УДК 004.8, 004.62

В.В. САВІН
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна, e-mail: *vladimyr.savin@gmail.com*.

О.О. КОЛОДЯЖНА
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна, e-mail: *kolodyazhna.lena@gmail.com*.

АДАПТАЦІЯ ТЕХНОЛОГІЇ NEURAL RADIANCE FIELDS (NeRF) ДЛЯ ЗАДАЧІ 3D-РЕКОНСТРУКЦІЇ СЦЕНИ В УМОВАХ ДИНАМІЧНОГО ОСВІТЛЕННЯ

Анотація. Розглянуто проблему синтезу нових зображень сцени з використанням технології Neural Radiance Fields (NeRF) для середовища з динамічним освітленням. Для навчання NeRF моделей використано фотометричну функцію втрат, тобто попіксельну різницю між значеннями інтенсивності зображень сцени та зображень, згенерованих за допомогою NeRF. Для відбивних поверхонь інтенсивність зображення залежить від кута спостереження. Цей ефект враховано шляхом використання напрямку променя як вхідного параметра моделі NeRF. Для сцен з динамічним освітленням інтенсивність зображення залежить не лише від позиції та напрямку спостереження, а й від часу. Показано, що зміна освітлення впливає на навчання NeRF із стандартною фотометричною функцією втрат і зумовлює зниження якості отриманих зображень та карт глибин. Щоб подолати цю проблему, запропоновано ввести час як додатковий вхідний аргумент до моделі NeRF. Експерименти, проведені на наборі даних ScanNet, свідчать про те, що NeRF зі зміненим входом перевершує оригінальну версію моделі та генерує більш послідовні й цілісні 3D-структури. Результати цього дослідження можна використати для покращення якості аугментації навчальних даних для навчання моделей передбачення відстані (наприклад, моделей depth-from-stereo, які забезпечують передбачення глибини/відстані на основі стереоданих) для сцен з нестатичним освітленням.

Ключові слова: комп'ютерний зір, технологія Neural Radiance Fields, динамічне освітлення, синтез даних, 3D-реконструкція сцени.

ВСТУП

Тривимірна реконструкція сцени є однією з фундаментальних проблем комп'ютерного зору. Суть цієї проблеми полягає в розумінні тривимірної структури сцени на основі її двовимірних зображень. 3D-реконструкцію сцени застосовують у різних галузях, зокрема у доповненій реальності (AR) та віртуальній реальності (VR). Наприклад, вона дає змогу обробляти зіткнення та перекриття між доповненими об'єктами і фізичним світом для реалізації природної та реалістичної взаємодії в AR. Для розв'язання задачі реконструкції 3D-сцени застосовують різноманітні методи та інструменти, зокрема марковські випадкові поля [1], алгоритми локального стереозіставлення [2, 3] та глибокі нейронні мережі [4]. Ця задача є складною, оскільки необхідно одночасно реконструювати глобальні структури сцени та їхні локальні деталі. Це зумовлює потребу у масивних обчисленнях та великій кількості даних. Наявність точних і достовірних даних є ключовим фактором для навчання глибоких нейронних мереж.

Ручний збір даних з подальшим анотуванням та генерування синтетичних даних є поширеними підходами для збору навчальних наборів даних. Процес ручного збору дуже трудомісткий і дорогий. Потрібно одночасно зібрати зображення, істинні дані про глибину та точні положення камери. Отже, цей процес потребує додаткового специфічного обладнання, наприклад, камери глибини та системи позиціонування. Повністю синтетичні дані не можуть повноцінно замінити реальні дані в навчанні нейронних мереж. Для того, щоб гарантувати адекватну якість роботи мережі в реальних умовах експлуатації, моделі, отри-

мані після навчання на таких повністю синтетичних даних, мають бути додатково налаштовані на реальних даних з цільової області [5].

Одним із останніх досягнень у синтезі зображень з надійними результатами є Neural Radiance Fields (NeRF) [6]. Це повнозв'язна мережа, яка може генерувати нові ракурси сцени, приймаючи під час навчання обмежену кількість знімків сцени з відповідними положеннями камер. Мережа NeRF оптимізує базову неперервну об'ємну функцію за допомогою розрідженого набору вхідних даних [6]. Цей метод можна використати для генерування нових даних та аугментації навчальних наборів даних для нейронних мереж передбачення глибини. Наприклад, у [7] автори використовують синтетичні зображення, згенеровані за допомогою NeRF, для розв'язання задач локалізації. Вони демонструють, що додаткові синтетичні дані покращують точність регресії положення камери. Одна з переваг цього багатошарового перцептрона полягає в тому, що він може синтезувати не лише кольорові зображення, а й карти глибини, важливі для мереж передбачення відстані. Нагадаємо, що 3D-реконструкція сцени має бути наближеною до оригінальних структур середовища, тому для цього потрібні достовірні карти глибини.

Під час навчання NeRF оптимізує фотометричну функцію втрат, яка дорівнює попіксельній різниці між інтенсивностями оригінального та згенерованого зображень. Однак, вона має свої обмеження в динамічних сценах і сценах зі змінним освітленням. Через залежність від інтенсивності зображення, ці дані порушують припущення узгодженості яскравості, яке є важливим для таких фотометричних функцій втрат. Зазначена функція втрат також є поширеною. Її мінімізують у мережах передбачення глибини/відстані, що навчаються в режимі навчання без учителя (наприклад у [8, 9]). Автори [10] аналізують проблему використання фотометричних функцій втрат для наборів даних із поганим або динамічним освітленням. Вони демонструють, що стандартна фотометрична функція втрат призводить до погіршення якості реконструкції 3D-сцени для таких даних.

Модифікацію NeRF для розв'язання задачі 3D-реконструкції сцени в умовах змінного освітлення представлено у роботі [11]. На відміну від [11], у цій статті більш детально розглянуто проблематику моделювання середовища у разі динамічного освітлення та представлено результати експериментів, які підтверджують запропоновану модифікацію оригінальної моделі.

Основним внеском цієї роботи в задачу 3D-реконструкції сцени є покращення її якості в умовах змінного освітлення за рахунок використання часової змінної як послідовного індексу зображення, доданого до вхідних параметрів моделі NeRF під час реконструкції статичних сцен, але з динамічним освітленням. У цій модифікації моделі враховано деякі зміни освітлення в усіх наборах даних, що дає змогу здійснити їхню компенсацію. Щоб оцінити вплив модифікованої NeRF моделі на реконструкцію сцени, якість генерування карт глибин за допомогою NeRF виміряно шляхом обчислення відносних похибок глибин на двох ScanNet [12] сценах з деякими змінами освітлення та за наявності віддзеркалень.

Стаття має таку структуру. У розд. 1 розглянуто роботи, що стосуються генерування зображень, карт глибин та реконструкції сцени зі змінами освітлення. У розд. 2 описано проблему навчання NeRF для сцен з динамічним освітленням і запропоновано модифікації архітектури NeRF та відповідної функції втрат. У розд. 3 розглянуто експерименти, які демонструють ефект запропонованих модифікацій. У розділі «Висновки» підсумовано основні наукові результати статті та майбутні напрями досліджень.

1. ОГЛЯД ЛІТЕРАТУРИ ТА СУЧАСНИХ ПІДХОДІВ ГЕНЕРУВАННЯ ЗОБРАЖЕНЬ

Маючи певну кількість зображень сцени, знятих з різних положень камери, мережа NeRF реконструює всю її 3D-структуру та дає змогу синтезувати нові зображення сцени. Цей метод не є єдиним для розв'язання таких задач. Інші підходи створення нових зображень сцени з інших положень камери включають, наприклад, оптимізацію представлення сцени на основі тривимірної сітки [13], або методи, оптимізовані глибокими нейронними мережами, які відображають координати *XYZ* у функції знакової відстані (sign distance functions, SDF) [14, 15]. Проте через свою обчислювальну складність ці методи потребують високоякісних істинних 3D-даних і мають низьку здатність до масштабування під час генерування зображень з високою роздільною здатністю. З іншого боку, NeRF оптимізує представлення сцени у формі неперервної диференційованої функції, що дає змогу навчати модель наскрізним способом (end-to-end approach).

Однак, оригінальна модель NeRF [6] також має недоліки, зокрема через витрати часу на оптимізацію та вимоги щодо забезпечення точного положення камери. Автори моделі «Bundle-adjusting neural radiance fields» (BARF) [16] пом'якшують вимоги до точних положень камери, додаючи їх до процесу оптимізації. Моделі NeRF і BARF мають спільне обмеження щодо підтримки лише однієї сцени. Обидві моделі не можна узагальнити на інші середовища. Вони дають змогу синтезувати зображення лише для конкретної сцени, використаної у процесі навчання.

Зауважимо, що є такі моделі NeRF, як MVSNeRF [17] та NerfingMVS [18], які долають зазначену проблему за допомогою підходів на основі даних з довільних точок спостереження (multi-view stereo, MVS).

У разі застосування MVSNeRF [17] є можливість не навчати модель для кожного набору даних окремо від самого початку, а навчити її на одному наборі даних, а потім донавчити на іншому. Отже, на відміну від інших робіт, пов'язаних із синтезом зображень за допомогою NeRF, тут показано, що ця модель використовує переваги MVS-архітектури. Це дає змогу більш ефективно знаходити різні відповідності між зображеннями. До того ж, завдяки поєднанню MVS та 3D-згорткових нейронних мереж можна узагальнити модель на інші набори даних.

Основна мета застосування моделі NerfingMVS [18] — це покращення отриманих карт глибин. Також автори стверджують, що у такий спосіб можна покращити спроможність NeRF генерувати нові зображення. Модель спочатку оптимізує монокулярну мережу для оцінювання карт глибин на цільовій сцені шляхом донавчання її на розріджених картах відстаней, отриманих з використанням підходу реконст- рукції структури за допомогою руху (structure from motion — SfM) з COLMAP [19]. Процедура вибору точок на променях під час оптимізації NeRF ґрунтується на отриманій карті помилок синтезованих зображень. У такий спосіб підвищують якість карти глибин і отриманий результат є більш точним та чітким.

Однією з проблем NeRF є велика кількість зображень, які потрібно подавати на вхід мережі. Автори мережі DDPNeRF [20] розв'язують цю проблему, використовуючи COLMAP [19] та генеруючи щільні карти глибин разом із картами невизначеності, які далі застосовують для оптимізації NeRF. Цей метод дає змогу генерувати нові зображення сцени, використовуючи як вхід лише 18–36 зображень. Мережа DDPNeRF подібна до NerfingMVS [18], проте використовує заповнення карт відстаней, отриманих з COLMAP, а також обчислює та використовує карти невизначеності.

Розглянуті методи реконструкції 3D-середовища та генерування зображень можна застосовувати лише в умовах статичної сцени. Їхнім суттєвим недоліком є відсутність можливості моделювання сцен зі змінним освітленням.

Загалом, моделі, що ґрунтуються на NeRF, працюють за припущення, що сцени є статичними, але це не завжди так. Модель Neural Scene Flow Fields (NSFF) [21] є одним з варіантів NeRF, який забезпечує оптимізацію для динамічних сцен. Автори цієї моделі модифікували NeRF представлення сцени, враховуючи динамічні умови середовища. У результаті отримано нове представлення сцени. Мережа NSFF моделює динамічну сцену як змінну в часі неперервну функцію представлення середовища, геометрії та руху 3D-сцени. Такий підхід дає змогу інтерполювати зміни як у просторі, так і в часі. На відміну від зазначеної роботи, у цій статті розглянуто часову залежність NeRF для статичних сцен з динамічним освітленням.

Розглядаючи NeRF як інструмент доповнення даних для подальшого навчання моделей розуміння сцени (наприклад, передбачення глибини), варто також розглянути альтернативні підходи, які ґрунтуються на фотореалістичному рендерингу. У [22] запропоновано фотореалістичний синтетичний набір даних Hypersim для задач розуміння сцени та метод, за допомогою якого отримано цей набір даних. Оскільки використаний процес рендерингу ґрунтується на фізичних принципах освітлення та розсіювання світла, суттєвою перевагою цього методу є якість згенерованих зображень. Недоліком Hypersim є потреба у використані фотореалістичних 3D-моделей середовища та окремих об'єктів, наявність яких обмежена у відкритому доступі. У цьому контексті перевагою NeRF є можливість реконструювання та генерування даних на основі обмеженої кількості зображень, отриманих в лабораторних умовах.

2. МЕТОДОЛОГІЯ ПОКРАЩЕННЯ ЯКОСТІ МОДЕЛІ NeRF

2.1. Технологія NeRF. Функція втрат для навчання. Нещодавно науковці досягли успіху у створенні нових зображень для складних сцен з використанням NeRF [6]. Цей підхід забезпечує представлення сцени у вигляді повнозв'язної глибокої нейронної мережі. Вхідними даними для моделі є п'ятивимірна векторна функція, аргументами якої є просторове розташування (x, y, z) та напрямок спостереження (θ, φ). Результатом є щільність об'єму σ та RGB-колір $\vec{c}$ для кожного пікселя. Модель можна представити так [6]:

$$F_w : (\vec{x}, \vec{d}) \to (\vec{c}, \sigma). \tag{1}$$

Щоб навчити NeRF, потрібно мати набір даних із RGB-зображеннями сцени, який відтворює одну локацію з різних кутів спостереження, положення камери та внутрішні параметри камери. Процес навчання включає рендеринг відповідних ракурсів сцени та мінімізацію фотометричної похибки між синтезованими зображеннями та спостережуваними зображеннями.

Процес рендерингу зображено на рис. 1. Спочатку знаходять промені камери, які проходять через кожен піксель зображення, та вибирають деякі точки на

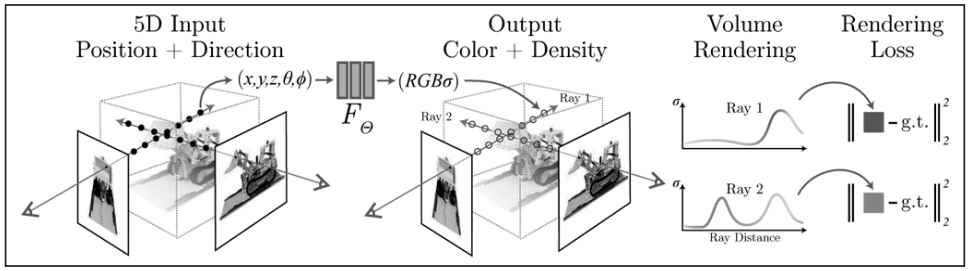


Рис. 1. Процес рендерингу за допомогою NeRF [6]

цих променях. Потім ці точки передають у мережу багатошарового перцептрона, яка передбачає колір і щільність для кожної з них. На останньому етапі використовують класичну об'ємну візуалізацію [23] для агрегації усіх кольорів та щільностей для кожної точки вибірки та отримання кінцевого результату для кожного пікселя.

Під час навчання NeRF мінімізує фотометричну функцію втрат. Зазвичай цей метод використовує дві мережі: грубу та точну [6]. Для простоти розглянемо лише першу, грубу підмережу. Нехай маємо M зображень $(I_1, \ldots, I_M)$. Метою навчання NeRF є оптимізація такої цільової функції на основі синтезу:

$$L_F = \sum_{i=1}^{M} \sum_{u} || \hat{I}_i(u; w) - I_i(u) ||_2^2, \tag{2}$$

де w — параметри мережі, які також залежать від напрямків спостереження, u — координати пікселів, $\hat{I}(u; w)$ — синтезоване значення RGB-кольору у пікселі.

2.2. Обмеження фотометричної функції втрат. Однією з характеристик реальних даних є динамічне освітлення. Зазвичай наявні набори даних містять статичні сцени без змін освітлення. Зміни освітлення можуть бути спричинені певними факторами, наприклад, зовнішніми джерелами, такими як сонце, або фарами автомобіля. Крім того, ці зміни можуть бути пов'язані з експозицією камери. Моделі NeRF враховують позиції та напрямки вхідних променів, які можуть компенсувати ефекти, спричинені відбивними поверхнями, але вони не враховують змінну часу. Залежність від часу є важливою для сцен з динамічним освітленням. Отже, стандартні моделі NeRF можуть здійснювати рендеринг неправильних зображень із таких наборів даних. Крім того, як показано в [10], мережі передбачення глибини, які використовують фотометричні функції втрат, не можуть відновити структуру 3D-сцени на наборах даних із поганим або динамічним освітленням. Моделі на основі NeRF можуть мати ті самі проблеми, оскільки вони також використовують таку саму функцію втрат.

2.3. Модифікації оригінальної моделі NeRF для врахування умов динамічного освітлення. Функція втрат за глибиною. Оригінальна модель NeRF та її модифікації передбачають колір разом зі щільністю σ, яку можна інтерпретувати як непрозорість об'єктів. Використовуючи отриману щільність, можна розрахувати відстані до об'єктів.

Під час навчання NeRF не здійснюється додаткова оптимізація передбачених карт глибини. Це призводить до погіршення передбачених відстаней до об'єктів і, як наслідок, до зниження якості реконструкції 3D-сцени. Набори даних, зібрані із застосуванням RGB-D камер, мають (неповні) карти глибини, які можна використовувати для покращення якості NeRF. Для цього запропоновано ввести додаткову функцію втрат, яку визначають як середньоквадратичну похибку між істинними і передбаченими значеннями карт відстаней:

$$L_D = || D - \hat{D} ||_2^2, \tag{3}$$

де D — істинні карти глибини, $\hat{D}$ — передбачені карти глибини.

Загальна функція втрат має такий вигляд:

$$L = \sum_{i=1}^{M} \sum_{u} || \hat{I}_i(u; w) - I_i(u) ||_2^2 + || D - \hat{D} ||_2^2, \tag{4}$$

де M — кількість зображень, u — координати пікселя, $\hat{I}_i(u; w)$ — синтезований RGB-колір у пікселі u для зображення i, $I_i(u)$ — істинний RGB-колір

у пікселі u для зображення i, D — істинна карта глибини, $\hat{D}$ — карта глибини, передбачена за допомогою NeRF.

2.4. Модифікації оригінальної моделі NeRF для врахування умов динамічного освітлення. Час як додаткова вхідна змінна моделі NeRF. Щоб змоделювати динамічне освітлення, запропоновано використовувати час t як додаткову вхідну змінну для мережі. Кожен елемент t відповідає послідовному індексу вхідного зображення для навчання NeRF. Індекси часу та індекси зображень пов'язані лінійною залежністю. Це можна пояснити тим, що оброблюваний набір даних є відеопослідовністю з фіксованим значенням кадрів за секунду (FPS). Змінну t додатково нормалізують на інтервалі $[0, 1]$ і застосовують до неї позиційне кодування. Цей аргумент можна додати до моделі у двох варіантах. Перший варіант відповідає додаванню часу t до 3D-координат, його позначимо як $(\vec{x}, t)$. Другий варіант — це додавання часу t до напрямків вхідних променів, його позначимо як $(\vec{d}, t)$.

Тоді модифіковану модель NeRF можна визначити так:

$$F_w: (\vec{x}, \vec{d}, t) \to (\vec{c}, \sigma). \tag{5}$$

3. ЕКСПЕРИМЕНТИ

3.1. Опис набору даних. Набір ScanNet [11] — це набір кольорових даних та відповідних карт глибин (RGB-D), який містить 2.5 мільйона зображень, зібраних у понад 1500 різних приміщеннях. Внутрішні та зовнішні характеристики камери (положення камери) надано для кожної сцени. У цій статті використано дві ScanNet сцени: scene0000_00 і scene0005_01. Для цих сцен через наявність відбивних поверхонь, а також через зміну часу, освітлення змінюється залежно від кута спостереження (рис. 2). Часова динаміка, найімовірніше, спричинена автоматичною експозицією камери. Сцена Scene0000_00 містить 5577 зображень, а scene0005_01 — 1449 зображень. Для експериментів використано лише фрагмент сцени scene0000_00, яка містить 500 зображень. Вибраний фрагмент набору даних містить зображення зі змінним освітленням.

3.2. Опис експерименту. Метрики. Щоб продемонструвати ефект запропонованих модифікацій, проведено такі експерименти:

1. Навчання оригінальної NeRF моделі.
2. Навчання NeRF моделі з доданою функцією втрат для оптимізації карт глибин.
3. Навчання NeRF моделі з додатковим вхідним параметром часу.
4. Навчання NeRF моделі з доданими функцією втрат за глибиною та часовою змінною.

Архітектура моделей відповідає архітектурі оригінальної моделі NeRF [6] зі 128 прихованими нейронами у кожному шарі MLP. Для навчання моделей дані кожної сцени розподілено у співвідношенні 90 % : 10 % як навчальні та

Рис. 2. Приклади сцен зі зміною освітлення поверхонь залежно від кута спостереження (*а*) та за наявності ефектів відбиття світла (*б*)

валідаційні відповідно. Під час навчання використано кожну другу картинку з набору даних. Розмір зображень змінено до 426×560 пікселів. З кожного зображення вибрано 2048 променів у випадковий спосіб. З кожного променя вибрано 128 точок для числового інтегрування. Для передбачення моделей вибрано діапазон глибини 0–8 м. Під час навчання використано оптимізатор Adam зі швидкістю навчання, яка дорівнює 1e–3 та експоненційно зменшується до 1e–4. Навчання кожної моделі здійснено протягом 200 тисяч ітерацій.

Оцінювання моделей проведено за допомогою двох метрик: середньої абсолютної похибки (mean absolute error, MAE) та середньої абсолютної відносної похибки (mean absolute relative error, MARE). Виміряно похибки між передбаченою та істинною глибинами. Ці метрики є поширеними для вимірювання похибок між картами глибин. Наприклад, у [18, 24] MARE розглядають як одну з величин для порівняння карт глибин.

MAE можна розрахувати у такий спосіб:

$$\mathrm{MAE} = \frac{1}{N}\frac{1}{n}\frac{1}{m}\sum_{k=1}^{N}\sum_{i,j}^{n,m} | y_{k,i,j} - \hat{y}_{k,i,j} |, \tag{6}$$

де $y_{k,i,j}$ — істинне значення глибини для (i, j) пікселя, $\hat{y}_{k,i,j}$ — передбачене значення глибини для (i, j) пікселя, N — кількість зображень у наборі даних, n — висота зображення, m — ширина зображення.

MARE можна розрахувати за такою формулою:

$$\mathrm{MARE} = \frac{1}{N}\frac{1}{n}\frac{1}{m}\sum_{k=1}^{N}\sum_{i,j}^{n,m}\frac{| y_{k,i,j} - \hat{y}_{k,i,j} |}{y_{k,i,j}} \cdot 100\,\%, \tag{7}$$

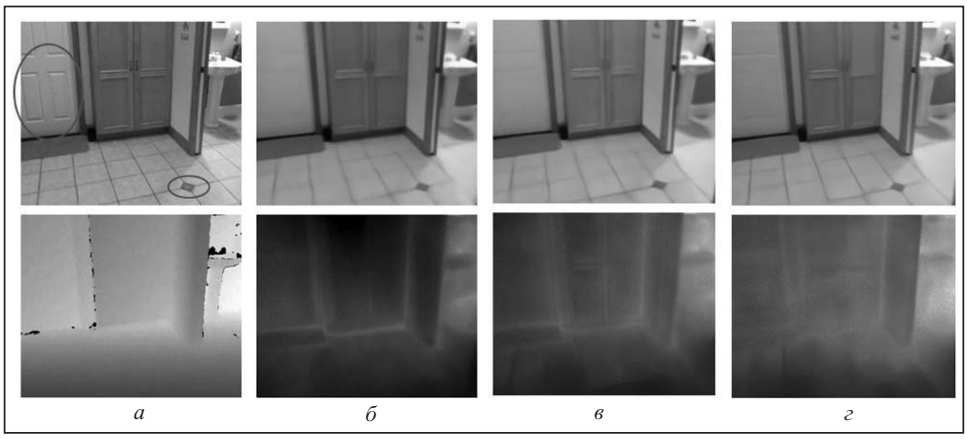
де $y_{k,i,j}$ — істинне значення глибини для (i, j) пікселя, $\hat{y}_{k,i,j}$ — передбачене значення глибини для (i, j) пікселя, N — кількість зображень у наборі даних, n — висота зображення, m — ширина зображення.

Для оцінювання навчених моделей згенеровано всі картинки кожної з двох використаних сцен разом з відповідними картами глибин. Підрахунок MAE та MARE здійснено на кожному наборі даних окремо, як і для кожної навченої моделі. Під час оцінювання якості результатів відкинуто нульові значення карт глибин, що містяться в істинних даних. Ці значення є наслідком використання специфічних сенсорів під час отримання карт глибин.

3.3. Результати експерименту. У табл. 1 наведено метрики для моделей, навчених у різних режимах. Для scene0000_00 моделі NeRF навчено без додавання функції втрат за глибиною. Базова модель показує низьку якість відносної похибки глибини на рівні 28.1 %. Додавання часової змінної дає змогу покращити якість моделі на 8–11 %. Якісні результати для цього набору даних зображено на рис. 3, де перший рядок відповідає RGB-зображенням, а другий рядок — картам глибини. Усі наведені карти глибин відображено в одному діапазоні 1114–3737 мм, де темні кольори відповідають меншим значенням глибин, а світлі — їхнім більшим значенням. Для кожної карти глибин наведено відповідний діапазон у мм. Пропущені значення в істинних картах глибин, які дорівнюють нулю, відкинуто у представленому діапазоні. З рис. 3. видно, що додавання часової змінної до NeRF моделі покращує якість як RGB-зображень (показано у виділених ділянках), так і карт глибин (діапазон передбачених значень наближається до істинного).

Таблиця 1. Метрики для набору даних ScanNet

Набір даних	Режим навчання	MAE, м	MARE, %
scene0000_00	Базова модель	0.686	28.1
	$(\vec{d},t)$	0.498	20
	$(\vec{x},t)$	0.429	17.3
scene0005_01	Базова модель	0.577	30
	Функція втрат за глибиною	0.03	1.881
	Функція втрат за глибиною $+(\vec{d},t)$	0.03	1.875
	Функція втрат за глибиною $+(\vec{x},t)$	0.015	0.93



а *б* *в* *г*

Рис. 3. Якісні результати порівняння для scene0000_00: істинні дані, діапазон глибини 1605–3737 мм (*а*); NeRF, діапазон глибини 1114–2943 мм (*б*); NeRF + $(\vec{d},t)$, діапазон глибини 1392–3660 мм (*в*); NeRF + $(\vec{x},t)$, діапазон глибини 1510–3699 мм (*г*)

Для набору даних scene0005_01 проаналізовано вплив обох модифікацій: із функцією втрат за глибиною та з додатковою часовою змінною t. У разі модифікації з функцією втрат за глибиною відносна похибка глибини зменшується з 30 % до 1.88 %. Додавання часової змінної до того ж покращує якість як синтезованих кольорових зображень, так і карт глибини. Це найбільше проявляється у разі додавання змінної часу до вхідних 3D-координат $(\vec{x},t)$. Відносна похибка за глибиною в цьому режимі дорівнює 0.93 %. Якісні результати для набору даних scene0005_01 зображено на рис. 4, де перший рядок відповідає RGB-зображенням, а другий рядок — картам глибини. Усі наведені карти глибин відображено в одному діапазоні 826–3472 мм, де темні кольори відповідають меншим значенням глибин, а світлі — їхнім більшим значенням. Для кожної карти глибин наведено відповідний діапазон у мм. Пропущені значення в істинних картах глибин, які дорівнюють нулю, відкинуто у представленому діапазоні. Візуальне порівняння згенерованих зображень та глибин на рис. 4. свідчить про покращення якості синтезу зображень завдяки запропонованим модифікаціям, та дає моделі змогу краще генерувати деякі структури сцени (виділені ділянки).

Також проаналізовано вигляд 3D-реконструкції сцени для різних версій моделі. Результати цього аналізу представлено на рис. 5. Розширення базової моделі NeRF додатковим вхідним аргументом, а саме часовою змінною, дає змогу отримати 3D-сітку для сцени, що містить більше структурних деталей та більш точно зіставляється з істинною 3D-реконструкцією. Еліпсами виділено ділянки з покращенням.

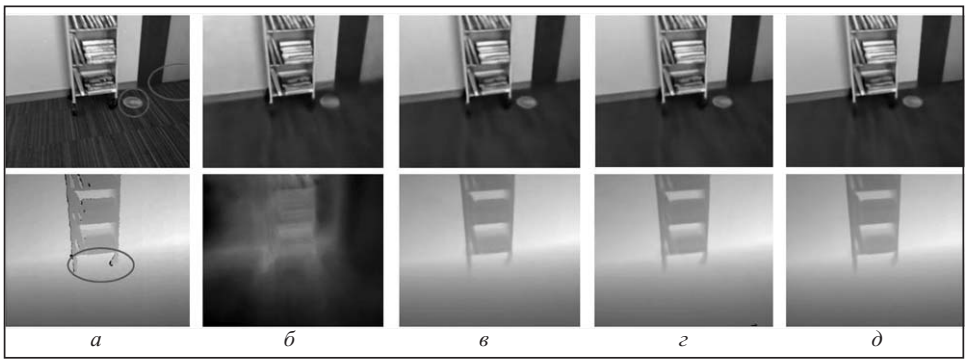


Рис. 4. Якісні результати порівняння для scene0005_01: істинні дані, діапазон глибини 1501–2612 мм (*а*); NeRF, діапазон глибини 826–1909 мм (*б*); NeRF + функція втрат за глибиною, діапазон глибини 1443–2656 мм (*в*); NeRF + функція втрат за глибиною + $(\vec{d},t)$, діапазон глибини 1451–2653 мм (*г*); NeRF + функція втрат за глибиною + $(\vec{x},t)$, діапазон глибини 1502–2610 мм (*д*)

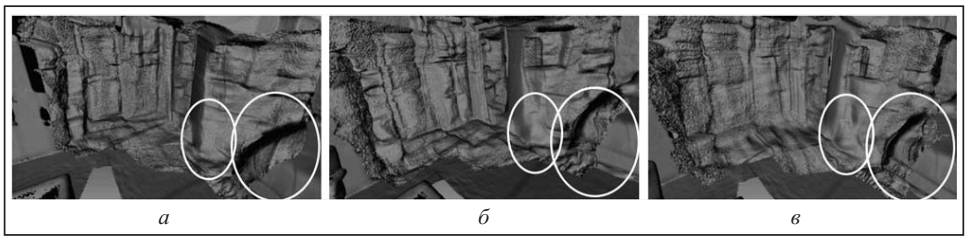


Рис. 5. Порівняння 3D-реконструкції сцени для scene0000_00: NeRF (*а*); NeRF + $(\vec{d},t)$ (*б*); NeRF + $(\vec{x},t)$ (*в*)

ВИСНОВКИ

У цій статті досліджено, як динамічне освітлення впливає на якість представлення сцени за допомогою NeRF моделі. Динамічне освітлення може бути спричинене джерелами освітлення, інтенсивність світла яких змінюється в часі (сонячне світло в похмуру погоду), джерелами світла, які вмикають або вимикають під час знімання сцени, або автоматичною експозицією камери. Показано, що ці зміни не можна змоделювати за допомогою стандартного NeRF-підходу, використовуючи позицію та напрямок кута спостереження як вхідні дані. Це призводить до погіршення якості генерування карт глибини. Для розв'язання цієї проблеми запропоновано розширити вхід NeRF додатковою часовою змінною. Цю ідею попередньо використано для моделювання сцен із динамічними об'єктами. Показано, що такий самий підхід доцільно застосовувати у разі статичних сцен із динамічним освітленням. Експерименти на наборі даних ScanNet свідчать про те, що розширення вхідних даних NeRF за допомогою часової змінної зумовлює покращення якості синтезованих зображень (наприклад, для тонких структур) і зменшення відносної похибки за глибиною на 10–28 %. Оскільки запропонований метод ґрунтується на технології NeRF, він має ті самі недоліки та обмеження. Це, насамперед, потреба у навчанні окремої моделі під кожну сцену та суттєвий час навчання. До того ж, ще одним недоліком запропонованого методу є погана здатність до реконструкції динамічної сцени, де є рухомі об'єкти. З практичного погляду, результати цієї роботи можна використати для покращення якості доповнення даних для навчання моделей передбачення відстані, де дуже важливою є якість візуалізації як зображення, так і глибини.

СПИСОК ЛІТЕРАТУРИ

1. Tardon L., Barbancho I., Alberola-Lopez C. Markov random fields in the context of stereo vision. In: Advances in Theory and Applications of Stereo Vision. Bhatti A. (Ed.). IntechOpen, 2011. 366 p. https://doi.org/10.5772/12953.

2. Zhu S., Yan L. Local stereo matching algorithm with efficient matching cost and adaptive guided image filter. *The Visual Computer*. 2017. Vol. 33, Iss. 9. P. 1087–1102. https://doi.org/10.1007/s00371-016-1264-6.
3. Bleyer M., Rhemann C., Rother C. PatchMatch stereo - stereo matching with slanted support windows. *Proc. 2011 British Machine Vision Conference (BMVC 2011)* (29 August – 2 September 2011, Dundee, UK). Dundee, 2011. 11 p. URL: https://api.semanticscholar.org/CorpusID:1798946.
4. Laga H., Jospin L.V., Boussaid F., Bennamoun M. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020. Vol. 44, N 4. P. 1738–1764. http://dx.doi.org/10.1109/TPAMI.2020.3032602.
5. Watson J., Aodha O.M., Turmukhambetov D., Brostow G.J., Firman M. Learning stereo from single images. *Proc. 16th European Conference on Computer Vision* (ECCV 2020) (23–28 August 2020, Glasgow, UK). Glasgow, 2020. Lecture Notes in Computer Science. Vol. 12346. P. 722–740. https://doi.org/10.1007/978-3-030-58452-8_42.
6. Mildenhall B., Srinivasan P.P., Tancik M., Barron J.T., Ramamoorthi R., Ng R. NERF: Representing scenes as neural radiance fields for view synthesis. *Proc. 16th European Conference on Computer Vision (ECCV 2020)* (23–28 August 2020, virtual). Lecture Notes in Computer Science. 2020. Vol. 12346. P. 405–421. https://doi.org/10.1007/978-3-030-58452-8_24.
7. Moreau A., Piasco N., Tsishkou D., Stanciulescu B., de La Fortelle A. LENS: Localization enhanced by NeRF synthesis. *Proc. 5th Conference on Robot Learning* (8–11 November 2021, London, UK). London, 2021. Vol. 164, P. 1347–1356. https://doi.org/10.48550/arXiv.2110.06558.
8. Godard C., Aodha O.M., Brostow G.J. Unsupervised monocular depth estimation with left-right consistency. *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition* (21–26 July 2017, Honolulu, HI, USA). Honolulu, 2017. P. 6602–6611. http://dx.doi.org/10.1109/CVPR.2017.699.
9. Li H., Gordon A., Zhao H., Casser V., Angelova A. Unsupervised monocular depth learning in dynamic scenes. *Proc. 2020 Conference on Robot Learning* (16–18 November 2020, virtual). 2021. Vol. 155. P. 1908–1917. https://doi.org/10.48550/arXiv.2010.16404.
10. Wang K., Zhang Z., Yan Z., Li X., Xu B., Li J., Yang J. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (10–17 October 2021, Montreal, QC, Canada). Montreal, 2021. P. 16035–16044. https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01575.
11. Kolodiazhna O., Savin V., Uss M., Kussul N. 3D scene reconstruction with neural radiance fields (NeRF) considering dynamic illumination conditions. *Proc. 11th International Conference on Applied Innovations in IT (ICAIIT 2023)* (9 March 2023, Kothen, Germany). Kothen, 2023. Vol. 11. P. 233–238. https://doi.org/10.25673/101943.
12. Dai A., Chang A.X., Savva M., Halber M., Funkhouser T., Nießner M. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (21–26 July 2017, Honolulu, HI, USA). Honolulu, 2017. P. 2432–2443. https://doi.org/10.48550/arXiv.1702.04405.
13. Waechter M., Moehrle N., Goesele M. Let there be color! Large-scale texturing of 3D reconstructions. *Proc. 13th European Conference on Computer Vision (ECCV 2014)* (6–12 September 2014, Zurich, Switzerland). Zurich, 2014. *Lecture Notes in Computer Science*. Vol. 8693. P. 836–850. https://doi.org/10.1007/978-3-319-10602-1_54.
14. Jiang C.M., Sud A., Makadia A., Huang J., Nießner M., Funkhouser T. Local implicit grid representations for 3D scenes. *Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (13–19 June 2020, Seattle, WA, USA). Seattle, 2020. P. 6000–6009. https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00604.
15. Penner E., Zhang L. Soft 3D reconstruction for view synthesis. *ACM Transactions on Graphics*. 2017. Vol. 36, Iss. 6. P. 1–11. https://doi.org/10.1145/3130800.3130855.

16. Lin C.-H., Ma W.-C., Torralba A., Lucey S. BARF: Bundle-adjusting neural radiance fields. *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (10–17 October 2021, Montreal, QC, Canada). Montreal, 2021. P. 5721–5731. http://dx.doi.org/10.1109/ICCV48922.2021.00569.
17. Chen A., Xu Z., Zhao F., Zhang X., Xiang F., Yu J., Su H. MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo. *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (10–17 October 2021, Montreal, QC, Canada). Montreal, 2021. P. 14104–14113. https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01386.
18. Wei Y., Liu S., Rao Y., Zhao W., Lu J., Zhou J. NerfingMVS: Guided optimization of neural radiance fields for indoor multi-view stereo. *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (10–17 October 2021, Montreal, QC, Canada). Montreal, 2021. P. 5590–5599. https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00556.
19. Schönberger J.L., Frahm J.-M. Structure-from-motion revisited. *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (27–30 June 2016, Las Vegas, NV, USA). Las Vegas, 2016. P. 4104–4113. https://doi.org/10.1109/CVPR.2016.445.
20. Roessle B., Barron J.T., Mildenhall B., Srinivasan P.P., Nieβner M. Dense depth priors for neural radiance fields from sparse input views. *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (18–24 June 2022, New Orleans, LA, USA). New Orleans, 2022. P. 12882–12891. https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01255.
21. Li Z., Niklaus S., Snavely N., Wang O. Neural scene flow fields for space-time view synthesis of dynamic scenes. *Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (*20–25 June 2021, Nashville, TN, USA). Nashville, 2021. P. 6498–6508. https://doi.org/10.1109/CVPR46437.2021.00643.
22. Roberts M., Ramapuram J., Ranjan A., Kunar A., Bautista M.A., Paczan N., Webb R., Susskind J.M. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (10–17 October 2021, Montreal, QC, Canada). Montreal, 2021. P. 10892–10902. https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01073.
23. Kajiya J.T., Herzen B.V. Ray tracing volume densities. *ACM SIGGRAPH Computer Graphics.* 1984. Vol. 18, N 3. P. 165–174. https://doi.org/10.1145/964965.808594.
24. Kusupati U., Cheng S., Chen R., Su H. Normal assisted stereo depth estimation. *Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (13–19 June 2020, Seattle, WA, USA). Seattle, 2020. P. 2186–2196. http://dx.doi.org/10.1109/CVPR42600.2020.00226.

V. Savin, O. Kolodiazhna

ADAPTING NEURAL RADIANCE FIELDS (NeRF) FOR 3D SCENE RECONSTRUCTION UNDER DYNAMIC ILLUMINATION CONDITIONS

Abstract. We address the problem of novel view synthesis using Neural Radiance Fields (NeRF) for scenes with dynamic illumination. NeRF training utilizes photometric consistency loss that is pixel-wise consistency between a set of scene images and intensity values rendered by NeRF. For reflective surfaces, image intensity depends on the viewing angle, and this effect is taken into account by using ray direction as the NeRF input. For scenes with dynamic illumination, image intensity depends not only on the position and viewing direction but also on time. We show that this factor affects NeRF training with standard photometric loss function and decreases the quality of both image and depth rendering. To cope with this problem, we propose to add time as additional NeRF input. Experiments on ScanNet dataset demonstrate that NeRF with modified input outperforms the original model version and renders more consistent 3D structures. The results of this study could be used to improve the quality of training data augmentation for depth prediction models (e.g., depth-from-stereo models) for scenes with non-static illumination.

Keywords: computer vision, neural radiance fields, dynamic illumination, view synthesis, 3D scene reconstruction..

Надійшла до редакції 24.04.2023