.

3 ESTAT DE L'ART

En aquesta fase inicial, és important el context i una visió general del panorama actual dels models de generació d'imatges, especialment SD. És això el que es desenvolupa a continuació.

Dins dels models de generació d'imatges s'hi poden trobar els *Generative Adversarial Networks* (GANs)[10], a trets generals, aquests estan formats per dues xarxes: un generador i un discriminador. El generador crea imatges i el discriminador intenta distingir si les imatges que li arriben han estat creades per aquest generador o no. Mitjançant l'entrenament adversari, el generador aprèn a produir imatges més realistes que puguin enganyar al discriminador.

Un altre tipus de models que permet crear imatges són els *Variational Autoencoders* (VAEs)[11], aquests consisteixen en un *encoder* que genera una representació de les imatges d'entrada en un espai latent, i un *decoder* que reconstrueix imatges a partir d'aquesta representació en l'espai latent. D'aquesta manera, mostrejant de l'espai latent el *decoder* pot generar imatges.

A diferència d'aquests, els models SD han mostrat millors resultats quan es tracta de crear imatges amb textures més complexes, detalls petits i cantonades afilades. Un dels trets responsables d'aquesta millora en els resultats és el procés iteratiu, aquest permet anar refinant la sortida del model i perfeccionar fins a l'últim detall.

El funcionament bàsic dels models de difusió consisteix a afegir un soroll conegut a les imatges d'entrenament per posteriorment fer el procés invers i aprendre a recuperar les imatges a partir del soroll. D'aquesta manera, un cop entrenats són capaços de crear imatges coherents partint de soroll.
En termes generals, es pot dividir la funcionalitat d'aquest model en dues categories: la creació d'imatges de zero i la modificació d'imatges ja existents. La creació d'imatges des de zero es pot condicionar utilitzant una *prompt* textual. Donat un fragment de text, aquest es codifica en el mateix espai latent i s'incorpora en el procés d'eliminació el soroll. En el cas de modificació d'imatges, no es parteix de només soroll, sinó que hi ha una imatge original que condiciona la sortida. El que es pretén és aplicar certes transformacions a aquesta imatge, però mantenint fidelitat a la imatge d'entrada, un exemple en seria l'eliminació de la pluja o de les ombres d'una imatge.

En aquest treball, l'objectiu és modificar imatges existents, concretament alterant les seves propietats físiques, però no sempre de la mateixa manera. Per tant, a més, es vol fer servir una guia o instrucció específica, en el cas de la tasca de *relighting*, aquesta hauria d'indicar les condicions concretes d'il·luminació que s'esperen en el resultat.

4 METODOLOGIA

La metodologia adoptada per aquest treball està basada en el plantejament *Agile*, aquest és un mètode seqüencial i d'adaptació continua. És cert que inicialment s'estableix un pla de treball concret, no obstant això, aquest es revisa sovint i s'adapta a les necessitats i característiques reals del projecte. És per aquest motiu que és important una comunicació freqüent i eficient que permeti traçar el millor camí per avançar de cara a l'objectiu final.

En aquest treball, les característiques esmentades pel que fa a la metodologia es veuen reflectides en una planificació setmanal, que es pot veure en detall a la següent secció *(secció: 5 Planificació)*, i per objectius, ja descrits en una de les seccions anteriors *(secció: 2 Objectius)*. Tot això s'acompanya de reunions setmanals amb el tutor on es discuteixen els resultats i/o problemes que puguin sorgir i es determina la millor manera de procedir. A més a més, s'avalua si s'han aconseguit els objectius establerts i es determina la millor manera de procedir. Aquestes trobades són clau per evitar l'estancament del projecte i optimitzar el procés de treball, ja que permeten identificar i abordar els problemes de manera ràpida, minimitzant l'impacte que poden tenir a llarg termini.

En resum, la metodologia adoptada combina una planificació flexible i adaptativa amb una comunicació i col·laboració constants, així com una avaluació periòdica del progrés, per garantir un desenvolupament eficient i satisfactori del projecte. Aquest enfocament proporciona una estructura coherent i a la

vegada permet una resposta àgil als canvis i les necessitats que sorgeixen durant el procés de treball.

Fins ara es parla de la metodologia de desenvolupament establerta, però en una fase posterior, s'explorarà amb més detall la metodologia tècnica, que inclourà una descripció exhaustiva de les eines utilitzades en el codi, com ara llibreries específiques i altres recursos rellevants per a l'execució del projecte. Aquesta anàlisi tècnica permetrà una comprensió més profunda de les tecnologies emprades que contribueixen a la implementació i execució dels objectius.

5 PLANIFICACIÓ

En aquesta secció es descriu minuciosament el pla de treball establert d'acord amb la metodologia descrita prèviament, aquest ens permet tenir una perspectiva general de l'evolució prevista del projecte al llarg del temps. Per una visió més clara es mostra en format tabular *(figura 16, annex)*, aquesta taula conté quatre columnes on s'especifiquen les setmanes, juntament amb les seves dates corresponents, els identificadors prèviament assignats als objectius previstos a assolir en aquests períodes. Aquests apareixen marcats amb un color o un altre en funció de si s'han assolit, si s'ha fet de manera parcial o si directament no s'ha aconseguit arribar-hi, els colors són verd, taronja i vermell respectivament. I finalment, en la darrera columna, una pinzellada dels assoliments reals, que es detallen en més profunditat en la secció 7 on es tracta amb detall i visualitzacions tot el desenvolupament del projecte fins a la data. Addicionalment, les files en negre destaquen les dates on hi ha entregues previstes.

6 CONCEPTES RELLEVANTS

Abans de prosseguir amb l'anàlisi de les proves i el treball realitzat, és essencial establir una base teòrica sobre el funcionament del model utilitzat i altres conceptes importants que juguen un paper fonamental en aquest projecte. Aquesta comprensió dels fonaments teòrics proporcionarà un marc essencial per a l'avaluació dels resultats i la interpretació del treball realitzat.

IR-SDE (Image Restoration via Stochastic Differential Equation)

Aquest és un mètode de restauració d'imatges que es basa en l'ús de l'Equació Diferencial Estocàstica (SDE) per recuperar imatges d'alta qualitat. Aquest mètode modela implícitament el procés de degradació de la imatge mitjançant una SDE que té tendència a revertir a la mitjana. En altres paraules, busca estimar la imatge original a partir de la imatge degradada tenint en compte una dinàmica estocàstica que simula els canvis en la imatge al llarg del temps.

Per recuperar la imatge original, la imatge degradada passa per un procés de denoising. Aquest parteix d'una imatge LowQuality (LQ) barrejada amb més soroll, aquesta LQ és la imatge que es passa al model i que es vol alterar perquè pateix algun tipus de degradació. És un procés de T iteracions, que comença en l'instant T amb aquesta imatge i acaba en l'instant 0 amb la imatge HighQuality (HQ) o GT. De manera que, en cada volta es prediu el soroll de la imatge per aquell instant *t* i es treu a la imatge obtinguda en l'instant anterior. És a dir, es comença amb la imatge LQ barrejada amb molt soroll i a poc a poc se li va traient fins a assolir la imatge sense soroll, GT.

Per determinar el soroll que hi ha en la imatge en un instant *t*, s'utilitza la ConditionalUNet, que donada la imatge en l'instant anterior, la imatge LQ i *t,* en fa la predicció. Per entrenar aquesta xarxa, en aquest article acadèmic que s'usa com a punt de partida, es proposa una nova loss (1)[12].

$$J_\gamma(\phi) := \sum_{i=1}^{T} \gamma_i \, \mathbb{E}\Big[\big\| \underbrace{x_i - (\mathrm{d}x_i)_{\tilde{\epsilon}_\phi}}_{\text{reversed } x_{i-1}} - x^*_{i-1} \big\|\Big], \qquad (1)$$

Aquesta vol minimitzar la diferència entre: la imatge en l'instant *t* menys el soroll predit, que seria la imatge que ens quedaria per l'instant *t-1* i a la que es refereix a la fórmula com a *reversed* x_{t-1}. Aquesta és la imatge òptima en

l'instant *t-1*, i ve determinada per la fórmula (2), que sorgeix de buscar la trajectòria òptima $x_{1:T}$ donada la imatge $GT(x_0)$ [12].

$$x^*_{i-1} = \frac{1-\mathrm{e}^{-2\,\bar{\theta}_{i-1}}}{1-\mathrm{e}^{-2\,\bar{\theta}_i}}\mathrm{e}^{-\theta'_i}(x_i-\mu) + \frac{1-\mathrm{e}^{-2\,\theta'_i}}{1-\mathrm{e}^{-2\,\bar{\theta}_i}}\mathrm{e}^{-\bar{\theta}_{i-1}}(x_0-\mu)+\mu. \quad (2)$$

Per poder fer aquest procés d'entrenament s'han de generar, donades unes imatges d'entrada, el valor de cadascuna per diferents instants *t* aleatoris, entre 0 i T. Per obtenir-les, es multiplica soroll Gaussià $\varepsilon_t \sim N(0, I)$ per la variancia en aquest mateix instant, v_t, i se suma a la mitjana també per aquest instant, m_t, aquestes es calculen a partir de les fórmules (3) [12].

$$m_t(x) \coloneqq \mu + (x(0)-\mu)\,\mathrm{e}^{-\bar{\theta}_t}, \qquad v_t \coloneqq \lambda^2\left(1-\mathrm{e}^{-2\,\bar{\theta}_t}\right). \quad (3)$$

7 PROGRÉS PER SETMANES

- ## Setmanes 1-4

  Objectiu: Guanyar una comprensió exhaustiva del funcionament dels models de difusió i en concret del codi escollit.

  Aquest projecte parteix d'un coneixement pràcticament nul sobre els models SD, per la qual cosa les primeres setmanes es dediquen a una recerca exhaustiva. Inicialment, es consolida una base de coneixements teòrics sobre el funcionament estàndard dels models, que es desenvolupa de manera general a la secció anterior (*secció: 6 Conceptes Rellevants*). Posteriorment, s'investiguen els repositoris més populars i se'n fan diverses proves fins a seleccionar el que millor s'ajusta a les necessitats del treball; en aquest cas, "*Refusion: Enabling large-size realistic image restoration with latent-space diffusion models"*[6][8]. Finalment, a través del debug, es dedica temps a comprendre el procés exacte que té lloc, mitjançant visualització de les imatges després de cada part del procés.

  Tot i aconseguir uns bons fonaments per començar a fer experiments, l'aprenentatge i la recerca no acaben aquí. Des d'aquest punt, aquestes tasques es desenvolupen de manera paral·lela als experiments, amb l'objectiu no només d'aprofundir en els conceptes ja tractats, sinó també en nous conceptes, per exemple els diferents mètodes per condicionar el model.

  Es pot concloure que, durant aquestes setmanes, s'assoleix l'objectiu plantejat, tot i que no es dóna per finalitzat, com s'esmenta anteriorment, és important destacar que aquest procés de recerca i aprenentatge és continu.

- ## Setmana 5

  Objectiu: fer un primer entrenament amb el dataset donat, des de scratch i partint dels pesos donats.

  En un primer pas, es tria un conjunt de dades amb per a les proves. En aquest cas, s'utilitza el conjunt de dades conegut com a rainy, que es proporciona des del repositori GitHub associat al codi de punt de partida del projecte. Aquest conjunt consta de 1440 parelles d'imatges, cadascuna formada per una imatge amb l'efecte de pluja i una sense, el groundtruth(GT). D'aquestes, un 15% es fa servir per a la validació i la resta per a l'entrenament.

  Fins aquí, tot el plantejament se centra en l'entrenament des de zero. No obstant això, també es vol explorar l'ús de pesos preentrenats per a una tasca diferent per veure si poden facilitar el procés. En el repositori de punt de partida es proporcionen pesos per a dues tasques: deraining, eliminació de pluja, i deshadow, eliminació d'ombres. Tot i això, els pesos per a la tasca de deshadow contenen errors, per tant, no podem aprofitar el conjunt de dades descrit anteriorment, ja que no ens interessa reentrenar uns pesos per la mateixa tasca que ja han estat entrenats. Així doncs,

s'opta per utilitzar els pesos preentrenats per a la tasca de deraining i es busca un nou conjunt de dades per a la tasca de deshadow. En aquest cas, es tria el conjunt de dades conegut com a Shadow Removal Dataset, SRD, que inclou 3960 imatges. Com abans, es destina el 15% d'aquestes imatges a la validació, mentre que la resta s'usa per a l'entrenament. Les *figures 1 i 2* mostren exemples d'una parella d'imatges de cadascun dels conjunts de dades.

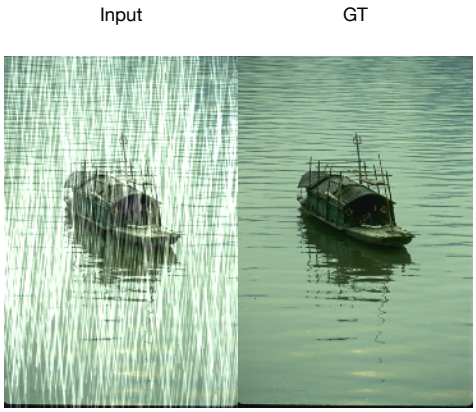

Figura 1. Parella d'imatges dataset *rainy*.

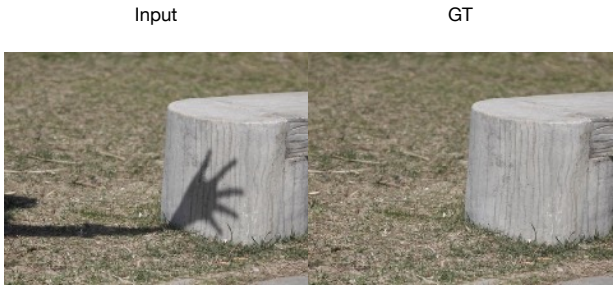

Figura 2. Parella d'imatges dataset SRD.

Els resultats que s'obtenen en la tasca de deraining, entrenant des de zero, són del tot abstractes, el model no arriba a reconstruir res semblant al que es veu en la imatge d'entrada *(figura 3)*, és evident que hi ha algun error important. Pel que fa a la tasca deshadow, partint de pesos preentrenats, la loss i el PSNR que s'assoleixen fan pensar que el model aprèn alguna cosa, ja que milloren al llarg de les iteracions *(figures 4 i 5)*. No obstant quan s'observa la sortida del model, els resultats no són gaire coherents. Es reconstrueix la imatge d'entrada, però apareixen per tota la superfície, taques de colors *(figura 6)*.

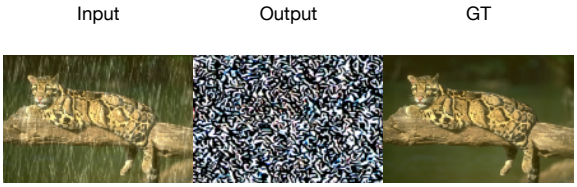

Figura 3. Seqüència d'una imatge abans de passar-la al model (Input), la sortida del model (Output), i la sortida esperada (GT). Tasca deraining.

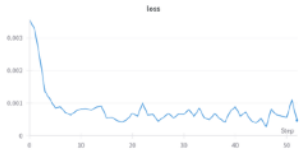

Figura 4. Loss validació cada 5000 iteracions.

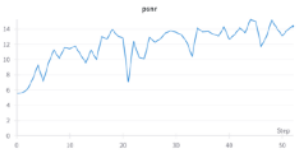

Figura 5. PSNR validació cada 5000 iteracions.

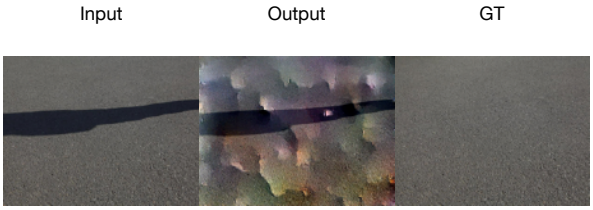

Figura 6. Seqüència Input, Output, i GT per una imatge. Tasca deshadow.

Tot i així, l'objectiu d'aquesta setmana es considera assolit, ja que s'ha completat el procés d'entrenament. Davant del fracàs, es pren la decisió de definir una tasca més senzilla per a properes proves, amb la finalitat de comprendre millor les capacitats del model i identificar els possibles problemes amb més claredat.

- ## Setmana 6

Objectiu: establir ordre i fer un entrenament per a una tasca més senzilla.

Aquesta setmana, abans d'entrenar una nova tasca s'opta per preparar el codi per recopilar totes les dades necessàries, ja que la loss i el psnr no són suficients per determinar gran cosa. Amb l'eina WandB, que es prefereix sobre el Tensorboard ja implementat, es recullen noves mètriques juntament amb imatges durant el procés d'entrenament. Això permet tenir una visió més precisa del desenvolupament en tot moment. Pel que fa a les mètriques

seleccionades, a continuació es mostra la llista juntament amb els motius de la seva inclusió. Encara que no deixen de ser de les més típiques per a tasques de reconstrucció.

Mètriques seleccionades [13][14]:

- MSE (Mean Squared Error): És una mètrica bàsica que mesura la mitjana de la diferència quadràtica entre els valors de píxel de dues imatges. Funciona bé només si volem generar una imatge amb la millor conformitat dels colors dels píxels respecte la imatge de referència, que en aquest cas ja interessa.

Donada una imatge lliure de soroll I de m×n píxels, i la seva aproximació amb soroll K, l'MSE es defineix com[15]:

$$MSE = \frac{1}{m\,n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2.$$

- PSNR (Peak Signal-to-Noise Ratio): És una mètrica amplament utilitzada per avaluar la qualitat de les imatges. S'escull perquè és fàcil de calcular, d'interpretar i mesura el soroll que conté la imatge. Ve definida de la següent manera[15].

$$\begin{aligned} PSNR &= 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \\ &= 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \\ &= 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(MSE). \end{aligned}$$

- SSIM (Structural Similarity Index Measure): Es fa servir per avaluar com són de semblants dues imatges en termes de la seva estructura visual, tenint en compte aspectes com el contrast, la brillantor i la similitud de les textures.

La mesura entre dues finestres x i y de mida comuna N×N és[]:

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

amb μ_x la mitjana de x ; μ_y la mitjana de y ; σ_x^2 la variància de x ; σ_y^2 la variància de y ; σ_{xy} la covariància de x i y ; $c_1=(k_1 L)^2$, $c_2=(k_2 L)^2$ dues variables per estabilitzar la divisió quan el denominador és petit; L el rang dinàmic dels valors de píxel (normalment és $2^{\text{nombre de bits per píxel}}-1$); $k_1=0{,}01$ i $K_2=0{,}03$ per defecte.

- LPIPS (Learned Perceptual Image Patch Similarity): És una mètrica que s'ha après a partir de dades i models de percepció visual humana. Utilitza xarxes neuronals per calcular la distància perceptiva entre dues imatges, tenint en compte característiques de baix nivell com la textura, el color i la forma.

Encara que aquesta setmana no s'ha assolit l'objectiu previst al complet, a causa del temps addicional que ha requerit l'adaptació del codi i la selecció de les mètriques, la feina feta és crucial. Aquesta facilitarà significativament tasques futures i permetrà una millor organització dels diferents experiments.

• Setmana 7

Objectiu: entrenar una tasca senzilla i identificar possibles problemes.

Per començar, cal determinar la tasca, la qual ha de ser relativament senzilla per a poder acotar possibles problemes, com es va mencionar anteriorment. S'opta per aplicar una divisió a un dels canals, en aquest cas el roig (R), amb l'objectiu que el model, donada la imatge original, pugui aprendre a reconstruir aquesta transformació.

Per fer aquesta prova, s'aprofita el dataset SRD creant noves imatges de GT. Aquestes corresponen a la imatge original, però amb la transformació aplicada al canal R on els valors d'aquest són la meitat del valor original, es divideix el canal R entre 2. A la *figura 7* hi podem observar un exemple, la parella d'imatges que formen aquest nou dataset són l'Input i el GT.

En la mateixa figura es pot veure també la sortida del model, que és molt precisa. Aquesta millora en els resultats no es deu només a la simplicitat de la tasca, sinó

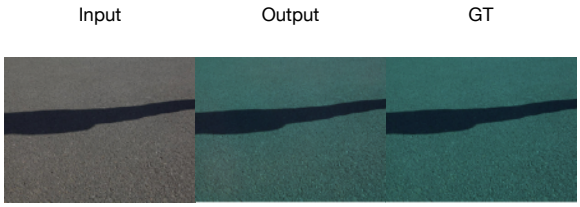

Figura 7. Seqüència Input, Output, i GT per una imatge. Tasca senzilla, transformació canal R.

també a la detecció prèvia d'errors de compatibilitat entre les versions de les llibreries i els drivers dels clústers on s'han executat algunes de les proves. Aquests errors s'han determinat com la causa de les imatges abstractes generades pel model en les primeres proves *(figura 3)*. No obstant això, aquests problemes no són la causa dels mals resultats de la prova amb pesos preentrenats *(figura 6)*, ja que l'entorn on s'ha executat aquesta última no presentava aquestes incompatibilitats.

Exposant aquestes proves al tutor del projecte és quan sorgeix la hipòtesi dels problemes d'incompatibilitat que s'acaben demostrant. Però no només això, també es planteja un malentès pel que fa a aquesta nova tasca, que s'havia acordat conjuntament al principi de la setmana. La tasca que es duu a terme no és realment la que s'havia acordat, havia de ser una mica més complexa i la idea no era que el model l'aprengués a replicar, sinó tot el contrari, que aprengués a reconstruir la imatge original donada una imatge amb transformacions. Per aquests motius, tot i que en certa manera s'ha complert l'objectiu, es continuarà treballant en la mateixa línia la pròxima setmana, tot i que la tasca serà una mica més complexa.

- ## Setmana 8

Objectiu: entrenar una tasca senzilla i identificar possibles problemes.

Com ja s'ha esmentat, aquesta setmana es manté el mateix objectiu, però amb una tasca una mica més complexa per poder detectar possibles errors. La tasca en qüestió, consisteix a fer que el model aprengui a reconstruir la imatge original donada una imatge d'entrada amb transformacions relativament aleatòries. Aquestes consistiran a dividir un dels canals, com ja es feia anteriorment, però en aquest cas, ja no serà sempre el vermell, se selecciona el canal de manera aleatòria i la divisió ja no és entre dos sinó entre un nombre també aleatori entre el 2 i el 4, ambdós inclosos. Aquest interval està limitat per no complicar excessivament la tasca, però es pot ampliar si cal. Aquestes transformacions s'apliquen cada cop que es vol obtenir un ítem del dataloader, permetent al model veure una mateixa imatge amb diverses transformacions.

Com que les imatges del dataset anterior no presenten molta variació de colors on es pugui apreciar amb detall la precisió de les reconstruccions del model, s'opta per canviar de dataset. El nou conjunt de dades és el dataset de PyTorch anomenat Flowers102 que està pensat originalment per a tasques de classificació, ja que conté diferents tipus de flors. Per a aquesta tasca, no s'utilitzen totes les imatges proporcionades; s'agafen algunes de les diferents tipologies de flors fins a aconseguir 8189 imatges, aquestes seran el GT.

En aquest punt del projecte les proves que es fan són de 20000 iteracions amb validació cada 1000. A la *figura 8*, es mostren els resultats de la validació per les iteracions 18000 i 19000 respectivament, ja que en les últimes iteracions hi ha sorgit un problema amb el guardat d'imatges, on

Figura 8. Seqüència Input, Output, i Ground Truth per dos imatges. Reconstrucció imatge amb un dels canals alterat.

apareixen desordenades i, en alguns casos, no apareixen.

Més enllà del problema en guardar les imatges sembla que tot funciona correctament, el model aprèn i torna uns resultats coherents. Tanmateix, els resultats encara són millorables, tot i que van pel bon camí. Serà necessari ajustar alguns paràmetres per aconseguir resoldre de manera més eficient una tasca tan senzilla com aquesta.

- **Setmana 9**

Objectiu: ajustar paràmetres per millorar el procés d'entrenament.

Després de les proves de la setmana anterior, es realitza un ajust al codi per aplicar sempre la mateixa transformació a les imatges durant la validació, permetent així una millor comparació de l'impacte de canviar alguns paràmetres. En l'entrenament continuaran essent aleatoris, dins dels llindars ja esmentats.

A més, es modifica el codi perquè la validació es pugui fer també amb un batch size superior a 1, en l'ajustar el càlcul de les mètriques perquè continuï sent correcte amb aquests canvis, es detecta que hi havia errors en algunes de les mètriques. És per això que en les proves anteriors no es mostren aquests valors. Es fan també uns últims canvis en el codi per evitar que se sobreescriguin les dades que es van guardant del model quan es fan execucions simultànies i s'arregla la manera en què es guarden imatges per no perdre'n cap i que no es desordenin com passava en les proves anteriors.

Un cop s'ha organitzat el codi, s'executen diverses proves variant el valor de T, que representa el nombre d'iteracions que el model realitza per generar la sortida durant la validació. En cadascuna d'aquestes iteracions es prediu el soroll que hi ha en la imatge i se li treu, el procés s'explica amb més detall en la secció 6. Es fan proves amb valors de T iguals a 10, 100 i 100, a més amb T=100 es prova entrenat des de zero i entrenat partint dels pesos de deraining. Es fa per 50000 iteracions amb validació cada 5000.

En la *figura 9,* es mostra l'últim batch de validació per les diferents proves, cada columna correspon a una mostra del batch. Pel que fa a les files, la primera són els inputs, la segona són els outputs i la darrera són els GTs. D'altra banda, en la *figura 10*, s'hi pot veure l'evolució de les mètriques en cada punt de validació al llarg de les iteracions.

Tot i no destacar un resultat de manera molt clara per sobre de la resta, el model amb T=10 sembla ser la millor elecció. En el darrer batch, aquest model demostra una precisió lleugerament superior i les mètriques també es mantenen sovint, lleugerament per sobre de la resta. Excepte la loss, que en aquest cas, encara que sigui superior no s'ha de tenir en compte, ja que en funció del valor de T aquesta es troba en rangs diferents, s'hauria de comparar el grau de disminució proporcionalment. Com més petit el valor de T, més grans són els valors de la loss, en aquest cas 10 i 100 vegades més grans. A més, amb aquest valor de T, el model és dos i tres vegades més ràpid respecte als altres. Pel que fa a la diferència entre entrenar des de scratch i partint d'uns pesos preentrenats, tot i que siguin per una

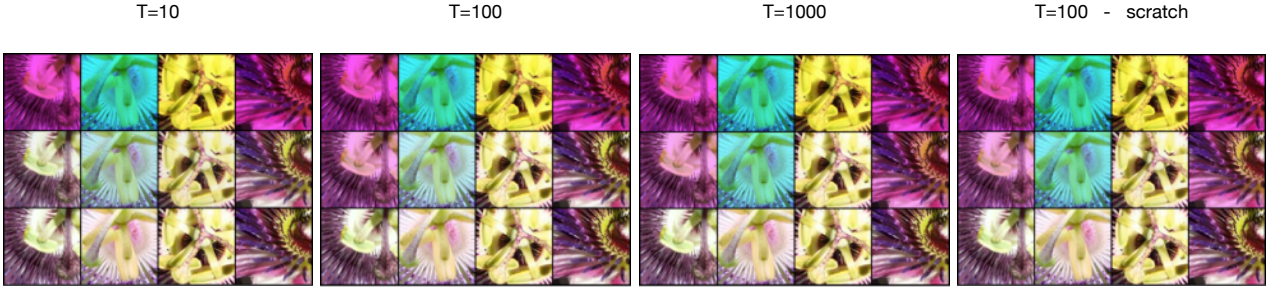

Figura 9. Batch de 4 imatges per cada valor de T. Cada columna correspon a una mostra, la primera fila correspon a l'Input, la segona a l'Output i la tercera al GT.

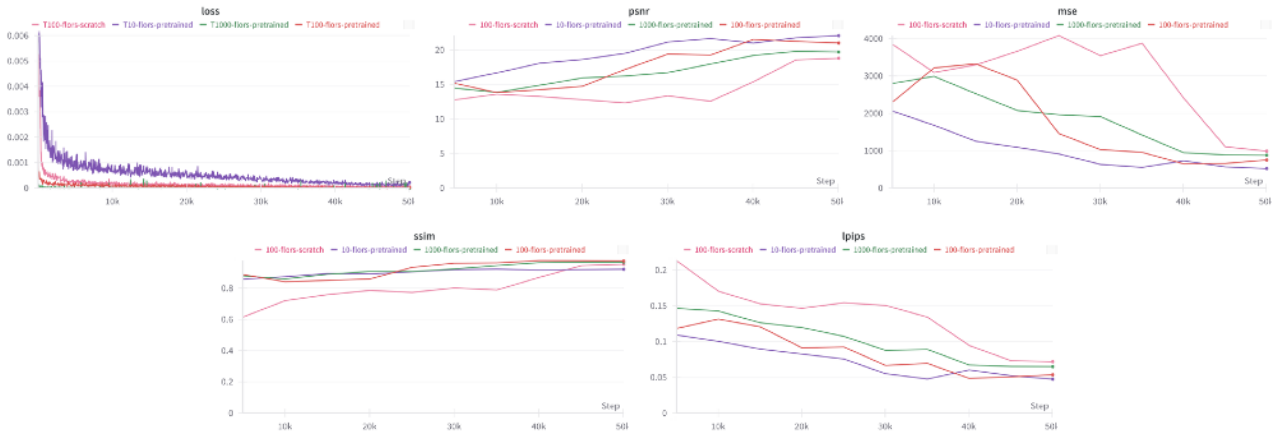

Figura 10. Mètriques de les diferents proves al llarg de les 50k iteracions.

tasca completament diferent, s'aprecia que per 50000 iteracions els resultats acaben essent pràcticament els mateixos. Tanmateix, és cert que al model que comença des de zero li costa més arribar-hi i amb menys iteracions la diferència hauria estat notable.

No obstant això, i que els resultats encara poden acabar de millorar, pràcticament tots han aconseguit resoldre la tasca amb èxit.

Es pot afirmar que s'ha assolit l'objectiu de la setmana, ja que es fan diverses proves i es veu l'efecte que tenen. També s'assoleix un entrenament que es pot considerar efectiu.

- ## Setmana 10

Objectiu: entrenar una tasca de relighting sense condicionants.

Després d'obtenir bons resultats en l'entrenament d'una nova tasca, ara es vol dirigir l'entrenament cap a l'objectiu final. En aquest cas, es busca entrenar el model per a una tasca de relighting sense cap condicionant, és a dir, donades diverses escenes amb diferents il·luminacions, es vol que el model retorni les mateixes escenes amb una il·luminació específica.

Per dur a terme aquesta tasca es fa servir el dataset NeRF[17], aquest consta de 20 escenes, cadascuna sota diferents condicions d'il·luminació i des de diferents angles. En aquest cas, s'utilitza sempre el mateix angle, i es fa l'entrenament amb totes les condicions d'il·luminació disponibles excepte una, que s'utilitza com a ground truth.

En les primeres proves, els resultats obtinguts són poc satisfactoris, com es pot veure a la *figura 11*. Es detecta que el model sembla estar perdut. Una de les hipòtesis és que aquesta situació pot ser causada pel crop que s'aplica a les imatges, ja que pot provocar una pèrdua important de context que és essencial per comprendre les ombres i reflexos. Per solucionar aquest problema, es decideix substituir el crop per un resize, i s'aconsegueixen els resultats de la *figura 12*.

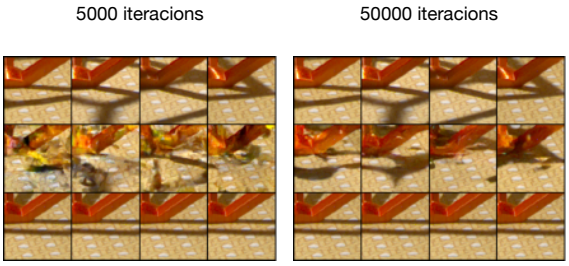

Figura 11. Batch de 4 imatges quan el model portava 5000 i 50000 iteracions. Cada columna correspon a una mostra, la primera fila correspon a l'Input, la segona a l'Output i la tercera al GT.

Amb el resize, els resultats canvien dràsticament, tal com s'esperava. Amb només 5000 iteracions, el model ja pot dur a terme la tasca de manera molt precisa. No obstant això, donades les poques imatges disponibles i les característiques del problema, hi ha sospites que el model pugui estar fent overfitting.

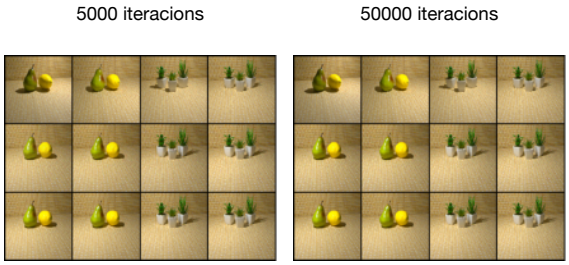

Figura 12. Batch de 4 imatges quan el model portava 5000 i 50000 iteracions. Cada columna correspon a una mostra, la primera fila correspon a l'Input, la segona a l'Output i la tercera al GT.

L'objectiu plantejat per aquesta setmana s'assoleix parcialment, ja que s'ha entrenat una tasca de relighting sense condicionants. Així i tot, com s'ha comentat, hi ha indicis clars d'overfitting, per la qual cosa cal continuar treballant per aconseguir realment aquesta meta.

- **Setmana 11**

Objectiu: entrenar una tasca de relighting sense condicionants.

Per la feina d'aquesta darrera setmana contemplada en aquest informe, es persegueix el mateix objectiu de la setmana anterior, es busca solucionar els problemes que van sorgir.

Com s'ha mencionat la possibilitat d'overfitting, es pren la decisió de verificar com es comporta el model quan se li presenta una escena que no ha vist prèviament. Com que s'han utilitzat totes les escenes en l'entrenament se'n crea una de nova només amb aquesta finalitat, els resultats obtinguts es mostren en la *figura 13*.

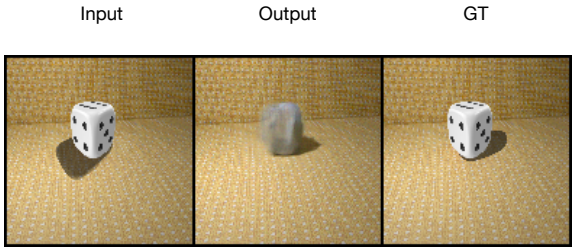

Figura 13. Seqüència Input, Output, i Ground Truth. Escena no vista en l'entrenament.

En aquesta figura, es pot observar com el model té una comprensió general sobre com canviar les ombres, però no és capaç de reconstruir objectes que no ha vist abans. Per tal d'analitzar millor el comportament del model davant d'escenes desconegudes i com evoluciona amb més iteracions, es repeteix l'entrenament deixant de banda dues de les escenes del dataset, s'obté el que es mostra en la *figura 14*.

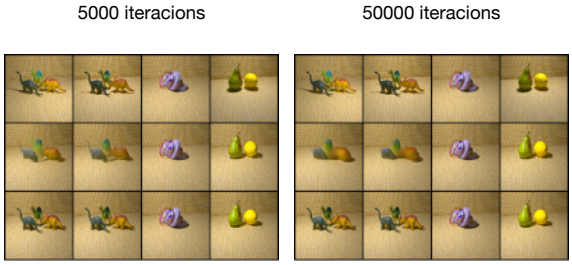

Figura 14. Batch de 4 imatges quan el model portava 5000 i 50000 iteracions. Cada columna correspon a una mostra, la primera fila correspon a l'Input, la segona a l'Output i la tercera al GT.

Es constata que a mesura que augmenten les iteracions, també ho fa l'overfitting i la dificultat del model per recrear escenes noves. Per mitigar aquesta situació, calen més dades que ajudin al model a generalitzar, tot i que trobar un gran volum de dades amb aquestes característiques no és una tasca senzilla. Per tant, de cara a les pròximes proves i de manera paral·lela a la busqueda d'altres datasets s'opta per buscar maneres de forçar la xarxa a generalitzar.

En aquest punt, no es pot considerar que s'hagi aconseguit l'objectiu, ja que encara no s'han resolt completament els problemes detectats. Es pren la decisió de continuar treballant en aquest objectiu durant almenys una altra setmana.

8 REFERÈNCIES

[1] Yixiong Yang, Hassan Ahmed Sial, Ramon Baldrich, Maria Vanrell (Juliol 2023) "*Relighting from a Single Image: Datasets and Deep Intrinsic-based Architecture*".

[2] Steins (2023) "*Stable diffusion clearly explained!*", *Medium*. Disponible a: https://

medium.com/@steinsfu/stable-diffusion-clearly-explained-ed008044e07e#0b5c (Accés: 05 Febrer 2024).

[3] Jamil, U. (2023) "*Coding stable diffusion from scratch in Pytorch"*, *YouTube*. Disponible a: https://www.youtube.com/watch?v=ZBKpAp_6TGI&t=2185s (Accés: 08 Febrer 2024).

[4] Adaloglou, N. and Karagiannakos, S. (2022) "*How diffusion models work: The math from scratch*, *AI Summer*." Disponible a: https://theaisummer.com/diffusion-models/ (Accés: 07 Febrer 2024).

[5] Lin, X. *et al.* (2023) "*Diffbir: Towards blind image restoration with generative diffusion prior"*, *arXiv.org*. Disponible a: https://arxiv.org/abs/2308.15070 (Accés: 13 Febrer 2024).

[6] Luo, Z. et al. (2023) "*Refusion: Enabling large-size realistic image restoration with latent-space diffusion models*", arXiv.org. Disponible a: https://arxiv.org/abs/2304.08291 (Accés: 15 Febrer 2024).

[7] Lin, X. *et al.* (2023) "*XPixelGroup/Diffbir: Official Codes of diffbir: Towards blind image restoration with generative diffusion prior"*, GitHub. Disponible a: https://github.com/XPixelGroup/DiffBIR/tree/main (Accés: 20 Febrer 2024).

[8] Luo, Z. et al. (2023) "*ALGOLZW/Image-Restoration-SDE: Image Restoration with mean-reverting stochastic differential equations, ICML 2023. winning solution of the NTIRE 2023 image shadow removal challenge."*, GitHub. Disponible a: https://github.com/Algolzw/image-restoration-sde?tab=readme-ov-file (Accés: 21 Febrer 2024).

[9] Rombach, R. *et al.* (2021) "*COMPVIS/stable-diffusion: A latent text-to-image diffusion model"*, GitHub. Disponible a: https://github.com/CompVis/stable-diffusion (Accés: 09 Març 2024).

[10]Goyal, S. (2019) "*Gans - a brief introduction to generative Adversarial Networks*, *Medium*." Disponible a: https://medium.com/analytics-vidhya/gans-a-brief-introduction-to-generative-adversarial-networks-f06216c7200e (Accés: 18 Febrer 2024).

[11]Zhu, D. (2020) "*Generate images using variational Autoencoder (VAE)*, *Medium*." Disponible a: https://medium.com/@judyyes10/generate-images-using-variational-autoencoder-vae-4d429d9bdb5 (Accés: 18 Febrer 2024).

[12]Luo, Z. et al. (2023) *"Image restoration with mean-reverting stochastic differential equations",* arXiv.org. Disponible a: https://arxiv.org/abs/2301.11699 (Accés: 26 Febrer 2024).

[13]Sara, U., Akter, M. and Uddin, M.S. (2019) "*Image quality assessment through FSIM, SSIM, MSE and PSNR-A comparative study, SCIRP"*. Disponible a: https://www.scirp.org/journal/paperinformation?paperid=90911 (Accés: 12 Març 2024).

[14]Monsters, D. (2020) A quick overview of methods to measure the similarity between images, Medium. Disponible a: https://medium.com/@datamonsters/a-quick-overview-of-methods-to-measure-the-similarity-between-images-f907166694ee (Accés: 12 Març 2024).

[15]Peak signal-to-noise ratio (2024) Wikipedia. Disponible a: https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio (Accés: 12 Març 2024).

[16]Structural similarity index measure (2024) Wikipedia. Disponible a: https://en.wikipedia.org/wiki/Structural_similarity_index_measure (Accés: 12 Març 2024).

[17]Toschi, M. et al. (2023) Relight my nerf, ReNé. Available at: https://eyecan-ai.github.io/rene/ (Accessed: 08 April 2024).

ANNEX

ID	Objectiu
0	Guanyar una comprensió exhaustiva del funcionament dels models de difusió i en concret del codi escollit.
1	Fer un primer entrenament amb el dataset donat, des de scratch i partint dels pesos donats.
2	Establir ordre i fer un entrenament per a una tasca més senzilla.
3	Entrenar una tasca senzilla i identificar possibles problemes.
4	Ajustar paràmetres per millorar el procés d'entrenament.
5	Entrenar una tasca de relighting sense condicionants.
6	Entrenar una tasca de relighting amb condicionants.
7	Anàlisi de resultats i possibles millores.

Figura 15. Taula d'objectius.

Setmana	Data/es	ID Objectiu/s	Assoliments
1	05/02 - 11/02	0	Familiarització amb conceptes bàsics de SD [2], [3], [4] i l'article de relighting [1].
2	12/02 - 18/02	0	Millor comprensió dels models SD. Comprensió del context actual de SD i concretament dels models SD de restauració d'imatges [5], [6].
3	19/02 - 25/02	0	Consolidació de coneixements base en SD. Proves dels codis més rellevants i afins als projectes trobats. [7], [8], [9] Tria del codi més adequat de cara a la tasca final.
4	26/02 - 03/02	0	Comprensió a trets generals del codi escollit. Primeres proves del funcionament del codi. Detecció i millora dels primers errors trobats.
5	04/03 - 10/03	0, 1	Millor comprensió del procés d'entrenament en el codi. Primer entrenament complet.
10/03/2024		***Informe Inicial***	

Setmana	Data/es	ID Objectiu/s	Assoliments
6	11/03 - 17/03	0, 2	Tria de mètriques a recollir. Wandb en funcionament per recollir aquestes mètriques.
7	18/03 - 24/03	0, 3	Detecció de problemes importants. Entrenament d'una tasca molt senzilla amb èxit.
8	25/03 - 31/03	0, 3	Entrenament d'una tasca una mica més complexa. Detecció d'errors en el WandB.
9	01/04 - 07/04	0, 4	Canvis en els paràmetres i anàlisi de l'impacte en l'entrenament. Millores pràctiques de WandB. Entrenament de la tasca més fructífer.
10	08/04 - 14/04	0, 5	Detecció de problemes en l'entrament d'una tasca relighting sense condicionants
11	15/04 - 21/04	0, 5	Constatació d'un problema d'overfitting.
21/04/2024		***Informe de Progrés I***	
12	22/04 - 28/04	0, 5	-
13	29/04 - 05/05	0, 5	-
14	06/05 - 12/05	0, 6	-
15	13/05 - 19/05	0, 6	-
16	20/05 - 26/05	0, 6	-
26/05/2024		***Informe de Progrés II***	
17	27/05 - 02/06	0, 6	-
18	03/06 - 09/06	0, 7	-
19	10/06 - 16/06	0, 7	-
16/06/2024		***Proposta Informe Final***	
20	17/06 - 23/06	0, 7	-
21	24/06 - 30/06	0,7	-
30/06/2024		***Proposta de Presentació***	
30/06/2024		***Dossier i Informe Final***	
01/07/2024		***Pòster***	
03/07/2024-10/07/2024		***Defensa TFG***	

Figura 16. Taula de planificació del projecte.