УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

МАСТЕР РАД

на катедри за

Рачунарство и информатику

на тему

Примена неуронских поља зрачења у рендеровању

Коста Грујчић

Београд, 11. децембар 2022.

Ментор:
проф. др Младен Николић
Универзитет у Београду, Математички факултет

Чланови комисије:
проф. др Младен Николић
Универзитет у Београду, Математички факултет

проф. др Младен Николић
Универзитет у Београду, Математички факултет

проф. др Младен Николић
Универзитет у Београду, Математички факултет

Датум одбране: 11. децембар 2022.

Посвета

Наслов мастер рада: Примена неуронских поља зрачења у рендеровању
Резиме:
Кључне речи: машинско учење, неуронска поља, рендеровање

Списак слика

Списак табела

Садржај

Глава 1

Увод

Један од споредних циљева овог рада је превођење израза који су се безразложно одомаћили у српској научној заједници као англицизми. С друге стране, скраћенице остају у изворном облику и то на латиници, што читаоцу пружа могућност за непосредно претраживање навода.

Глава 2

Основни појмови рачунарске графике

2.1 Светлост и боја

Светлост представља електромагнетно зрачење чија је таласна дужина у сегменту од око 350 до 700 nm, које побуђује визуелни систем човека. То значи да људи *не виде* светлост осталих таласних дужина, тако да ћемо убудуће под видљивим спектром светлости мислити управо на овај, видљив човеку.

Потребно је увести физичке величине којима се могу квантификовати основна физичка својства светлости. Прво ћемо дефинисати радиометријске величине, а онда одговарајуће фотометријске.

Дефиниција 2.1.1. *Фотон је квант електромагнетног зрачења. Енергија фотона је:*

$$E = \frac{hc}{\lambda}\,[\mathrm{J}]\,,$$

где је h Планкова константа, c брзина светлотси, а λ таласна дужина фотона.

Дефиниција 2.1.2. *Укупна енергија зрачења извора зрачења је:*

$$Q_e = \int_{S^2} E d\Omega\,[\mathrm{J}]\,.$$

Дефиниција 2.1.3. *Флукс зрачења је укупна енергија зрачења која доспе на неку површину по јединици времена:*

$$\Phi_e = \frac{\partial Q_e}{\partial t}\,[\mathrm{W}]\,.$$

Дефиниција 2.1.4. *Озрачење је укупан флукс зрачења по јединици површине:*

$$E_e = \frac{\partial \Phi_e}{\partial A}\left[\frac{\mathrm{W}}{\mathrm{m}^2}\right].$$

Дефиниција 2.1.5. *Јачина зрачења је укупан флукс зрачења у неком смеру по јединичном просторном углу:*

$$I_{e,\Omega} = \frac{\partial \Phi_e}{\partial \Omega}\left[\frac{\mathrm{W}}{\mathrm{sr}}\right].$$

Дефиниција 2.1.6. *Зрачење је је укупан флукс зрачења у неком смеру по јединици површине и јединичном просторном углу:*

$$L_{e,\Omega} = \frac{\partial^2 \Phi_e}{\partial \Omega \partial A} \left[\frac{\mathrm{W}}{\mathrm{sr} \cdot \mathrm{m}^2} \right].$$

Спектралне величине се дефинишу у односу на таласну дужину. На пример, спектрално зрачење је $L_{e,\Omega,\lambda} = \frac{\partial L_{e,\Omega}}{\partial \lambda}$. Аналогно и за остале.

Након дефинисања мноштва физичких величина, коначно долазимо и до фотометријских, или како се још називају и *визуелне*. Кључна разлика у односу на радиометријске, или како се још називају и *енергетске*, је што се у овом случају у обзир узима и спектрална осетљивост посматрача. То одговара интуитивном поимању светлости - човек светлост разликује на основу боје, што је у директној коресподенцији са таласном дужином.

У основи спектралне зависности посматрача је ***функција релативне светлосне осетљивости*** V. Помоћу ове функције изражавамо просечну осетљивост човека на светлост одређене таласне дужине. Просечна је у смислу да може варирати у популацији, али представља врло добру апроксимацију у општем случају, поготово имајући у виду да је стандардизована од стране Међународног комитета за осветљење.

Дефиниција 2.1.7. *Светлосни флукс је укупна енергија која протекне кроз неку површину у јединици времена:*

$$\Phi_v = K \int_0^\infty \Phi_{e,\lambda} V(\lambda) d\lambda \, [\mathrm{lm}] .$$

За вредност онстанте K из 2.1.7 се узима $683.002 \frac{\mathrm{lm}}{\mathrm{W}}$. Реч је о још једној примеру стандардизације у области фотометрије.

Дефиниција 2.1.8. *Јачина светлости је укупна снага коју емитује извор светлости у одређеном смеру по јединичном просторном углу:*

$$I_v = K \int_0^\infty I_{e,\lambda} V(\lambda) d\lambda \, [\mathrm{cd}] .$$

Дефиниција 2.1.9. *Осветљеност је укупан светлосни флукс на некој површини:*

$$E_v = K \int_0^\infty I_{e,\lambda} E_{e,\lambda} V(\lambda) d\lambda \, [\mathrm{lx}] .$$

Дефиниција 2.1.10. *Сјајност је укупан светлосні флукс који напушта, пролази или пада на површину по јединичном просторном углу и по ортогоналној пројекцији јединичне површине:*

$$L_v = \frac{\partial^2 \Phi_v}{\partial \Omega \partial A \cos \theta} \left[\frac{\mathrm{cd}}{\mathrm{m}^2} \right].$$

Сјајност је једина фотометријска величина коју човек непосредно опажа. Она представља мерило за субјективни утисак о мањој или већој сјајности светлеће или осветљене површине.

Већ смо поменули да људи разликују неке таласне дужине светлости, односно виде светлост одређене боје. Ту способност нам дају три врсте чепића који се налазе у жутој мрљи мрежњаче ока. Према томе, кроз ове рецепторе, наш мозак прима свега три врсте сигнала за сваки очни надражај. Међутим, људско око није идеалан спектрометар, па поједине таласе просто види као светлост исте боје.

Зато се боја на рачунару представља тројкама (R, G, B) где су координате удео црвене, плаве и зелене боје редом, које се узимају за основне. Постоје и други системи боја попут HSV или CMYK, али о њима неће бити речи јер ћемо надаље користити искључиво RGB систем.

На овај начин је могуће чувати тачно 255^3 боја, што је и више него довољно будући да људско око разликује до 10 милиона боја. Даља практична ограничења се тичу квалитета монитора као и графичког процесора, али се тиме нећемо бавити.

2.2 Камера

Можемо сматрати да нам је појам камере познат из стварног света. У овој глави ћемо строго дефинисати камеру и дати један од многобројних начина њеног моделовања у домену рачунарске графике.

Дефиниција 2.2.1. *Камера је пресликавање из* $\mathbb{R}^3$ *у* $\mathbb{R}^2$.

У питању је врло општа, готово бескорисна дефиниција. Међутим, ако размислимо о томе да се стваран свет врло добро може представити као простор димензије 3, а да се фотографија може схватити као раван, можемо увидети да је у питању заиста исправна формулација. Оно што овом дефиницијом није обухваћено јесте како се тачно од света око нас долази до слике. Зато има смисла говорити о моделима камере као различитим парадигмама изведбе поменутог пресликавања. У овом раду ће највише бити коришћен тачкасти модел камере тако да ћемо у наставку поглавља дати његову прецизну дефиницију.

2.2.1 Тачкасти модел камере

Посматрајмо канонски Еуклидски простор димензије 3 и раван $z = f$ коју ћемо звати ***раван слике***. У овом моделу камере се произвољна тачка $\mathbf{x_w} = (x, y, z)$ из простора пресликава у тачку $\mathbf{x_p} = (u, v)$ која је тачка пресека равни слике и праве која спаја $\mathbf{x_w}$ и центар камере $\mathbf{c}$, који ћемо за сад поставити у координатни почетак. Другим речима, у питању је централна пројекција са центром у координатном почетку. Тривијалном применом сличности троуглова долазимо до

$$(x, y, z) \mapsto (fx/z, fy/z). \tag{2.1}$$

Права која пролази кроз центар камере и нормална је на раван слике називамо ***главном осом***, а тачка у којој се главна оса и раван слике секу називамо ***главном тачком***.

Приметимо још једну особину централног пројектовања - све тачке праве која пролази кроз центар камере се пројектују у исту тачку равни слике. Зато ћемо увести хомогене координате. Отуда можемо писати $\mathbf{x_p} = P\mathbf{x_w}$. Овакво пресликавање се може преписати и у матричном облику

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \mapsto \begin{bmatrix} fx \\ fy \\ z \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}. \tag{2.2}$$

Ради конзистентности са наставком, растављећемо P на прозивод

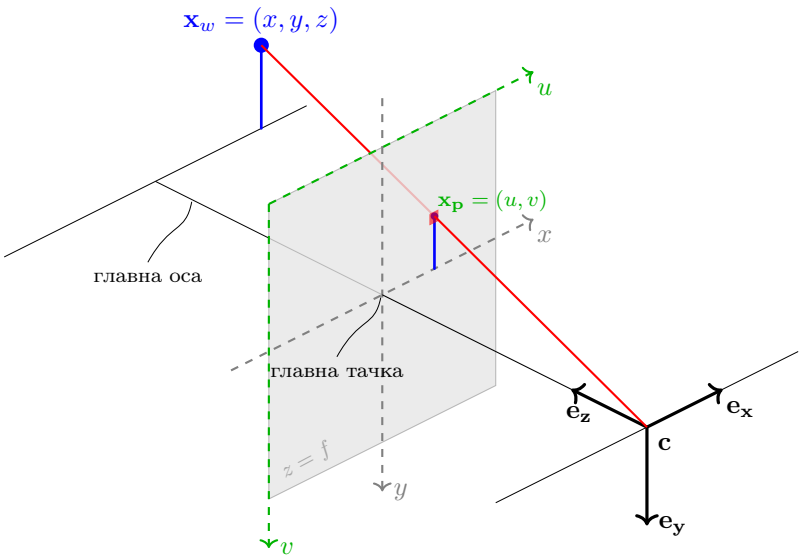
$$K \begin{bmatrix} R & t \end{bmatrix}, \tag{2.3}$$

где је $K \in M_{3,4}(\mathbb{R})$, $R \in M_{3,3}(\mathbb{R})$ и $t \in M_{3,1}(\mathbb{R})$. Уклањајући последњу колону из 2.2 добијамо матрицу K из овако измењеног пресликавања.

Уколико уопштимо положај главне тачке и њене координате означимо са (p_x, p_y), матрица K поприма облик

$$\begin{bmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.4}$$

Коначно, можемо уопштити и положај центра камере. Тада уочавамо два координатна система - онај с почетка, канонски, у ком нам је лако да баратамо и други, са центром камере као координатним почетком. Није тешко увидети да се кретањем (ротацијом и транслацијом) репер камере може довести до канонског. На тај начин употпуњујемо пресликавање P. На слици 2.1 је дата интепретација овог пресликавања.



Слика 2.1: Интерпретација тачкастог модела камере

Постоји укупно 9 слободних параметара - по 3 за сваку од матрица K, R и t. Матрица K се назива ***матрица калибрације***, а њене вредности ***унутрашњим параметрима*** камере. Вредности матрице $\begin{bmatrix} R & t \end{bmatrix}$ се називају ***спољашњим параметрима*** камере.

Поменимо још једну терминолошку конвенцију. Координатни систем из којег вршимо пројекцију се назива и ***светским координатним системом***, а координатни систем индукован положајем камере ***координатни систем камере***. ***Координатни систем слике*** је онај везан за раван слике.

Корисно је дати додатни коментар у вези са нулом у првом реду матрице калибрације. У питању је коецифијент смицања којим је могуће уопштити раван слике – правоугаоник постаје паралелограм.

2.3 Рендеровање

Када смо говорили о камери, видели смо да слика није ништа друго до дводимензиона пројекција стварног света. Слично је и када комплексну сцену, овог пута виртуелног, света желимо да прикажемо на монитору. Мора доћи до пројектовања света, а слику коју видимо је уствари слика добијена посматрањем света кроз виртуелну камеру. Тај процес се назива ***рендеровање***. Дакле, рендеровањем се од сцене долази до конкретне репрезентације слике у пикселима.

Пре него што се упустимо у рендеровање, вратимо се корак уназад. Светлост и камеру смо, на први поглед, увели врло неповезано. Светлост смо дефинисали више експериментално ослањајући се на физичке законе, док смо камеру увели строго геометријски. Сада ћемо оба појма ставити у контекст и објаснити њихову везу.

Праве које смо уочавали у моделу камере су светлосни зраци. Раван слике у пракси није бесконачна, већ има своју ширину и дужину којима се одређују димензије слике. Такође, услед физичких ограничења слика се мора дискретизовати. Посматрамо ли слику на тај начин, видећемо да је она матрица чије елементе називамо ***пикселима***. Како ће који пиксел да изгледа, односно које ће боје бити, одређује количина зрачења која допре до камере. Зрачење смо увели у зависности од правца, што у комбинацији са пројективним особинама модела камере одређује који ће пиксел *бити погођен*. Такође, природно је увести ограничење по питању дела простора који камера може да опажа. Величину слике смо већ поменули, али не и параметар по z оси. Ради једноставности геометрије, уводимо ***предњу и задњу раван одсецања***. Све испред предње и иза задње равни одсецања не утиче на резултујућу слику.

2.3.1 Запреминско рендеровање

Посматрајмо шта се дешава са светлости приликом проласка кроз неку средину. Може доћи до:

- упијања – средина упија фотоне светлосног зрака приликом чега се ослобађа топлота или неки други вид енерије.
- емисије – како је светлост зрачење, пролазак светла кроз средину је загрева. Када средина достигне одрђену температуру, може доћи до емитовања светлости.

- расипања – део фотона напушта правац зрака што доводи до мешања фотона са различитих праваца.

Према томе, више чинилаца утиче на коначно зрачење које ће путем неког светлосног зрака доћи до камере. Тачке неког светлосног зрака с почетком у $\mathbf{o}$ и правцем $\mathbf{d}$ једнозначно одређујемо као $\mathbf{r}_{\mathbf{o},\mathbf{d}}(t) = \mathbf{o} + t\mathbf{d}$.

Нека су предња и задња раван одсецања редом $z = t_n$ и $z = t_f$, а светлосне зраке посматрамо из положаја камере $\mathbf{c}$. Према томе, до камере дуж зрака $\mathbf{r}_{\mathbf{c},\mathbf{d}}$ допире следећа количина зрачења

$$\int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}_{\mathbf{c},\mathbf{d}}(t))L_{e,\Omega}\big|_{\mathbf{r}_{\mathbf{c},\mathbf{d}}(t)}dt, \tag{2.5}$$

где је $T(t) = \exp\left(-\int_{t_n}^{t}\sigma(\mathbf{r}_{\mathbf{c},\mathbf{d}}(s)ds)\right)$ акумулирана пропусност зрака од t_n до t. Ова величина представља вероватноћу да светлосни зрак на путу од t_n до t не удари нити у једну препреку. Напоменимо да се интеграција врши искључиво по фотонима који се налазе на светлосном зраку од интереса, или другим речима, интеграли се само по расутој светлости.

Како нам је боја пиксела крајњи циљ, једначину 2.5 ћемо преписати у колориметријском облику

$$C(\Pi_{\mathbf{r}_{\mathbf{c},\mathbf{d}}}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}_{\mathbf{c},\mathbf{d}}(t))C(\mathbf{r}_{\mathbf{c},\mathbf{d}}(t))dt, \tag{2.6}$$

где је $\Pi_{\mathbf{r}_{\mathbf{c},\mathbf{d}}}$ тачка пресека светлосног зрака и равни слике, а C поље које сваку тачку пресликава у њену RGB боју.

Приказани поступак одређивања боје пиксела се назива ***запреминско рендеровање***. У питању је само један од многобројних алгоритама за рендеровање који ће бити коришћен у наставку.

Глава 3

Основни појмови машинског учења

У овој глави ћемо формулисати теоријски оквир неопходан за разматрање и примену машинског учења.

Машинско учење је област *вештачке интелигенције*. Неформално говорећи, машинско учење обухвата алгоритме изведене из података или како се то често говори - научених из података. То значи да нема експлицитног програмирања, а врло често ни контроле процеса учења, већ се алгоритми дефинишу низом операција параметризованих на основу података који су изложени процесу обучавања. Процес обучавања, дакле, представља одређивање поменутих параметара којима се операције од значаја израчунавају на што исправнији начин. Приметимо да је та исправност прилично неодређен појам и зависи од примене, података и циља обучавања. Такође, врло се често не може ни квантификовати што проблем обучавања чини утолико тежим.

Имајући у виду досег овог рада, потребно је увести појам надгледаног учења. Реч је о врсти учења у којој су уз податке присутни и додатни подаци којима је директно могуће утврдити да ли је излаз алгоритма исправан или није. На пример, ако је потребно утврдити да ли је на датој слици мачка, уз слику би постојао једнобитни податак који то недвосмислено потврђује. Према томе, скуп података можемо посматрати као скуп одбирака $\mathcal{D} = \{x_i, y_i\}$, где x_i представља улаз, а y_i одговарајући излаз. То нам омогућава да посматрамо расподелу таквих одбирака, односно густину расподеле $p(x, y)$. Природно се намеће потреба за што приближнијем одређивању поменуте расподеле, што подразумева одређивање функције f којом успостављамо везу између одговарајућих парова x и y. Кандидата за f има несагледиво много, а нама је потребна она *најбоља*, при чему се овог пута то мора формално дефинисати.

Претпоставимо да су y_i узорковани из метричког простора. Зато можемо дефинисати *функцију грешке* $\mathcal{L}$ којом меримо квалитет апроксимације y_i вредношћу $f(x_i)$. Вреднујући одбирке у складу са својом густином дефинишемо *ризик*

$$R(f) = \mathbb{E}(\mathcal{L}(y, f(x))) = \int \mathcal{L}(y, f(x))p(x, y)dxy. \tag{3.1}$$

Проблем надгледаног учења покушава да дође до f за које се ризик минимизује. Начелно, обучавање се може извести по свим могућим функцијама f. Како је то практично немогуће, претрага функција се усмерава увођењем додатних претпоставки, односно рестрикција, скупа претраге. Ми ћемо посматрати функције параметризоване скупом параметара Θ, тако да минимизовање ризика посматрамо као

$$\min_{\Theta} R(f_\Theta). \tag{3.2}$$

У општем случају не мора постојати само један исправан излаз за конкретан улаз, што оправдава дефинисање ризика коришћењем заједничке расподеле $p(x, y)$. То за наше потребе неће бити неопходно. Наиме, посматраћемо условну расподелу $p(y|x)$.

Расподела $p(y|x)$ је ретко кад позната. Зато се спроводи стандардни статистички третман - ризик се мења емпиријским ризиком:

$$ER(f_\Theta, \mathcal{D}) = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} \mathcal{L}(y_i, f_\Theta(x_i)), \tag{3.3}$$

где је $\mathcal{S}$ узорак скупа $\mathcal{D}$.

Претпоставимо да се скуп параметара састоји од само једног вектора димензије свега 2, тј. $\Theta = \{\mathbf{w}\} = \{(w_1, w_2)\}$, а да је $x_i \in \mathbb{R}^1$. Имајући у виду израз 3.2, за кандидате функције f, између осталог, имамо $\sin(x_i w_1 + w_2)$, $\log(x_i w_1 w_2)$, $\exp(w_2 x_i^{w_1})$. Иако је проблем постављен као веома једноставан, а параметарском рестрикцијом начињен још једноставнијим, и даље се чини као веома тежак будући да немамо начин да за разумно коначно времена претражимо све такве кандидате. Зато је потребно посматрати тачно одређене класе функција f за које је то могуће, а међу најпознатијим су свакако *неуронске мреже*. Заправо, потребно је фиксирати конкретну архитектуру модела, а потом вршити оптимизацију параметара у складу са одабраном метриком.

3.1 Неуронске мреже

Дефиниција 3.1.1. *Неуронска мрежа је функција облика*

$$f_\Theta(\mathbf{x}) = \left(\prod_{i=1}^{L} a_i(\beta_i + W_i) \right)(x),$$

где је $\Theta = \{W_i\}_{i=1}^L$ *скуп параметара који се обучава,* $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$ *при чему је* $d_0 = \dim(x)$ *и* $\{a_i\}_{i=1}^L$ *функције које се примењују члан по члан. Функције* $\{a_i\}_{i=1}^L$ *називамо* ***активационим функцијама****. Параметар* L *називамо* ***бројем слојева неуронске мреже****.*

Активационе функције могу бити произвољне, докле год поштују услове димензионалности. Међутим, уколико су оне линеарне, тада неуронска мрежа постаје линеарна, односно постаје линеарна регресија. У циљу добијања што разноврснијих модела, за функције активације се узимају нелинеарне функције. Један од популарних избора је $ReLU(x) = \max\{0, x\}$ [4].

Неуронске мреже су познате и под именом *вишеслојни перцептрони*, па ће убудуће бити коришћена и скраћеница MLP (енг. *multilayer perceptron*).

Значајан теоријски резултат даје следећа теорема [5]

Теорема 3.1.1. *За сваку функцију* $F : \mathbb{R}^n \to \mathbb{R}^m$ *интеграбилну у Бохнеровом смислу и свако* $\epsilon > 0$*, постоји неуронска мрежа* f_Θ *са* $ReLU$ *активационим функцијама, тако да је* $\{d_i = \max\{n + 1, m\}\}_{i=1}^L$ *и важи*

$$\int_{\mathbb{R}^n} \|f_\Theta(x) - F(x)\| \, dx < \epsilon.$$

Ово за последицу има да се готово свака функција може произвољно добро апроксимирати неуронском мрежом, али како доказ није конструктиван није очигледно како таква неуронска мрежа изгледа.

Уколико су активационе функције диференцијабилне функције, то ће бити и неуронска мрежа. Зато се поступак минимизације емпиријског ризика спроводи градијентним спустом. Поступак се спроводи у две етапе - прво се неуронска мрежа примени над подацима пропагацијом унапред, а потом се пропагацијом уназад користећи правило ланца изврши ажурирање параметара.

Како је градијентни спуст оптимизација првог реда, није гарантовано да ће пронађени локални минимум бити и глобални. У пракси се показује да градијентни спуст даје веома добра решења.

Када је реч о имплементацији неуронских мрежа и конкретим детаљима њиховог обучавања, важно је разматрати хардверска ограничења. Наиме, ови модели могу бити веома комплексни где се ред величине броја параметара креће и до неколико милиона, па чак и до неколико милијарди. Зато се приликом обучавања прибегава разнородним техникама, често нумеричке или хеуристичке природе, којима се смањује укупно време обучавања или смањује утрошак меморије. Еклатантан пример је *стохастички градијентни спус*т [6] - параметри се ажурирају на основу само једног одбирка уместо целог скупа података. Овај метод је сушта супротност конвенционалном градијентном спусту, тако да се углавном врши компромис.

Глава 4

Неуронска поља зрачења

Ову главу почињемо са две важне дефиниције.

Дефиниција 4.1. *Поље је пресликавање $F : \mathbb{R}^n \to \mathbb{R}^m$. Специјално, поље је скаларно за $m = 1$.*

Дефиниција 4.2. *Неуронско поље је поље макар делимично параметризовано неуронском мрежом.*

Неуронско поље зрачења је посебан случај неуронског поља за $n = 5$ и $m = 3$. У питању је параметризација пресликавања које свакој тачки (x, y, z) придружује зрачење и пропусност и то за сваки правац одређен угловима θ и ϕ. Пропусност се може видети и као вероватноћа да се зрак *заустави* у тој тачки.

Размотримо шта добијамо оваквом поставком. Претпоставимо да је неуронско поље савршено обучено. То значи да можемо утврдити боју пиксела из сваког могућег угла гледања. Тако се сцена рендерује и из погледа који се нису нашли у скупу за обучавање.

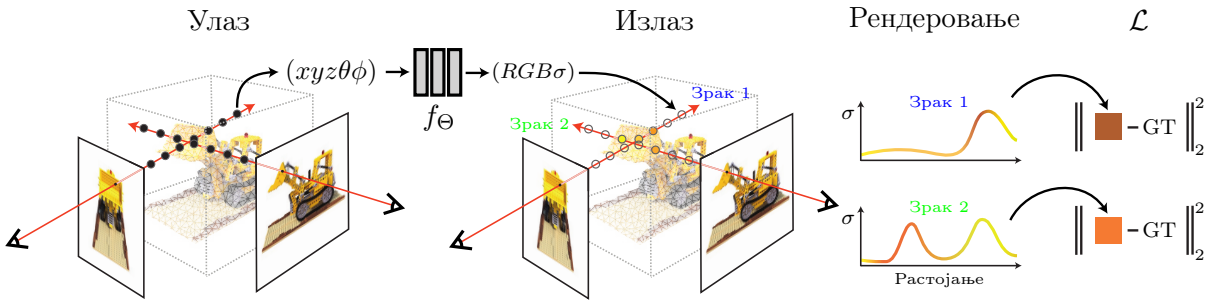
Како је таква неуронска параметризација непозната *a priori*, имамо слободу у виду дизајнирања архитектуре параметризације. Подробно ћемо обрадити оригинални NeRF модел и његово проширење Mip-NeRF, а потом се осврнути на изузетно оптимизован модел InstantNeRF, који се намеће као оптималан избор када је реч о имплементацији неуронских поља зрачења.

Оно што на први поглед издваја неуронска поља зрачења од осталих параметризација неуронским мрежама јесте обучавање. Узмимо за потребе илустрације ImageNet [7] скуп података. Поменути скуп је сачињен од великог броја троканалних слика, тако да је сваки узорак изворног скупа једноставан за представити. Међутим, скуп података којим ми баратамо је сачињен од различитих погледа, а уз сваки поглед је придружена одговарајућа изрендерована слика. Имајући у виду улаз неуронског поља зрачења, скупу података је неопходно увести још један слој гранулације – зрак са сваки угао гледања у одређеном погледу. Дакле, један одбирак из скупа података над којим се обучава неуронско поље зрачења се може видети као један светлосни зрак и то за један поглед уз пратећи рендер.

4.1 NeRF

NeRF (енг. *Neural Radiance Field*) је најстарији модел у фамилији модела који ћемо у овом раду обрадити. Заснива се на запреминском рендеровању и прва је изведба неуронских поља зрачења овог типа.

На слици 4.1 се може видети илустрација овог модела.



Слика 4.1: Илустрација NeRF модела према [1]

4.1.1 Параметризација

Неуронско поље зрачења се параметризује вишеслојним перцептроном, али на благо неуобичајен начин. Јасно је да σ не зависи од угла гледања већ искључиво од положаја $\mathbf{x}$. С друге стране, боја пиксела зависи од положаја као и од пропусности. Зато ReLU MLP од 8 слојева са по 256 неурона на улазу добија само $\mathbf{x}$, док је излаз предвиђена вредност σ и вектор димензије 256. Тај вектор се спаја са параметрима угла гледања и пропушта кроз један слој са 128 неурона и ReLU активационом функцијом. На овај начин се добија боја пиксела која је условљена погледом.

Чак и уз овако дефинисан поступак добијања коначног излаза, улаз и даље има малу димензију. Да би се превазишао тај проблем, улази се прво пресликају у простор веће димензије. Према томе, неуронску мрежу F_Θ можемо видети као композицију $F'_\theta \circ \gamma$ где се γ не обучава. У овом случају је $\gamma : \mathbb{R} \to \mathbb{R}^{2D}$ и то конкретно

$$\gamma(x) = (\sin(2^0\pi x), \cos(2^0\pi x), ..., \sin(2^{D-1}\pi x), \cos(2^{D-1}\pi x)). \tag{4.1}$$

Исто пресликавање се користи за све улазе које MLP добија, с тим што се за положај узима $D = 10$, а за поглед $D = 4$. Овај начин *подизања* у простор веће димензије није случајан и инспирисан је [8], а могуће варијације су објашњене у [9].

4.1.2 Рендеровање

У основи је једначина 2.6. За потребе апроксимације интеграла биће коришћен Гаусов метод. Уколико користимо детерминистички приступ, одбирци који учествују у апроксимацији ће увек бити исти. То је неповољан приступ будући да би у том случају неуронско поље зрачења било увек узорковано за исти скуп параметара. Уместо

тога, увешћемо еквидистантну поделу сегмента $[t_n, t_f]$, а потом насумице одабрати по један број из сваког елемента поделе. То формално исказујемо наредним изразом

$$t_i \sim \mathcal{U}\left[t_n + \frac{i-1}{N}(t_f - t_n), t_n + \frac{i}{N}(t_f - t_n)\right]. \tag{4.2}$$

Препишимо једначину 2.6 у дискретном облику, у складу са 4.2

$$\hat{C}(\Pi_{\mathbf{r_{c,d}}}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\Delta_i))\mathbf{c}_i, \tag{4.3}$$

где је $T_i = \exp\left(-\sum_{j=1}^{i-1}\sigma_j\Delta_j\right)$, парови $(\mathbf{c}_i, \sigma_i)$ одговарајући у складу са избором t_i, а Δ_i ширина елемента поделе.

Једначине 2.6 и 4.3 се видно разликују по начину на који учествује пропусност. Наиме, уколико би Гаусов метод применили директно, појавио би се линеаран члан $\sigma_i\Delta_i$. Тај члан је је у апроксимацији замењен експоненцијалним $1 - \exp(-\sigma\Delta_i)$. У наставку ћемо објаснити и зашто [10].

Погледајмо како произвољни елемент поделе учествује у апроксимацији. Како узимамо да су на сваком елементу поделе боја и пропусност константни, једначину 2.6 расписујемо у складу са тим

$$\begin{aligned}
\hat{C}_i(\Pi_{\mathbf{r_{c,d}}}) &= \int_{t_i}^{t_{i+1}} T(t)\sigma_i c_i dt \\
&= \int_{t_i}^{t_{i+1}} \sigma_i c_i \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r_{c,d}}(s)ds)\right) dt \\
&= \sigma_i c_i \int_{t_i}^{t_{i+1}} \exp\left(-\int_{t_n}^{t_i} \sigma(\mathbf{r_{c,d}}(s)ds)\right) \exp\left(-\int_{t_i}^{t_{i+1}} \sigma_i ds\right) dt \\
&= T_i\sigma_i c_i \int_{t_i}^{t_{i+1}} \exp(-\sigma_i(t - t_i))dt \\
&= T_i\sigma_i c_i \frac{\exp(-\sigma_i(t - t_i))}{-\sigma_i}\Bigg|_{t_i}^{t_{i+1}} \\
&= T_i(1 - \exp(-\sigma_i\Delta_i))c_i.
\end{aligned} \tag{4.4}$$

4.1.3 Обучавање

Имплементација инспирисана описаним поступком рендеровања је врло неефикасна – празан простор и прикривени региони не доприносе рендерованој слици, а непрестано се узоркују. Из тог разлога се користе два неуронска поља, за *фина* и *груба* предвиђања. Прво се узоркује N_c одбирака на начин објашњен у 4.2 који се потом дају грубом пољу. На основу грубих предвиђања се врши још једно узорковање, али овог пута боље навођено. Фино узорковање је пристрасније релевантним деловима простора.

Да би то спровели у дело, излаз грубог неуронског поља записујемо као збир боја дуж зрака пондерисаног непрозирношћу

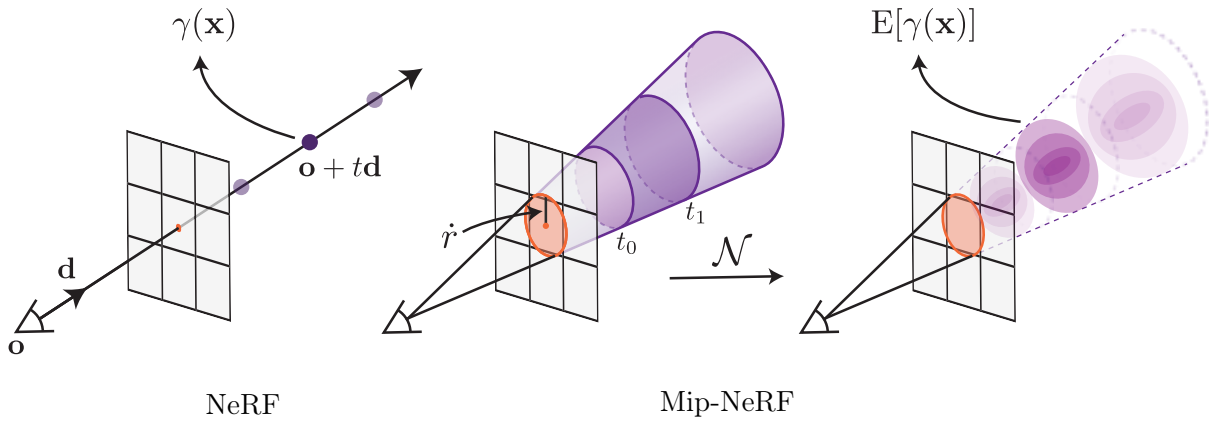
$$\hat{C}_c(\Pi_{\mathbf{r_{c,d}}}) = \sum_{i=1}^{N_c} w_i c_i, \quad w_i = T_i(1 - \exp(-\sigma_i \Delta_i)). \tag{4.5}$$

Скалирањем тежина $w_i = w_i / \sum_{j=1}^{N_c} w_j$, добија се део-по-део константна густина расподеле дуж зрака. Из ове расподеле се (инверзно) узоркује N_f одбирака, а потом фино неуронско поље израчуна у свих $N_c + N_f$ одбирака на начин описан у једначини 4.3. У свим експериментима се за N_c узима 64, а за N_f 128.

Функција губитка у случају NeRF-a је

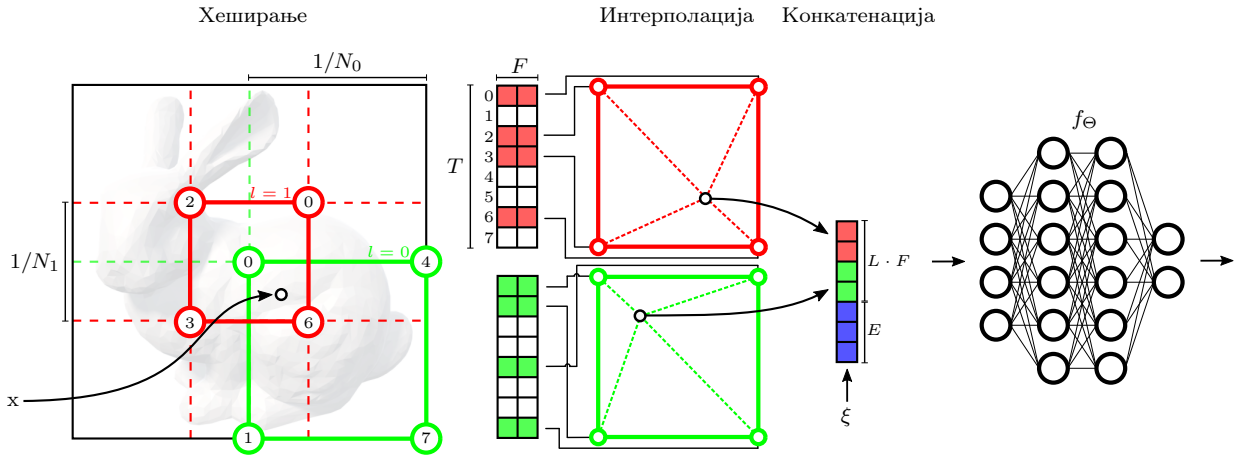
$$\mathcal{L} = \sum_{\mathbf{r_{c,d}} \in \mathcal{B}} \left(\left\| \hat{C}_c(\Pi_{\mathbf{r_{c,d}}}) - C(\Pi_{\mathbf{r_{c,d}}}) \right\|_2^2 + \left\| \hat{C}_f(\Pi_{\mathbf{r_{c,d}}}) - C(\Pi_{\mathbf{r_{c,d}}}) \right\|_2^2 \right) \tag{4.6}$$

4.2 Mip-NeRF



Слика 4.2: Поређење NeRF и Mip-NeRF модела према [2]

4.3 Instant-NGP



Слика 4.3: Илустрација Instant-NGP модела према [3]

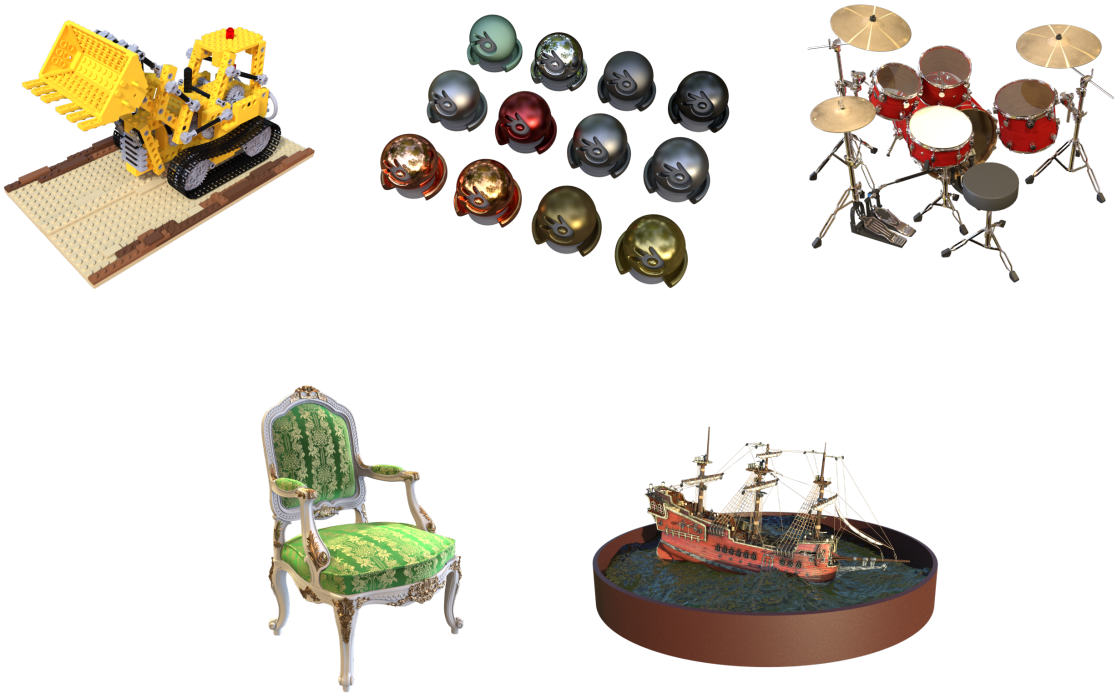
Глава 5

Скупови података

Скупови података у области неуронских поља зрачења се угрубо могу поделити на две групе – да ли је позната калибрација камере или не. У овом раду смо се опредили за прву групу. Разлози су пре свега техничке природе. Проблем калибрације камере је један од централних проблема рачунарског вида и као такав је добро испитан. Може се рећи да је овај проблем *лоше условљен*. Готово незнатне разлике могу довести до значајних грешака у пројекцијама, а врло прецизна калибрација изискује и скупу опрему. Олакшавајућа околност је да су познати поступци одређивања унутрашњих и спољашњих параметара камере, али јасно је да они не могу бити ни близу прецизни као што је, на пример, роботска калибрација. С тим у вези, мишљења смо да је оправдано користити синтетички генерисане скупове података у којима ће камера бити постављена на унапред одређено место, а сви њени параметри ће бити познати по дефиницији.

За потребе овог рада, коришћени су Lego, Materials, Drums, Chair и Ship. У питању су познати, референтни скупови података. Модели ће бити упоређени на сваком од њих у истом окружењу. Конкретно, у питању су редом багер од Лего коцки, 16 лоптица различитих материјала, бубњеви, столица и брод. На слици 5.1 је приказан по један поглед за сваки од скупова.

Сваки скуп података је подељен на три подскупа – за обучавање, проверу и тестирање. У подскуповима за обучавање и проверу се налази тачно 100 различитих погледа, док се за тестирање користи 200. Рендери су синтетички, резолуције 800×800 пиксела без радијалне и тангенцијалне дисторзије.



Слика 5.1: По један поглед из сваког скупа података

Глава 6

Експерименти

Модели ће бити упоређени по свакој од метрика које ћемо навести и објаснити у наставку. Такође, како је дужина трајања обучавања од великог значаја, посебну пажњу ћемо посветити и том аспекту.

Конфигурација рачунара на ком су сви модели обучени се састоји од Intel Xeon W-2665 централног процесора са 24 језгра на фреквенцији 3.5GHz, 64GB радне меморије и Nvidia Titan Xp графичке карте са 12GB GDDR5X сопствене меморије. Обучавање је вршено у локалу са избором параметара у складу са изворним радовима. Имплементација је изведена у програмском језику Python користећи PyTorch [11], док је радно окружење Windows 11.

6.1 Метрике

PSNR (енг. *Peak Signal-to-Noise Ratio*). Нека је дата монохроматска слика I и њена апроксимација K. Средњеквадратно одступање ове две слике је

$$\mathrm{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (I_{i,j} - K_{i,j})^2. \tag{6.1}$$

Однос сигнала и шума је

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{I_{\max}^2}{\mathrm{MSE}} \right) [\mathrm{dB}], \tag{6.2}$$

где је $I_{\max}$ највећа могућа вредност коју пиксел на слици може имати. Како се канал слике обично представља једним бајтом, ова вредност у већини случајева износи 255.

Више вредности ове метрике указују на бољу апроксимацију.

LPIPS (енг. *Learned Perceptual Image Patch Similarity*) [12]. Идеја је имати метрику која опонаша људску процену сличности две слике. У ту сврху се користи модел $\mathcal{F}$ обучен на сликовном скупу података. Не постоје никаква ограничења у погледу архитектуре, па чак ни скупа података, али се показује да VGG [13] и AlexNet [14] у комбинацији са ImageNet-ом врло добро раде у пракси.

Означимо са I троканалну слику, а са K њену апроксимацију. Иако слојеви модела $\mathcal{F}$ не морају имати ни улаз ни излаз димензије 2, у овом тренутку ће бити лакше

да претпоставимо да то јесте случај. Пре свега из рачунских разлога. Уколико $\mathcal{F}$ захтева тензоре неке друге димензије, увек се томе можемо прилагодити једноставним преобличавањем.

Нека је $\boldsymbol{\alpha}^I_{i,j,k}$ резултат примене активационе функције у слоју k у врсти i и колони j за улазну слику I. Ове вредности су нормализоване канал по канал. Аналогно имамо и $\boldsymbol{\alpha}^K_{i,j,k}$.

$$\text{LPIPS} = \sum_{k=1}^{L} \frac{1}{N_i N_j} \sum_{i=0}^{N_i-1} \sum_{j=0}^{N_j-1} \left\| w_k \cdot (\boldsymbol{\alpha}^I_{i,j,k} - \boldsymbol{\alpha}^K_{i,j,k}) \right\|^2. \tag{6.3}$$

Ниже вредности ове метрике указују на бољу апроксимацију.

SSIM (енг. *Structual Similarity Index Measure*) [15]. Нека је дата монохроматска слика I и њена апроксимација K.

$$\text{SSIM} = \frac{(2\mu_I\mu_K + c_1)(2\sigma_{I,K} + c_2)}{(\mu_I^2 + \mu_K^2 + c_1)(\sigma_I^2 + \sigma_K^2 + c_2)}, \tag{6.4}$$

где су μ_I и μ_K узорачке средине, σ_I^2 и σ_K^2 узорачке дисперзије, а $\sigma_{I,K}$ узорачка коваријанса између I и K. Вредности c_i су дефинисане као $(k_i I_{\max})^2$, где је су k_i константе и то обично редом 0.01 и 0.03.

У пракси се показује да овако дефинисана метрика неће дати увек задовољавајуће резултате. Зато се она ретко примењује на целој слици, већ се слика дели на мање делове, а резултати метрике потом упросече.

Више вредност ове метрике указују на бољу апроксимацију.

6.2 Време обучавања

6.3 Квалитет резултата

Глава 7

Закључак

Библиографија

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[2] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," *ICCV*, 2021.

[3] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, pp. 102:1–102:15, July 2022.

[4] K. Fukushima, "Cognitron: A self-organizing multilayered neural network," *Biological Cybernetics*, vol. 20, pp. 121–136, 1975.

[5] J. L. Sejun Park, Chulhee Yun and J. Shin, "Minimum width for universal approximation," in *Proceedings of American Federation of Information Processing Societies: 1977 National Computer Conference*, ICLR, 2021.

[6] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, vol. 23, pp. 462–466, 1952.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[9] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," 2020.

[10] N. L. Max and M. Chen, "Local and global illumination in the volume rendering integral," in *Scientific Visualization: Advanced Concepts*, 2010.

[11] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[15] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.