האוניברסיטה הפתוחה
המחלקה למתמטיקה ומדעי המחשב

NeRF-ים של סצנות גדולות:
תאוריה וישומים עכשוויים

עבודה מסכמת זו הוגשה כחלק מהדרישות לקבלת תואר
"מוסמך למדעים" M.Sc במדעי המחשב
באוניברסיטה הפתוחה
המחלקה למתמטיקה ולמדעי המחשב

על ידי
אביב סבר

העבודה הוכנה בהדרכתה של ד"ר מיריי אביגל

07/2024

תוכן עניינים

טבלת איורים

תקציר

שחזור תלת-ממד של סצנות אורבניות רחבות היקף באמצעות NeRF

עבודת גמר זו מתמקדת בשחזור תלת-ממדי של סצנות אורבניות גדולות בהתבסס על רשתות NeRF.
שחזור תלת מימד - גם חזותית וגם גאומטרית - מקלט דו מימדי היא בעיה מורכבת, וסצנות חיצוניות גדולות מוסיפות על המורכבות הקיימת התמודדות עם מורכבות גיאומטרית גבוהה, חוסר אחידות בתאורה והיקף נתוני קלט עצום. במשך שנים היה תחום שחזור תלת מימדי חזקתם של אלגוריתמים קלסיים שונים, כגון MVS (Multi View Stereo) ו-SLAM (Simultanious Localization and Mapping).
בשנים האחרונות פרצו אליו אלגוריתמי למידה עמוקה ובראשם NeRF (Neural Radiance Fields) ונגזרותיו הרבות. NeRF-ים הן רשתות מבוססות למידה עמוקה המאפשרות ייצוג מרומז ומתוצמת של מבנה סצנה תלת מימדית - הן מבחינת גאומטריה והן מבחינת צבע - ומאפשרות שחזור חוזי ותמונות עומק של סצנה שכזו מנקודות מבט חדשות . למרות ריבוי העבודות העוסקות ב- NeRF , רק מעטות עוסקות במתארים המכסים שטח נרחב - סצנות גדולות, ובפרט סצנות אורבניות גדולות. מעטות עוד יותר עוסקות במוצהר בסצנות שלא ניתן לייצג ב-NeRF בודד.
עבודה זו סוקרת את הבסיס התיאורטי והאלגוריתמי של NeRF ו-MipNeRF, הנדרש להבנת הפתרונות השונים המוצגים, את התכונות והאתגרים המייחדים את תת הבעיה של סצנות גדולות ,ואת הפתרונות המובילים בה.
בין השיטות הנחקרות נכללות Bungee-NeRF, Block-NeRF, Mega-NeRF, URF, Instant-NGP ו-Neuralangelo. העבודה מציגה את הפתרונות האלגוריתמיים השונים, דנה בנקודות העבודה השונות שכל פיתרון מתמודד איתן, בוחנת את סט הבעיות שכל פיתרון בא לפתור ומעלה אפשרויות לשילוב בין העבודות השונות. בנוסף העבודה דנה בשאלות יסודיות הקשורות לשחזור סצנות גדולות כגון מהי סצנה גדולה, חלוקה של סצנה לתתי חלקים, איסוף נתונים והתמודדות עם מורכבות ויזואלית.

כבסיס לדיון, העבודה גם כוללת רקע תיאורטי קצר על שיטות וכלים קלסיים להתמודדות עם בעיית השיחזור התלת מימדי. שיטות וכלים אלו משמעותיים לדיון בנושא, בין אם ככלים לטיפול מקדים, כהשראה לחלקים מהאלגוריתמיקה המוצגת, או כדרך להסביר תהליכים חלופיים המוצגים בעבודות הנסקרות.

התוצאות המוצגות בעבודה מדגימות את ההתקדמות המשמעותית שהושגה בשחזור סצנות אורבניות גדולות באמצעות NeRF. מתוך גוף הפעילות הנרחב הנכלל תחת "אלגוריתמי NeRF", עבודה זו מרכזת את הפתרונות המתמודדים עם תת-הבעיה של סצנות גדולות, מציגה את היתרונות והחסרונות של כל פיתרון -הן יחסית לאחרים והן באופן עצמאי - ומציעה כיווני מחקר עתידיים בנושא.

1. מבוא

בעולם הראייה הממוחשבת, שחזור מודלים תלת מימדיים פוטוראליסטיים מאיסופים דו מימדיים - דהיינו תמונות - הוא בעיה וותיקה ומורכבת. צרכים ואילוצים שונים העלו כמובן פתרונות שונים, שנזכיר בקצרה בהמשך העבודה. שאלת הבסיס שמובילה את העבודה הזו היא: בהנתן אוסף תמונות מתאימות מאיזור גאוגרפי כלשהו A, וספציפית כש-A היא סצנה גדולה ו/או אורבנית, מה תראה מצלמה כלשהי C, שתמוקם בנקודה P ובאוריינטציה O - מבחינת תמונת RGB, ולעיתים גם מבחינת מפת העומקים המתאימה. המיקום של C לאו דווקא נכלל באיסוף הקלט, ואין שום אילוץ לגבי הפרמטרים הפנימיים שלה ביחס למצלמות האוספות.

NeRFs, או בשמם המלא Neural Radiance Fields [1, 1.1], הוא תחום שפרץ בלמידה עמוקה ב 2020, והציג גישה חדשה, מבוססת MLP, ליצירת תמונות מרונדרות של סצנה מתוך אוסף דליל של תמונות, מנקודות מבט השונות מתמונות האיסוף.

הגישה מאפשרת רינדור בפירוט גבוה, ומדגימה יכולות טובות יותר מפתרונות קלסיים להתמודדות עם השתקפויות, אפקטים לא למברטיים ומשטחים חלקים.
מכיוון שהסצנה מיוצגת באופן מרומז במשקולות המודל המאומן, ולא באופן מפורש כייצוג ווקסלי דיסקרטי - (discrete grids / voxel space/ triangle meshes) - דרישות הזיכרון קטנות בהרבה בפירוט הגיאומטרי וברזולוציה הפיקסלית המדוברת.

מצד שני - NeRF-ים לא חפים גם מחסרונות: גם אימון המודל וגם הרינדור פר תמונה דורשים זמן רב יחסית - כתלות גם בגודל הסצנה **וגם** בגודל כל תמונה מרונדרת, הרשת המאומנת מתאימה ומתארת סצנה ספציפית, לרוב תמונות הקלט הן בסקלה בודדת, ובמרחק ממוצע דומה מהסצנה - ומכאן עולה חיסרון נוסף לנקודת העבודה הנדונה - הגישה הבסיסית מתאימה לסצנות יחסית קטנות וממוקדות, כגון אובייקט ספציפי, חדר, מבנה בודד וכו', ופחות לסצנות נרחבות.

ייצוג סצנות גדולות בכלל, וגדולות ועירוניות בפרט, מעלה אתגרים ספציפיים נוספים: אחוז שטח הכיסוי של הסצנה בתמונה בודדת קטן יותר, בסיס הנתונים לאימון לרוב כולל הרבה יותר תמונות, קנה המידה והרזולוציה בין התמונות השונות עלול להיות שונה מאוד (לעיתים בסדרי גודל).

מאז פירסום [1.1] ב-2020 המאמר (או גרסאות שלו) צוטט מעל 3500 פעם, ובהערכה זהירה - עד לכתיבת שורות אלו - פורסמו כבר מעל 200 מאמרים שעוסקים ספציפית בתחום ה-NeRF-ים, שבאים לתת מענה, במשולב או בנפרד, לבעיות שמנינו ולאחרות שעלו.
רק בודדים מתוכם מתייחסים מפורשות לסצנות גדולות ולבעיות המייחדות אותם. בודדים אחרים מציעים שיפורים, הרחבות ועבודות המשך שיכולות להיות מהותיות לייצוג סצנות גדולות, אף שהמאמרים עצמם לא מתמקדים בתת בעיה זו.

מטרת עבודה זו היא לסקור את המחקר העכשווי בתחום העוסק ב- NeRF-ים, הבא לתת מענה לשחזור פוטוראליסטי תלת מימדי של סצנות גדולות בכלל, ושל סצנות אורבניות גדולות בפרט.

לצורך כך נציג את הרקע התיאורטי הנדרש, נציין את סט הבעיות הכללי ל NeRF-ים והפרטני לאופי הסצנה היחודי, נסקור מספר פתרונות הבאים, כל אחד בדרכו, לתת מענה לתת-סט כלשהו של הבעיות הנ"ל, ונסיים בדיון על שילובים אפשריים בין הפיתרונות המוצגים ועל עתיד המחקר בתחום. בנוסף, נכלול ברקע התיאורטי סקירה קצרה של כלים ושיטות קלסיות לשחזור תלת מימד, שהבנה שלהם משמעותיים לדיון בעבודות המוזכרות, בין אם ככלים ששימשו לטיפול מקדים בקלט, או כהשראה לאלגוריתמיקה המוזכרת.

2. רקע תיאורטי

כדי לדון ב- NeRF-ים ובשחזור תלת מימד פוטוראליסטי (קלאסי או מבוסס למידה עמוקה), יש צורך לדון במספר נושאים ומושגים, כדי לבנות את הבסיס ההגדרתי שעליו נקיים את הדיון.

בפרק זה נעסוק במספר תחומים: הגדרות בסיסיות בעולם הלמידה העמוקה, הצגה מעמיקה של NeRF-ים, כולל היתרונות והחסרונות המייחדים אותם, הרחבה של NeRF בשם Mip-NeRF - שלא נותנת מענה ישירות לסצנות גדולות, אך היוותה השראה למספר פתרונות שנדון בהם - וסקירה קצרה של מושגים מעולם השחזור התלת מימדי הקלסי, בעיקר בנושאים שמהווים כלי טיפול מקדים בחלק מהרשתות שנציג, או שעיקרון השימוש היווה השראה לפתרון כלשהו.

2.1 למידה עמוקה:

למידה עמוקה הוא ענף בלמידת מכונה, העוסק בלמידה מבוססת רשתות נוירונים. ה"עומק" בלמידה עמוקה מגיע ממספר השכבות הנסתרות שבין שכבת הקלט ושכבת הפלט. התחום כולל בתוכו מודלים מתחומי עניין רבים ומגוונים, מניתוח נפחי מידע גדולים (big-data), דרך בעיות שונות ומרובות מעולם הראייה הממוחשבת - סגמנטציה, קלסיפיקציה, רשתות קונבולוציה ומודלים גנרטיביים רק כדוגמאות אנקדוטליות - וכלה בניתוח שפה טבעית (NLP) ומודלי שפה גדולים (LLMs). עבודה זאת מניחה הכרות בסיסית עם מושגים ועקרונות של התחום, בלעדיהם יהיה קשה לעקוב אחרי פתרונות תת הבעיה הספציפית שבה עסקינן.

2.1.1 *MLP - פרספטרון רב-שכבתי*

פרספטרון רב-שכבתי (Multi Layer Perceptron) הוא מבנה רשת בסיסי, המורכב כולו ממספר שכבות בעלות קישוריות מלאה. מוכר גם כ-
Fully Connected Neural Network (FCNN). נהוג לאפיין MLP בעומק הרשת (כמות השכבות) ומספר הערוצים בכל שכבה.

2.1.2 *NeRF: Neural Radiance Fields*

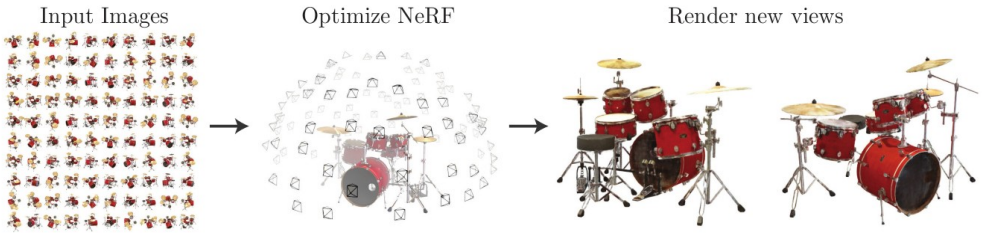
כהגדרה ראשונית, NeRF הוא ייצוג סצנה נוירוני מבוסס מיקומים, המתבסס על כלי רינדור נפחי על מנת לבצע אופטימיזציה ע"י שימוש בפונקציית מחיר הניתנת לגזירה, כדי לשחזר סט של תמונות קלט שנאספו ממיקומים ידועים. ברמה הנמוכה ביותר, לאחר האופטימיזציה, מודל NeRF יכול להחזיר צבע ורמת אטימות לפיקסל נתון מכל כיוון הסתכלות נדרש. יכולת זו יכולה לשמש לשיחזור מבטים חדשים על הסצנה, מנקודות מבט חדשות יחסית לקלט - לדוגמה בכיוון ההסתכלות, במרחק מהסצנה, או ב- FOV (Field Of View). (ראו לינק לאתר המאמר בסוף תת הפרק).

המונח NeRF כקיצור לשדות קירון נוירוניים (Neural Radiance Fields) נטבע על ידי Mildenhall *et al.* [1.1][1]. עבודתם עוסקת בחיזוי תמונה של סצנות מורכבות, מנקודת מבט חדשה, שלא נכללה בסט האימון. חיזוי זה משתמש באופטימיזציה של פונקצייה רציפה, המתארת את הסצנה מבחינה נפחית. לכל קואורדינטה במרחב הסצנה וכיוון הסתכלות נתון, הפונקצייה מחזירה צבע ורמת צפיפות - תוך שימוש בסט דליל של מבטי קלט. בזמן שייצוג קלסי של מרחב תלת מימדי דורש משטחים או גופים דיסקרטיים לשם הייצוג, סצנה נפחית רציפה מסוגלת לתאר כל נקודה במרחב מבחינת צבע, צפיפות ואטימות. האלגוריתם מייצג את הסצנה ע"י שימוש ברשת עמוקה בעלת קישוריות מלאה (MLP), שהקלט אליה הוא וקטור 5D - קואורדינטת מיקום תלת מימדית X=(x,y,z) וכיוון מבט (θ,ϕ) כקואורדינטות קוטביות כדוריות (spherical polar coordinates). באופן מעשי, מאחר וכיוון המבט מתורגם בכניסה למערכת מוקטור קוטבי כדורי לוקטור כיוון קרטזי תלת מימדי מנורמל d (כווקטור יחידה) - הקלט **מעשית** הוא וקטור 6D קרטזי המייצג מיקום וכיוון מבט.
הפלט של הרשת המאומנת הוא ערך אטימות (opacity) כצפיפות נפחית בנקודה X, ורמת הקירון המוחזרת מ-X לכיוון המבט הנתון. רמת הקירון מיוצגת על ידי צבע (RGB) ויכולה להשתנות בהתאם לכיוון המבט המבוקש, ומדד האטימות מתאר את רמת ההשפעה שיש לדגימה בנקודה על כל הדגימות שמאחוריה לאורך כל הקרניים העוברות דרך X - מאטום לחלוטין, שאז לכל הערכים המרוחקים יותר מנקודת המבט אין כל השפעה, ועד שקוף לחלוטין, שאז אותה דגימה נתונה לא משנה כלל את סכימת כל הדגימות המרוחקות יותר על הקרן.
הרכבת המבטים החדשים נעשית על ידי תחקור קוארדינטות 5D לאורך קרני מצלמה ושימוש בשיטות קלסיות של רינדור צפיפות נפחית (volume densities ray tracing)[8] על מנת להטיל את פלט הצבעים והצפיפות על תמונה - קרן עבור כל פיקסל . רינדור נפחי הוא אינטגרטיבי במהותו, כשערך כל פיקסל בודד נסכם מבחינת צבע וצפיפות לאורך קרן נתונה, ולכן גם גזיר באופן טבעי. תכונה זאת מאפשרת שחזור כשהקלט היחיד הנדרש הוא סט תמונות, עם מיקומי מצלמות ידועים.

---

1 כדאי לשים לב שהגרסה המצוטטת יותר [1] מאוחרת יותר, ומשמיטה מספר פרטי מימוש, תוך הפניה למאמר המקורי [1.1] להשלמת הפרטים.

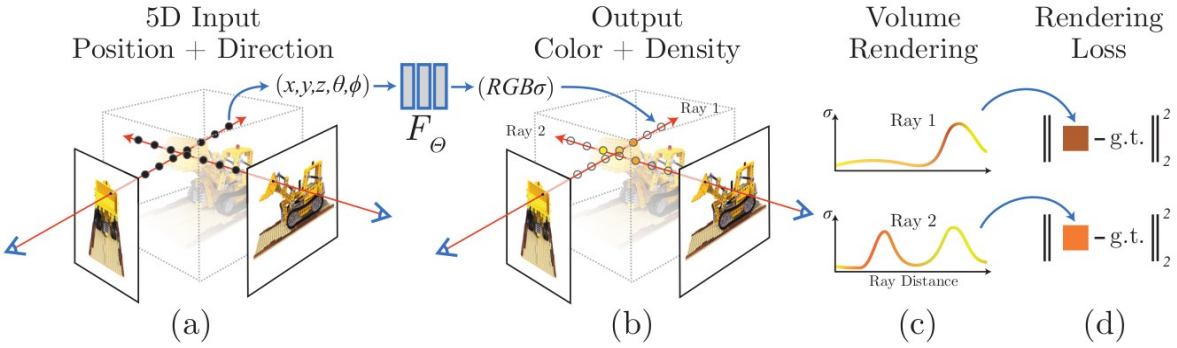
באיור 1 ניתן לראות הפשטה של תיאור התהליך: מקלט של תמונות ומיקומי מצלמות (ראשון משמאל), דרך אימון רשת הנרף (שני משמאל), ועד שני מבטים חדשים שלא נכללו בתמונות הקלט (תמונה ראשונה ושנייה מימין).



איור 1: NeRF - הצגה מפושטת של התהליך.
מתוך [1.1]

משערכים את ייצוג הסצנה ה-5D הרציפה ע"י רשת MLP הבאה: $F_\Theta : (x, d) \rightarrow (c, \sigma)$, כאשר x הוא מיקום מצלמה תלת מימדי, d מייצג כיוון מבט בייצוג כדורי קוטבי, c הוא ערך צבע ב RGB, ואילו σ מייצג את מקדם הצפיפות (או השקיפות) בנקודה. מבצעים אופטמיזציה על סט המשקולות Θ על ידי דגימה של קואורדינטות 5D לאורך קרני המצלמות השונות, הזנה של קואורדינטות אלה ל F_Θ, כדי לקבל ערכי צבע וצפיפות. על ידי שימוש בטכניקות רינדור נפחי מייצרים מתוצרים אלה תמונת פלט, שבה משתמשים כדי לבצע מינימיזציה על ההפרשים בין תמונת הפלט והמקור.

טכניקות רינדור נפחי קלסיות סובלות מבעיית סקלביליות לתמונות ברזולוציה גבוהה, בגלל הצורך האינהרנטי בדגימה דיסקרטית צפופה יותר בדגימת המרחב התלת מימדי. NeRF-ים עוקפים בעיה זו ע"י קידוד נפח רציף בפרמטרים של הרשת המאומנת. פיתרון זה לא רק מייצר תמונות ברמת פירוט גבוהה משמעותית, אלא גם מציגה דרישות זיכרון זעירות יחסית לנפח האכסון הנדרש לייצוג של מרחב דגום מתאים.

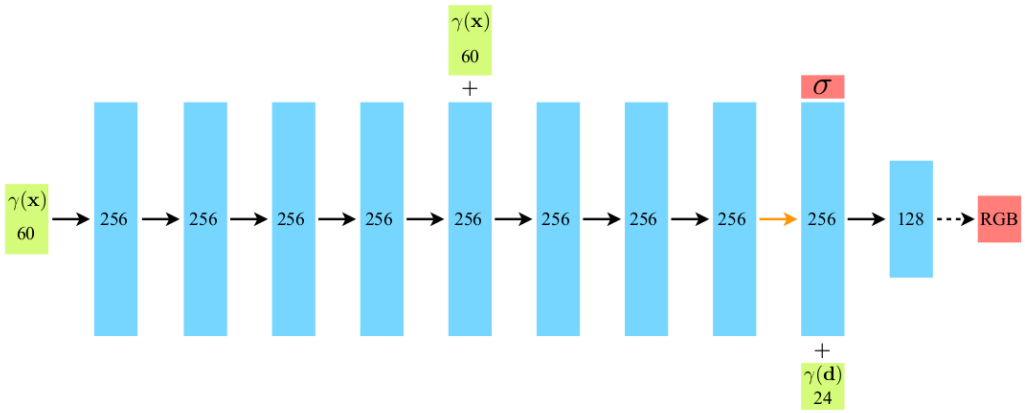


איור 2: תאור כללי של תהליך האימון.
מתוך[1.1]

באיור 2 מוצגים שלבי האימון:
a) קלט האימון - מכל תמונת קלט מגרילים נקודות 5D (מיקום וכיוון הסתכלות) לאורך קרני המצלמות.
b) פלט האימון - מזינים את הקלט לרשת F_Θ ומקבלים כפלט צבע וצפיפות. משתמשים בערכי הפלט ובטכניקות רינדור נפחיות כדי להרכיב תמונת פלט.
c) רינדור נפחי של הפיקסל התואם לקרן - מבצעים אינטגרציה של הדגימות לאורך הקרן
d) חישוב פונקציית מחיר - פונקציית רינדור (c) גזירה ולכן ניתן לבצע מינימיזציה על פונקציית מחיר של ההפרש בין תמונת הפלט c לתמונת הקלט.

מבנה הרשת: כזכור, הקלט בכל דגימה מורכב ממיקום תלת מימדי X וכיוון הסתכלות d, והפלט הוא צפיפות נפחית σ וערך צבע **c**=(r,g,b). על מנת לעודד את המודל לעקביות בריבוי מבטים, הרשת מתוכננת לחיזוי σ על בסיס X בלבד, בזמן שאת הצבע **c** הרשת מאומנת על בסיס X, d ביחד. מזינים את X (אחרי קידוד מיקום) ל-MLP המורכב מ-8 שכבות בעלות קישוריות מלאה, עם אקטיבציית ReLU בינהן ו-256 ערוצים, שמוציא את σ ווקטור מתאר באורך 256

מימדים. אל ווקטור זה משרשרים את כיוון ההסתכלות d (שוב אחרי קידוד מיקום), ומזינים אותו לשכבה בעלת קישוריות מלאה נוספת (עם אקטיבציית ReLU ו-128 ערוצים) שמוציאה את ערך הצבע תלוי כיוון ההסתכלות **c**.



איור 3: סכמת מבנה רשת NeRF.
מתוך [1.1]

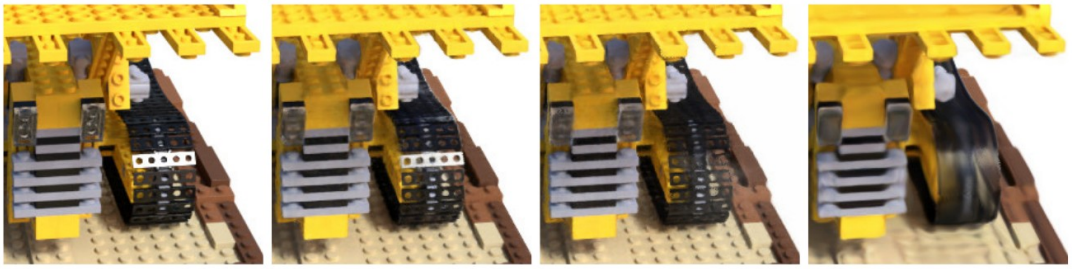
באיור 3 מוצגת סכימת מבנה רשת ה- NeRF. ווקטורי קלט מצויינים בירוק, שכבות נסתרות בכחול, פלטים באדום. המספר בתוך כל בלוק מציין את המימד. כל השכבות הן בעלות קישוריות מלאה, חיצים שחורים מסמלים אקטיבציית ReLU, חיצים כתומים שכבות ללא אקטיבציה, חיצים מקווקווים מסמלים אקטיבציית סיגמואיד, ו "+" מסמל שרשור וקטורים. γ(..) מייצג את פונקציית קידוד המיקום (ראו בהמשך- נוסחא (1)).

קידוד מיקום (positional encoding-PE): לרשתות עמוקות קיימת העדפה ללימוד פונקציות בתדר נמוך, כפי שהוצג ע"י Ahaman *et al.* [9]. נסיונות ב NeRF לאמן את F_Θ ישירות על קלט של $xyz\theta\phi$ העלו שהרשת מתקשה בייצוג שינוי בתדר גבוה בצבע ובגאומטריה, בגלל מרחב הפרמטרים הנמוך. בנוסף, [9] מראים שמיפוי הקלט למרחב ממימד גבוה יותר על ידי פונקציות מתדר גבוה מאפשר לרשתות להתאמן על דאטא המכיל שינויים בתדר גבוה. במאמר Attention is All you need [10], בו הוצגו לראשונה טרנספורמרים[2], המשתמשים בפונקציה סינוסיאודלית כדי להזין לכל טוקן מיקום במשפט, בכדי לאפשר הזנה של משפט שלם לרשת.
ב- NeRF [1.1] בחרו בהשראתם למפות את $\mathbb{R}\to\mathbb{R}^{2L}$ על ידי מתארי פורייה

$$\gamma(p) = \left(\sin\left(2^0\pi p\right), \cos\left(2^0\pi p\right), \cdots, \sin\left(2^{L-1}\pi p\right), \cos\left(2^{L-1}\pi p\right)\right). \quad (1)_{[1]}$$

L תוחם את המעלה המקסימלית המשמשת למתאר הפורייה. מעשית במאמר L=10 עבור $\gamma(\mathbf{x})$ כש- x הוא מיקום תלת מימדי, ו-L=4 עבור $\gamma(\mathbf{d})$, כש-d הוא כיוון ההסתכלות.
המיפוי נעשה עבור כל אחד מערכי הקואורדינטות בנפרד, על מנת לאפשר לMLP לשערך פונקציית תדר גבוה יותר. קידוד זה מוזכר בהמשך גם כ- **PE** (**P**ositional **E**ncoding).

2 טרנספורמרים הם סוג של רשת נוירונים עמוקה, שפותח במסגרת תחום עיבוד שפה טבעית (למרות שכיום משמש במגוון שימושים נוספים). הטרנספורמרים הציגו את תהליך תשומת הלב (Attention), בו מחושבים הקשרים של מילה נתונה (או טוקן) לחלקים אחרים במשפט. כדי להמנע מהצורך להזין את המשפט מילה מילה כמו ברשתות ישנות יותר (למשל RNN), ולאפשר בניית מפת תשומת לב לכל טוקן במקביל, כל מילה קודדה עם מתאר מקום סינוסיאדלי, בדומה ל (1), כך שכל המשפט מוזן לרשת ביחד, ועדיין למילים בחלקים שונים של המשפט יש השפעה שונה על חלקים אחרים.



איור 4: וויזואליזציה של השפעת קידוד המיקום והתלות בכיוון המבט.
מתוך [1.1]

באיור 4 ניתן לראות את השפעות החסרת קידוד המקום וכן חיסור שימוש בכיוון ההסתכלות.
משבצת ראשונה משמאל - תמונת המקור,
שנייה משמאל - שחזור ע"י המודל המלא,
שנייה מימין - ללא שימוש בכיוון הסתכלות (מזינים לרשת רק את X ללא d)
ראשונה מימין - ללא קידוד מיקום.
ניתן לראות שללא כיוון ההסתכלות, המודל לא מייצג את ההחזר הספקולרי על הזחל, בעוד ללא קידוד המיקום התוצאה מראה בברור את חוסר יכולת המודל לייצג פרטים בתדר גבוה גם בגאומטריה וגם בטקסטורה, המביאה לתמונה מוחלקת יתר על המידה.

דגימת נפח היררכית: כדי לאמן את הרשת, יש לדגום מספר דוגמאות לאורך כל קרן מצלמה נבחרת מכל תמונות הקלט. דגימה רנדומלית ללא תכנון מביאה לדגימה לא יעילה: חללים פתוחים מחד ואזורים נסתרים מאידך , שלא תורמים לתמונה המרונדרת, יידגמו שוב ושוב. כפיתרון מוצע תהליך בו מפזרים את הדגימות באופן יחסי לציפייה מהשפעתם על הרינדור. הרשת מאמנת בו זמנית שני MLP, אחת ברזולוצייה גסה MLP_c ואחת עדינה MLP_f. דוגמים את MLP_c תחילה, על ידי חלוקה של כל קרן ל N_c מקטעים,ומכל מקטע מוגרלת בהתפלגות אחידה נקודה רנדומלית. כעת דוגמים סט נוסף של N_f דגימות , אבל דוגמים אותן בהתחשב ב-N_c הדגימות הקודמות, כך שבחירת הדגימות מוטה כלפי אזורים רלוונטיים יותר בסצנה. בסופו של דבר משערכים את MLP_f עם האיחוד של N_c+N_f הדגימות כדי לקבל את הצבע עבור אותה קרן,תוך דיסקרטיזציה לא אחידה (ויותר מתאימה למבנה הסצנה) של ייצוג מרחב האינטגרציה.
לשם ההבהרה - עבור הרינדורים המוצגים במאמר [1.1 ,1] $N_c=64$, $N_f=128$, כך שבסופו של דבר נדרשו $N_c+(N_c+N_f)=256$ שאילתות לרשת פר קרן.

תיחום הנפח: השיחזורים המוצגים במאמר הם גם של סצנות סינטטיות, בעלות תיחום ברור, וגם סצנות טבעיות, שבהן לא קיים כל תיחום, ותיאורטית אפשרי שיהיה שהסצנה תיפרס ממישור המצלמה האוספת ועד האופק (האינסוף מבחינת גאומטריה פרוייקטיבית - projective geometry).

בסצנות סינטטיות מומר קנה המידה כך שתיחום הנפח יהיה תיבה שאורך כל פאה שלה 2 יחידות, ומרכזה במרכז הסצנה, כך שהמרכז הוא (0,0,0) מהתיבה נמשכת בכל ציר בין [1, -1]. הרשת דוגמת רק בתוך נפח תחום זה.

בסצנות טבעיות משתמשים בקואורדינטות מסך מנורמלות (normalized device coordinates), כדי למפות את טווח העומקים במצלמה לטווח [1,-1]. זה מביא לכך שכל מקורות הקרניים יוצאים ממישור החיתוך הקרוב (near plane), הקרניים הפרספקטיביות בקואורדינטות מצלמה (camera coordinates) הופכות לקריים מקבילות במרחב המחושב, ומשתמש בעומק משוחלף (inverse depth / disparity) במקום בעומק מטרי, כך שכל הקואורדינטות הופכות לתחומות (bounded). נירמול טווח זה מביא לכך שהנפח המשוחזר תחום בקוביה $[1, -1]^3$. תיחום זה מכיל בתוכו הנחות ניסתרות על הסצנה ועל המצלמות האוספות, שנדון בהן בהמשך.

זמנים ונפחי זיכרון: זמן אימון (על Nvidia V100 יחיד) הוא כ 100-300k איטרציות לסצנה בודדת, ובהערכה גסה 1-2 ימים לסצנה. מצד שני, משקל הזיכרון הנדרש לייצוג הרשת ומשקלותיה נמוך בסדרי גודל מגריד[3] ווקסלי תלת מימדי דיסקרטי המקביל הנדרש - בהשוואה לשיטה מתחרה שהוצגה במאמר [1.1] נדרשו 15 GB לייצוג גריד מפורש דיסקרטי למרחב הסצנה, לעומת רק 5 MB לייצוג המשקולות של רשת ה- NeRF. כפי שהוצג יפה במאמר [1.1] - דרישות הזיכרון של NeRF נמוכות מסך נפח תמונות הקלט הנדרשות לאימון הרשת לבדן.

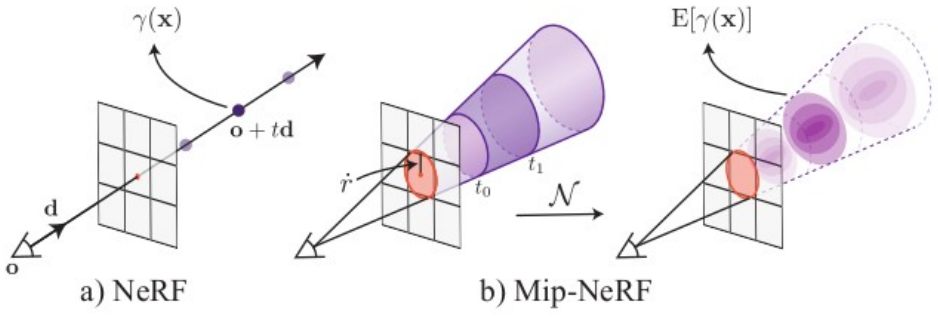
לגבי רינדור - יצירת תמונה אחת מסצנה טבעית שהוצגה במאמר דרשה כ 762k קרניים , או בין 150 ל 200 מיליון שאילתות לרשת המאומנת. על Nvidia V100 (כמוצג במאמר) מדובר על כ- 30 שניות לתמונה. זה המקום לציין שרינדור בשיטות קלסיות ממודל משולשים בעל טקסטורה מובצע כבר שנים בקצב גבוה של 30-60 FPS ואף יותר, גם על חומרות חלשות יותר.

מיקומי מצלמות: בסצנות הסינטטיות, מיקומי המצלמות בשימוש היו מיקומי ה GT. בסצנות הטבעיות הריצו חבילת **SFM** (ראו ב- 2.3) על תמונות הקלט, ומיקומי המצלמות המתוקנות הוזנו לNeRF. חבילת ה SFM שהורץ במאמר (כמו גם בחלק המאמרים הנוספים שיעלו לדיון בהמשך) היא Colmap [14,15].

סרטי הדגמה: אם תמונה שווה אלף מילים, אז ווידאו שווה אלף תמונות (אם הוא 40 שניות ב 25 FPS). כותבי [1.1] פרסמו גם דף שמרכז סרטוני וידאו מתוך עבודתם, שמסבירים את עבודתם, מדגימים יכולות שונות ומדגישים גם יכולות שפחות רלוונטיות לנושא עבודה זו ולכן לא הוזכרו כמו הארה מחדש ושילוב במציאות רבודה. יש צדדים שונים בראייה ממוחשבת שמודגמים יותר טוב בתנועה ושינוי מאשר בתמונה סטטית, ולכן מומלץ לקורא הרוצה להתעמק לצפות: https://www.matthewtancik.com/nerf .

2.1.3 *Mip-NeRF*

Barron *et al.* - חלקם בין כותבי [1.1] - פירסמו מאמר המשך [11], העוסק בהרחבת יכולות למאמר המקורי. במאמר, שגם שמו וגם האלגוריתם עצמו שואבים השראה מאלגורימי mipmap בגרפיקה ממוחשבת משנות ה-80 וה-90, מוצגות שתי הרחבות לאלגוריתם ה-NeRF המקורי: עקיבת קרניים קונית (Cone tracing) ובהתאמה קידוד מיקום משולב (IPE - Integrated Positional Encoding). השימוש בשני כלים אלה משפר את יכולת ההתמודדות עם בעיות דגימה שעשויות לצוף ב NeRF - עיוות (aliasing) וקושי להתמודד עם תמונות ברזולוציות שונות. יישום ההרחבות הנ"ל הציג שיפור במהירות החישוב של כ- 7%, על סט מבחן מרובה רזולוציות הוצגה הפחתת קצב השגיאה של 60%, ובגלל מבנה הרשת המודע לקנה המידה, ניתן לבטל ב-Mip-NeRF את הצורך בהרצה על שתי רשתות, גסה ועדינה (ראה דגימת נפח היררכית), מה שהביא לצימצום מספר הפרמטרים בחצי יחסית ל- NeRF המקורי.



איור 5: דגימה ב- NeRF לעומת דגימה ב-Mip-NeRF.
מתוך [11]

באיור 5 מוצג אחד ההבדלים העיקריים בין NeRF [1.1] ו-MipNeRF [11]. ב-a) מוצג אופן דגימת הקרן ב-NeRF - מאחר והקרן אידיאלית, אין משמעות למרחק המצלמה מהסצנה, ובאופן משמעותי יותר - קלט מטווחים שונים לא יכול לייצג את הרזולוציות הקרקעיות השונות בסצנה. ב-b) לעומת זאת הדגימה מתרחשת בפרוסטום (Frustum) קוני המשקף את השפעת השינוי בטווח, והשימוש בקידוד מיקום אינטגרטיבי (IPE - ראו בהמשך) מאפשר לייצג זאת.

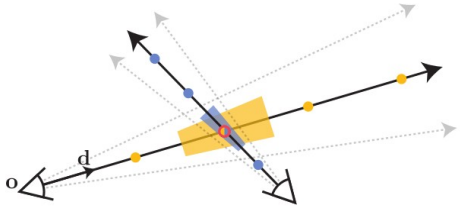
3 התרגום לעברית של grid הוא רשת או רישות, הגדרה שמתנגשת עם תרגום Network לעברית בסוגה שבה אנו עוסקים. אי לכך וכדי למנוע בלבול - מכאן ולהבא grid הוא "גריד" ו-Network היא רשת.

עקיבת קרניים קונית (Cone tracing): מתוך הבנה שערך פיקסל מושפע מאיזור גדול יותר מרוחק מהמצלמה, וקטן יותר קרוב למצלמה, Mip-NeRF בוחרים לא לדגום נקודות בדידות לאורך כל קרן מוטלת, אלא מגדירים קונוס קטום (conical frustum), הבנוי מחיתוך ע"י שני מישורים מקבילים למישור המצלמה (t_0, t_1 באיור 5) עם קונוס שראשו במיקום המצלמה o ורדיוס הבסיס שלו במישור התמונה o+d (d הוא גובה הקונוס) מוגדר כ- r , המוערך (לצורך פישוט חישובי) כרוחב הפיקסל באותה קואורדינטת עולם, כפול $2/\sqrt{12}$. לא הוסבר במאמר מקור הפקטור הנ"ל, מלבד שפישוט זה מביא לחתך קונוס עם שונות ב-x וב-y התואמת לרוחב הרצוי של הפיקסל בקואורדינטות עולם.

בנוסחה (2) [11] מוצגת קבוצת הנקודות הקיימות סביב x בתוך הקונוס הקטום בין t_0, t_1 כך ש :

$$\mathrm{F}(\mathbf{x},\mathbf{o},\mathbf{d},\dot{r},t_0,t_1)=\mathbb{1}\left\{\left(t_0<\frac{\mathbf{d}^{\mathrm{T}}(\mathbf{x}-\mathbf{o})}{\|\mathbf{d}\|_2^2}<t_1\right)\right. \quad (2)\ [11]$$
$$\left.\wedge\left(\frac{\mathbf{d}^{\mathrm{T}}(\mathbf{x}-\mathbf{o})}{\|\mathbf{d}\|_2\|\mathbf{x}-\mathbf{o}\|_2}>\frac{1}{\sqrt{1+(\dot{r}/\|\mathbf{d}\|_2)^2}}\right)\right\},$$

כאשר הסימון - $\mathbb{1}\{\cdot\}$ הוא פונקציה מייצגת : $F(X,\cdot)=1$ אם"ם x נמצא בתוך הפרוסטום הקוני המוגדר על ידי $(o,d,\dot{r},t_0,t_1)$, ו-0 אם x מחוץ לו.



איור 6:עקיבת קרניים קונית - פרוסטומים מצטלבים. מתוך [11]

באיור 6 ניתן לראות איך אותה נקודה, במרחקים וכיוונים שונים בין שתי מצלמות, מחזירה פרוסטומים קוניים שונים בתכלית.

קידוד מיקום משולב (IPE - Integrated Positional Encoding):

21

היות ובשיטה המוצעת דגימה היא כבר לא נקודה בודדת, אלא אינטגרציה של תיחום, גם קידוד המיקום PE כבר לא מספק, ונידרשת הגדרה אינטגרטיבית. הגישה הכי יעילה שמצאו Barron *et al.* [11] היא לחשב את סכימת ה PE של כל הקואורדינטות המוכלות בתוך הפרוסטום הקוני:

מתוך (1),(2) ניתן להסיק את (3) [11] $\gamma^*(\mathbf{o},\mathbf{d},\dot{r},t_0,t_1)=\frac{\int\gamma(\mathbf{x})\,\mathrm{F}(\mathbf{x},\mathbf{o},\mathbf{d},\dot{r},t_0,t_1)\,d\mathbf{x}}{\int\mathrm{F}(\mathbf{x},\mathbf{o},\mathbf{d},\dot{r},t_0,t_1)\,d\mathbf{x}}$, כלומר אינטגרל של קידוד פוריה של קבוצת הנקודות F(x...) (הנקודות הנכללות בקונוס הקטום) חלקי האינטגרל של F(x...),

אבל גישה זו בעייתית מאחר ולאינטגרל שבמונה אין פיתרון בתצורה סגורה. אם נבחן את (2) נראה שמאופן הגדרתה, (2) איננה רציפה ומאוד "מדרגתית", מה שלא מאפשר פיתרון סגור ואנליטי של השטח שתחת הקמור. במקום זאת, משערכים את הפרוסטום הקוני בעזרת התפלגות נורמלית רב מימדית (multivariate Gaussian). בהנחה ורוצים לייצג את הפרוסטום הקוני שסביב הקרן במקטע שבין t_0 ו-t_1 , וניתן להניח שהקונוס מעגלי, (מאחר והוא סימטרי סביב הקרן), ניתן לייצג גאוסיין שכזה בעזרת **d** , **o** ו- 3 פרמטרים נוספים: המרחק הממוצע לאורך הקרן μ_t , השונות לאורך הקרן σ_t^2, והשונות הניצבת לקרן σ_r^2 :

$$\mu_t = t_\mu + \frac{2t_\mu t_\delta^2}{3t_\mu^2 + t_\delta^2}, \quad \sigma_t^2 = \frac{t_\delta^2}{3} - \frac{4t_\delta^4(12t_\mu^2 - t_\delta^2)}{15(3t_\mu^2 + t_\delta^2)^2}, \quad \sigma_r^2 = \dot{r}^2\left(\frac{t_\mu^2}{4} + \frac{5t_\delta^2}{12} - \frac{4t_\delta^4}{15(3t_\mu^2 + t_\delta^2)}\right). \tag{4}$$

יש לשים לב שלצורך יציבות נומרית , ערכים אלה (μ_t , σ_t^2, σ_r^2) עוברים פרמטריזציה בהתייחס לנקודת אמצע $t_u=(t_0+t_1)/2$ ולחצי רוחב $t_\delta=(t_1-t_0)/2$. ממירים את הגאוסיאן הנ"ל ממערכת קואורדינטות של הפרוסטום הקוני למערכת עולם :

$$\boldsymbol{\mu} = \mathbf{o} + \mu_t \mathbf{d}, \quad \boldsymbol{\Sigma} = \sigma_t^2(\mathbf{d}\mathbf{d}^{\mathrm{T}}) + \sigma_r^2\left(\mathbf{I} - \frac{\mathbf{d}\mathbf{d}^{\mathrm{T}}}{\|\mathbf{d}\|_2^2}\right) \tag{5}$$

מכאן, נוכל לחלץ את ה- IPE, שהוא ההתפלגות הצפויה של קואורדינטה מקודדת מיקום המשקף את הגאוסיאן האמור.

נציג מחדש את נוסחת PE (1) עבור קואורדינטה $\mathbf{x}$ כך:

$$\gamma(\mathbf{x}) = \left[\sin(\mathbf{x}), \cos(\mathbf{x}), \ldots, \sin\left(2^{L-1}\mathbf{x}\right), \cos\left(2^{L-1}\mathbf{x}\right)\right]^{\mathrm{T}}, \tag{6}$$

ונשכתב אותה כמתאר פורייה

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 2 & 0 & 0 & & 2^{L-1} & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 & 0 & \cdots & 0 & 2^{L-1} & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 & & 0 & 0 & 2^{L-1} \end{bmatrix}^{\mathrm{T}}, \quad \gamma(\mathbf{x}) = \begin{bmatrix} \sin(\mathbf{P}\mathbf{x}) \\ \cos(\mathbf{P}\mathbf{x}) \end{bmatrix}. \tag{7}$$

מה שמאפשר לנו לחשב תצורה סגורה ל- IPE. נתבסס על העובדה שהשונות המשותפת (covariance) של טרנספורמציה לינארית של משתנה היא טרנספורמציה לינארית של השונות המשותפת של המשתנה (Cov [Ax, By]=ACov[x,y]B^{T}) כדי לבודד את הממוצע והשונות המשותפת של הפרוסטום הקוני **לאחר** המרתו ל PE בבסיס P:

$$\boldsymbol{\mu}_\gamma = \mathbf{P}\boldsymbol{\mu}, \quad \boldsymbol{\Sigma}_\gamma = \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}^{\mathrm{T}}. \tag{8}$$

לבסוף נאפנן את מתאר ה-IPE על ידי סינוסים וקוסינוסים:

$$\mathrm{E}_{x\sim\mathcal{N}(\mu,\sigma^2)}[\sin(x)] = \sin(\mu)\exp\left(-(1/2)\sigma^2\right) \tag{9}$$

$$\mathrm{E}_{x\sim\mathcal{N}(\mu,\sigma^2)}[\cos(x)] = \cos(\mu)\exp\left(-(1/2)\sigma^2\right) \tag{10}$$

ומכאן נוכל לחשב את מתאר ה-IPE כסינוסים וקוסינוסים של הממוצע והאלכסון של מטריצת השונות המשותפת:

$$\gamma(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathrm{E}_{\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma)}[\gamma(\mathbf{x})] \tag{11}$$

$$= \begin{bmatrix} \sin(\boldsymbol{\mu}_\gamma) \circ \exp(-(1/2)\operatorname{diag}(\boldsymbol{\Sigma}_\gamma)) \\ \cos(\boldsymbol{\mu}_\gamma) \circ \exp(-(1/2)\operatorname{diag}(\boldsymbol{\Sigma}_\gamma)) \end{bmatrix} \tag{12}$$

כשהסימן $\circ$ מייצג מכפלה איבר-איבר. היות ונדרש לקודד כל מימד באופן עצמאי, קידוד זה מסתמך רק על ההתפלגות השולית של $\gamma(\mathbf{x})$ ועל אלכסון מטריצת השונות המשותפת. בהתחשב בכך שלהפיק את מטריצת $\boldsymbol{\Sigma}_\gamma$ כולה יקר מאוד חישובית, עדיף לחשב ישירות רק את האלכסון

$$\operatorname{diag}(\boldsymbol{\Sigma}_\gamma) = \left[\operatorname{diag}(\boldsymbol{\Sigma}), 4\operatorname{diag}(\boldsymbol{\Sigma}), \ldots, 4^{L-1}\operatorname{diag}(\boldsymbol{\Sigma})\right]^{\mathrm{T}}, \tag{13}$$

שבתורו תלוי רק באלכסון של השונות המשותפת של המיקום התלת מימדי Σ, שניתן לחשבו כ:

$$\operatorname{diag}(\boldsymbol{\Sigma}) = \sigma_t^2(\mathbf{d}\circ\mathbf{d}) + \sigma_r^2\left(\mathbf{1} - \frac{\mathbf{d}\circ\mathbf{d}}{\|\mathbf{d}\|_2^2}\right) \tag{14}$$

חישוב ישיר כזה הופך את חישוב המתארים של IPE ליקר חישובית להפקה בערך כמו אלה של PE.

נסכם ונציג את הנוסחה המפושטת ל IPE כ:

(חזור ל24)

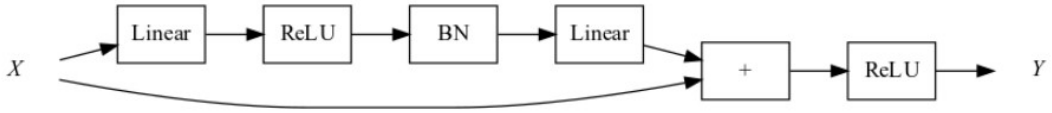
$$\gamma(\mu, \boldsymbol{\Sigma}) = \left\{ \begin{bmatrix} \sin(2^m \mu) \exp(-2^{2m-1} \operatorname{diag}(\boldsymbol{\Sigma})) \\ \cos(2^m \mu) \exp(-2^{2m-1} \operatorname{diag}(\boldsymbol{\Sigma})) \end{bmatrix} \right\}_{m=0}^{M-1} \quad (15)$$

[11]

2.1.4 רשתות שיוריות, חיבורי דילוג ובלוקים שיוריים

נהוג להניח, שרשתות עמוקות יותר מסוגלות להכליל טוב יותר מרשתות שטוחות (בעיקר לגבי רשתות **קונבולוציה** עמוקות), אבל רשתות עמוקות מאוד מציבות אתגר לגבי חישוב הגרדיאנט - מאחר והתפשטות הגרדיאנט לאחור (back propagation) מבוסס על כפל נגזרת שכבה נתונה בנגזרת השכבה שלפניה, ערכים גדולים או קטנים מדי גורמים לאפקט הגרדיאנט המתפוצץ או הנעלם בהתאמה, ומקשים על התכנסות ואימון.
משפחת רשתות ResNet [40] הציגו לראשונה מנגנון בשם **חיבור דילוג** (skip connection), העוקף קבוצה של שכבות והופך את לימוד פונקציית הזהות לברירת המחדל. בשימוש ברכיב כזה, השכבות הנעקפות צריכות ללמוד רק את התוספת לפונקציית הזהות, אם נדרשת כזו, או שאיפוס המשקולות בשכבות אלה יביא להתפשטות לפנים של פונקציית הזהות.

בלוק שיורי הוא מבנה בסיסי בהגדרת הרשת, הכולל בתוכו חיבור דילוג שכזה. לדוגמה ברשת בעלת קישוריות מלאה:



איור 7: דוגמה פשטנית לבלוק שיורי ברשת בעלת קישוריות מלאה. מתוך מדריך הלמידה בקורס למידה עמוקה, האוני' הפתוחה

אפשר להגדיר מספר סוגי בלוקים שיוריים שכאלה, ולשרשר אותם בצירופים שונים על מנת לבנות רשת עמוקה, מבלי לחשוש שבהתאמה לקלט הנלמד האימון ייכשל רק עקב עומק רב מדי.

מעבר לאיפשור של בניית ואימון רשתות עמוקות יותר, בלוקים שיוריים מאפשרים לשכבות מאוחרות יותר להיות מושפעים מרמות קודמות / לוקליות יותר / מפורטות יותר (למשל ברשתות קונבולוציה, שבהן מירב הפרטים בקלט) - והכל בהתאם למרחב הבעיה והמימוש.
השימוש בעיקרון של בלוק שיורי וחיבורי דילוג אומנם התחילו בתחום הראייה הממוחשבת ורשתות הקונבולוציה, אבל כיום משמש (או משמש השראה) בתחומים רבים, מ NLP עם LSTM ועד MLP ו-NeRF (ראה Bungee-NeRF בהמשך[21]).

2.2 שחזור תלת מימד - שיטות קלסיות:

עבודה זו עוסקת כולה באלגוריתמיקה מבוססת למידה עמוקה, המתמקדת בבעיות היחודיות לסצנות גדולות. יחד עם זאת, חלק מהמימושים שיוצגו בהמשך שואבים רעיונות מפתרונות קלסיים לתת בעיות, או משתמשים בכלים קלסיים כשלב מקדים, המטייב את הקלט. לטובת הקורא שלא מעורה בתחום הקלסי, נציג בחלק זה מושגים ומטה-אלגוריתמים מראייה ממוחשבת קלסית. ניתן לדלג על החלק במקרה של הכרות טובה עם עולם הבעיה, או לחילופין לחזור לסעיף זה לריענון, אם במהלך המשך הקריאה יוזכרו כלים, מושגים או אלגוריתמים לא מוכרים.

Baseline 2.2.1

למושג יש משמעויות שונות בדיסיפלינות שונות, אבל בגאומטריה פרוייקטיבית הכוונה היא למרחק בין מוקדי מצלמות שונות הצופות על אותה סצנה. משמעותי מאוד בחישוב נקודות תלת מימדיות על סמך עקיבה בין שתי מצלמות (ראו איור 9 בהמשך). ראו [33]

Volume density Ray Tracing 2.2.2

עקיבת קרניים (ray tracing) [8] היא טכניקת מידול לבחינת השפעה של מקורות תאורה על סצנה תלת מימדית מורכבת, מעולם הגרפיקה התלת מימדית. לדוגמה - במידול צללים ניתן להוציא קרניים מכל מקור אור בסצנה מורכבת, ולסכום את השפעת אותו מקור אור על נקודת הפגיעה במודל מבחינת הסתרות והצללות עצמיות. כמובן שצפיפות הקרניים והמרחק ממקור האור משפיעים על הדיסקרטיזציה של התוצאה, ונדרש איזון בין דרישות חישוב וארטיפקטים בולטים לעין.

טכניקות עקיבת קרניים העוסקות בצפיפות נפחית נועדו במקור לתת מענה למידול סצנות שקופות חלקית, כגון עננים, תוך התייחסות לפיזור ולהצללות עצמיות בתוך הענן. בהנתן נפח נתון שאותו רוצים למדל, מחלקים את הנפח לגריד תלת מימדי סדור, שעבור כל תא בו (voxel) מוגדרת האטימות שלו - כמה מהאור הנכנס לווקסל כזה יוצא בצד השני. בהנתן קרן מנקודת המבט החוצה את הנפח הממודל, מוצאים את כל הווקסלים שהקרן עוברת דרכם, ומחשבים לאחור כמה אור יגיע לפיקסל ספציפי בנקודת המבט. בהנחה שווקסל מסויים אטום (מכיל נפח מוצק) כל החישוב ממנו והלאה לא רלוונטי, ורק רמת הקירון של הווקסל הספציפי, והאטימות והקירון של כל הווקסלים שבינו לבין נקודת המבט, משפיעים על ערך הפיקסל הסופי.

הבנת הדיסקרטיזציה שרזולוציית אותו גריד תלת מימדי סדור יוצרת חיונית להבנת סיבוכיות החישוב, ומשפיעה גם על נפח הזיכרון הנדרש לייצוג וגם על כמות החישובים הנדרשים עבור כל קרן. ב NeRF-ים עקפו את בעיית נפח הזיכרון הנדרש ע"י הפיכת ייצוג פונקציית הצפיפות לרציפה במקום דיסקרטית.

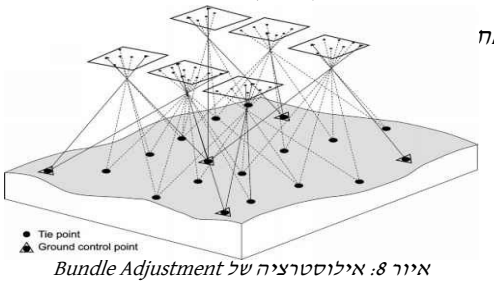
Simoultamous Localization And Mapping - SLAM 2.2.3

מציאת מיקומי מצלמה ומיפוי בו זמנית: שם כולל למשפחה של אלגוריתמי אופטימיזציה, שבהנתן סט תמונות ומיקומים, מחזירים ענן נקודות דליל של הסביבה, ובו זמנית משוערכים תיקונים למיקומי המצלמות, ואופציונלית גם פרמטרי מצלמה פנימיים שונים (Intrinsic parameters). השונות במרחב הפתרונות רחבה, ונובעת בעיקרה מאופי קלט שונה מחד, וצרכים שונים לשימוש בפלט מאידך. ברמת הקלט - קיים הבדל משמעותי האם מבוצע SLAM אינקרמנטלי (למשל ממקור וידאו) או בלוק (כל התמונות זמינות מראש בעת החישוב), האם נקודות הקשר בין התמונות (להלן "עקיבה") היא בצמדי תמונות או משורשרת על מספר תמונות גבוה יותר, האם העקיבה נתונה כבר או מחושבת תוך כדי, והאם משוערכים גם פרמטרי מצלמה פנימיים כגון FOV, principal point או עיוותי מצלמה רדיאליים. לגבי מיקומי המצלמות חשוב לציין, שפיתרון המצלמות המוחזר הוא קרטזי ויחסי בין המצלמות, ולכן לא ישקף הבדלים בפתרונות התלויים בקנה מידה או טרנספורמציות ריג'ידיות על כל הפתרון. רק אם סופקו ניחושים ראשוניים למיקומי המצלמות הכוללים מידע גאודזי (למשל מקור המיקום הראשוני הוא GPS), או יסופקו נקודות עוגן מתאימות, תהיה ההתכנסות מייצגת בסקלה ובמיקום מוחלט. ראו לדוגמא [36]

Bundle Adjustment - BA 2.2.4

Bundle adjustment הוא לב תהליך ה-SLAM, בו מתבצע חישוב מיזעור השגיאה בין מיקומי המצלמות ונקודות העקיבה בינהן - ביצוע האופטימיזציה עצמה. בהנתן סט מצלמות (מיקומים + פרמטרים פנימיים) וקישורים בין המצלמות השונות (עקיבה), ממזער את שגיאת ההטלה בין

המצלמות, ומבצע אופטימיזציה בו זמנית על מיקומי המצלמות והנקודות התלת מימדיות המוטלות מהעקיבה. לצורך זה ניתן לשנות את שיערוך מיקומי המצלמות, לשנות את שיערוכי הנקודות התלת מימדיות, ובהתאם לשחרור פרמטרים נוספים לשיערוך - גם לשערך עיוותי עדשה או פרמטרים פנימיים נוספים. זהו תהליך יקר חישובית, ולכן קיים איזון עדין בין צרכי נקודת העבודה לרמת החופש המוגדרת או מספר הפעמים שהתהליך ייקרא. ראו [33]



איור 8: אילוסטרציה של Bundle Adjustment

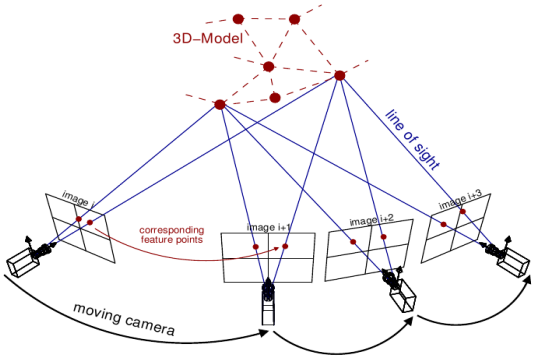
באיור 8 ניתן לראות אילוסטרציה של התהליך, המסבירה גם את מקור השם: בהנתן סט מצלמות אוויריות הצופות על שטח משותף, נמתח קרניים מכל אחת מהמצלמות אל נקודות הקשר שעל הקרקע.
תהליך מזעור השגיאה ינסה להזיז את מיקום נקודות הקשר ומנח המצלמות על מנת "להדק" את אלומות הקרניים, תוך התייחסות למשקולות השונים המוגדרים כמחיר להזזתם, והתוצאה הסופית תהייה מיקומי מצלמות מטוייבים וענן נקודות תלת מימדי דליל של הסצנה.

Structure From Motion - SFM 2.2.5

43,41,38,35,34,25

SFM (לעיתים נכתב גם כ- SfM) [37] הוא שם שניתן לתהליך קצה לקצה (pipeline) המקבל תמונות ומיקומים תואמים, מבצע SLAM ומחזיר שיערוך של ענן נקודות תלת מימדיות, ביחד עם מיקומי מצלמות מתוקנים. גם פה קיימת התייחסות מימושית לקלט אינקרמנטלי לעומת גלובלי. חלק מהחבילות הקיימות מסוגלות להתמודד עם מיקומי מצלמות חסרים בקלט גלובלי, ולבצע שיערוך מיקום מצלמות ראשוני לפני התהליך המלא.

באופן כללי, ובגרסתו הבסיסית ביותר, התהליך כולל: חילוץ מאפיינים מכל אחת מהתמונות, מציאת התאמות בין זוגות של תמונות, שיערוך מיקום ומנח ראשוניים למצלמות (אם נדרש), קליברציה לנתוני המצלמות (ממקור חיצוני או משיערוך) ובאופן איטרטיבי - סינון חריגים (בשיטות שונות), BA - עד להתכנסות לסף רצוי.



איור 9: אילוסטרציה של תהליך מבנה מתוך תנועה אינקרמנטלי.
נלקח מ http://theia-sfm.org/sfm.html

מאחר ופיתרון BA נאיבי יכול להגיע ל $O(N^3)$ במספר המצלמות, וכיסוי סצנה גדולה יכול להגיע לאלפי ועשרות אלפי תמונות, ברור שSFM איננו סקלבילי לעד, ואילוצי נפח זיכרון וזמן חישוב תוחמים את יחס שטח הכיסוי מול רמת הפירוט הנדרש. ב SFM אינקרמנטלי (למשל ממקור וידאו) נהוג לבצע חישוב בתדר גבוה רק על "חלון" זמן של מספר תמונות אחרונות, עם חישוב מקיף יותר על תמונות מפתח היסטוריות (keyframes) בתדר נמוך, שבא לתקן סחיפות, זיהוי מעגלים במסלול וכו'. בפיתרון בלוק (SFM גלובלי) ניתן לפתור תחילה סט שלדי (skelatal set) או לחשב עץ פורש של מצלמות עוגן. קיימים פתרונות המאפשרים למקבל את התהליך, אבל לא בזאת עסקינן בעבודה זו.

נהוג להשתמש במינוח SFM כשהרכיב המעניין את משתמש הקצה הוא ענן הנקודות, ואילו תיקון נתוני המצלמות הוא תוצר לוואי, ולהשתמש במינוח SLAM במצב ההפוך,שבו התוצר העיקרי הרצוי הוא נתוני המצלמות המתוקנים, אבל אין חובה. במאמרים שבהם דנה עבודה זו מרבים להשתמש במושג SFM, כשהכוונה היא דווקא לתוצרי מיקומי המצלמות המתוקנות / מטוייבות (refined camera poses), ללא התייחסות לתוצר ענן הנקודות. חשוב להבהיר ש-

NeRF קלסי לא משתמש בענני נקודות כקלט. במספר מאמרים שבהם דנה עבודה זו מוזכר שימוש ב SFM pipeline של חבילת קוד בשם Colmap [14,15].

זה גם המקום לציין שפורסמו מספר מאמרים על חבילות SLAM מבוססות DL, כגון deepSLAM [29], DynamicSLAM [30], וכן מודלי NeRF הכוללים טיוב מיקומי מצלמות, כדוגמת - BARF Bundle Adjustment for NeRF [31] או iNeRF [32], אבל רק באחד מהמאמרים שנסקרים בעבודה זו בחרו להשתמש בגישה שכזו.

Multi-View Stereo – MVS 2.2.6

38,34,26

Multi-View Stereo - MVS [33,38,39] הוא שם כולל לתחום שחזור תלת מימד מתוך דו מימד (תמונות). בקונוטציה של כלים או תהליכים, מדובר (לרוב) בתהליך שבהנתן תוצרי SFM (ענן נקודות ומיקומי מצלמות מתוקנים), מחלץ משטח (surface) ומשלש אותו לכדי סריג (mesh), ומטיל טקסטורה נבחרת לכל משולש (או קבוצת משולשים) מתוך תמונות הקלט. בחירת הטקסטורה משתמשת במיקומי המצלמות המתוקנים, כדי למזער את הארטיפקטים הנובעים משגיאות הטלה. לרוב כלי MVS לא ידרשו מהמשתמש תוצרי SFM, אלא יצפו לקלט בסיסי (תמונות ומיקומי מצלמות, כאמור), ויעטפו את תהליך ה- SFM פנימית.

תוצר של תהליך כזה הוא מודל משולשים בעל טקסטורות, שעל ידי שימוש בכלי גרפיקה ממוחשבת והאצת חומרה ניתן לרנדר ממנו תמונות בקצב אינטראקטיבי, מנקודות מבט שלא נראו בסט האיסוף.

מבחינת סצנות גדולות אלמנטים נוספים שמקורם בגרפיקת 3D ושאפשר לנצל היא יצירת רמות פירוט שונות (Level Of Detail -LOD) , שמאפשרים רינדור מהיר יותר פר נקודת מבט, על חשבון אחסון כמות גדולה יותר של קבצי מודל. שינויי הפירוט בין רמות הפירוט השונות הן כמובן דיסקרטיים, ומכאן שגם רמת הפרוט בזמן הרינדור דיסקרטי בטווח מנקודת המבט.

חלק מהכלים המוכוונים לסצנות גדולות מבצעים אופטימיזציה בזמן ובמשאבים ע"י חלוקת אזור העניין לתת חלקים (tiles), פיתרון BA או MVS לכל אחד מהם בנפרד (תלוי מימוש) ואיחוד המודלים לאחר מכן. בחירת חלוקת תמונות הקלט לתתי קבוצות מתאימות לרוב נעשה ע"פ הערכת עקבה קרקעית (footprint), ובחירה מטוייבת יכולה להקל על תהליך האיחוד בשלב מאוחר יותר.

Colmap 2.2.7

38,34,17,10

חבילת קוד פתוח [14,15] המספקת כלי SLAM, כגון מסלול פיתרון SFM גלובלי, וכן מסלול פיתרון MVS. במספר מאמרים שיידונו בהמשך שימשה על מנת לחלץ פרמטרי מצלמה פנימיים ומיקומי מצלמות מתוקנים של קלט in the wild (להבדיל מקלט סינטטי), ובאחרים שימש מסלול ה-MVS לשחזור קלסי של הסצנה לשם השוואה של הרינדור.

3. הגדרת מרחב הבעיה - מה מיוחד בסצנות גדולות

לשימוש ב-NeRF-ים לצורך שיחזור סצנות גדולות קיימים מספר אתגרים ספציפיים, שחלקם אף מחמיר כשמדובר בסצנות גדולות *אורבניות*. בסעיף זה נדון ונפרט מהם:

I אופי האיסוף ואופי הסצנה הטיפוסיים ל- NeRF הם סצנה ממוקדת אובייקט, כשטווח המצלמה ממוקד הסצנה פחות או יותר שווה, במעין כיפה מסביב לאובייקט.

I.i. הנחה של תמונות בסקלה דומה.

I.i.i בגלל ההטלה לנפח תחום בגודל $[-1,1]^3$ בסצנות טבעיות, תמונות בסקלה שונה משמעותית עלולות להשפיע מאוד על פרוט רמת הפרטים הנלמדת, או לייצר ארטיפקטים כתוצאה מעיוות (aliasing).

I.ii. הבדל משמעתי ברזולוציה הקרקעית בחלקים שונים בתמונה.

I.ii.i אופי איסוף כיפתי פחות מתאים לסצנות גדולות, ויגרום להבדלים גדולים מאוד בין רמת הפירוט הנלמדת מכל תמונה ומחלקים שונים באותה תמונה. פיזור הקרניים ואופי הדגימה המוגרלת עשוי להשתנות משמעותית בחלקים שונים של אותה תמונה, אם מדובר בעקבה קרקעית גדולה.

I.iii. יש הנחה נסתרת על אחוז חפיפה גבוה מבחינת שטח הכיסוי הנראה בין תמונות שונות. בסצנות גדולות עולה הסיכוי שתמונות לא יחלקו בינהן מספיק פרטים כדי להיות משמעותיות באימון. בסצנה אורבנית הבעיה משמעותית יותר, בגלל ריבוי הסתרות .

I.iv. ברוב הדוגמאות המוצגות במחקר על NeRF-ים תמונות הקלט צופות מחוץ לסצנה אל תוכה, ופחות דוגמאות עוסקות במצבים שבהם תמונת קלט נאספה **בתוך** הסצנה ורואה רק חלק קטן מכלל הסצנה - ארוע שבהחלט סביר באיסוף קרקעי באזור אורבני נרחב לדוגמה.

I.v. גודל הסצנה ניתן להגדרה במספר דרכים. לדוגמא ניתן להשתמש ברדיוס הכדור החוסם המינימלי המכיל את אזור העניין. ניתן להשתמש גם ברדיוס הכדור החוסם את מיקומי המצלמות המשתתפות. בסצנה מוכוונת אובייקט כמו באיור 1 ההבדל בין תוצאת שתי ההגדרות זניח. בסצנה אורבנית מאיסוף אווירי - ההבדל עלול להמדד בקילומטרים.

II סצנות גדולות הן... ובכן... גדולות.

II.i. אמנם NeRF חסכוני בזיכרון, אבל מבנה הרשת איננו סקלאבילי עד אין סוף. ייצוג סצנות גדולות מדי יגררו פיזור דגימות דליל מדי לאורך כל קרן, ויוריד את היכולת לתאר סצנה בפירוט הרצוי, או לחליפין - יידרוש כמות דגימות גבוהה משמעותית, ויעלה את זמני האימון והרינדור עוד יותר מהביצועים (האיטיים גם כך) המוצגים ב[1.1].

II.ii. כיסוי שטח גאודזי נרחב בצורה איכותית מבחינת צילום עשוי לדרוש אלפי ועשרות אלפי תמונות. בשלב מסויים זמן האימון על GPU בודד הופך ללא ישים, וגודל הרשת הופך את רוב ה- GPUs הזמינים בשוק, מלבד החזקים ביותר, לחלשים מכדי אפילו לטעון רשת מאומנת על מנת לרנדר מתוכה תמונה.

II.iii. גם זמן רינדור, לא רק אימון, גדל ביחס ישיר לגודל הסצנה. אמנם קצב הרינדור ב-NeRF איטי ורחוק מאינטראקטיבי (ראה Instant-NGP בחלקה האחרון של העבודה לשבירה של הנחה זו), אבל ככל שהסצנה גדולה יותר, כך גדל הפער.

II.iv. תמונות שהעקבה הקרקעית שלהן לא נחתכות מספקות הרבה פחות מידע ל NeRF בודד מחד, אבל מאחר וNeRF בודד מגדיר נפח תחום (bounded volume) לעבוד בו, מורידות את הרזולוציה האפקטיבית של דגימות על כל קרן עבור **כל** המצלמות המשתתפות בו.

II.v. קיימת הנחה סמוייה ב-NeRF שכל האיסוף של הקלט התרחש בחתימת זמן בודדת - מבחינת תנאי תאורה, מזג אוויר, אובייקטים נעים בסצנה, מקורות אור ואפקטים לא למברטיים כגון השתקפויות. כיסוי של סצנה גדולה יכול לגרור שינוי של שעות עד ימים ואף חודשים בין זמני האיסוף של תמונות שונות, ואובייקטים נעים הופכים בה לבעיה משמעותית.

II.vi. ככל שסצנה גדולה יותר, כך יש סיכוי גבוה יותר של שינוי גיאומטרי (סטטי) משמעותי במרחב, בהפרשי זמני איסוף מספיק גדולים - למשל מבנים בבנייה או צמחיה שגדלה או משנה צבע או נושרת. שינוי כזה עשוי לדרוש איסוף ואימון מחדש של כל הרשת.

II.vii. ככל שמדובר ביותר תמונות ובשטח גדול יותר,הסיכוי לסחיפה ולחוסר דיוק במיקומי המצלמות האוספות עולה. סיבוכיות תיקון מיקומי מצלמות (SLAM\SFM) יכול להגיע ל $O(n^3)$ במספר המצלמות, ופיתרון SFM כמובן פשוט יותר כשאחוז החפיפה בכיסוי בין כל המצלמות גבוה.

II.viii. סצנות גדולות הן ברובן המוחלט סצנות טבעיות חיצוניות, עם כל בעיות התאורה וההצללה האופייניות למתאר הנ"ל. אם נציין רשימה חלקית: ריבוי אפשרויות למקורות אור, הצללה - גם עצמית וגם על ידי חלקים אחרים בסצנה, שינויי תאורה בין תמונות שונות הנובעות מזוויות איסוף מול מקור אור (כגון שמש), שינויי תאורה בין תמונות שונות הנובעות ממנגנון מצלמת האיסוף כגון Automatic Gain Control - AGC או מנגנוני white balance שונים. בנוסף, כפי שהוזכר ב II.vi, II.vii , גודל הסצנה עלול להשפיע על הפרש זמן באיסוף תמונות הצופות על אותו חלק מהסצנה מזוויות שונות, דבר היכול להביא לשינוי בתנאי התאורה החיצוניים (שינוי תאורה הנובע מתנועת השמש לדוגמא) בין תמונות שונות.

4.גישות לפיתרון

בפרק זה נדון בארבע גישות שונות הבאות לתת מענה לבעיית שחזור מבט מנקודת ראייה חדשה (Novel View Synthesis) בסצנות גדולות - חלקן מניחות ומתוכננות למתן מענה לסצנות *אורבניות* גדולות - וכן בשתי עבודות שמציגות שיפורים משמעותיים ל-NeRF שרלוונטיים למרחב הבעייה או מדגימים יכולות במרחב זה מבלי להגדיר את הסצנה הגדולה כמטרה ספציפית. עבור כל אחת מהגישות שיוצגו: נסקור את תת הבעיה או תתי הבעיות שאיתן הגישה מתמודדת, נזכיר את תרחיש הייחוס שעבורו פותחה הגישה, נסקור טיפול קדם שבוצע על הקלט, דרישות או מגבלות חומרה אם קיימות, שיפורים למודל ה-NeRF, אלגוריתמיקה נוספת שפותחה, וצורת ההערכה שבוצעה.

הגישות שנסקור:

4.1 :Bungee-NeRF - מדגימה שילוב של קלט ממקורות בגבהים שונים משמעותית - מלווין ועד רחפן.

- גרסה קודמת פורסמה תחת השם CityNeRF

4.2 :Block-NeRF - מדגימה פיתרון מלא, כולל התמודדות עם הבדלי תאורה משמעותיים באיסוף הקלט לסצנה אורבנית רחבת ידים, על בסיס פלטפורמת איסוף **קרקעית**.

4.3 :Mega-NeRF - מדגימה פיתרון מהיר לאימון לסצנה אורבנית רחבת ידים על בסיס רחפן במתאר חילוץ והצלה.

4.4 :Urban Radiance fields (URF) - מדגימה פיתרון על בסיס איסוף רגלי קטן יחסית לשאר הגישות, הכולל דגימות LIDAR.

4.5 :Instant Neural Graphics Primitives (Instant-NGP) - מציגה מהפכה באופן קידוד המיקום, שיפור בסדרי גודל בזמנים ובביצועים, ושימוש בעקרונות זהים לחישוב פרימיטיבים גרפיים שונים - NeRF-ים, SDF-ים, תמונות גיגהפיקסל והצללות סצנה.

4.6 :Neurolangelo - מדגימה ניצול של עקרונות מ-Instant NGP על מנת להחליף את השימוש במדד אטימות עם ייצוג SDF כדי להדגים שחזור מפורט מאוד של מבנה הגאומטריה בסצנה.

בפרק 5 ניתן למצוא דיון באלמנטים משותפים לכל הפתרונות, שוני או ייחוד של פתרון זה או אחר לעומת השאר, ודגשים מיוחדים הראויים לציון לגבי פתרונות מסויימים.

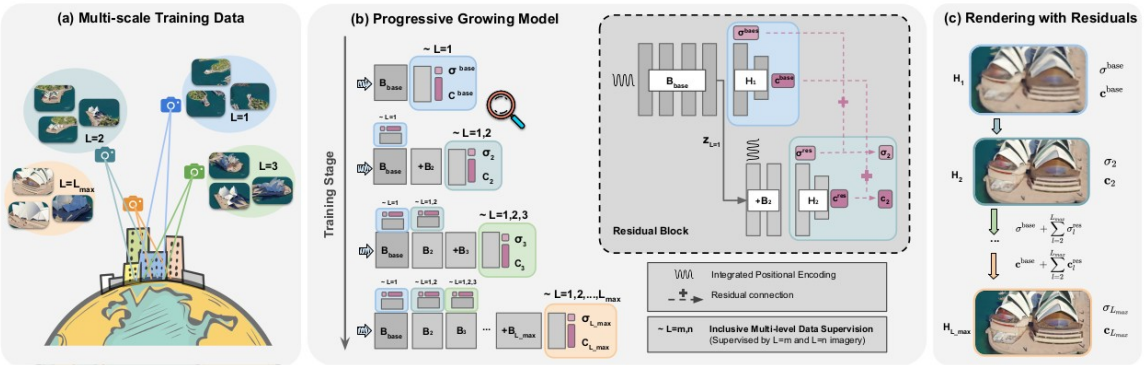
Bungee-NeRF 4.1

כפי ששם המאמר [6] -
BungeeNeRF: Progressive Neural Radiance Field for Extreme Multi-scale Scene Rendering
רומז, גישה זו מתרכזת בבעיה שהוצגה כ- 3.i - קרי, הפרשי קנה מידה משמעותיים בין תמונות באיסופים לסצנות תלת מימדיות בעולם האמיתי, וספציפית בסצנות עירוניות. שינויי קנה מידה אלה יכולים להגיע מקלט מתמונות לווין ברמת ערים שלמות, דרך איסופים אוויריים ממטוסי צילום, דרך רחפנים באיסופים אוויריים קרובים, ועד איסופים קרקעיים המציגים פרטים מורכבים בארכיטקטורה. מגוון רחב באופי נקודות המבט הנאספות יכול גם הוא להציג שונות גבוהה ברמת הפירוט.

איור 10: Bungee-NeRF - קני מידה וטווחים. מתוך [6]
באיור 10 ניתן לראות משמאל ציור סכמטי של 3 גבהים טיפוסיים, מ L=1 מרוחק (לווייין) עד L=3 קרוב (גובה גגות). צבע המעגלים משמש גם כצבע מסגרת התמונות המשוחזרות באותו גובה. בכל תמונה מדד PSNR לשחזור - אפשר לראות שה- PSNR נשאר עקבי בין סקאלות. בתחתית רצף שחזור מסקאלת כדה"א עד גובה גגות.

בבסיסו, בנג'י-נרף מציג גישה חדשנית לרעיון אימון "מ-גס למפורט" המשתמש בשתי הרחבות ייחודיות. הראשונה היא מודל מבוסס מבנה בלוק שיורי , הגדל פרוגרסיבית תוך כדי שלבי האימון. השניה היא תהליך אימון מפוקח, המנהל את אופי וסדר קנה המידה המוצג למודל המתאמן בכל שלב אימון. בנוסף, בנג'י-נרף שואל אלמנטים מ-Mip-NeRF, ספציפית את קידוד המיקום המשולב (IPE) , תוך שימוש מושכל בקומפוננטות גבוהות התדר בהתקדמות האימון בין הבלוקים.

על מנת להתרכז אך ורק בתת הבעיה הספציפית של ריבוי קני מידה בקלט, בנג'י-נרף מפרידים את שאלת הכיסוי האופקי בקנה מידה בודד (גודל השטח הגאודזי המכוסה על ידי מודל NeRF בודד) מבעיית ריבוי קנה המידה בקלט - ובמוצהר לא מתמודדים איתה. בנוסף, כדי להמנע מבעיות הנובעות משיערוך מיקום קשה בין תמונות מקנה מידה שונה, עיקר הקלט מופק באופן סינטטי דרך Google Earth Studio, המאפשר להגדיר מיקומי מצלמה ולקבל קלט באיכות מספקת. הקלט מופק במסלולים מעגליים קונסנטריים, כשרדיוס המעגל גדל ממעגל נמוך לגבוה. כן מוצגים תוצרים של איסופי מל"טים ורחפנים, כך שהקלט הוא לא רק סינטטי באופיו.



איור 11: תאור כללי של Bungee-NeRF. מתוך [6]

באיור 11 משמאל (a): אילוסטרציה של מגוון קני המידה הזמינים. האיסוף הנמוך ביותר (המפורט ביותר) מוגדר כ L=L_max ושאר הרמות ממוספרות בהתאם. במרכז (b): בכל שלב באימון מתווסף בלוק שיורי נוסף, בעוד הקלט מורחב ברמת קנה מידה נוספת. מימין (c): רינדור על בסיס בלוקים שיוריים שונים מאפשר רינדור ברמת פירוט (LOD) שונה.

מודל פרוגרסיבי מבוסס מבנה בלוק שיורי

בדומה לקונספט הפירמידה בעיבוד תמונה ולרמות פירוט (LOD) בגרפיקה תלת מימדית, בנג'י-נרף מציע מבנה רשת פרוגריסבי, בו מתחילים לאמן את המודל על קלט מקנה מידה מרוחק, ושלב אחר שלב מרחיבים את סט האימון בקלט מהטווח הבא בכל שלב אימון. כתוצאה, מודל ה-בנג'י-נרף לומד היררכיה של יצוגים בכל הסקאלות הזמינות עבור הסצנה. בכל שלב אימון המודל מוציא כפלט פונקציית צפיפות ופונקציית צבע נפרדות לאותו בלוק, כך שאפשר להתייחס לכל בלוק אימון כאנלוגיה לרמה שונה ב-LOD.

הקלט מחולק למספר רמות שונות, המוגדר מראש. המצלמות הקרובות ביותר (בעלות הרזולוציה הקרקעית הטובה ביותר) מוגדרות כרמה $L=L_{max}$, ואילו הרחוקה ביותר כ-$L=L_1$. בניסויים המוצגים כל רמה הוגדרה כהתרחקות בזום לפי פקטור של $\{2^1,2^2,2^3,\dots\}$. עבור כל תמונת קלט מוגדר סמן שלב I_l , כש $\{I_l=L\}$ מגדיר את אוסף כל תמונות הקלט השייכות לקנה מידה L. בבסיסו, ככל שגובה המצלמות עולה, אובייקטים על הקרקע מוצגים ברזולוציה פיקסלית מופחתת, עם פירוט גאומטרי נמוך יותר (ראו רזולוציה קרקעית בנספח). בו בזמן, ככל שגובה האיסוף עולה נכנסים לשולי התמונה אובייקטים חדשים, ומרחב הכיסוי גדל.

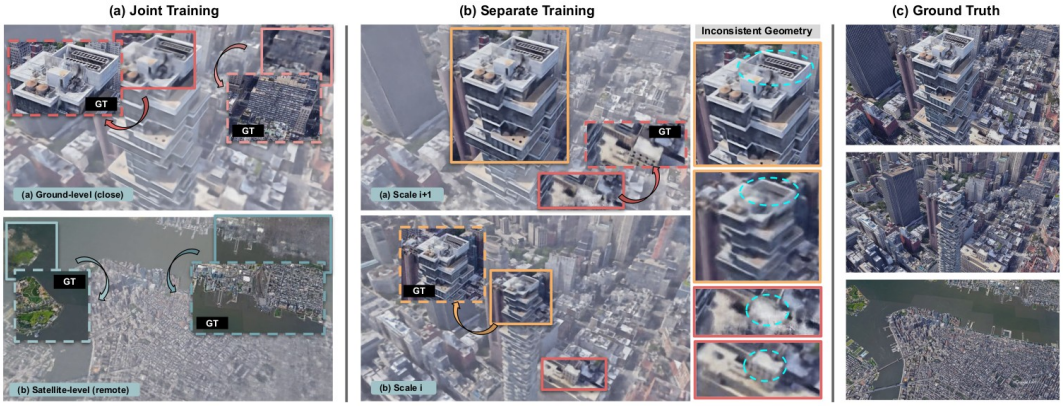
במודל 3 סוגי בלוקים: בלוק הבסיס B_{base} הוא MLP בעומק 4 שכבות נסתרות, כל אחת בעלת 256 ערוצים. בלוק ראש פלט המפיק את ערכי האטימות σ והצבע c זהה למוגדר ב-NeRF [7], ובלוק שיורי B_L המורכב מ 2 שכבות FC עם חיבור דילוג.

בשלב האימון הראשון המודל מורכב מ-B_{base} וראש פלט המפיק את σ_{base} ואת c_{base} (ובסך הכל 6 שכבות). שלב זה מאומן על קלט מקנה מידה L=1 בלבד. בשלב האימון הבא מוסיפים חיבור דילוג מבלוק הבסיס לבלוק שיורי B_L, יחד עם קידוד המיקום IPE , ומוציאים ראש פלט נפרד המוציא את הפלט **השיורי** של הבלוק החדש, כך שפונקציות הצפיפות והצבע הן סכימה של תוצרי ראשי הפלט בשכבות עד אותו שלב. לצורך העניין, ב-L=4, מבנה הרשת הוא MLP של 10 שכבות, עם חיבורי דילוג בשכבות 4,6,8, ו-4 ראשי פלט היוצאים משכבות 4,6,8,10. פלט הצבע והצפיפות של כל שכבה הם סכום של כל ראשי הפלט של הבלוקים הקודמים, עד לנוכחי (כולל) ומוגדרים כדלקמן:

$$\mathbf{c}_L = \mathbf{c}^{base} + \sum_{l=2}^{L} \mathbf{c}_l^{res}, \quad \sigma_L = \sigma^{base} + \sum_{l=2}^{L} \sigma_l^{res} \quad (16)$$

ראו איור 11(b) להבהרה.

אימון סדור, המשתמש באופן מושכל בבחירת קני המידה הזמינים לרשת בשלבי אימון שונים

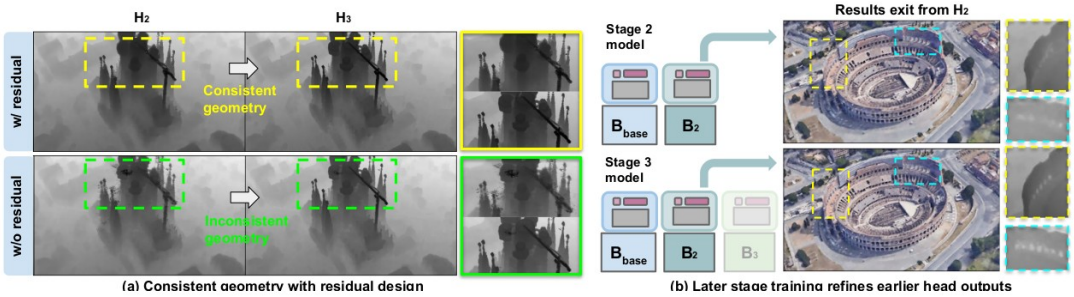


איור 12:דוגמאות למגבלות האימון של NeRF בקני מידה שונים. מתוך [6]

איור 12: ב-(a) אומנו כל תמונות הקלט מכל קני המידה יחד, מה שגרם לשימוש רק בתדרים נמוכים מידי וחוסר פרטים בשחזור. ב-(b) אומנו כל קנה מידה בנפרד, מה שהביא לחוסר עיקביות בשחזור הגאומטריה.

ערוצי התדר היעילים גם ב- PE וגם ב-IPE שונים בקני מידה שונים. מתארי פורייה הנדרשים לייצוג מספק בטווח קרוב (לדוגמא L=3) הם מתדר גבוה יותר מאשר במבט מרוחק יותר (למשל L=1). באימון משותף על כל קני המידה בבת אחת, בעקבות מיעוט יחסי של מבטים מטווח קרוב, NeRF ו-Mip-NeRF מתעדפים פיתרון המנצל בעיקר תדרים נמוכים יותר במתארי פורייה.

משטר האימון המוצע ב- Bungee-NeRF סדור יותר: האימון מתחלק לשלבים, בהתאמה למספר שכבות הקלט המוגדרות. בשלב הראשון, מאומן בלוק הבסיס B_{base} על הקלט המרוחק ביותר - L=1. בשלב הבא יאומן מודל בלוק הבסיס בתוספת בלוק שיורי אחד על שתי רמות הקלט הראשונות - L=1,2. בשלב הבא יאומן מודל על בלוק בסיס B_{base} ושני בלוקים שיוריים B_{res} על שלוש רמות הקלט L=1,2,3, וכך הלאה עד $L=L_{max}$. כל שכבות המודל זמינות לאימון לאורך כל שלבי האימון, וכתוצאה מכך גם שכבות הפלט ברמות הפירוט הנמוכות יותר מתעדכנות ומשתפרות משלב לשלב.

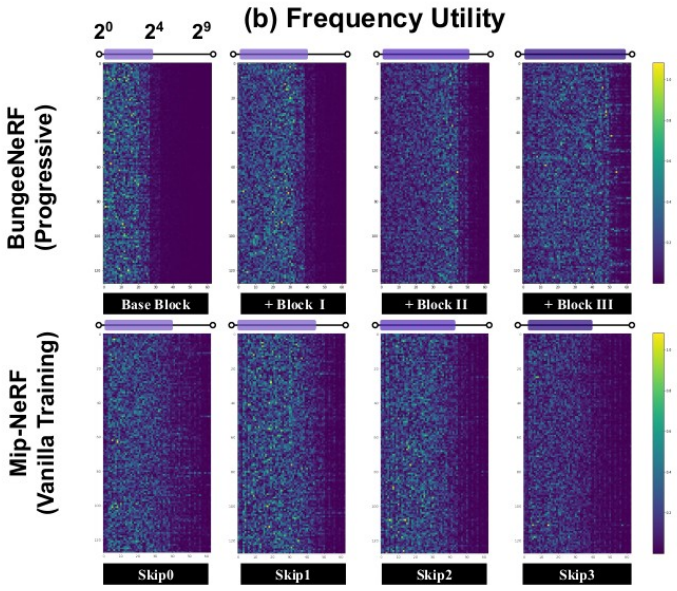


איור 13: דוגמאות ליתרונות החיבור השיורי. מתוך [6]

איור 13: מבנה החיבור השיורי מאפשר לבלוקים מאוחרים יותר לעדן פלטים מראשי פלט של בלוקים קודמים. משמאל (a) דוגמה לעיקביות בעומק (גאומטריה המופקת מתוך פונקציית הצפיפות הנפחית) בין שכבות, בעוד מימין (b) פלט משופר מבלוק 2 לאחר הוספה של אימון של בלוק 3

שימוש ב IPE ופרוסטום קוני

מחקרים קודמים ובדיקות במהלך הפיתוח העלו שגישת Mip-NeRF לקידוד מיקום על בסיס פרוסטום קוני (IPE) משיבה תוצאות טובות יותר מאשר PE בכל מרחב המקרים שבדקו, ולכן אומצה גישת ה-IPE באופן גורף. בהשוואה ל- Mip-NeRF, ניצול מרחב התדרים במתארי פורייה של IPE יעיל באופן משמעותי יותר בשימוש במתאר האימון הסדור שהוצג. באיור 14 מוצג ניצול מרחב התדר בכל שכבת קנה מידה בקלט ב- Mip-NeRF לעומת Bungee-NeRF.



איור 14:איכות הניצול של מרחב מתארי פורייה בין Mip-NeRF ו-BungeeNeRF

באיור 14: ניתן לראות השוואה בשימוש בתדרים גבוהים במתארי הפוריה של IPE בקני מידה שונים בין Mip-NeRF שאומן על כל הקלט בו זמנית, לבין Bungee-NeRF שאומן במשטר אימון פרוגרסיבי. ניתן לראות ש-Bungee-NeRF מעורר באופן אפקטיבי שימוש בתדרים גבוהים יותר בבלוקים מתקדמים יותר (קנה מידה קרוב יותר), בעוד Mip-NeRF מתעדף תדרים נמוכים יותר בכל השכבות באופן יחסית אחיד.

תוצאות ודוגמאות

עיקר הניסויים בוצעו על סצנות מ- 12 ערים מרחבי העולם, שקלט ב-4 גבהי איסוף הופק עבורן מתוך Google Earth Studio. שימוש במקור זה גורר עיגון הדדי טוב בין התמונות השונות, הן בתוך כל קנה מידה והן בינהן, היות והקלט מופק באופן סינטטי ממערכת פתורה כללית - הדמאות הלווין עברו עיגון הדדי ופיתרון מיקום מצלמות בתהליך ההכנסה ל-Google Earth Studio. השוואה נעשתה מול שתי גירסאות של NeRF - קלאסית וגרסה המשתמשת בקידוד מיקום תחום (Windowed PE- WPE), וכן מול גירסה של Mip-NeRF. כדי לאפשר השוואה, מודלי רשתות הבסיס שמולם בוצעה ההרחבה שונו מעט - הוסף חיבור דילוג אחרי 4 שכבות ב MLP, כדי לדמות את חיבור הדילוג הקיים ב Bungee-NeRF אחרי בלוק הבסיס. בכל אפשרויות קידוד המקום (PE / WPE / IPE) התדר הגבוה בייצוג קידוד המיקום (נוסחה 15) נבחר להיות M=10. באיסופים הסינטטיים שהופקו מ- Google Earth Studio נבחרו 4 רמות קנה מידה (L_{max}=4). מוצגים גם איסופי רחפן שנאספו, אבל המודל דליל יותר (L_{max}=2).

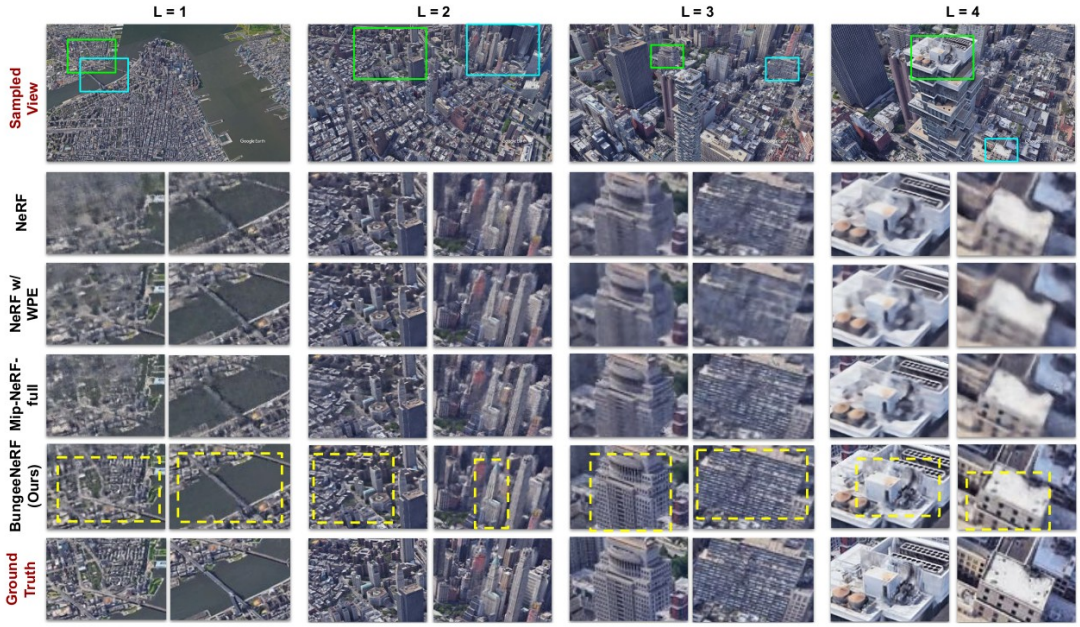
בטבלה 1 (בהמשך) ניתן לראות את הפרשי הגבהים המשמעותיים בין האיסופים, בעוד טבלה 2 מציגה השוואת ביצועים בין BungeeNeRF לאלגוריתמים מתחרים.

	Main Experiments		Extra Tests: Used for various illustrations throughout the paper and additional comparisons in Supplementary										
City Scene	New York	San Francisco	Sydney	Seattle	Chicago	Quebec	Amsterdam	Barcelona	Rome	Los Angeles	Bilbao	Paris	
Alt (m)	56 Leonard	Transamerica	Opera House	Space Needle	Pritzker Pavilion	Château Frontenac	New Church	Sagrada Familia	Colosseum	Hollywood	Guggenheim	Pompidou	
Lowest	290	326	115	271	365	166	95	299	130	660	163	159	
Highest	3,389	2,962	2,136	18,091	6,511	3,390	2,3509	8,524	8,225	12,642	7,260	2,710	
# Images	463	455	220										

טבלה 1: סצנות מ-12 ערים מרחבי העולם, שהופקו בעזרת Google Earth Studio. גבהים מינימליים ומקסימליים, מספר תמונות כללי בתרחישים העיקריים (ניו יורק וסאן פרנסיסקו). מתוך [6]

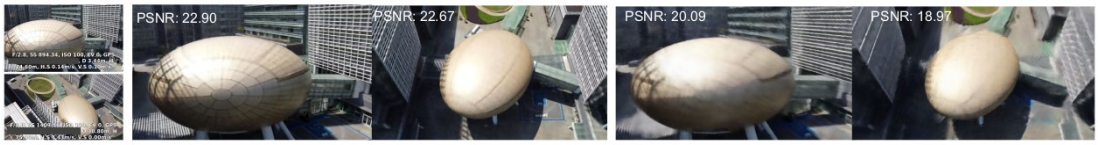
	56 Leonard (PSNR ↑)				*Transamerica* (PSNR ↑)				*56 Leonard* (Avg.)			*Transamerica* (Avg.)		
	Stage I.	Stage II.	Stage III.	Stage IV.	Stage I.	Stage II.	Stage III.	Stage IV.	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑
NeRF (D=8, Skip=4)	21.279	22.053	22.100	21.853	22.711	22.811	22.976	21.581	21.702	0.320	0.636	22.642	0.318	0.690
NeRF w/ WPE (D=8, Skip=4)	22.022	22.097	21.799	21.439	23.382	23.171	22.450	20.806	21.672	0.365	0.633	22.352	0.331	0.680
Mip-NeRF-small (D=8, Skip=4)	21.899	22.179	22.045	21.763	23.346	23.030	22.645	20.937	21.975	0.344	0.648	22.692	0.327	0.687
Mip-NeRF-large (D=10, Skip=4)	22.020	22.403	22.277	21.975	23.196	22.900	22.439	20.743	22.227	0.318	0.666	22.525	0.330	0.686
Mip-NeRF-full (D=10, Skip=4,6,8)	22.043	22.484	22.690	22.355	23.377	23.156	22.854	21.152	22.312	0.266	0.689	22.828	0.314	0.699
BungeeNeRF (same iter as baselines)	**23.145**	**23.548**	**23.744**	**23.015**	**24.259**	**23.911**	**23.507**	**22.942**	**23.481**	**0.235**	**0.739**	**23.606**	**0.265**	**0.749**
BungeeNeRF (until convergence)	**24.120**	**24.345**	**25.382**	**25.112**	**24.608**	**24.350**	**24.357**	**24.608**	**24.513**	**0.160**	**0.815**	**24.415**	**0.192**	**0.801**

*טבלה 2: השוואה כמותית על סצנות ניו יורק וסאן פרנסיסקו. D מייצג עומק הרשת הסופי, Skip מייצג את השכבות שבהן הוספו חיבורי דילוג. כל מודלי הבסיס אומנו עד התכנסות, BungeeNeRF מציג 2 תצורות - כל שלב מאומן לאורך 100K מחזורים, וכל שלב מאומן עד התכנסות ללא הגבלת מס' מחזורים (תוצאות טובות יותר). **ציון מיטבי**/מקום שני. מתוך [6]*



איור 15: תוצאות איכותיות מסצנת ניו יורק (רח' לאונרד 56). מתוך [6]

באיור 15 ניתן לראות את הקושי ש-NeRF סובל ממנו בסצנה מרובת קנה מידה - ב רזולוציה קרקעית נמוכה (טווח רחוק) מריחות וחוסר בפרטים, ואילו ברזולוציות קרקעיות גבוהות יותר (בטווחים הקצרים יותר) - טישטוש וחלקות. Mip-NeRF מציג שיפור, אך עדיין סובל מטישטוש וחלקות, בעוד Bungee-NeRF מציג פירוט יחסית אחיד בכל הרזולוציות הקרקעיות.



איור 16: איסוף בעזרת רחפן של בניין אודיטוריום. מתוך [6].

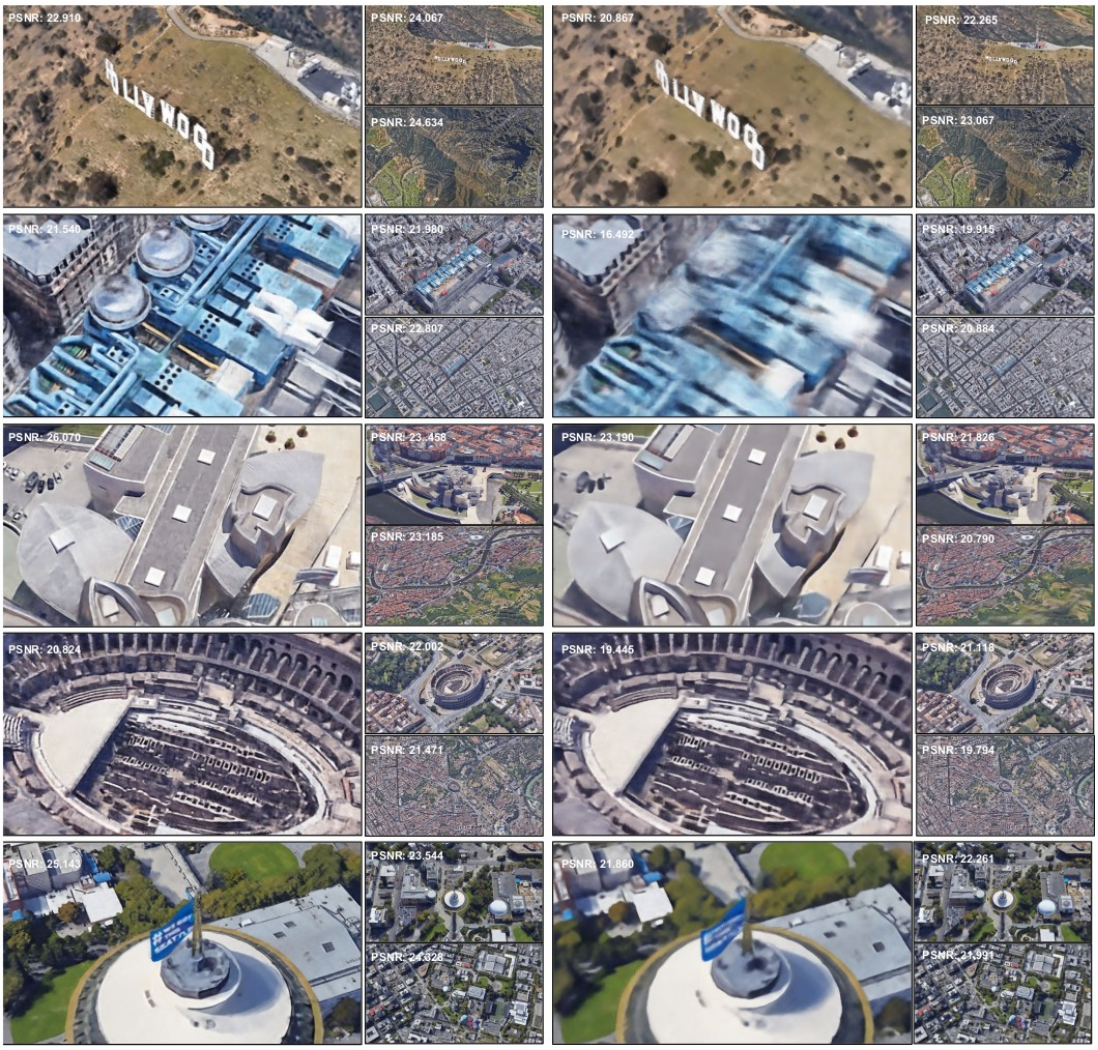
בתחתית איור 10 ניתן לראות שחזור בקנה מידה של כל כדה"א. האימון התבצע ב-5 שלבים (L=5).

באיור 16 ניתן לראות שיחזור in the wild של בניין אודיטוריום בעזרת רחפן. המודל אומן בשני שלבים בלבד (L=2), ומיקומי המצלמות תוקנו בעזרת SFM (Colmap). בתוסף (supplement) למאמר[4] פורסם עוד מידע על ניסוי זה, ונטען שנדרש איסוף עם הרבה

4 Bungee-NeRF supplement - פורסם בעמוד ה-github של הפרוייקט. ראו https://city-super.github.io/citynerf/img/supp.pdf .

פחות מבטים ממה שהיה נדרש לתוצאת MVS קלסית איכותית. לא סופקו נתונים כמותיים.

באיור 17 (למטה) ניתן לראות השוואה מפורטת יותר בין Bungee-NeRF ו- Mip-NeRF. ברמת הפירוט הגבוהה ההבדל ניכר לעין, אבל גם בקנה המידה הנמוך ניתן לראות ארטיפקטים (בולט לעין בעיקר במוזאון גוגנהיים).



איור 17: השוואה בין תוצאות על 5 סצנות נוספות מ-3 קני מידה.
מתוך חומר נוסף באתר הפרוייקט.

באיור 17 ניתן לראות השוואה בין תוצאות של 5 סצנות נוספות: שלט הוליווד בלוס אנג'לס, מרכז פומפידו בפריס, מוזאון גוגנהיים בבילבאו, הקוליסאום ברומא וה-space needle בסיאטל. משמאל תוצאות Bungee-NeRF, מימין Mip-NeRF

איור 18: תוצאות מראשי פלט משכבות שונות (H_2, H_3,H_4) לתמונות בקני מידה שונים.
מתוך [6]

כפי שצויין בהקדמה - שם הפרוייקט הוחלף במהלך הפיתוח מ-CityNeRF ל- BungeeNeRF.

איור 18 מציג תוצרים מראשי פלט בשכבות השונות ($H_{x=\{2,3,4\}}$) בתיחקור תמונות ברזולוציות קרקעיות שונות. מסומן במסגרת ירוקה פלט באיכות מספקת. קל לראות שברזולוציה קרקעית גבוהה התוצרים מהשכבות הנמוכות לא מספקים, וכוללים טישטושים והחסרת פרטים,בעוד בטווח רחוק (רזולוציה קרקעית נמוכה) אפילו H_2 כבר מחזיר תוצאות מספקות. חשוב לשקול שימוש בראשים מוקדמים בעת שחזור אזורים ברזולוציה קרקעית נמוכה, לשם חיסכון בזמן חישוב ונפחי איכסון. קל לראות את האנלוגיה לשיחזור ממודל LOD גדול, בו רמות פירוט שונות משמשות לרנדר אזורים שונים בתמונה.

סיכום

Bungee-NeRF היא הגישה היחידה בעבודה זו, העוסקת במקורות איסוף שונים מהותית (לווין לעומת רחפן), ואופי הפיתרון מרתק. הגישה לא עוסקת - מבחירה ובאופן מוצהר- בענייני פריסה אופקית וכיסוי שטחי קרקע נרחבים, כמו למשל על ידי חלוקה לתת NeRF-ים (כפי שנראה בפתרונות בהמשך). למרות זאת השיטה מתאימה מאין כמוה לפיתרון חלוקה שכזה, כשניתן לדמיין חלוקה גאודזית לתתי מודלים המייצגים את השכבות המתקדמות (והקרובות לקרקע), בעוד שכבת הבסיס- המכסה אזור גאודזי נרחב בכל מקרה - משותפת ביניהם. שילוב עם גישות שכן עוסקות בניהול תת מודלים וחלוקת אזור העניין לחלקים נפרדים, עם ניהול התפר, מתבקש. אם הגדרת גודל הסצנה תלויה גם במיקומי המצלמות האוספות - כפי שנראה בהגדרות של פתרונות נוספים בהמשך - חלק מהסצנות המוצגות פרוסות על פני 20 ק"מ...

היכולת להפיק תוצרים ברמות פירוט שונות, כפונקציה של טווח נקודת המבט מהאזור המשוחזר, מתאימה גם היא באופן טבעי לשחזור שטח גדול מנקודת מבט אלכסונית. כמו במודלי משולשים בעלי רמות פירוט (LOD) קלסיים, אזורים קרובים לנקודת המבט יוכלו להנות מפרוט גבוה, בעוד מבטים מרוחקים יותר יוכלו לחסוך משאבים ולשחזר מרמת פירוט נמוכה יותר. חשוב לזכור שעיגון הדדי של תמונות בקני מידה שונים בצורה חריגה כפי שהוצג היא בעיה משמעותית (מאוד) בפני עצמה, שהמודל המוצג לא מתמודד איתה. עיגון לקוי או לא מדוייק בין איסופי רחפן ממשיים לאיסופי לווין עלולים להשפיע קשות על ביצועי המודל המוצע. מצד שני, קיימים תרחישי שימוש רבים שבהם מקור קלט סינטטי כמו Google Earth Studio יכול להיות מספק, ומצד שני - ניתן לחשוב על מתארים שבהם איסוף באופי של מספר מעגלים קונסנטריים בגבהים קבועים מראש - שמתאים מאוד לבניית תוכנית איסוף אווירית - תוך שילוב של מספק כלי איסוף, לדוגמא על ידי כלי טיס (מאוייש או לא) בשכבות הגבוהות, ושכבות נמוכות על בסיס רחפנים - בהחלט מציע תרחישי שימוש אפשריים, כשקלט כזה הוא בהחלט סביר להפעלת SFM קלסי כמקור למיקומי מצלמות מטוייבות.

מידע נוסף

עמוד הבית של הפרויקט נמצא ב- https://city-super.github.io/citynerf, שם ניתן למצוא את המאמר, תוסף למאמר המפרט חלק מהניסויים שלא פורטו במאמר המלא, סרטי וידאו של תוצאות שונות, וגישה לקוד שפורסם במקביל למאמר.

איור 19: Block-NeRF . בציור ובתמונות: שכונת אלאמו סקוור בסאן פרנסיסקו. מתוך [4]

איור 19: Block-NeRF מאפשר שיחזור של סצנות אורבניות גדולות על ידי ייצוג הסצנה ע"י הרבהNeRF-ים מקומיים, ואיחוד חלק של NeRF-ים רלוונטיים בזמן חיזוי. מאפשר איסוף קלט לאורך זמן רב ועידכון רק חלק מהסצנה במקרה של בנייה (מימין)

Block-NeRF, עבודתם של Tancik *et al.* [4], הוא המשך ישיר ל-NeRF [1] ול-MipNeRF [11]. שניים מהכותבים - Tancik ו-Mildenhall - חתומים על [1], וארבעה מהם - Tancik, Mildenhall, Srinivasan, Barron - חתומים על [11]. Block-NeRF נולד מההכרה ש-NeRF בודד אינו סקלבילי בשום צורה בניסיון לייצג סצנות גדולות ברמת ערים. זהו וריאנט מבוסס מיפ-נרף שמיועד לייצג סצנות גדולות, ובמקרה זה - ספציפית סצנות אורבניות בסגנון בנייה אמריקאי (קונספט הבניה שהביא לנו את מרחק מנהטן). בלב הפיתרון מוצגת גישה, המחלקת את הסצנה הגדולה (המכילה מספר "בלוקים" או צבירי בניינים, המופרדים על ידי רשת של רחובות) לתת-NeRF-ים על כל בלוק, המאומנים כל אחד בנפרד, ותוצריהם מאוחדים בזמן חיזוי. פתרון זה מאפשר להפריד את אימון הבלוקים הבודדים (למשל על חומרות נפרדות), להרחיב את איזור הכיסוי במדורג, וכן מאפשר עידכון נפרד פר בלוק. בנוסף - ולא פחות חשוב - הפתרון מנתק את זמן הרינדור מגודל הסצנה, ומאפשר לרנדר חוזי מסצנות גדולות כרצוננו. הצורך לאחד תוצרים מ-נרפים שונים, שפוטנציאלית נאספו בתנאי סביבה ותאורה שונים, דחף גם לשילוב כלים שמאפשרים הפרדה של תנאי הסביבה מהשיחזור, ומכך האפשרות לקבוע את תנאי הסביבה המשוחזרים לפי הצורך.

בלוק-נרף בא לתת מענה ספציפי למתאר איסוף של סצנה אורבנית נרחבת, ולכן סקירה כללית **באיזה** תת סוגי הבעיות שהצגנו ב-3 רלוונטיות מיותרת - העבודה נוגעת בכולן.
תרחיש הייחוס שאליו מתכוננת העבודה הוא סימולציות במעגל סגור למערכות רובוטיות בכללי, ובעיקר למערכות נהיגה אוטונומית. מערכות אלה מאומנות לעיתים קרובות על סמך הרצה מחדש על תרחישים שהמערכת נתקלה בהם בעבר, אבל כל שינוי בהתנהגות המערכת עלול לגרור שינוי במסלול הרכב, שגורר צורך שחזור תמונות חדשות ברזולוציה גבוהה, מנקודות מבט חדשות לאורך המסלול המעודכן.

על מנת להשלים את מטרתו, בלוק-נרף דורש מספר שינויים ושיפורים ארכיטקטוניים:

- חלוקה אזורית של תת רשתות
- קידוד מראה / תנאי סביבה (appearance embedding).
- חיזוי נראות (visability prediction)
- איסוף איכותי תוך טיוב נלמד של מיקומי מצלמה.
- התייחסות בזמן אימון לאוביקטים נעים מחד, ושינויים גאומטריים תלויי זמן מאידך.
- מדד חשיפה נשלט ונפרד לכל בלוק-נרף פרטני.
- מנגנון השוואת נראות בין בלוק-נרפים שכנים.
- בחירה דינמית של בלוק-נרפים רלוונטיים בעת חיזוי, שמהם יאוחדו הפלטים באופן חלק להחזרת תוצאה משולבת.

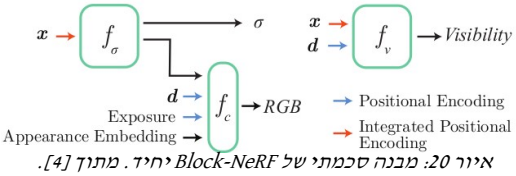
*שתי הנקודות האחרונות מאפשרות להרחיב את יכולת הסימולציה על ידי שינוי נשלט של תנאי הסצנה, כגון שינוי תנאי התאורה, זמן חשיפה, שעת יום ומזג אוויר.

חלוקה אזורית של תת רשתות ואימון נפרד

במאמר [4] מראים ניצול של הגריד הסדור - שמביא התיכנון האורבני הקיים בהרבה ערים אמריקאיות - להגדרה של פריסת סט מודלי בלוק-נרף על ידי מיקום מרכז בלוק-נרף בכל צומת X בין רחובות. יש לציין שמדובר באיסוף קרקעי על בסיס רכב, ולכן הבחירה בחלוקה תלויית צירים הגיונית מאוד מחד, ומאידך אופי איסוף כזה מייצר בהכרח שטח "מת" במרכז כל בלוק, וכיסוי אחיד ויעיל לאו דווקא היה אפשרי אלמלא אופי הארכיטקטורה.

מאחר ובאופן כללי, המטרה היא לפרוס את ה-בלוק-נרפים הפרטניים באופן שיגיע לכיסוי מלא של אזור העניין, הגדרה של מרכז בלוק בכל צומת, עם כיסוי לאורך 75% מהכבישים היוצאים מאותו צומת לכיוון הצומת הבאה, מביאה לכיסוי מלא תוך חפיפה של 50% בין כל שני בלוקים צמודים. תכונה זו מאפשרת השוואת נראות קלה יותר (ראו בהמשך). כל בלוק-נרף מאומן על קלט התחום לאזור גאוגרפי ספציפי, מאחר וחלוקה כזו ניתנת לאוטומציה בקלות ומסתמכת רק על מידע מיפוי בסיסי מאוד (Open Street Map כדוגמא) .

קידוד מראה / תנאי סביבה (appearance embedding).



איור 20: מבנה סכמתי של Block-NeRF יחיד. מתוך [4].

איור 20: לאחר שה-MLP הראשון f_σ מחשב את פונקציית הצפיפות הנפחית, משרשרים לוקטור המתארים היוצא את כיוון המבט d ואת פרמטר החשיפה (שניהם מקודדים ע"ב PE) וכן את קידוד המראה ל-MLP השני f_c, המחזיר את הצבע RGB בנקודה. בנוסף מאמנים רשת נראות f_v, המשערכת האם נקודה במרחב נראתה במבטי הקלט, שישמש בעת חיזוי לסינון בלוקים משתתפים.

כפי שכבר צויין, חלקים שונים של הקלט עשויים להאסף בתנאי ראות, תאורה ושעת יום שונים זה מזה משמעותית. בלוק-נרף מתבסס על עבודות קודמות (בעיקר NeRF-W [12]) כדי להפריד את תנאי התאורה והסביבה מחיזוי הצבע כקידוד עבור כל תמונת קלט בנפרד. קידוד זה מאפשר למודל ה-נרף הבסיסי להתייחס למספר תכונות משנות נראות בין תמונות קלט, כגון מזג אוויר ותאורה. מניפולציה יזומה של קידוד זה מאפשרת לבצע אינטרפולציה בין תנאים שונים הנראים בקלט האימון. בזמן חיזוי, קידוד זה מאפשר לבצע התאמה בנראות של בלוקים צמודים עבור רינדור תוצרים ממבטים המכוסים על ידי יותר מבלוק אחד.

איור 21: קידוד מראה מאפשר למודל לייצג תנאי תאורה, חשיפה ומזג אוויר שונים. מתוך [4]

איסוף איכותי תוך טיוב נלמד של מיקומי מצלמה

תחילה יש לציין, שמערך האיסוף המוצג בעבודה אינו טריוויאלי כלל, ואיכותו מונעת חלק מהקשיים הנובעים ממיקומי מצלמה לא מדוייקים. המערך המוצג כולל 8 מצלמות על גג הרכב האוסף, המספקות כיסוי 360^o, ו-4 מצלמות בחזית הרכב, האוספות קדימה ולצדדים. כל המצלמות אוספות ב-10 הרץ, ושומרות גם ערך חשיפה סקלרי. המצלמות מכויילות ופתורות מבחינת מיקומים מול הרכב ואחת מול השנייה, ומיקום המצלמות מבוצע על בסיס אודומטריה איכותית ע"ב מספר סנסורים בזמן האיסוף. מאחר ומדובר ברכב ואין בעיית משקל - סביר שמדובר ב-IMU איכותי ו- GPS או DGPS, אבל במאמר אין פירוט מדוייק. למרות ההנחה של מיקומים מדוייקים המודל גם לומד עידוני מנח (pose refinement) לכל מצלמה - הזזה וזוויות (t,r) - שנלמדים במשולב עם ה-נרף. בקרה הדוקה על חופש הפעולה של עידונים אלה בשלבים הראשונים מאפשר למודל

ללמוד מבנה גס של הסצנה, לפני שחרור תיקוני המנח. התהליך המשולב בבלוק- נרף מבוסס על עבודות קודמות כגון:

- Barf - Bundle-adjusting neural radiance fields [18]
- A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose [35]
- iNeRF: Inverting neural radiance fields for pose estimation [32]

ועוד מספר עבודות, אך אין כל פירוט במאמר על דרך השילוב במודל או רמת ההשפעה של כל אחת מעבודות קודמות אלה.

אובייקטים נעים ושינויים גאומטריים תלויי זמן

בכל שחזור תלת מימד קיימת הנחה שהסצנה סטטית ולא השתנתה בין צילום מבטים שונים עליה. אובייקטים נעים בתוך הסצנה אינם מקיימים הנחה זו, ויכולים לגרום לשגיאות משמעותיות בשחזור (ד"א נכון באותה מידה לשחזור MVS קלסי). בלוק-נרף מריצים מודל סגמנטציה סמנטית על הקלט, המסנן מחלקות נפוצות של אובייקטים נעים (רכבים, אנשים, אופניים וכו), וממסכים אותם כבר בקלט. שינויים לאורך זמן באובייקטים סטטיים - כדוגמת שלבי בנייה שונים באיסוף מתמשך, או שינויים בצמחיה - לא מקבל מענה במיסוד זה, ועלולים להכניס שגיאות או טישטושים, אבל החלוקה לתת מקטעים מאפשרת אימון מחדש של אזור מצומצם במקרה כזה.

חיזוי נראות (Visibility Prediction)

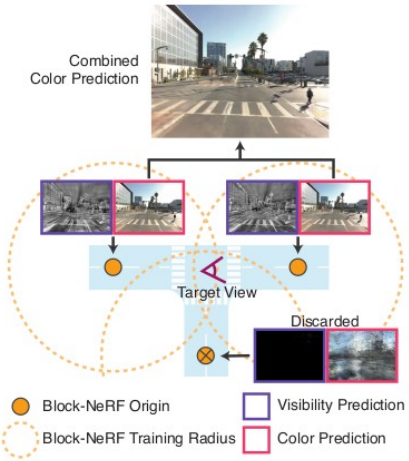
במקביל ל-MLP המחשב אטימות (f_σ) מאומנת רשת MLP קטנה f_v במקביל (כרשת נפרדת). רשת זו מקבלת גם היא מיקום (מקודד IPE) וכיוון הסתכלות (מקודד PE), ומשערכת את נראות הנקודה עבור כל דגימה לאורך קרן באימון. נראות במקרה זה מייצגת עד כמה נקודה ספציפית ניתנת לצפיה ממצלמת קלט נתונה. נקודות בחלל החופשי או קרוב לפני השטח יקבלו ערך קרוב ל-1, בעוד נקודות בתוך או מאחורי אובייקטים יקבלו ערך קרוב ל-0. נקודות שנראות מחלק ממצלמות הקלט, בעוד שאינן נראות ממצלמות אחרות, יקבלו את הערך הממוצע על פני כל מצלמות האימון, שיהיה בין 0-1 וייצג נקודה נראית חלקית. פלט מרשת זו משמש בזמן חיזוי בבחירת איזה בלוק-נרפים לאחד בעת שיחזור נקודת מבט החוצה בלוקים (ראו בהמשך), וכן מסייעת בהשוואת הנראות בין בלוקים (ראו בהמשך).

קידוד מדד חשיפה

איור 22: דוגמה לתנאי חשיפה שונים מאוד באותו מיקום בין איסופים שונים. מתוך [4]

מאחר והקלט עשוי להאסף לאורך זמן ובמשרעת רחבה מאוד של רמות חשיפה, אימון מודל שלא מתייחס למימד זה בקלט עלול להיפגע באופן משמעותי. הזנה של נתוני החשיפה של המצלמה כחלק מקידוד הנראות מאפשר למודל לפצות על הבדלים באיסוף הסצנה. מקודדים את מהירות הצמצם כפול הגברה אנלוגית חלקי פקטור קנה מידה (מעשית, פקטור זה הוא 1000), על בסיס PE סינוסואידלי בעל 4 רמות.

בחירה דינמית בעת חיזוי

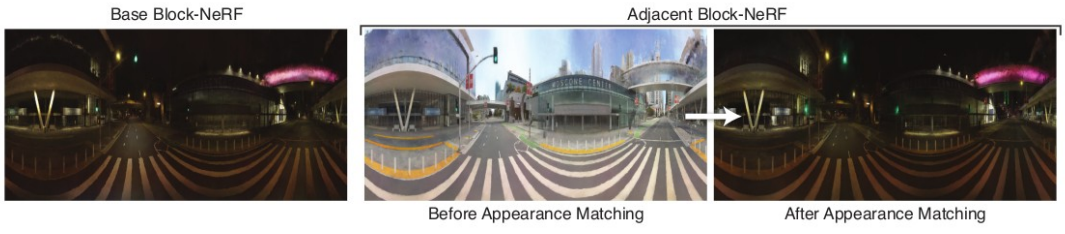


איור 23: דוגמה לסינון בבחירת בלוקים להשתתפות בשחזור תמונה מנקודת מבט נדרשת על בסיס נראות. מתוך [4]

הסצנה הכללית עלולה להכיל מספר כלשהו של בלוק-נרפים. לצורך יעילות, בוחרים את הבלוקים הרלוונטיים לשחזור נקודת מבט ספציפית על ידי 2 פילטרים שונים. הראשון - רק בלוקים שמרכזיהם נכללים ברדיוס קבוע מנקודת המבט נשקלים לשימוש. בהמשך, משתמשים ברשת הנראות f_v של כל אחד מהבלוקים המועמדים כדי לחשב את ממוצע הנראות מנקודת המבט הנתונה. אם ממוצע זה מתחת לסף, הבלוק ייפסל לשימוש ב בשחזור זה. נראות ניתנת לחישוב מהיר, כי זו רשת נפרדת מחיזוי הצבע, ולא נדרשת לרינדור ברזולוציית היעד. לאחר סינון זה, נותרים בדרך כלל 1-3 בלוקים שישתתפו באיחוד. באיור 23 ניתן לראות דוגמה, כששלושה בלוקים נמצאים ברדיוס החיפוש של נקודת המבט לרינדור. בלוק-נרף עם הערכת נראות נמוכה (התחתון באיור) נפסל, ואיחוד המבטים מתבסס על שחזור תוצר הצבע משני הבלוקים הנותרים, ואינטרפולציה בינהם לפי משקול מרחק הופכי (inverse distance) ממרכזי הבלוק למיקום נקודת המבט המשוחזרת.

מנגנון השוואת נראות בין בלוק-נרפים שכנים בעת חיזוי

כפי שכבר הוסבר, כל בלוק-נרף מאמן בנפרד קידוד מראה לכל תמונת קלט. מאחר וקידוד זה מאותחל רנדומלית במהלך האימון, אותו קידוד יוביל לתוצאות שונות אם יוזן לבלוק-נרפים שונים. תוצאה זו אינה רצויה כשמאחדים מספר בלוקים שונים , מאחר וזה עלול להביא לחוסר עקביות בשחזור. בהנתן מראה נתון בבלוק מסויים (בלוק הבסיס להשוואה), מבוצעת השוואת נראות בינו לבין הבלוקים האחרים המשמשים בשחזור. לצורך כך, בוחרים תחילה מיקום תלת מימדי לביצוע ההתאמה בין זוג בלוקים. מיקום זה צריך לקבל ציון נראות גבוה בשני הבלוקים - על בסיס f_v של כל אחד מהבלוקים. בהנתן מיקום זה, מקפיאים את משקולות הבלוק-נרפים, ומבצעים אופטימיזציה רק של קידוד המראה של בלוק המטרה, כדי להוריד את מחיר l_2 בין שני המרנדרים. תהליך זה מהיר, ומתכנס תוך כ-100 איטרציות. תהליך זה לא מביא תוצאות מושלמות, אבל מתקן את רוב תכונות הסצנה הגלובליות, כגון שעה ביום, איזון צבע ומזג אוויר. באיור 24 (למטה) ניתן לראות איך - בהנתן בלוק בסיס (שמאל) שנאסף בלילה - בלוק סמוך שנאסף ביום עובר השוואת מראה.



איור 24: דוגמא להשוואת מראה. מתוך [4]

איסופים, תוצאות ודוגמאות

מאגרי מידע קיימים לא נתנו מענה מספק לצורך שחזור נקודות מבט חדשות בסצנות גדולות. חלקם - למשל KITTI ו-Cityscapes - לא מכילים כיסוי מספק, בעוד אחרים - כגון NuScene ו-Waymo Open Project - מתרכזים בגיוון מבטים ולא בתצפיות חוזרות של איזור מטרה בודד, ומיועדים יותר למחקר בתחומי עקיבה וזיהוי אובייקטים.
בלוק-נרף אספו שני מאגרי מידע. אחד באזור כיכר אלאמו בסאן פרנסיסקו, המכסה שטח של כ- 960X570 מטר, שנאסף לאורך 3 חודשים, עם כ- 40 איסופים שונים , 18-28 דקות כל אחד, ושימש לבדיקות הסקלביליות. אזור זה שוחזר ע"י פריסה של 35 בלוק-נרפים במרחב, שכל אחד מהם אומן על 64,575 - 108,216 תמונות (אחרי סינון תמונות יתירות -למשל מאיסוף במצב עצירה). מאגר המידע הכללי מורכב מ-13.4 שעות איסוף, לאורך 1330 איסופים שונים, ומכיל כ- 2.8 מיליון תמונות לאימון.
מאגר המידע השני נאסף מאזור שכונת mission bay, גם היא בסאן פרנסיסקו. האזור הוא אורבני, עם גאומטריה מאתגרת ומשטחים רפלקטיביים (מתאר מאתגר מאוד ל-MVS למשל). מאגר זה נאסף בנסיעה בודדת של 100 שניות ו-1.08 ק"מ, לצורך הבטחת תנאי סביבה אחידים. סט זה מכיל 12K תמונות מ-12 מצלמות, כולל מטא-מידע (מיקום, אורכי מוקד ונתוני חשיפה), ואף שוחרר לשימוש (ראו בהמשך במידע נוסף).

איור 25: הערכה איכותית של השוואה ל MipNeRF, תוך השוואות חסר שונות. מתוך [4]

	NeRFs	PSNR↑	SSIM↑	LPIPS↓
	mip-NeRF	17.86	0.563	0.509
Ours	-Appearance	20.13	0.611	0.458
	-Exposure	23.55	**0.649**	0.418
	-Pose Opt.	23.05	0.625	0.442
	Full	**23.60**	**0.649**	**0.417**

טבלה 3: הערכה מספרית להשוואה בין Block-NeRF ל-MipNeRF. כולל השוואה של החסרת יכולות שונות והשפעתן.

באיור 25 ניתן לראות את השפעת השיפורים השונים על תוצאה סופית (וכן השוואה ל-Mip-NeRF). ניתן לראות שקידוד המראה מאפשר לרשת להתייחס ולתקן שינויי סביבה כמזג אוויר ותאורה, קידוד נתוני חשיפה משפר קלות דיוק, שיערוך המיקום מחדד ומסיר כפילויות (קל לראות לאורך עמוד הטלפון בשורה העליונה). מעניין לראות את "הסרת" כל הרכבים המופיעים ב- ground truth.

עיקר הדגמת היכולות מועברת טוב יותר בסרטי וידאו ולא בתמונות, ולכן מומלץ מאוד לגשת לאתר הפרוייקט (ראה במידע נוסף בהמשך), ולצפות בסרטים המדגימים יכולות שונות.

סיכום

Block-NeRF מציגים עקרונות וכיוונים לפיתרון כולל לבעיה של שחזור שטח נרחב, ומציעים מענה לשורה של בעיות העולות בשחזור סצנות גדולות אמיתיות - כיסוי, קושי באיסוף בתנאי סביבה ופרמטרי מצלמה שונים, שחזור באזורי תפר בין תת בלוקים, התמודדות עם אובייקטים נעים בסצנה "חיה" (להבדיל מסינטטית), וכן הוספת יכולות של שינוי הסצנה בעת שיחזור לתנאי סביבה שונים מזמן האיסוף, שמשפר מאוד את השימושיות עבור סימולציות. מנגד - מערך האיסוף המוצג איננו גנרי או פשוט, לא פורסם קוד ולא הוצג פירוט מבנה הרשתות המשופר או האלגוריתמיקה הנוספת. אופי האיסוף הקרקעי מאפשר שחזור של מבטים חדשים מנקודת מבט של רכב או של רחפן מנמיך טוס, אך לא של רחפן מגובה מעל הביניינים, מאחר והגגות לא ניצפו. מומלץ לצפות בסרטון מרחוב לומברד (רחוב מפורסם בסאן פרנסיסקו - משופע מאוד ומפותל) שבו שוחזר יעף רחפן (למרות שהאיסוף רכוב) ובסופו מגביה טוס כדי להציג את גבול השחזור מעל הגגות. באותה מידה- איסוף במתאר של רבי קומות או איסופים במתאר שלא מאפשר נסיעה סדורה , עלולים לאתגר את אופי הפיתרון הנ"ל. איכות האודומטריה הקיימת לרחפנים, ומספר המבטים הרב הנדרש - שהאיסוף הקרקעי מאפשר - יאתגר ניסיון להפעיל את הפיתרון הנ"ל על איסוף אווירי קרוב. לנקודת העבודה המוצעת - הפיתרון מרשים בהקפו וביכולותיו. חשוב לציין שמרבית השיפורים והיכולות ש-Block-NeRF מציגים באים לידי ביטוי רק בהתמודדות עם סצנות גדולות

הדורשות מספר בלוקים. בטיפול בסצנה "קטנה" הניתנת לכיסוי על ידי בלוק בודד - רוב היכולות לא באות לידי ביטוי, ותוצאת הפיתרון יהיו דומות לפיתרון Mip-NeRF שעליו הוא מבוסס. עם זאת, מדד הנראות וקידוד מדד חשיפה משפרים ביצועים ומתמודדים עם בעיות שנובעות מאופי הסצנה החיצוני -לא מגודלה - כך שזהו שיפור על פני MipMap המשמעותי גם עבור בלוק בודד.

מידע נוסף

אתר הפרוייקט נמצא ב- https://waymo.com/research/block-nerf/ , וכולל גישה למאמר, גישה למאגר המידע שפורסם (Mission Bay Dataset), הכולל כ-12K תמונות ומטאדאטא תואם, ו-9 סרטוני הדגמה המראים יכולות שונות של המערכת (מומלץ!). לא פורסם קוד במקביל למאמר.

Mega-NeRF 4.3

עבודתם של Turki *et al.* [3], מתרכזת בשחזור נרפי של סצנות גדולות על בסיס איסופי רחפן, באופן סקלבילי. ספציפית, מגה-נרף בא לפתור 3 בעיות מרכזיות, הנוגעות לשחזור סצנות גדולות. הראשונה היא הצורך למדל אלפי תמונות, עם נתוני תאורה וחשיפה שונים, כשכל אחת מהן צופה רק על מקטע קטן מהסצנה. בעיה שנייה היא הצורך להכיל את המודל, שבתצורה בדידה גדול בצורה חריגה, ושהופך את אפשרות האימון על GPU בודד בלתי ישימה. הבעיה השלישית היא האתגר המשמעותי שבנסיון לרנדר באופן מהיר מספיק לצורך רינדור יעפים אינטראקטיביים.

תרחיש הייחוס העיקרי הוא מתאר חיפוש והצלה בסביבה אורבנית, על בסיס איסופי רחפן, שהוא פיתרון זול וזמין לסקירה של שטח נתון, ללא תלות בעבירות קרקעית, המאפשר סקירה מהירה ותיעדוף משאבי חילוץ והצלה - כלים וכוח אדם (קרקעיים). בגלל מגבלות זמני סוללה ורוחב פס, פתרונות אלו כיום לרוב יכללו בניה של תמונה אורטוגרפית דו מימדית לאחר האיסוף, וניתוח שלה בדיעבד. מגה-נרף מציעים להפוך את הסקירה הזו לתלת מימדית, תוך מתן אפשרות לכוחות הסורקים לבחון את כל השטח כאילו הטיסו רחפן בדיוק לנקודות המבט המעניינות אותם, ברמת פירוט שלא אפשרית מבחינת SFM או MVS קלסי. תרחיש הייחוס הופך את זמן האימון לקריטי - בהנתן סטטיסטיקות ההישרדות של ניצולים בהריסות. הדרישה לאינטראקטיביות מאתגרת את פתרונות הרינדור המהיר מ-נרף הקיימים (בעת כתיבת המאמר).

מגה-נרף מציעים את הפתרונות הבאים לסט הבעיות הנתון:

- פירוק מרחב הסצנה התלת מימדי, תוך פיצול תמונות הקלט (ואף **חלקי** תמונות) בין תת מודלים שונים.
- קידוד מראה ותנאי סביבה
- שיפור בתהליך האימון, על מנת לנצל תכונות של מקומיות במרחב המאומן, תוך אימון ממוקבל שמאפשר שיפור 3X תוך שיפור איכות השחזור (מול גישות קיימות).
- גישת רינדור חדשה, המנצלת תכונות רציפות טמפורלית הקיימת באופי מסלול השחזור הנובע מתרחיש היחוס, תוך בחינה של פתרונות רינדור מהירים קיימים.

בנוסף, מוצג מאגר מידע חדש, שנאסף מיעפי רחפן, המכיל אלפי תמונות של אזור בגודל כ- 100K m^2.

פתרונות ברמת ארכיטקטורה:

קידוד מראה ותנאי סביבה

בדומה ל- NeRF-W [12] - ול-בלוק-נרף [4] שאימצו את אותה גישה - מגה-נרף משייכים וקטור קידוד מראה $l^{(a)}$ עבור כל תמונה a, מה שמאפשר למודל המאומן להתמודד עם תנאי תאורה שונים, תכונה שנדרשת בהתחשב בגודל הסצנה וזמן הכיסוי הנדרש.

חלוקה מרחבית של הסצנה

מגה-נרף מחלקים את המרחב התלת מימדי לרשת (פוטנציאלית תלת מימדית) של תאים, שמרכזיהם $n_{\in N}=(n_x, n_y, n_z)$, ומאתחלים לכל תא מודל מבוסס NeRF של משקולות f^n. בגלל אופי הכיסוי האווירי, נשקלו מגוון דרכים לכסות את המרחב, כולל קיבוץ מבוסס k-means, אך לבסוף נבחרה גישה פשוטה של רשת 2D של תאים, המכסה את הסצנה, בגלל יחסי הגדלים בין הגובה לאורך ורוחב השטח המכוסה. בסצנות המוצגות שונות הגבהים של מרכזי התאים הנדרשים הייתה זניחה, ולכן כל התאים ממורכזים סביב אותו גובה, אבל הגישה מאפשרת התייחסות ברורה למתארי איסוף אחרים - צלע הר, הפרשים משמעותיים בגובה בין תאים שונים, או מתאר שדורש כיסוי תאים תלת מימדי לא אחיד (גורד שחקים במרכז אזור שטוח לדוגמא). בעת תשאול, מגה-נרף יחזיר ערכים על בסיס מודל f^n שמרכז התא שאליו הוא משוייך קרוב ביותר:

$$f^n(x)=\sigma \quad [3](17)$$

$$f^n(x,d,l^{(a)})=c \quad [3](18)$$

$$where \quad n=argmin_{n\in N}\|n-x\|^2 \quad [3](19).$$

פירוק לחזית ורקע הסצנה

איור 26: חזית הסצנה לפי Mega-NeRF (ימין) ו- NeRF++ (שמאל). מתוך [3]

NeRF++ [13] - הרחבה של נרף שעוסקת בסצנות לא תחומות (unbound) - מציעים לחלק את המרחב לחזית הסצנה (scene foreground) הכוללת את הכדור העוטף את כל מיקומי המצלמות המשתתפים ב-נרף ספציפי, ואת רקע הסצנה (scene background), שהוא האזור המשלים מחוץ לכדור זה ועלול להמשך לטווח ארוך בהתאם למתאר האיסוף. רקע הסצנה ממודל על ידי MLP נפרד, אחרי פרמטריזציה מחדש. מגה-נרף משתמשים באותה חלוקה, אבל במקום כדור מגדירים אליפסואיד הדוק יותר, כדי להשיג תיחום הדוק יותר לאזור העניין. בנוסף, נתוני הגובה של המצלמות משמשים גם הם לתיחום דגימות הקרניים, כדי להימנע מדגימת אזורים תת קרקעיים.

פיתרון מיקום מצלמות

מגה-נרף משתמשים בחבילת SFM בשם PixSFM [17] על מנת לפתור ולעדן את מיקומי המצלמות שמקורן ב-GPS ו-IMU של הרחפן, מאחר ותוצאות ישירות ללא עידון יצאו מאוד מטושטשות. נבחנו גם כלים מבוססי למידה כ-BaRF [18], וכן כלים מסחריים לפיתרון פוטוגרמטרי של יעפי רחפן כ-Pix4DMapper. PixSFM הביא באופן עקבי לתוצאות טובות יותר, עם הפרש של מעל 6db באיכות התוצאות מהפיתרון הבא באיכותו (Pix4DMapper). פיתרון SFM זה מורץ כתהליך קדם.

פתרונות ברמת מתודולוגיית אימון:

מקבול מידע מרחבי לתת מודלים

כל תת מודל של מגה-נרף (f^n) הוא MLP בפני עצמו, המאומן במקביל לתתי מודל אחרים ללא קשר בינהם. מאחר וכל תמונה מכסה רק חלק קטן מהסצנה הכוללת, מגבילים את שיוך התמונות ההתחלתי לסטי אימון כך שפיקסל נדגם יישוייך רק לתתי המודל שהקרן שלו חותכת. מוגדרת חפיפה של 15% בין תאים צמודים, כדי להמנע מארטיפקטים בשחזור. חלוקה זו מתקיימת בשלב ההתחלתי, כשאין מידע על הגאומטריה הספציפית לסצנה, ומביאה לצימצום סט האימון של כל תא לכ- 10% מהסט המלא. כמובן שאחוז זה יורד ככל שגודל הסצנה המכוסה עולה.

דילול מידע מרחבי לתת מודלים

לאחר שלב האימון הראשוני - כ-100K איטרציות - כשהתאים השונים לומדים את הגאומטריה הגסה, ניתן לדלל פיקסלים נוספים בתמונות על בסיס הסתרות מתאים קרובים יותר לאורך הקרן. בניסויים שערכו כותבי [3], דילול נוסף זה לבדו הקטין פי שניים את גודל סטי האימון לתאים השונים .

איור 27: Mega-NeRF - חלוקת קלט ראשונית ולאחר דילול.
מתוך [3]

באיור 27 ניתן לראות (שורה עליונה) חלוקה ראשונית של תמונות לעומת דילול לפי גאומטריה (שורה תחתונה). אותה תמונה שבאופן ראשוני נכללה כמעט כולה ב-3 תאים שונים, לאחר חלוקה פיקסלית לתאים ניתן לראות שרוב הפיקסלים בתמונה מתחלקים בין 2 תאים (אדום וכחול), בעוד עדיין נשמרת השפעה קלה על תא מרוחק נוסף (צהוב), וכמות הפיקסלים שמשתתפים באימון כל תא יורדת משמעותית.

שיפורים ברמת רינדור בשאיפה לרינדור אינטראקטיבי:

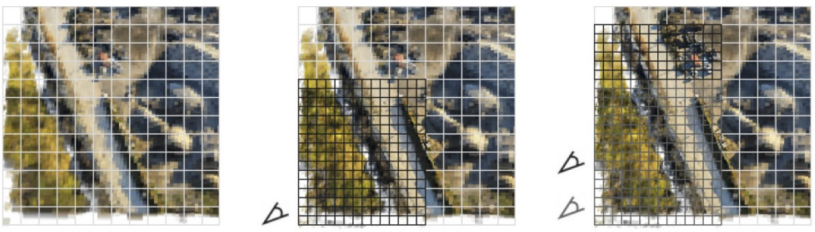
תרחיש הייחוס של מגה-נרף הוא אימון על שטח גדול בזמן קצר מספיק שכוחות הצלה יוכלו לנצל את המודל המאומן לתיחקור הסצנה כוידאו של מעוף אינטראקטיבי. כפי שצויין בעבר, זמני רינדור תמונה ב-NeRF אינם קרובים לקצב אינטראקטיבי, כשהמרנדר של NeRF [1.1] דורש מעל 2 דקות לייצר תמונת 720p בודדת. כדי לעמוד בהגדרה של רינדור אינטראקטיבי במתאר הייחוס, מגה-נרף נדרש ל-3 משימות: לשמר את נאמנות השיחזור, לצמצם זמני חישוב נוספים מעבר לאימון עצמו, ולהאיץ את הרינדור עצמו. במסגרת המחקר נבחנו מספר מימושים של מרנדרים מהירים קיימים (ראו פירוט בתוצאות), בנוסף לפיתרון ספציפי של מגה-נרף.

טיפול קדם והטמנה (caching)

רוב המרנדרים המהירים הקיימים ל-NeRF (בעת כתיבת המאמר), עושים שימוש בחישובים מקדימים על הסצנה ובניית זיכרון מטמון, תהליך שלא מתיישב בקנה אחד עם גודל הסצנות המדוברות. לדוגמא Plenoctree [16] בונה עץ חלוקה ווקסלי (voxel octree) דליל מנתוני הצפיפות והמקדמים ההרמוניים. בנייה של עץ מלא ב-8 רמות לקחה שעת חישוב ובין 1 ל- 12 GB זיכרון (בהתאם לפורמט הקידוד של הקירון). להוספת רמה בודדת נוספת כבר נדרשו 10 שעות ועץ החלוקה המתקבל הגיע ל-55 GB - עובדה שמונעת שימוש בכל GPU פרט לחזקים ביותר כיום. מצד שני - שימוש בעץ מפורט פחות מוריד את רמת הפירוט בפלט, ומעקר את יתרון האיכות שמציע שימוש ב-נרפים. מתאר הייחוס של מגה-נרף לא מאפשר כמובן הנחה של חישובי קדם כאלה, גם אם קיימת חומרה מתאימה.

רציפות טמפורלית ברינדור ועץ ייצוג דינמי

מגה-נרף מציעים מרנדר נרף שמנצל את ההבנה שבמתאר אינטראקטיבי, יש חפיפה משמעותית בין פריימים עוקבים. לאחר שחושב המידע הנדרש לרינדור מבט מסויים, חלק גדול ממידע זה נכון גם לפריים העוקב. כמו Plenoctree, מחושב עץ יצוג גס של צבעים וצפיפויות, כך עץ זה לא מחושב כולו מראש, אלא גדל באופן דינמי בהתאם למסלול המשוחזר. במהלך השחזור, המרנדר משתמש בעץ הקיים כדי לייצר מבט ראשוני, תוך כדי דגימות נוספות של המודל כדי להוסיף פרוט, דגימות שמוספות דינמית לעץ הייצוג הקיים - כהכנה לפריים הבא. מאחר וכל פריים מקיים חפיפה גבוהה עם הפריים הקודם לו, מדובר רק בכמות קטנה יחסית של עבודה אינקרמנטלית הנדרשת כדי לשמר איכות שיחזור מספקת.

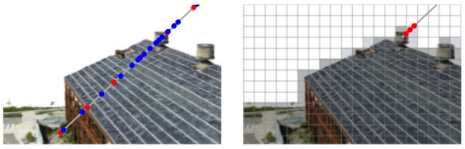


איור 28:המרנדר הדינמי של Mega-NeRF. מתוך [3]

כדי להמנע מייצור עץ ייצוג סטטי, שיסבול או מרזולוציה גסה מדי או מזמן חישוב גדול מדי לייצור, המרנדר הדינמי של מגה-נרף (ראו איור 28) מרחיב את עץ הייצוג בהתאם למיקום הנוכחי. פריימים עוקבים יוכלו להשתמש בחלק גדול מהמידע, ורק חלק קטן יידרש לחישוב נוסף.

דגימה מונחית

מגה-נרף מבצעים סבב בודד של דגימות קרניים אחרי עידון עץ הייצוג, על מנת לשפר את איכות הרינדור. להבדיל מהגישה הדו שלבית של NeRF (דגימה גסה ועדינה), מגה-נרף משתמשים בערכים השמורים בעץ הייצוג במקום הדגימה הגסה, ומאחר ועץ הייצוג נותן שיערוך איכותי של פני המשטח, רק כמות קטנה של דגימות, הממוקמות קרוב לפני המשטח באיזורי העניין. בדומה למרנדרים מהירים אחרים שמולם נערכה ההשוואה. ערך הקירון נצבר לאורך הקרן ונמנעים מביצוע דגימות (מיותרות) מעבר לסף נתון.



איור 29:דגימה רגילה ב-NeRF לעומת דגימה מונחית ב Mega-NeRF. מתוך [3]

באיור 29 : ב-NeRF (משמאל) מתבצעת דגימה רגילה הדוגמת דגימה גסה בצעד קבוע, ומשפרת את מיקומי הדגימה העדינה בהתאם. דגימה מונחית ב-Mega-NeRF משתמשת בעץ הייצוג במקום בדגימה גסה כדי לדלג על אזורים ריקים ולהוסיף מספר קטן של דגימות ליד משטחים ידועים.

נתוני מודל הרשת ופרטי אימון

נעשו מספר בדיקות לגבי אופן חלוקת תתי המודלים במגה-נרף. המבנה שמולו שוערכו התוצאות הוא של 8 תת מודלים - כל אחד מהם MLP עם 8 שכבות ו- 256 ערוצים לחישוב אטימות σ, ושכבה סופית עם 128 ערוצים לחישוב צבע. משתמשים בדגימה היררכית (כמו NeRF) שבה 256 דגימות גסות ו-512 דגימות מעודנות פר קרן בחזית הסצנה, ו 128/256 פר קרן ברקע הסצנה. להבדיל מ-NeRF דגימות גסות ומפורטות מתבצעות כולן על אותו MLP, כדי לחסוך בנפח המודל ולהוריד 25% מכמות הדגימות פר קרן. תהליך עידון מקודד המראה הוא כפי שמתואר ב NeRF-W [12].

מאגרי מידע

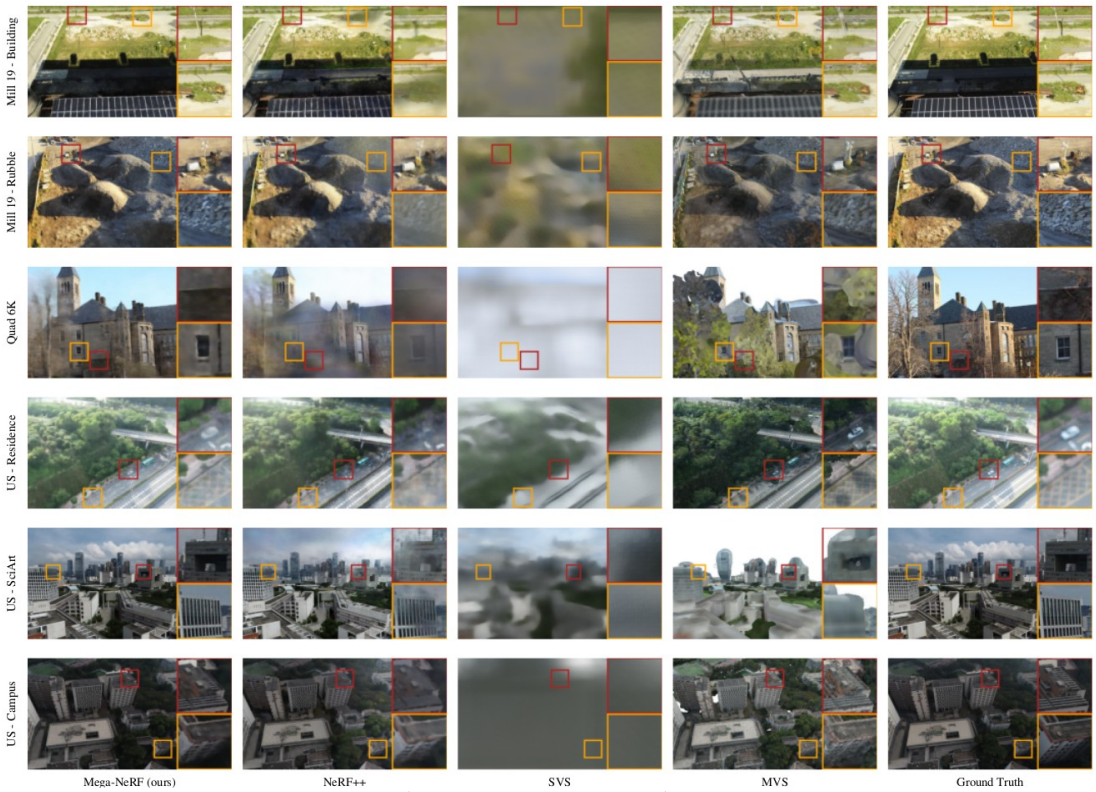
איור 30: גריד איסוף מאגר Mill 19. מתוך [3]

ההערכה נעשית על 2 מאגרי מידע קיימים ואחד שנאסף במיוחד לצורך מחקר זה ופורסם (ראו במידע נוסף). מאגר Quad 6k [19], שנאסף בתחומי אוניברסיטת קורנל לצורך מחקר ב-SFM, ומספר סצנות מתוך UrbanScene3D [20], המכילות איסופי רחפן של אזורים עירוניים נרחבים. בנוסף בוצע איסוף רחפן על שטח תעשייתי לשעבר (Mill 19), ומספק 2 סצנות: מבנה תעשייתי גדול וסביבתו הכולל שטח של 250 x 500 מטרים רבועים, ושטח הריסות קרוב שבו הוסתרו בובות אדם לדימוי איזור אסון (שהוא כזכור מתאר היחוס שהמאמר מנסה לתת לו מענה).

נתוני ה-GPS שכלולים ב- UrbanScene3D וב- Mill 19, וכן נתוני המיקום המשוערכים שב-Quad 6k עברו טיוב מנח ע"י שימוש ב-PixSFM [17]. על כל מאגרי המידע הופעל מודל סגמנטציה סמנטית מאומן מראש על מנת למסך אובייקטים נעים נפוצים, ומתעלמים מפיקסלים ממוסכים במהלך האימון. באיור 30 ניתן לראות את היעף הרשתי בו נאסף מאגר המידע Mill 19 . איסוף שטח של כ-200,000 מטרים רבועים לקח כשעתיים.

הערכה בוצעה בשני מישורים: איכות השיחזור - שנבדק מול מגוון משחזרים שונים ללא קשר לזמן הרינדור הנדרש, ואיכות הרינדור, שבו נערכה השוואה מול מימושי מרנדרים שונים מבחינת איכות ומהירות, כולם על בסיס שחזור של מגה-נרף.

מבחינת משחזרים, השוואה נערכה מול המשחזרים הבאים: NeRF [1.1], NeRF++ [13] - הרחבה של NeRF שמתמודדת עם סצנות לא תחומות, Stable View Synthesis (SVS) [21] - שיטת שיחזור היברידית בה מייצרים שלד גאומטרי על בסיס SFM ומאמנים רשת עמוקה לשערך את בניית הטקסטורה, DeepView - גישת שחזור מבוססת למידה עמוקה שונה מ-NeRF, ולבסוף שימוש בערוץ ה-MVS של חבילת Colmap כדי להשוות לשחזור קלסי.



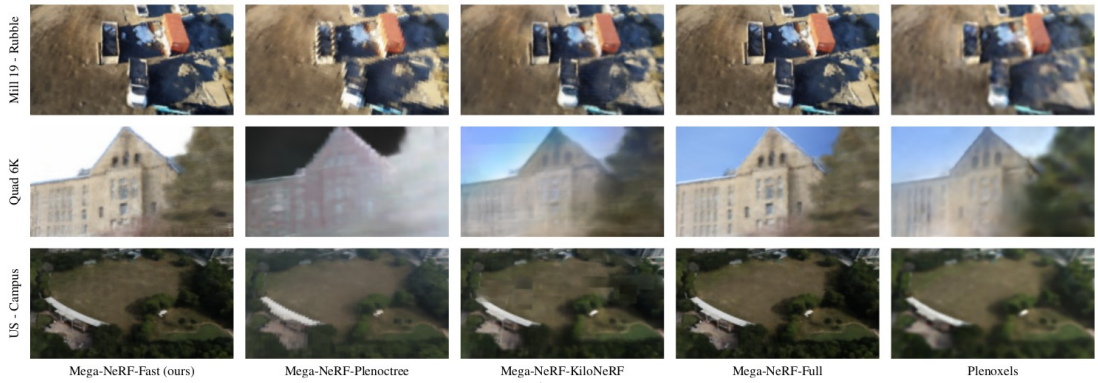
איור 31: השוואה איכותית של תוצרי המשחזרים השונים מול 6 סצנות שונות. מתוך [3]

איור 31 מציג השוואה איכותית של משחזרים שונים. NeRF++ מציג טשטושים, מריחות ואובדן פרטים דומים לדוגמאות קודמות שהוצגו בפתרונות אחרים במידול של אזור גדול מדי במודל בודד. SVS בברור לא מתמודד עם אופי הסצנה המשוחזרת. ההשוואה לפיתרון MVS מעניינת בגלל חוסר העקביות שבה: בשתי הסצנות מ-Mill-19 מוצג שחזור שמתחרה ב-MegaNeRF וטוב מכל האחרים, ב us-campus תוצר רועש יותר אך גם חד יותר מ-MegaNeRF, ב- US-Residence בברור ב-MVS נבחרה טקסטורה מכיוון מבט אחר, מה שגורם לשחזור חד, ברור ובעל יותר פרטים מה- ground-truth עצמו (מה שיגרום ל-PSNR נמוך יותר, למרות שמשתמש אנושי היה מגדיר את השחזור כטוב יותר). ב-US-sciArt בברור היה מספר מבטים נמוך מדי לשחזור איכותי ב-MVS, לעומת שחזור איכותי של Mega-NeRF, וב-Quad 6k ה-MVS משחזר צמחייה בעלת עלווה, למרות שבברור תמונת ה-ground-truth נאספה בעת שלכת - בלי לבחון לעומק את כל המבטים הזמינים במאגר המידע לא ניתן לקבוע מדוע שוחזרה צמחיה ירוקה יש מאין, אבל סט השוואות זה מדגים באופן מצויין את יתרונות גישת NeRF על פתרונות קלסיים בנקודות עבודה מסויימות. השוואה איכותית ל-DeepView לא הוצגה, אבל בהסתמך על טב' 4 (בדף הבא), בה ניתן לראות תוצאות כמותיות להשוואות בין האלגוריתמים המתחרים, נראה שהתוצאה של DeepView דומה באיכותה ל-SVS ולא ברת תחרות.

	Mill 19 - Building				Mill 19 - Rubble				Quad 6k			
	↑PSNR	↑SSIM	↓LPIPS	↓Time (h)	↑PSNR	↑SSIM	↓LPIPS	↓Time(h)	↑PSNR	↑SSIM	↓LPIPS	↓Time(h)
NeRF	19.54	0.525	0.512	59:51	21.14	0.522	0.546	60:21	16.75	0.559	0.616	62:48
NeRF++	19.48	0.520	0.514	89:02	20.90	0.519	0.548	90:42	16.73	0.560	0.611	90:34
SVS	12.59	0.299	0.778	38:17	13.97	0.323	0.788	37:33	11.45	0.504	0.637	29:48
DeepView	13.28	0.295	0.751	31:20	14.47	0.310	0.734	32:11	11.34	0.471	0.708	19:51
MVS	16.45	0.451	0.545	32:29	18.59	0.478	0.532	31:42	11.81	0.425	**0.594**	**18:55**
Mega-NeRF	**20.93**	**0.547**	**0.504**	**29:49**	**24.06**	**0.553**	**0.516**	**30:48**	**18.13**	**0.568**	0.602	39:43
	UrbanScene3D - Residence				UrbanScene3D - Sci-Art				UrbanScene3D - Campus			
	↑PSNR	↑SSIM	↓LPIPS	↓Time (h)	↑PSNR	↑SSIM	↓LPIPS	↓Time(h)	↑PSNR	↑SSIM	↓LPIPS	↓Time(h)
NeRF	19.01	0.593	0.488	62:40	20.70	0.727	0.418	60:15	21.83	0.521	0.630	61:56
NeRF++	18.99	0.586	0.493	90:48	20.83	0.755	0.393	95:00	21.81	0.520	0.630	93:50
SVS	16.55	0.388	0.704	77:15	15.05	0.493	0.716	59:58	13.45	0.356	0.773	105:01
DeepView	13.07	0.313	0.767	30:30	12.22	0.454	0.831	31:29	13.77	0.351	0.764	33:08
MVS	17.18	0.532	**0.429**	69:07	14.38	0.499	0.672	73:24	16.51	0.382	**0.581**	96:01
Mega-NeRF	**22.08**	**0.628**	0.489	**27:20**	**25.60**	**0.770**	**0.390**	**27:39**	**23.42**	**0.537**	0.618	**29:03**

טבלה 4: השוואה כמותית בין שיטות השחזור השונות. מתוך [3]

מבחינת מרנדרים, ההערכה מתבצעת בין מימושים של שני מרנדרי נרף מהירים קיימים - Plenoctree [16] ו- KiloNeRF [23], מימוש מרנדר NeRF מקורי (מכונה בהשוואות Mega-NeRF-Full), ומימוש מרנדר Mega-NeRF דינמי. כל אלה מומשו במסגרת Mega-NeRF ונבחנים על אותו מודל מאומן. בנוסף נבחן מרנדר בשם Plenoxels [22] שאומן על ווריאנט אחר, המשתמש בהרמוניות ספריות. Plenoxels מייצר גם הוא מבנה ווקסלי דליל, אבל מבצע אינטרפולציה טרי-ליניארית במקום NN (nearest neighbour).



איור 32: השוואה איכותית של מרנדרים שונים. מתוך [3]

באיור 32 ניתן לראות שביחס ל-Mega-NeRF : Plenoctree מייצר שגיאות ווקסליזציה בולטות, Plenoxel מרנדר מטושטש, KiloNeRF מחזיר תוצאות טובות אך מתקשה בפרטים עדינים ומכיל מספר ארטיפקטים נראים.

best second-best	Mill 19					Quad 6k					UrbanScene3D				
				Preprocess	Render				Preprocess	Render				Preprocess	Render
	↑PSNR	↑SSIM	↓LPIPS	Time (h)	Time (s)	↑PSNR	↑SSIM	↓LPIPS	Time (h)	Time (s)	↑PSNR	↑SSIM	↓LPIPS	Time (h)	Time (s)
Mega-NeRF-Plenoctree	16.27	0.430	0.621	1:26	**0.031**	13.88	0.589	0.427	1:33	**0.010**	16.41	0.498	0.530	**1:07**	**0.025**
Mega-NeRF-KiloNeRF	21.85	0.521	0.512	30:03	0.784	20.61	0.652	0.356	27:33	1.021	21.11	0.542	0.453	34:00	0.824
Mega-NeRF-Full	**22.96**	**0.588**	**0.452**	-	101	**21.52**	**0.676**	0.355	-	174	**24.92**	**0.710**	**0.393**	-	122
Plenoxels	19.32	0.476	0.592	-	0.482	18.61	0.645	0.411	-	0.194	20.06	0.608	0.503	-	0.531
Mega-NeRF-Initial	17.41	0.447	0.570	**1:08**	0.235	14.30	0.585	0.386	**1:31**	0.214	17.22	0.527	0.506	1:10	0.221
Mega-NeRF-Dynamic	22.34	0.573	0.464	**1:08**	3.96	20.84	0.658	**0.342**	**1:31**	2.91	23.99	0.691	0.408	1:10	3.219

טבלה 5: השוואה כמותית של המרנדרים השונים. ההשוואה מתייחסת לאיכות הרינדור ולביצועי זמנים במקביל. מתוך [3]

מאחר והמרנדר הדינמי מבצע רינדור ראשוני על בסיס עץ החלוקה הפנימי, ואז מוסיף פרטים עדינים, הוספה הערכה של שלב הביניים הנ"ל כ- Mega-NeRF-Initial.
כפי שניתן לראות בטבלה 5, מימוש המרנדר המלא מחזיר PSNR גבוה ביותר בכל בכל הסצנות, במחיר זמנים לא אינטראקטיבים בעליל, בעוד המרנדר הדינמי של מגה-נרף שומר על מרחק של 0.8db ממנו, בהאצה של פי 40. Plenoctree אמנם שומר על FPS גבוה, אבל ההתבססות שלו על בניית מבנה נתונים ווקסלי גורם לביצועיו לרדת משמעותית כתלות בגודל הסצנה. כדאי לשים לב גם לזמן ההכנה המקדים הנדרש עבור כל מרנדר

סיכום

MegaNeRF מציגים מנגנון חלוקה מרחבי דליל, המסוגל להתמודד עם גודל סצנה גדול כרצוננו, מאפשר חלוקת קלט האימון למספר תתי מודלים הניתנים לאימון בנפרד ובאופן ממוקבל, וכן סכמת בחירת תת מודל מרנדר פשוטה. השיפורים המוצעים מאפשרים האצה של עד פי 3 בזמן האימון, תוך העלאה של איכות השחזור. השוואת המרנדרים מקיפה ומשמעותית, אבל הפיתרון עוד דורש עבודה לאינטראקטיביות אמיתית. גישת הרציפות הטמפורלית שהמרנדר הדינמי הציג היא צעד מעניין בכיוון, מאחר והוא מצמצם חישובים לא נחוצים בין מבטים עוקבים. בחינה מחדש לאור כלים שיצאו במקביל או מעט מאוחר יותר - כגון Instant-NGP (ראו בהמשך) - קריטית.

מידע נוסף

אתר הפרויקט נמצא ב- https://meganerf.cmusatyalab.org/, ומוצגים בו המאמר, קישור לחבילת הקוד, קישור למאגר המידע שנאסף במהלך המחקר (Mill 19) ומספר סירטוני הדגמה.

עבודתם של Ramatas *et al*. [5], URF מציג הרחבה ל-NeRF, במטרה לבצע שחזור תלת מימד ושחזור נקודות מבט חדשות, מתוך מידע הנאסף על ידי פלטפורמות **מיפוי** (לדוגמא Google Street View). איסופים כאלה מכתיבים לרוב איסוף קרקעי של סצנות גדולות (מאות מטרים רבועים), בסביבות פחות מבוקרות מרוב הסצנות הנפוצות במחקר NeRF: מבנים ואובייקטים רבים ושונים, תאורה טבעית מהשמש (כולל הרבה הסתרות). מערך האיסוף נישא ע"י אדם - תיק גב שעליו מצלמת עין דג, GPS וחיישן לידר (LIDAR). המצלמה אוספת תמונות פנורמיות של הסצנה, בזמן שהלידר מייצר ענן נקודות תלת מימדי. מערך ה-Trakker משתמש ב-GPS ו-SFM כדי לחתום כל תמונה ואת דגימות הלידר עם מנח גלובלי - מיקום וזוויות. חשוב לציין שדגימות הלידר נאספות באופן א-סינכרוני מהתמונות. אופי האיסוף מוכוון מסלולי הליכה - כלומר קווים ישרים יחסית, בחלקים ספציפיים של הסצנה - מה שגורר כיסוי לא אחיד של הסצנה מבחינת כמות וכיוון מבטים. בנוסף, במערך ב Trakker של גוגל (שעל איסופים שלו התקיים המחקר) המצלמה דוגמת ב 2Hz, כך שמאגר המידע מכיל כמות קטנה משמעותית של מבטים ממאגרי מידע אחרים, כולל כאלה שהוצגו בחלקים קודמים בעבודה זו. נתונים אחרים שמייחדים את אופי האיסוף: כמות גדולה של פיקסלים צופים בשמיים, וכן כמות משמעותית של אובייקטים נעים בסצנה (בעיקר בני אדם), אופי התמונות משתנה מבחינת חשיפה (auto-exposure), ותאורה (auto white balance / AGC) תוך כדי האיסוף, כך שאותו מבנה נראה בצבעים שונים בין תמונות שונות באיסוף. בנוסף, צפיפות הנקודות מהלידר משתנה ביחס לטווח מהאוסף, והן עלולות להיות חסרות לחלוטין במקרה של משטחים שקופים או מבריקים. מידע מבוסס Street View מעניין בין השאר כי איסופים שלו קיימים בחלקים נרחבים של העולם, והוא מאפשר בחינה של סצנות נרחבות ומגוונות.
כדי לבחון את כלליות הפיתרון בוצעו גם שיחזורים של מאגר הנתונים Tanks and Temples, לאחר שהריצו Colmap על הקלט והזינו את ענן הנקודות המשוערך במקום דגימות לידר [5.1].

URF מציע 3 הרחבות של מודל ה-NeRF:

הרחבה ראשונה - שילוב מידע הנובע מהלידר, דבר שיכול לפצות על דלילות מספר המבטים ביחס לגודל הסצנה. שילוב זה מושג על ידי הוספת פונקציות מטרה מבוססות לידר.
הרחבה שנייה היא הוספת מידול נפרד לפיקסלי שמיים ופיקוח מוגדר היטב בעת האימון על השפעת קרניים הפונות לכיוון השמיים, והרחבה שלישית - תיקון ההבדלים בנראות על בסיס שיערוך טרנספורמציית צבע אפינית עבור כל מצלמה.

טיפול מקדים בקלט

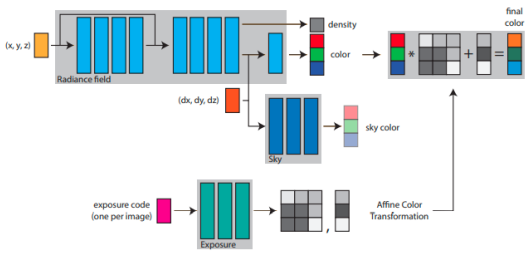
שני כלי מדף מורצים על תמונות הקלט: סגמנטציה סמנטית (שבעזרתה ממסכים פיקסלים המכילים אנשים) וזיהוי שמיים. בהנתן מקור מאגר המידע (google trakker), התבססו החוקרים על הנחה שמיקומי המצלמות ומיקומי מקורות דגימות הלידר עברו שיערוך מנח (pose estimation) בעזרת SFM באופן אוטומטי בזמן האיסוף כחלק מתהליך הקליטה, ולכן לא מורץ טיוב מנח נוסף בעת הטיפול של URF בקלט.

הגדרת המודל

URF מגדיר MLP בודד, עם סט משקולות θ עבור כל הסצנה וסט נתוני חשיפה $\{\beta_i\}$ עבור כל תמונה (ראו איור 33 בדף הבא). בהנתן תמונה ומידע הלידר עבור הסצנה, מטייבים את URF על ידי מינימיזציה של פונקציית המחיר הבאה:

$$\underset{\theta,\{\beta_i\}}{arg\,min} \quad L_{rgb}(\theta,\{\beta_i\})+L_{seg}(\theta)+L_{depth}(\theta)+L_{sight}(\theta) \quad [5] \quad (20)$$

המורכבת מרכיבים פוטומטריים (L_{rgb}, L_{seg}) ורכיבים מייצגי לידר (L_{depth}, L_{sight}).



איור 33: מבנה מודל URF. מתוך [5.1]

פונקציות מחיר פוטומטריות

$L_{rgb}(\theta,\beta_i)$ דומה לפונקציית המחיר המקורית של NeRF , מלבד התייחסות לפונקציית צבע מעודכנת, המשקללת את נתוני החשיפה של כל תמונה, ומתייחסת לזיהוי שמיים

$$C(r;\beta_i)=\int_{t_n}^{t_f} w(t)\cdot\Gamma(\beta_i)\cdot c(t)dt+c_{sky}(d) \quad [5](21)$$

כאשר $\Gamma(\beta_i)$ מייצג שיערוך של מטריצה 3x3 הממפה טרנספורמציה אפינית של הקירון המשוערך:

$$\Gamma(\beta_i)=\Gamma(\beta_i;\theta):\mathbb{R}^B\rightarrow\mathbb{R}^{3x3} \quad [5](22)$$

ייצוג זה נבחר במקום קידוד המראה (ראה עמ' 29) המשמש ב- NeRF-W (וכן ב-Block-NeRF וב- MegaNeRF שהצגנו) מתוך טענה שקידוד המראה מבצע פרמטריזציית-יתר, ומאפשר לרשת לפצות באמצעותו גם על שגיאות שאינן תלויות חשיפה, ואילו $\Gamma(\beta_i)$ ממפה את השינויים בחשיפה ובתיקוני תאורה גלובליים באופן הדוק יותר, ולכן צפוי פחות לערב בין שיערוך קירון הסצנה ושיערוך החשיפה הספציפית לכל תמונה בעת אימון משותף.

$c_{sky}(d)$ מייצג מיפוי קירון כדורי, המיוצג כרשת נוירונים מבוססת קואורדינטות, המשמש כמידול שמיים.

$$C_{sky}(d)=C_{sky}(d;\theta):\mathbb{R}^3\rightarrow\mathbb{R}^3 \quad [5](23)$$

המחזיר צבע רקע תלוי כיוון הסתכלות לאזורי השמיים, שבהם הרשת מתקשה לחזות רמת קירון.

בנוסף, פונקציית מחיר פוטומטרית שנייה L_{seg} מתייחסת גם היא לסגמנטציית השמיים שהורצה, ומעודדת את הרשת לתת צפיפות ואטימות אפסית לכל הדגימות על קרניים העוברות דרך פיקסל ששוערך כשמיים

פונקציות מחיר ללידר

גם פונקציית המחיר ללידר מחולקת לשני רכיבים: פיקוח על העומק בנקודה L_{depth}, ופיקוח על נראות הנקודה L_{sight}. בהנתן נקודה דגומה p ומקור o, ניתן לפקח על העומק המשוערך על ידי פונקציית הצפיפות הנפחית, שתשווה את ערך העומק בנקודה לטווח (p-o). בנוסף ניתן להניח שאין השפעה אטמוספרית על הצפיפות לאורך הקרן עד P, ושכל ערך הקירון המשוערך בפיקסל לאורך הקרן מרוכז ב-p. בנוסף - ערך שנדגם ב- p פרושו ששיערוך הצפיפות ב-p הוא 1, ואילו כל הדגימות לאורך קו הראייה **עד** p בעלות צפיפות אפסית.

תוצאות ודוגמאות

נבחרו מתוך מאגר Street View 10 סצנות מרחבי העולם. בכל אחת כ-20 תמונות פנורמה (שכל אחת מורכבת מ-7 תמונות) וכ-6 מליון נקודות לידר דגומות (בממוצע). כל סצנה מכסה מאות מטר רבוע ומייצגת סביבה עירונית שונה. מדווחים מדדים מספריים מהסצנות טאיפיי, צירידָ, ניו יורק ורומא. נערכים שני סוגי ניסויים - נקודות מבט חדשות, ושחזור תלת מימד. בראשון 20% מתמונות הקלט נבחרות באקראי ומוחסרות

מהאימון, כשהן משמשות כסט בדיקה, ואילו בשני בוחרים ידנית מבנה ומסירים את כל דגימות הלידר המסתיימות בו, והן משמשות כסט הבדיקה, כדי להעריך את איכות שחזור המשטח התלת מימדי.

מבחינת אלגוריתמים אליהם משווים - NeRF, Mip-NeRF, NeRF-W, שכבר נזכרו בעבודה זו, משתמשים כקלט בתמונות בלבד, וכן משווים ל- DS-NeRF [24]- וריאנט נרף היודע לקבל נקודות תלת מימדיות ששוערכו בעזרת SFM. נערכה התאמה כדי לאפשר לו לקבל נקודות לידר (זהות לאלה שבהן URF משתמש).

נקודות מבט חדשות

איור 34: השוואת תוצאות איכותית מול אלגוריתמים שונים. מתוך [5]

באיור 34 מוצגת השוואת תוצאות איכותית מול אלגוריתמים שונים. URF מדייק יותר, ולא סובל מאובייקטים מרחפים או מפגמים שמקורם בשגיאות שיערוך חשיפה

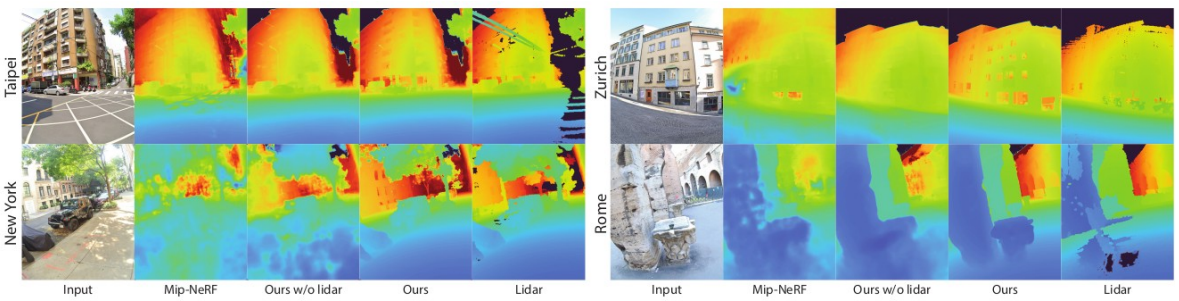
	Lidar	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [42]	☐	14.791	0.477	0.569
NeRF-W [40]	☐	17.156	0.455	0.620
Mip-NeRF [4]	☐	16.987	0.516	0.458
DS-NeRF [30]	☑	15.178	0.500	0.537
Ours w/o lidar	☐	19.638	0.541	0.447
Ours	☑	**20.421**	**0.563**	**0.409**

טבלה 6: השוואה כמותית של תוצאות שחזור מבטים חדשים (נקודות מבט מוחסרות). מתוך [5]

בסקירת התוצאות - mip-NeRF לא מתמודד טוב עם שינויי חשיפה בקלט האימון, ומייצר ארטיפקטים בניסיון להסביר תמונות עם נתוני חשיפה שונים. כל שיטות ההשוואה סבלו בגלל מיעוט המבטים הזמינים ושינויי החשיפה שבקלט. URF מציג תוצאות טובות הן משל NeRF-W, שתוכנן לסצנות חיצוניות ופתוחות, והן מ-DS-NeRF, שמסוגל לנצל את דגימות הלידר בתהליך הלמידה שלו. טבלה 6 מציגה ציונים במספר מדדי דמיון שונים בין האלגוריתמים השונים המשתתפים בהשוואה.

איכות שיחזור תלת מימד

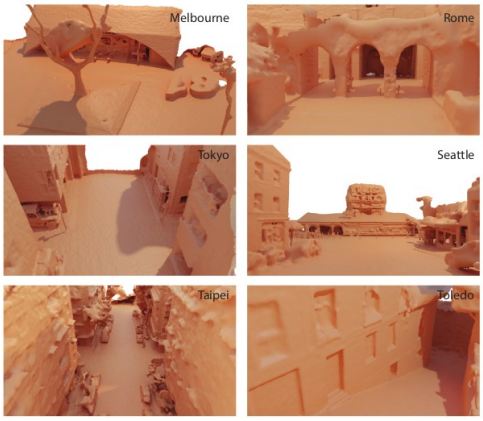
איכות שחזור תלת מימד נבחנת ב-2 מישורים: הערכת מפות עומקים ממיקום נתון, וכן נכונות ענני נקודות. את דיוק מפות העומק ניתן לבחון גם על ידי בחינת הסריגים (meshes) שמיוצרים מהם



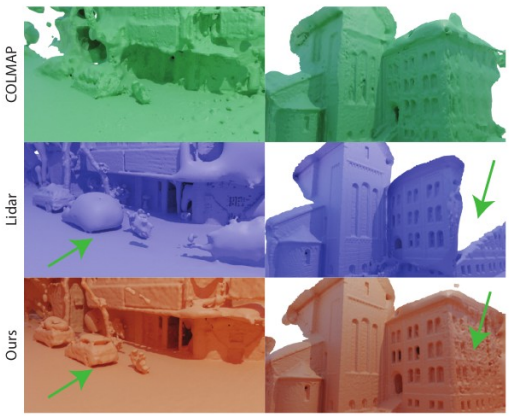
איור 35: דוגמאות איכותיות של מפות עומקים מ-4 סצנות. מתוך [5]

באיור 35 ניתן לראות דוגמאות איכותיות של מפות עומקים מ-4 סצנות, על בסיס Mip-NeRF, URF ללא לידר, URF מלא, ולידר לבדו כ- GT.

באיור 35 ניתן לראות ש- URF מחזיר תוצאה מפורטת יותר גם מ-Mip-NeRF וגם מהלידר לבדו. הדבר בולט בשחזור הפתחים בבניינים (חלונות ומרפסות), וכן בשחזור הרווחים שמסביב לראש העמוד הקורינתי בסצנת רומא. בטאיפיי מעניין לראות את שגיאות הכיסוי שיש בלידר כתוצאה מכבל חשמל נמוך בסצנה, בעוד URF משלימים את חזית המבנה בפירוט גבוה.



איור 37: דוגמאות נוספות לחילוץ מודל 3D (mesh) מסצנות אורבניות רחבות ידיים. מתוך [5]



איור 36: הצגה של שחזור מודל תלת מימדי על בסיס שיטות שונות. מתוך [5]

באיור 36 ניתן לראות שחזור רועש מאוד מ-MVS (כנראה עקב מיעוט מבטים) (Colmap - שורה עליונה), שגיאות חלקות כתוצאה ממשטחים שקופים או מבהיקים ב-Lidar (שורה אמצעית), ושחזור מלא יותר ומדוייק תחת URF בשורה תחתונה.

		Held-out Viewpoints				Held-out Building			
	Lidar	Avg Error (m) ↓	Acc↑	CD↓	F↑	Avg Error (m) ↓	Acc↑	CD↓	F↑
NeRF [42]	☐	1.582	0.264	3.045	0.528	1.423	0.274	2.857	0.535
NeRF-W [40]	☐	3.663	0.144	6.165	0.372	1.348	0.207	4.054	0.552
Mip-NeRF [4]	☐	1.596	0.133	2.812	0.363	1.417	0.132	2.508	0.427
DS-NeRF [30]	☑	1.502	0.259	2.571	0.526	1.367	0.294	2.720	0.558
Ours	☑	**0.463**	**0.742**	**0.272**	**0.880**	**0.770**	**0.363**	**2.312**	**0.687**

טבלה 7: הערכה כמותית לאיכות שיחזור. מתוך [5]

בטבלה 7 דנים בהשוואה בין גישות NeRF שונות, הנמדדות בשני מתארים: מבטים חסרים ומבנים חסרים. Ours כמובן מייצג את URF. התוצאות כוללות 2 מטריקות איכות שיערוך עומק - שגיאה ממוצעת ודיוק - ושתי מטריקות איכות ענן נקודות - מרחק שמפר (CD) ו-F-Score. הטבלה מציגה יתרון ברור לURF על פני האלגוריתמים המתחרים בכל המטריקות ובשני המתארים, כולל מתחרה המסוגל לנצל ענן נקודות (DS-NeRF).

(להסבר מפורט על מטריקות CD ו-F_1-score- ראו נספח א'-מושגים)

סיכום

URF מציגים גישה חדשה להרחבת NeRF לשחזור תלת מימד ונקודות מבט חדשות בסביבה אורבנית, תוך שימוש במעט תמונות מחד וניצול מידע Lidar מאידך. איכות ודיוק המשטח התלת מימדי המשוחזר טוב בהרבה מהמושג בטכנולוגיות אחרות או בגישות אחרות לNeRF. איכות המשטחים המשוחזרים טובה גם יחסית ל-NeRF עם יותר מבטים, ומאגר המידע שעליו הוכחה היכולת מכסה אזורים אורבניים נרחבים ברחבי העולם.

מצד שני - כמו ב- Bungee-NeRF - לא הוצגה התמודדות אמיתית עם כיסוי שטחים נרחבים מיכולת ההכלה של מודל בודד - כגון תפירה של URF-ים שכנים, והסצנות המוצגות הן רק תת-סט מהאיסופים שהיו זמינים. ההתבססות על פיתרון SFM קודם של המיקומים גרר לעיתים שימוש במיקומים רועשים יותר מהצפוי. כמו Bungee-NeRF, גם פה ההתמקדות הייתה בפן ספציפי של בעיית שחזור סצנות חיצוניות לא תחומות (outdoors), בדגש על סצנות אורבניות. לבסוף, עיקר הקלט - מאגר trakker של Google

Street View - לא זמין בקלות לקהל הרחב, והיה זמין לחוקרים רק בעקבות פנייה אישית.

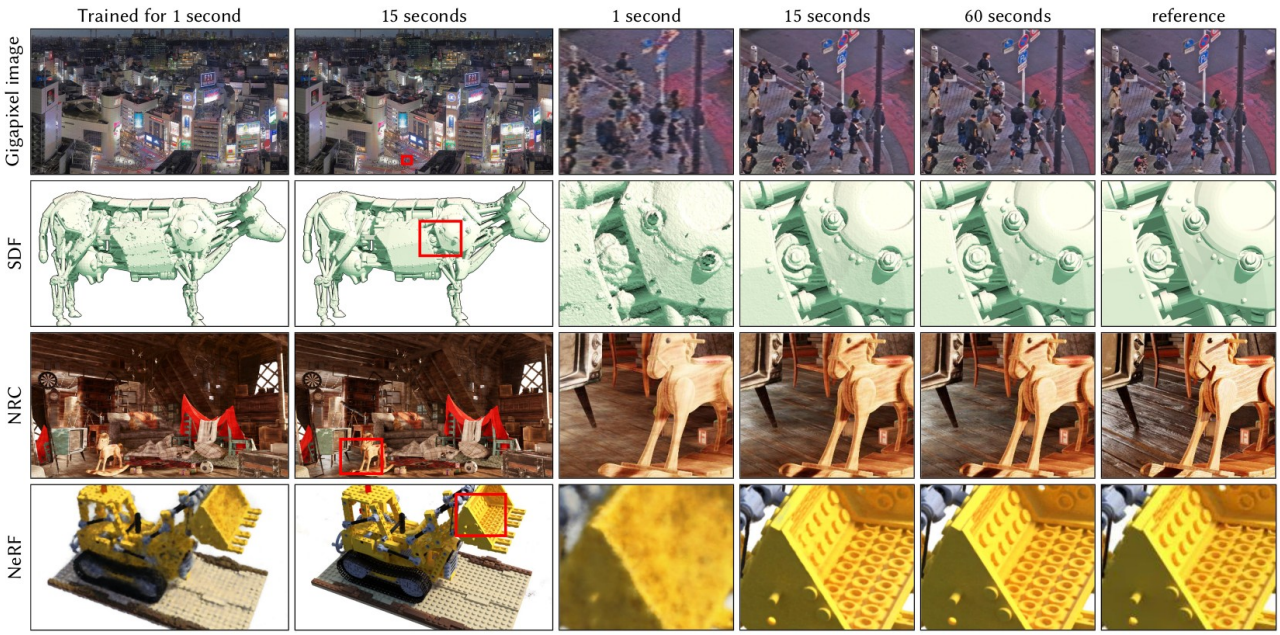
מידע נוסף

אתר הפרוייקט נמצא ב- https://urban-radiance-fields.github.io/. מכיל הפניות למאמר, מספר סרטונים והצגת יכולות. מידע נלווה למאמר פורסם באתר המאמר ב-CVPR 2022 [5] תחת "supp" ומציג פרטים נוספים כגון סכמת מבנה הרשת, תוצאות נוספות ממאגרי מידע אחרים ודוגמאות נוספות. לא פורסם קוד מלווה.

5 https://openaccess.thecvf.com/content/CVPR2022/html/Rematas_Urban_Radiance_Fields_CVPR_2022_paper.html

בנוסף לפתרונות המוצהרים, הבאים לשחזר מבטים חדשים וייצוג תלת מימדי בסצנות גדולות ו/או עירוניות, יש מקום לסקור שיפורים כלליים למודל ה-NeRF, שלא מציגים את גודל הסצנה או אופייה כמטרה, אבל מספקים הרחבות או פתרונות משלימים לשיטות שהוצגו, וראוי לדון בהם.

במאמר מאת Muller *et al.* [2] אמנם לא מוזכרות באופן ישיר סצנות גדולות או אורבניות, אבל הוא מציג שיפור דרמתי בביצועים, הן מבחינת זמנים והן מבחינת איכות גאומטרית. מוצגת בו שיטת קידוד שונה מ- PE או IPE שהוזכרו בעבר, המבוססת על קידוד בטבלאות גיבוב מרובות רמות פירוט (Multiresolution hash tables). גישה זו מאפשרת האצה משמעותית בביצועי הרשת - אימון וחיזוי. במאמר מוצגים 4 שימושים לאותו קידוד מוצע בפרימיטיבים גרפיים שונים: ייצוג תמונת גדולה (ג'יגה-פיקסל), ייצוג משטח תלת מימדי כ- SDF, ניהול שדה קירון (Neural Radiance Caching) לצרכי שליטה בתאורה ושינוי הצללה באופן מהיר, ואחרון חביב NeRF - החלפת קידוד ה-PE בייצוג מפות גיבוב רב מימדיות . פרימיטיבים גאומטריים אלה כולם ניתנים למימוש בעזרת MLP ופונקציית מחיר מתאימה, וכפי שכבר הזכרנו לגבי NeRF-ים, ייצוג הקלט במרחב רב מימדי משפר את יכולת הרשת לייצג פרטים לוקליים. בנוסף לקידוד פרמטרי דליל, החבילה מציעה שיפור נוסף למימוש NeRF - ניהול גריד איכלוס גס, המאפשר לשפר משמעותית את מהירות האימון על ידי סינון דגימות קרניים ממרחבים "ריקים"



איור 38: Instant-NGP- הדגמת יכולות. מתוך [2]

באיור 38 מוצגת הדגמת יכולות בארבעה תחומים: זמן אימון מול איכות תוצאה, בתמונה כוללת ובתקריבים.
אימון ב-NeRF ו-MipNeRF על סצנות מקבילות נמדד בשעות, ריינדור בדקות פר תמונה.
תמונת ג'יגהפיקסל של טוקיו ©(CC BY-NC-ND 2.0) Trevor Dobson, מודל הלגו של הבולדוזר©(CC BY-NC 2.0) Håvard Dalen.

קידודים - קידוד תדר וקידוד פרמטרי צפוף

כפי שכבר הוזכר, ב-NeRF [1.1] וב-MipNeRF [11] הוצגו קידודים מבוססי תדר (מתארי פורייה) שאיפשרו לייצג את הקלט במרחב גבוה-מימד ולהכיל אותו במשקולות ה-MLP.

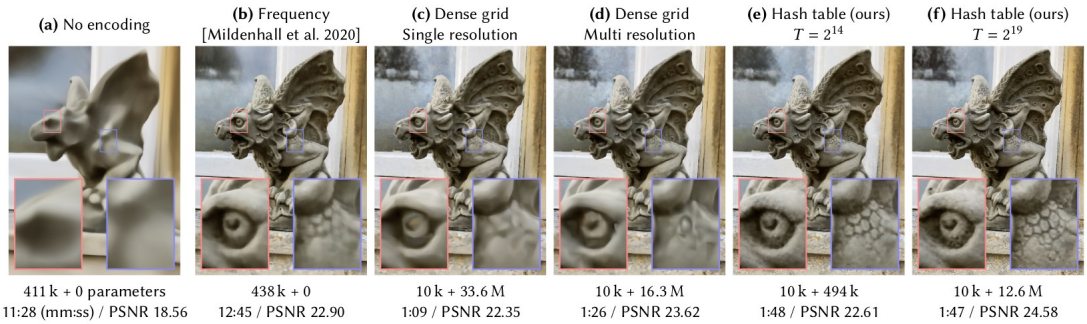
במקביל לפיתוח NeRF, בתחום קרוב - ייצוג נוירוני של משטחים תלת מימדיים (SDF כדוגמא, למרות שזהו תחום נרחב בפני עצמו), פותחו שיטות של קידוד פרמטרי - מבנה נתונים חיצוני ל-MLP, שמאפשר תשאול - ובמידת הצורך אינטרפולציה - של פרמטרים נלמדים נוספים, מעבר למשקולות וערכי ה-bias. פיתרון כזה ממיר את הסיבוכיות החישובית - עבור כל נגזרת המחלחלת לאחור ברשת, יש לעדכן את כל משקולות הרשת - בחתימת גודל זיכרון גדולה משמעותית בעת האימון, שבה עידכונים לוקליים משפיעים על מעט משתנים. בגריד תלת מימדי של מתארים, המתוכנן לאינטרפולציה טרי-לינארית, עדיין רק 8 קודקודים בגריד נדרשים לעדכון עבור כל דוגמה. באופן זה, למרות שמספר הפרמטרים הכללי בקידוד פרמטרי גדול בהרבה מקידוד מבוסס תדר, מבחינת מספר פעולות חישוב וגישות זיכרון במהלך אימון, השינוי לא גדול. בעקבות

הקטנת ה-MLP המייצג את הסצנה, זמן האימון עד התכנסות יכול להתקצר משמעותית, בלי אובדן איכות השיערוך.

לייצוג פרמטרי צפוף שכזה יש חסרונות - הרזולוציה המירבית נקבעת מצפיפות הגריד המוגדר, וכמות הזיכרון המוקצה זהה עבור ווקסל באיזור פתוח ולווקסל קרוב למשטח. בעוד מספר הפרמטרים גדל ב - $O(n^3)$, פני המשטח המיוצג גדלים רק ב- $o(n^2)$. כבר בגריד בגודל 128^3 כמות הווקסלים שבאמת נוגעים במשטח היא פחות מ-3%. בנוסף, סצנות טבעיות מציגות חלקות שמעודדת שימוש בפירוק לרזולוציות שונות, שכמובן מחמירות את בעיית חתימת הזיכרון שהוצגה. במקרה של SDF - שבו המשטח הכללי ידוע מראש - קיימים גם פיתרונות מבוססי עצים במקום גרידים, שמאפשר דילול תתי עצים מנוונים, או ייצוג בגריד דליל. ב-NeRF כמובן פיתרון זה בעייתי, מאחר והמשטח מחושב כחלק מתהליך האימון, ולא ידוע בתחילתו.

קידוד פרמטרי דליל

Instant-NGP מציעים שימוש בקידוד גמיש ויעיל על בסיס היררכיה של טבלאות גיבוב רב מימדיות, המוגדרים בעזרת 2 היפר פרמטרים בסך הכל - T ו- N_{max}. T הוא מספר הפרמטרים הכללי, ואילו N_{max} מגדיר את הרזולוציה המקסימלית הרצויה. ווקטורי המתאר הניתנים לאימון מאוחסנים בסדרת טבלאות גיבוב מרחביות המייצגות ריבוי רזולוציות.

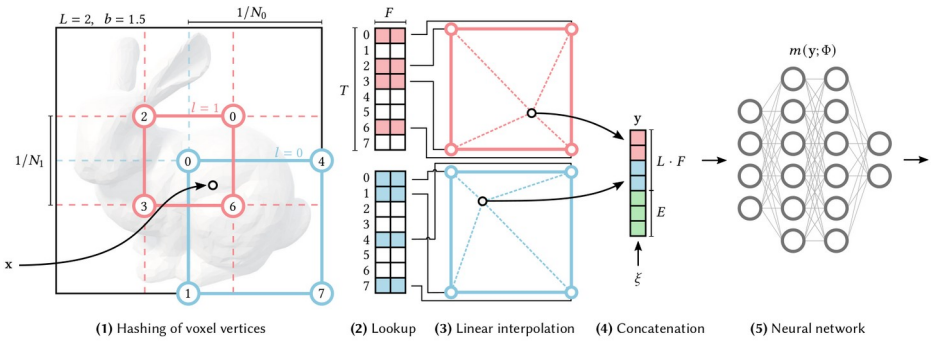


איור 39:הדגמת השפעת קידודים שונים על Instant-NGP NeRF. מתוך [2]

הדגמת השפעת קידודים שונים על ביצועי NeRF במימוש מואץ על ידי טבלת איכלוס מוצגת באיור 39. מתחת כל תמונה מצויין כמות הפרמטרים המאומנים (MLP+קידוד פרמטרי מאומן), זמן אימון ואיכות שיחזור (PSNR). משמאל לימין: ללא קידוד מיקום כלל, קידוד תדר (PE), גריד פרמטרי צפוף לרזולוציה בודדת, גריד פרמטרי צפוף מרובה רזולוציות, טבלת גיבוב עם $T=2^{14}$ פרמטרים (INGP), טבלת גיבוב עם $T=2^{19}$ פרמטרים (INGP). כל המודלים אומנו במשך 11K צעדים, עם שינוי רק בגודל ה-MLP וכמות הפרמטרים. דוגמה (e) סיימה אימון פי 8 מהר יותר מ-(b), למרות כמות דומה של משתנים מאומנים, בגלל דלילות הצורך בגישה לפרמטרים לצורך עדכון. הגדלת מספר הפרמטרים ב-(f) מאפשרת שיפור בציון ללא תוספת משמעותית בזמן חישוב.

ברזולוציות הגסות יש מיפוי ישיר 1:1 בין דגימות לבין וקטורי המתאר. הרזולוציות מפורטות יותר מתייחסים לערך בטבלה כאל קוד גיבוב (hash) כך שמספר דגימות מתייחסות לאותו ערך. התנגשויות גיבוב כאלה גורמות לגרדיאנטים המתנגשים להתמצע, באופן שהגרדיאנטים הגדולים ביותר - שהם אלה שהכי משמעותיים לפונקציית המחיר - הם אלה שישפיעו. תכונה זו מביאה לכך שטבלאות הגיבוב מתעדפות באופן אוטומטי את האזורים הדלילים המכילים את המידע החשוב של רמות הפירוט הגבוהות. להבדיל מבעבודות אחרות, באופן זה גם לא נדרשת התאמה או שינוי מבנה למבנה הנתונים במהלך האימון.

גישת גיבוב זו גם מאוד יעילה מבחינת שימוש על GPU - חיפושים בטבלאות הם ב- $O(1)$, ואין צורך ברדיפה אחרי סדרות פויינטרים, דבר שמאפיין התנהלות בעצים ומאוד לא מומלץ ב-GPU-ים. בנוסף, תחקור רמות שונות של טבלאות הגיבוב ניתן לביצוע במקביל ובאופן עצמאי.



איור 40: אילוסטרציה של תהליך הגיבוב בדו מימד. מתוך [2]

בהנתן MLP כלשהו $m(y;\Phi)$, כש- Φ הם משקולות הרשת ו-y הוא קידוד הקלט $y=enc(x,\theta)$, נגדיר את θ כפרמטרי קידוד ברי אימון, המשמשים לקידוד x. θ מאורגנים ב- L רמות, שבכל אחד עד T ווקטורי מתאר ממימד F

Parameter	Symbol	Value
Number of levels	L	16
Max. entries per level (hash table size)	T	2^{14} to 2^{24}
Number of feature dimensions per entry	F	2
Coarsest resolution	$N_{\min}$	16
Finest resolution	$N_{\max}$	512 to 524288

טבלה 8:פרמטרי קידוד וטווחים טיפוסיים. מתוך [2]

בטבלה 8 מוצגים ערכי פרמטרי הקידוד והטווחים התואמים הטיפוסיים. רק T ו- N_{max} דורשים התאמה בין המשימות השונות.

כל רמה $\lfloor$ היא בלתי תלויה, כשווקטורי המתאר מקודדים בקודקודי הגריד. רזולוציית הגריד היא סדרה הנדסית בין רזולוציית המינימום N_{min} ורזולוציית המקסימום N_{max} כך ש-

$$N_l := \left\lfloor N_{min} \cdot b^l \right\rfloor \quad [2](24)$$

$$b := \exp\left(\frac{\ln N_{max} - \ln N_{min}}{L-1}\right) \quad [2](25)$$

N_{max} נבחר על בסיס הפירוט הגבוה ביותר הנדרש בסצנה, ומאחר ו-L יחסית גבוה (ראו טבלה 8) פקטור הגידול בין רמות עוקבות יחסית קטן. במקרה המוצג $b \in [1.26, 4]$.

פונקציית הגיבוב המרחבית $h: \mathbb{Z}^d \to \mathbb{Z}_T$ בכל שכבה היא

$$h(x) := \exp\left(\bigoplus_{i=1}^{d} x_i \pi_i\right) \bmod T \quad [2](26)$$

כש-d הוא מימד הקלט, $\oplus$ מסמל bitwise-XOR ו- π_i הם מספרים ראשוניים גדולים וספציפיים שנבחרו.
מעשית, פונקציה (26) מבצעת xor על תוצאת פרמוטציה של קונגרואציה לינארית (פסאודו רנדומלית) [34] של כל מימד, ובכך מבצעת הפרדת צימוד (de-corrolation) בין ערכי ה=hash והמימד ממנו חולצו. ספציפית במאמר [2] הערכים הנבחרים של π_i הם π_1=1 , π_2=2,654,435,761 ו- π_3=805,459,861, מאחר ולמעשה, כדי להגיע להפרדה מספקת של d מימדים מספיק לקודד רק d-1 מהם.

על מנת לחשב את הפרמטר המשקל w_l, בהנתן קלט $x \in \mathbb{R}^d$ בוחרים בכל שכבה באופן בלתי תלוי את הווקסל המקיף את x, מתאימים את x לסקאלה של השכבה v_l ומבצעים אינטרפולציה d-לינארית על ווקטורי המתאר שבפינות הווקסל בהתאם למרחק היחסי של x מכל פינה, כך שמקדם האינטרפולציה הוא

$$w_l := x_l - \lfloor x_l \rfloor \quad [2](27)$$

תהליך זה מבוצע באופן עצמאי עבור כל אחת מ-L שכבות הפירוט, ואל תוצרי ווקטורי המתאר המשוקללים משרשרים לסופם קלט עזר נוסף ξ בהתאם ליישום (למשל כיוון הסתכלות ב-NeRF או כיוון הסתכלות מקודד וטקסטורות ב-NRC). תוצרי כל שכבה (כולל קלט העזר ξ) משורשים בתורם ומהווים את $y\in\mathbb{R}^{LF+E}$ שהוא הקלט המקודד ל-MLP: $\mathrm{enc}(x;\theta)$.$m(Y;\Phi)$

חשוב לציין, כי צורת קידוד זו, כפי שמוסבר יותר לעומק ב-[7],מייצרת בעיה בגזירה של גרדיאנטים אנליטיים לפי קידוד הגיבוב הנ"ל. בהנתן הגדרה (27) , ובהנתן שערכי העיגול $\lfloor x_{i,l}\rfloor,\lceil x_{i,l}\rceil$ מייצגים את פינות התא בגריד (ברזולוציה שונה בהתאם לשכבת הפירוט), ניתן לייצג את וקטור המתאר המגובב כ:

$$Y(x_{i,l})=Y(\lfloor x_{i,l}\rfloor)\cdot(1-w_l)+Y_l(\lceil x_{i,l}\rceil)\cdot w_l \quad [2](28)$$.

בגלל המיקומים המעוגלים $\lfloor\cdot\rfloor,\lceil\cdot\rceil$ (שאינם גזירים) , גרדיאנטים אינם רציפים במעבר בין תאים - הקידוד ישתנה באופן לא חלק במעבר בין תא לתא. כפי שיוצג בהמשך בסעיף 4.6, מאמר [7] מציג פיתרון גם למגבלה זו.

אילוסטרציה לדו מימד ו2 רמות פירוט ניתן לראות באיור 40.

הקידוד הפרמטרי הנל איננו תלוי משימה, והמשימות השונות ש-Instant-NGP מציג חולקות את אותם היפר-פרמטרים ואותו מימוש, עם שינוי רק בגודל טבלאות הגיבוב, לצורך איזון בין ביצועים נדרשים לאיכות. הקידוד מקונפג על ידי שני פרמטרים בלבד - מספר הפרמטרים T ורמת הפירוט המקסימלית הרצויה N_{max}, ומאפשר תוצאות באיכות גבוהה במהירות גבוהה בהרבה. ספציפית ל-NeRF, מדובר בהאצה של 20-60X.

NeRF ב- Instant-NGP

מימוש ה-NeRF בחבילה עבר שני שיפורים: הראשון הוא הקידוד הפרמטרי הדליל, שמאפשר הקטנה משמעותית של ה-MLP המאומן, והשני הוא ניהול גריד איכלוס גס, המאפשר בעת אימון פריסת דגימות קרניים בפיזור לא אחיד, ומוכוון איזורי עניין. בטבלה 9 אפשר לראות את תוצאות השוואה על אותן סצנות בין NeRF, Mip-NeRF ו-NSVF ומימושים של ה-INGP-NeRF עם גריד האיכלוס. כדי לספק השוואה מלאה מוצגות תוצאות של INGP-NeRF עם קידוד תדר (PE) ועם קידוד הגיבוב (Hash), מה שמאפשר השוואה להשפעת הקידוד על זמנים וביצועים. גריד האיכלוס עצמו עשוי להיות בעל מספר רמות, במאמר בסצנות הסינטטיות נבחרה רמה בודדת, בעוד בסצנות גדולות וטבעיות נבחרו 5 רמות פירוט לגריד האיכלוס. גריד האיכלוס מיוצג כשדה ביטים לשם חיסכון בנפח.

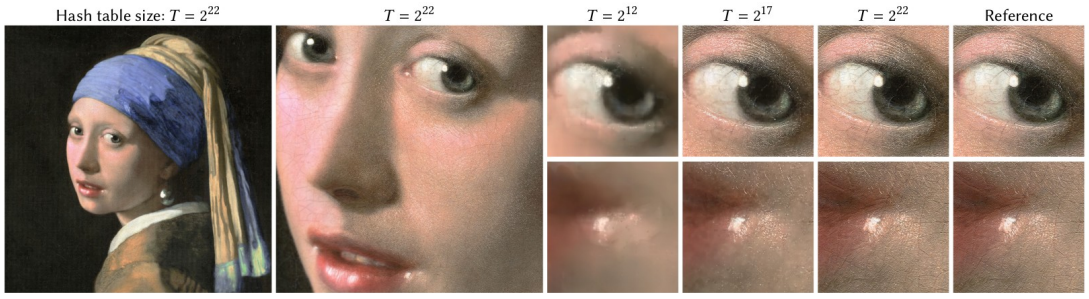
	Mic	Ficus	Chair	Hotdog	Materials	Drums	Ship	Lego	avg.
Ours: Hash (1 s)	26.09	21.30	21.55	21.63	22.07	17.76	20.38	18.83	21.202
Ours: Hash (5 s)	32.60	30.35	30.77	33.42	26.60	23.84	26.38	30.13	29.261
Ours: Hash (15 s)	34.76	32.26	32.95	35.56	28.25	25.23	28.56	33.68	31.407
Ours: Hash (1 min)	35.92	33.05	34.34	36.78	29.33	25.82	30.20	35.63	32.635
Ours: Hash (5 min)	36.22	33.51	35.00	37.40	29.78	26.02	31.10	36.39	33.176
mip-NeRF (~hours)	36.51	33.29	35.14	37.48	30.71	25.48	30.41	35.70	33.090
NSVF (~hours)	34.27	31.23	33.19	37.14	32.68	25.18	27.93	32.29	31.739
NeRF (~hours)	32.91	30.13	33.00	36.18	29.62	25.01	28.65	32.54	31.005
Ours: Frequency (5 min)	31.89	28.74	31.02	34.86	28.93	24.18	28.06	32.77	30.056
Ours: Frequency (1 min)	26.62	24.72	28.51	32.61	26.36	21.33	24.32	28.88	26.669

טבלה 9: ערכי PSNR לסצנות סינטטיות ששימשו במאמרי NeRF, Mip-NeRF, NSVF לשם השוואה. זמנים וערכי PSNR של אלגי אלה נלקח מהמאמרים עצמם ולא שוחזרו. מוצגות תוצאות על בסיס קידוד גיבוב אחרי שניה אחת עד 5 דקות, ועל בסיס קידוד PE אחרי דקה ו-5 דקות. בכל עמודה מדליות זהב, כסף וארד מציינות את מקום ראשון, שני ושלישי. מתוך [2]

מטבלה 9 ניתן לראות שהמימוש המהיר הוא בר השוואה מבחינת איכות ל-NeRF ול-NSVF לאחר 15 שניות בלבד, בעוד מול Mip-NeRF נדרשות 1-5 דקות כדי להגיע לאיכות קרובה או טובה יותר. בשלושת השיטות שמולן נערכת ההשוואה זמן האימון נמדד בשעות.

תמונת ג'יגה-פיקסל

אבן בסיס גרפית נוספת ש-Instant NGP מציג היא ייצוג של תמונות ענק (עד כמיליארד פיקסלים) באיכות ודיוק גבוהים גם בפרטים קטנים מאוד - כלומר רשת המקבלת קואורדינטה פיקסלית ב-2D ומחזירה ערך צבע RGB. שילוב הקידוד מבוסס הגיבוב מאפשר יכולת התאמה טבעית, שמאפשרת תוצאת PSNR אחרי אימון של דקות, שעד כה דרש שעות ועשרות שעות. באיור 38 (בתחילת תת הפרק) בשורה העליונה - תמונת טוקיו-אלגוריתם מתחרה בשם ACORN מאת Martel *et al.* (2021) הציג PSNR של 38.59 dB לאחר 36.9 שעות אימון. שימוש בקידוד טבלאות הגיבוב של INGP עם מספר דומה של פרמטרים ($T=2^{24}$) הציג את אותו PSNR לאחר 2.5 דקות, וציון מירבי של 41.9 dB לאחר 4 דקות.

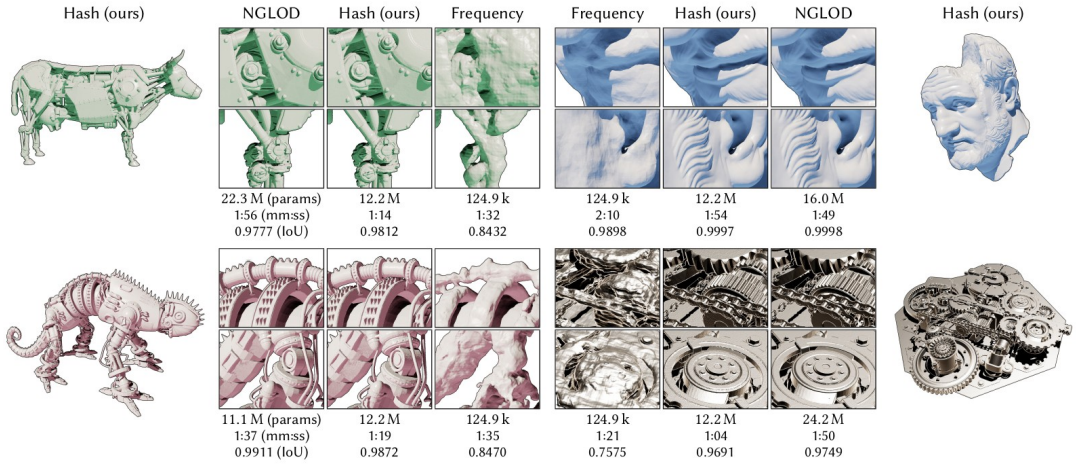


איור 41:הדגמת ייצוג תמונת ג'יגה פיקסל. מתוך [2]

ייצוג נוירוני של תמונת RGB בגודל 20,000x23,466 (469M פיקסלים), בעזרת קידוד גיבוב מרובה-רזולוציה. המודלים נבדלים בינהם רק בכמות הפרמטרים T. במודל המפורט ביותר ($T=2^{22}$) ציון ה-PSNR הוא 29.8 dB. התמונה היא חידוש של "נערה עם עגיל פנינה" © Koorosh Orooj (CC BY-SA 4.0).

SDF

INGP כולל מימוש מואץ ל-SDF, גם הוא על בסיס קידוד טבלאות הגיבוב. אלגוריתם ההשוואה הנבחר הוא NGLOD [25], המוגדר כ-state of the art, ובעצמו מנצל קידוד פרמטרי דינמי - ע"י שימוש בעץ אוקטלי העובר גיזום במהלך האימון. מאחר ומדובר בהשוואה של התאמה תלת ממדית, ולא נכונות פיקסלית של תמונה, המדד המשמש להשוואה הוא IoU. בהשוואה בין NGLOD ו- INGP-SDF (איור 42)ניתן לראות שבשימוש בכמות פרמטרים דומה או קטנה יותר, INGP-SDF משיג ציוני IoU קרובים בזמנים דומים, ולהבדיל מהקידוד הפרמטרי בעץ האוקטלי, המאפשר תחקור רק בסביבות המודל (בתוך איזור הכיסוי של תאי העץ), קידוד טבלאות הגיבוב מאפשר תחקור המודל מכל נקודה במרחב.



איור 42: השוואת יצוג מודל תלת מימדי על בסיס SDF. מתוך [2]

בהשוואה NGLOD, INGP-SDF (מסומן ב ours) ושימוש בקידוד תדר (PE) של NeRF קלסי. כל המודלים אומנו במשך 11,000 צעדים. מצויין מספר הפרמטרים המאומנים, זמן האימון וציון ה-IoU. Bearded Man ©Oliver Laric (CC BY-NC-SA 2.0).

סיכום

חבילת Instant-NGP מציעה 4 אבני בניין גרפיות על בסיס רשתות נוירונים וקידוד פרמטרי עם טבלאות גיבוב. לנושא העבודה הנוכחית - שחזור סצנות גדולות - רק מימוש

אלגוריתם ה-NeRF המוצע תורם ישירות. הפיתרון המוצע מציג ייצוג סקלבילי, רב מימדי, מבוסס טבלאות גיבוב, המאפשר הקטנת גודל ה-MLP מחד וייצוג בזיכרון שהוא לוג-לינארי ברזולוציה. הקטנת גודל ה- MLP וקיצור משמעותי מאוד בזמן האימון - עם חיזוי המאפשר רינדור בזמן אמת - מחד, והתמיכה בקידוד רזולוציות שונות בקלט מאידך, מאפשר לשקול אימון ב-NeRF בודד של סצנות גדולות משמעותית מאופי הדוגמאות הנפוץ עד כה. מעבר לכך, החלפת מימושי NeRF קיימים במימוש החדש יוכל לשפר מהותית ביצועים של אלגוריתמים שמייצרים תת מודולים לכיסוי הסצנה כגון Block-NeRF ו-Mega-NeRF מבלי לדרוש שינויים נרחבים.
אבן בניין נוספת - קידוד תמונות ענק (ג'יגה פיקסל) מציע שילוב ושימוש אפשרי בשחזור מקלט בקנה מידה שונה מהותית - כגון ב-Bungee-NeRF.
אבן בניין רלוונטית אחרונה היא מימוש ה- SDF ב-Instant NGP, שבעוד שאינו רלוונטי ישירות לנושא עבודה זו, מהווה חלק מהבסיס (יחד עם הקידוד הפרמטרי מבוסס הגיבוב) לאלגוריתם האחרון שייסקר בעבודה - Neurolangelo.

מידע נוסף

אתר הפרוייקט הוא ב-https://nvlabs.github.io/instant-ngp, מוצגים בו המאמר, סרטי הדגמה, מצגת וגישה לחבילת קוד. הקוד מבוסס חבילת tiny-cuda-nn (כלומר תלוי מעבדי Nvidia) ונגיש ב-https://github.com/nvlabs/instant-ngp

כזכור, מודל NeRF מורכב משתי רשתות MLP שמורצות בשירשור - הגאומטריה של הסצנה מיוצגת על ידי פונקציית אטימות (opacity), שלעיתים מתייחסים אליה כאל צפיפות מרחבית, בעוד ערך הצבע (RGB) מחושב בעזרת פלט האטימות ו- MLP קטן נוסף שמשורשר לאחר מכן [1, 1.1]. פונקציית אטימות זאת מצויינת כדי לייצג (או לנקות) אלמנטים שקופים למחצה בסצנה, כגון תנאי אטמוספירה (עשן, אובך) או רכיבים שקופים / שקופים למחצה, אבל מקשה על חילוץ של גאומטריה מדוייקת בפירוט גבוה, בעיקר מאחר והגישה הנוכחית מציעה שימוש בהגדרת סף ערכי לרמת האיטום הנסכמת לאורך קרן כדי לבחור את מיקום המשטח. מצד שני, כפי שהוצג בסקירה על Instant-NGP, SDF היא שיטה שמאפשרת תיאור גאומטריה מאוד מפורט, אבל קיימת הנחה נסתרת שכל הגאומטריה ידועה מראש - בעיקר עקב הצורך בייצוג פרמטרי של כל המרחב הווקסלי, ו"גיזום" הייצוג הזה בהתאם לגאומטריה המתוארת. נסיונות קודמים לייצג את הגאומטריה ב-NeRF כ-SDF דרשו בקרה של קלטים חיצוניים נוספים - לדוגמא ענן נקודות דליל מ-SFM, מה שגרם לאיכות השחזור להיות חסומה באיכות ה-SFM.

במאמר [7], Li, Muller *et al.* מציגים שיטה לשחזור סצנה מאיסוף מרובה מבטים, המבוססת על NeRF ועל קידוד טבלאות הגיבוב ותובנות נוספות מ-Instant-NGP. Neurolangelo משתמשת ב-SDF MLP (שממנו **מחושב** ערך אטימות) ואחריו MLP צבע -במקום MLP צפיפות נפחית ו-MLP צבע הנהוג ב-NeRF - כדי להפיק תמונות חדשות על ידי שימוש ברנדור נפחי מבוסס SDF, מתוך קלט של תמונות + מנח בלבד.

שני שיפורים נוספים שמאפשרים את הפיתרון המוצע הם שימוש בנגזרות מספריות (להבדיל מנגזרות אנליטיות), ושימוש בסכמת אופטימיזציה במהלך האימון, המאמנת מרמת פירוט גסה לעדינה.

רנדור נפחי מבוסס SDF

בהתבסס על עבודתם של wang *et al.* [26] על SDF-ים ניורונים, מוצעת גישה להמיר חיזוי צפיפות נפחית ב- NeRF על בסיס פונקציה לוגיסטית, על מנת לאפשר אופטימיזציה בעזרת רינדור נפחי. בהנתן נקודה תלת מימדית x_i וערך פונקצית SDF תואמת $f(x_i)$, ערך האטימות השקול α_i מחושב כ:

$$\alpha_i = max\left(\frac{\Phi_s(f(x_i)) - \Phi_s(f(x_{i+1}))}{\Phi_s(f(x_i))}, 0\right) \quad (29)_{[7]} \quad ([7])$$

כאשר Φ_s מייצג את פונקצית הסיגמויד.

קידוד פרמטרי רב מימדי מבוסס טבלאות גיבוב

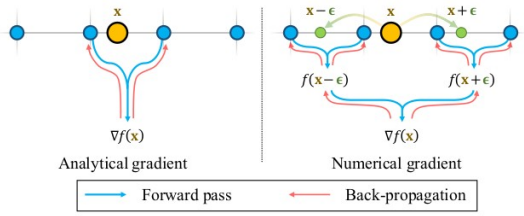
כפי שכבר הוצג בסקירה של Instant-NGP, קידוד פרמטרי רב מימדי מבוסס טבלאות גיבוב הציג יכולות סקלביליות טובות במספר משימות, ומאפשר קידוד של מידע גאומטרי מפורט מאוד. כפי שהוצג ב 4.5, ב- Instant-NGP הוא שימש הן לייצוג גאומטריות ידועות על בסיס SDF, והן לפיתרון חיזוי נקודות מבט חדשות על בסיס NeRF. ב-neurolangelo משתמשים בקידוד הנ"ל על מנת לשחזר משטחים תלת מימדיים ברמת פירוט גבוהה.

בכל רמת פירוט, כל פינה בכל ווקסל של הגריד ממופה לערך גיבוב, המצביע לווקטור מתאר של פינת הווקסל. בהנתן נקודה x, היא ממופה בנפרד לכל אחת מרמות הפירוט, שם ווקטור המתאר שלה מורכב מאינטרפולציה טרילינארית של ערכי הגיבוב השמורים בפינות הווקסל העוטף את x. ווקטור המתאר המלא של הנקודה בנוי משירשור כל המתארים מכל הרמות, והוא מוזן ל- MLP.

לשם ההבהרה, חלופה אפשרית לקידוד הגיבוב הוא מבנה ווקסלי דליל, המתאר את תלת המימד של מרחב המודל, וכך לכל פינה בגריד יש הגדרה עצמאית ויחידנית, ללא התנגשויות בגיבוב. אבל לייצוג זה שתי בעיות משמעותיות - תאור נפחי מלא שכזה דורש ייצוג היררכי כלשהו (כדוגמת עץ אוקטלי) על מנת שניתן יהיה לעקוב אחרי הנתונים שבו, אחרת, הזיכרון גדל במעלה שלישית יחסית לרזולוציה המרחבית. בהנתן היררכיה שכזו, רזולוציה ווקסלית גבוהה יותר לא תוכל לשחזר משטחים שיוצגו באופן שגוי על

ידי רמות גסות יותר בעץ. קידוד בעזרת מפות גיבוב לא מניח היררכיה מרחבית, ופותר התנגשויות בין רמות על בסיס מיצוע גרדיאנטים באימון.

שימוש בגרדיאנטים נומריים



איור 43: גרדיאנטים נומריים במקום אנליטיים. מתוך [7]
השימוש בחישוב גרדיאנטים נומריים במקום אנליטיים מפזר את עידכוני הגזירה לאחור אל מחוץ לתא המיידי בגריד שעוטף את x, והופך את החישוב להחלקה של התוצאה יחסית לנגזרות האנליטיות (דוגמא בדו מימד).

ל-SDF יש תכונה ייחודית שהיא גזירה עם נגזרת מנורמת היחידה, כלומר הגרדיאנט של ה SDF ממלא את משוואת אייקונל $\|\nabla f(x)\|_2=1$. כדי לשמור על אימון SDF תקין, נדרשת הוספת פונקציית מחיר אייקונלי

$$L_{eik}=\frac{1}{N}\sum_{i=1}^{N}\left(\|\nabla f(x_i)\|_2-1\right)^2 \quad [7](30)$$

כש-N הוא מספר הנקודות הדגומות הכולל.
לרוב ב-SDF, חישוב הנורמל למשטח $\nabla f(x)$ מחושב על בסיס נגזרות אנליטיות, אבל בשימוש עם קידוד גיבוב לגבי מיקום - עולה בעיה. אם נבחן את נוסחה (28) בסעיף 4.5, נראה שהקידוד איננו רציף במרחב לאחר האינטרפולציה הטרי-לינארית: אם x_i עובר לתא הסמוך, הקידוד בתא החדש - המתבסס על קידוד גיבוב שונה של פינות התא - יהיה שונה, בלי לשמור על רציפות יחסית לתא הקודם. נשים לב שעיגול המיקום $\lceil x_{i,l} \rceil$, $\lfloor x_{i,l} \rfloor$, המשמש להתאמת x_i לסקאלה של כל שכבה איננו גזיר, מה שיגרור פיעפוע לאחור של (30) ברמה מקומית בלבד על כל ווקסל בנפרד. לדוגמא בקירות חלקים, נורמל המשטח ייאומן בנפרד עבור דגימות בווקסלים סמוכים, וייצוג המשטח לא יהיה אחיד ופוטנציאלית רועש.

הפיתרון המוצע ב-[7] הוא שימוש בגרדיאנטים נומריים במקום אנליטיים. גודל צעד קטן מגודל קפיצה בגריד יביא לאותה תוצאה כמו גרדיאנט נומרי, אבל גודל צעד שמביא ליציאה מתא הגריד לתא סמוך יגרור השתתפות של תאים שכנים בתהליך חישוב הנורמלים למשטח והאופטימיזציה. דוגמה לפיתרון (בדו מימד) וליתרון של ניגזרת נומרית במקום אנליטית במהלך פיעפוע לאחור ניתן לראות באיור 43 בראשית הדף.

חישוב גרדיאנטים נומריים גורר צורך בדגימות נוספות מה-SDF לכל חישוב. בהנתן נקודה דגומה $x_i=(x_i,y_i,z_i)$, דוגמים עוד שתי נקודות לאורך כל אחד מהצירים בסביבת ϵ של x_i, כך שלדוגמא עבור x_i רכיב x של הנורמל למשטח יהיה:

$$\nabla_x f(x_i)=\frac{f(\gamma(x_i+\epsilon_x))-f(\gamma(x_i-\epsilon_x))}{2\epsilon}, \quad \epsilon_x=[\epsilon,0,0] \text{ כש-} \quad [7](31)$$

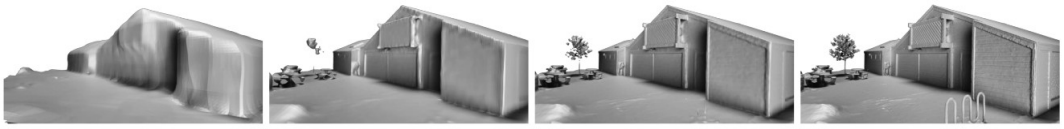
פיתרון זה דורש 6 דגימות SDF נוספות בסך הכל לכל נקודה, בתמורה לשחזור משטח פחות רועש ויותר מדוייק .

אימון הדרגתי של רמות פירוט

אופטימיזציה בשלבים מרמת פירוט גסה ועד עדינה מגדירה באופן איכותי יותר את פני השטח של פונקציית המחיר, ומאפשרת המנעות טובה יותר מעצירות במינימום לוקלי שגוי. זאת גישה מוכרת מהרבה פתרונות בראייה ממוחשבת, ובצורות שונות הוזכרו גם בעבודה זו - בתהליכי האימון של Bungee-NeRF (על ידי חלוקה לטווחי איסוף שונים, מהרחוק לקרוב) וכן ב-Mega-NeRF (ע"י טיוב שיוך פיקסלים לתת מודולים שונים במהלך התקדמות האימון).

שילוב הגישה ב-Neurolangelo מנצל את השימוש בגרדיאנטים נומריים, ומבצע אופטימיזציה הדרגתית על ידי שימוש בשני מנגנונים:

גודל צעד ϵ : כפי שהוזכר להלן, שימוש בגרדיאנטים נומריים מהווה בעצם תהליך החלקה של המשטח המשוחזר, כשגודל הצעד ϵ שולט על הרזולוציה וכמות הפרטים המשוחזרים. חישוב L_{eik} (30) עם ϵ גדול מבטיח חישוב נורמל למשטח שיהיה רציף ואחיד בקנה מידה גדול, בעוד ϵ קטן יותר ישפיע על אזור קטן יותר בגריד וימנע החלקה של פרטים. מעשית בתחילת האימון גודל הצעד ϵ מאותחל לגודל גריד הגיבוב הגס ביותר, ומוקטן באופן מעריכי במהלך האימון, כדי להתאים לשכבות פירוט שונות בגריד הגיבוב.



איור 44: תוצאות שיחזור איכותיות באימון שכבות פירוט שונות בגריד הגיבוב. מתוך [7]
חלק מהאלמנטים בסצנה לא שוחזרו בשכבות הגסות - העץ, השולחן ומעמד האופניים. שכבות מפורטות יותר יכולות להשלים את החסר בהדרגה. הדוגמא גם ממחישה את הצורך ברמת פירוט גבוהה גם למשטחים חלקים וגדולים, ולעדכונים הלא לוקליים שגרדיאנטים נומריים מביאים לנגזרות ממעלה גבוהה.
תמונת קלט מייצגת לסצנה ניתן לראות באיור 45.

רזולוצית גריד הגיבוב (V): הפעלה של כל רמות הפירוט בגריד הגיבוב בתחילת תהליך האימון ידרוש משכבות פירוט גבוה לשכוח מידע שנלמד בשלב אימון מוקדם (עם גודל צעד ϵ גדול) לפני שיוכלו ללמוד מחדש פרטים מדוייקים עם ϵ קטן. אם התהליך מתכנס לפני שהלמידה מחדש מתרחשת, פרטים עדינים יאבדו. לכן Neurolangelo מפעילים בתחילת תהליך האימון רק סט התחלתי של שכבות גסות בגריד, ומפעילים שכבות פירוט עדינות יותר במהלך האימון יחד עם הקטנת ϵ. באופן כזה נמנע הצורך ללמוד מחדש ושחזור פרטים עדינים משתפר. בנוסף, משתמשים בדעיכת משקולות על כל הפרמטרים, על מנת למנוע מפרטים ברזולוציה בודדת לשלוט בתוצאה הסופית.

קידוד מראה ותנאי סביבה

גם Neurolangelo, כמו Block-NeRF ו- Mega-NeRF שנסקרו לעיל, משתמשים בקידוד מראה ותנאי סביבה נלמד על בסיס המימוש ב- NeRF-W [12], על מנת לשחזר סצנות חיצוניות גדולות, שבהן תנאי תאורה והצללה שונים.

פונקציית מחיר קמירות

על מנת לעודד חלקות בשיחזור, מוספת פונקציית מחיר לקימור הממוצע של המשטח. זה מחושב מלפלסיאן דיסקרטי, בדומה לאופן שמחושב הנורמל למשטח בנוסחה (31). פנ' מחיר הקמירות מוגדר כ:

$$L_{curv}=\frac{1}{N}\sum_{i=1}^{N}|\nabla^2 f(x_i)| \quad (32)$$

הדגימות הנוספות מה-SDF שמבוצעות עבור חישוב L_{eik} (31) משמשות גם לחישוב (32)

מודל הרשת ופרטי מימוש

פונקציית המחיר הכוללת L לרשת היא סכום משקלים ממושקל, כך:

$$L=L_{RGB}+\omega_{eik}L_{eik}+\omega_{curv}L_{curv} \quad (33)$$

כש- L_{RGB} היא פונקציית מחיר הקירון, L_{eik} (30) היא פונקציית מחיר אייקונל, L_{curv} (32) היא פונקציית מחיר קמירות, ו-ω_{eik} ו-ω_{curv} הם מישקולי שתי פונקציות המחיר האחרונות, בהתאמה.

כל הפרמטרים, כולל ה MLP-ים ופרמטרי קידוד הגיבוב, מאומנים ביחד לאורך כל האימון.

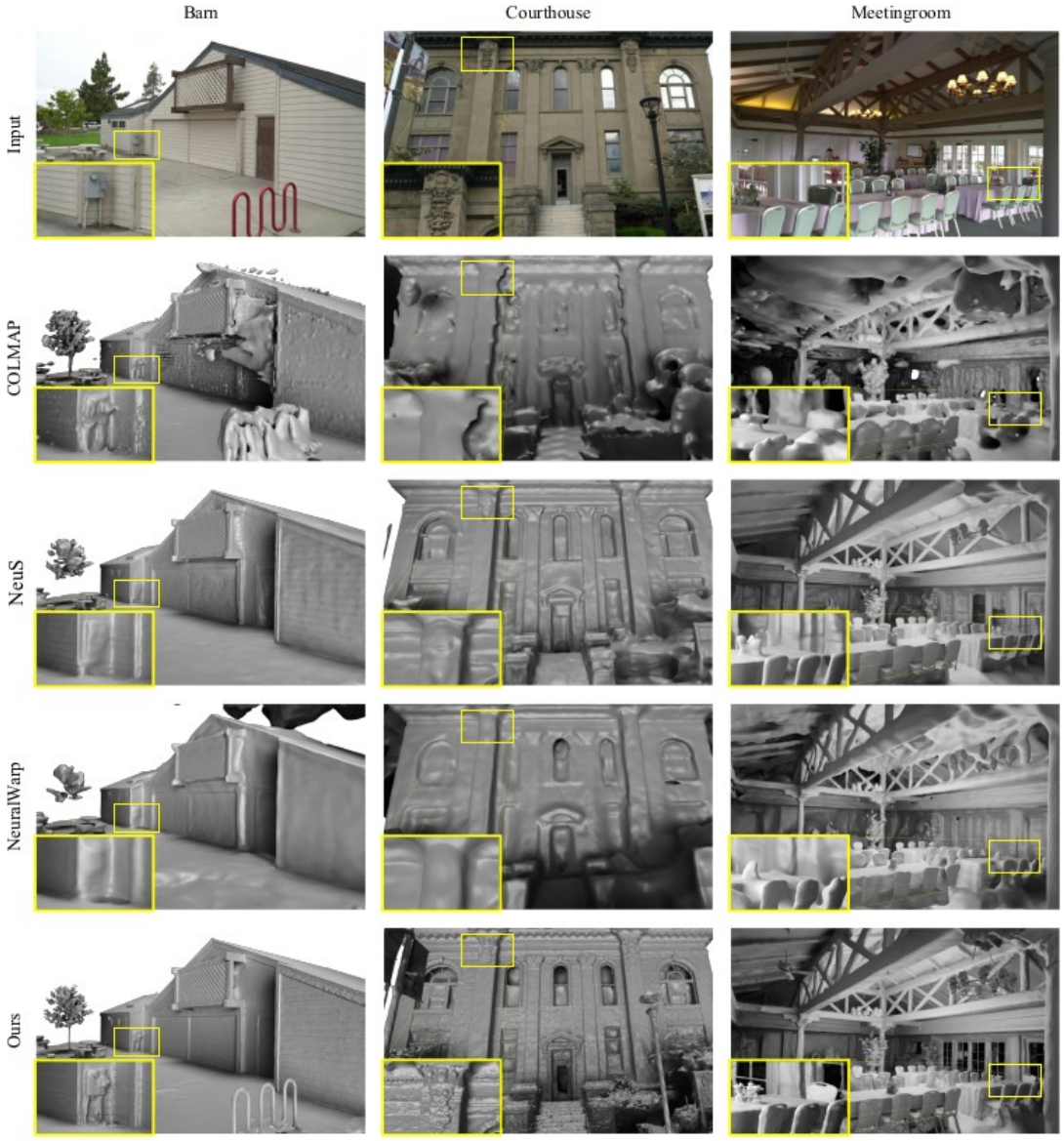
רזולוצית קידוד הגיבוב היא בין 2^5 ל-2^{11}, עם 16 רמות פירוט. כל ערך גיבוב בעל 8 ערוצים, וכל רמת פירוט יכולה לקודד עד 2^{22} ערכים. במהלך האימון מוכנסת לשימוש שכבת פירוט חדשה כל 5000 איטרציות, כשגודל הצעד ϵ שווה לגודל תא בשכבה.

מאגרי מידע

כדי להציג סצנות דומות לעבודות קודמות, בוצעו הרצות על מאגר DTU [28] -שבו כל סצנה בעלת 49 או 64 תמונות, ועל מאגר Tanks and Temples [27], כולל סצנות גדולות, שבהן 263-1107 תמונות לסצנה. בנוסף פורסמו בתוסף למאמר 2 סצנות גדולות שנאספו על בסיס רחפן - פארק מנהלת Nvidia ואוניברסיטת ג'ון הופקינס.

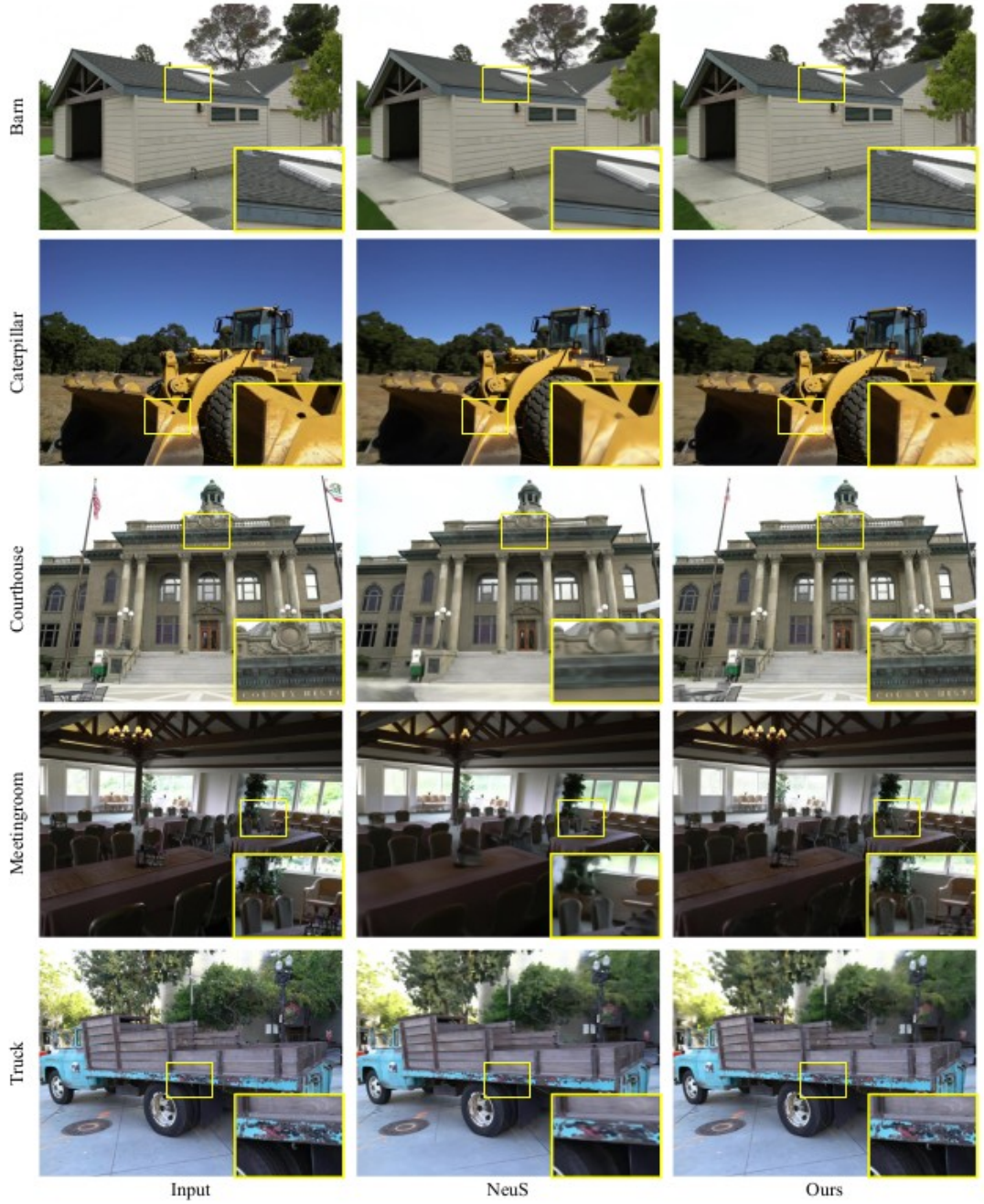
תוצאות ודוגמאות

התוצאות הוערכו על בסיס מספר מדדים: PSNR שימש להשוואות שחזור מבטים, ומטריקטות מרחק שמפר (chamfer distance) ו-F_1-score שימשו למדידת איכות הגאומטריה. (ראו נספח א' - מושגים, לפירוט על המדדים השונים)



איור 45: השוואה איכותית של שיחזור גאומטריה. מתוך [7]

השורה המוצגת באיור 45 כ- ours מייצג את Neurolangelo. תמונת הקלט של האסם (שמאל למעלה) מציגה את המבנה המוצג גם באיור 44.



איור 46: השוואה איכותית של שחזור מבטים מנקודות מבט חדשות. מתוך תוסף למאמר [7.1]

באיור 45 ניתן לראות השוואה איכותית של שחזור הגאומטריה. ניתן לראות את היתרון של Neurolangelo בהתמודדות עם נושא משטחים אחידים וחלקים מחד ועם פרטים מדוייקים ודקים מאידך (גזע העץ , שחזור קורות הברזל בגג, פרטי החיפוי על הקיר באסם ובגדר).

באיור 46 ניתן לראות השוואה איכותית של שחזור מבטים מנקודות מבט חדשות, על בסיס מאגר Tanks And Temples, כש-ours מייצג את התוצאות של Neuralangelo. ניתן לראות את השפעת שחזור הגאומטריה על איכות שחזור החוזי, ברמת הפירוט והחדות.

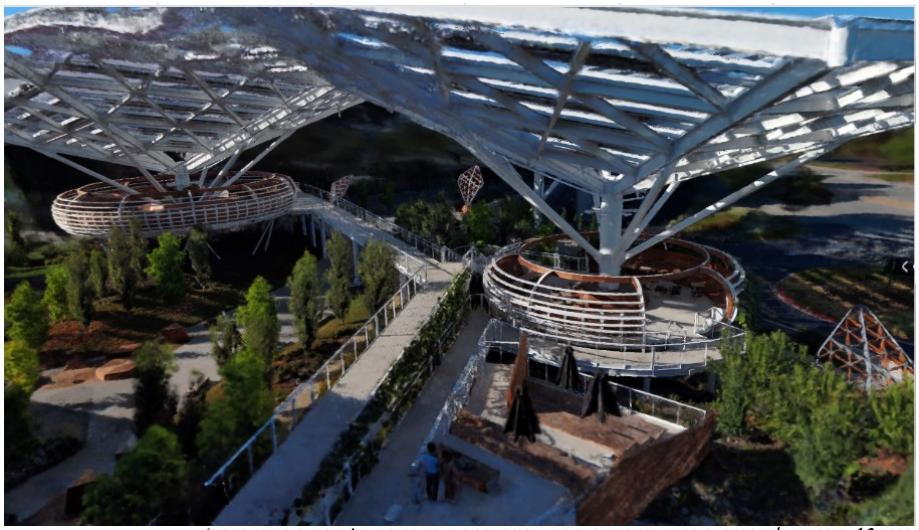
אמנם הסצנות ממאגר Tanks & Temples המופיעות במאמר [7] הן ממורכזות אובייקט והשטח הנאסף לאו דווקא מהווה תחרות מבחינת גודל הסצנה לנקודות העבודה במאמרים קודמים שנסקרו, אבל איכות השחזור ש- Neuralangelo מציגים בסצנות פתוחות וגדולות, כמו איור 46 ואף יותר מכך באיור 47, מייצג ומשמעותי לנושא עבודה זו, גם אם התמודדות עם סצנות כאלה לא הוגדרה כמטרה רשמית של האלגוריתם.

NVIDIA HQ Park

איור 47: דוגמאות משחזור סצנות NVidia וג'ון הופקינס . מתוך [7.1]

חוץ מזכוכיות והשתקפויות, צמחייה ורשתות/פסים דקים הם כנראה המתארים הקשים ביותר ל-MVS לשערך היטב.

בתוסף למאמר [7.1] פורסמו תוצאות נוספות, הכוללות שני איסופים עצמאיים - פארק מנהלת Nvidia ואוניברסיטת ג'ון הופקינס. בשני המקרים האיסוף היה על בסיס רחפן, ומיקום התמונות תוקן על ידי הרצת SFM (colmap). באיור 47ניתן לראות את גודל הסצנות ואיכות הגאומטריה המשוחזרת.



איור 48: דוגמא לשחזור תמונה מנקודת מבט. מתוך אתר הפרוייקט (ראה במידע נוסף)

מאחר ו-Neuralangelo בעיקרו מתרכז בשחזור המשטח, ופחות בשחזור חוזי ממבטים חדשים, לא נכללה במאמר או בתוסף דוגמא לשחזור מבט מסצנה חיצונית נרחבת. איור 48 נלקח מאתר הפרוייקט, ומציג איכות שחזור גבוהה. מבחינת שחזור קלסי (MVS) עצים וצמחייה, עמודים דקים וגדרות או רשתות הם מתאר קשה מאוד, ואילו Neuralangelo מפגין יכולת גבוהה.

	F1 Score ↑							PSNR ↑				
	NeuralWarp [3]	COLMAP [27]	NeuS [36]	AG	AG+P	NG	NG+P (Ours)	NeuS [36]	AG	AG+P	NG	NG+P (Ours)
Barn	0.22	0.55	0.29	0.22	0.31	**0.63**	**0.70**	26.36	**26.91**	26.69	26.14	**28.57**
Caterpillar	0.18	0.01	0.29	0.23	0.24	**0.30**	**0.36**	25.21	**26.04**	25.12	26.16	**27.81**
Courthouse	0.08	0.11	0.17	0.08	0.09	**0.24**	**0.28**	23.55	25.43	**25.63**	25.06	**27.23**
Ignatius	0.02	0.22	0.83	0.72	0.73	**0.85**	**0.89**	**23.27**	22.69	22.73	**23.78**	**23.67**
Meetingroom	0.08	0.19	0.24	0.04	0.05	**0.27**	**0.32**	25.38	**28.13**	28.05	27.44	**30.70**
Truck	0.35	0.19	**0.45**	0.33	0.37	0.44	**0.48**	23.71	23.89	**23.95**	22.99	**25.43**
Mean	0.15	0.21	0.38	0.27	0.30	**0.45**	**0.50**	24.58	**25.51**	25.36	25.26	**27.24**

טבלה 10: תוצאות כמותיות של מאגר Tanks and Temples. ב-COLMAP הכוונה לערוץ ה-MVS, המייצר mesh. AG=גרדיאנט אנליטי, NG=גרדיאנט נומרי, P=קידוד פרמטרי בטבלאות גיבוב, כך ש NG+P=גרדיאנט נומרי עם קידוד פרמטרי. מקום ראשון בירוק, מקום שני בכחול. מתוך [7]

בטבלה 10 מוצגת השוואה כמותית מול שני אלגוריתמים לומדים (NeuralWarp, NeuS) פיתרון MVS קלסי (Colmap), וכן צירופים שונים של מימושים על בסיס נגזרות אנליטיות (AG), נומריות (NG) ושימוש בקידוד הפרמטרי (P+).

מבחינת זמנים, הוצגו נתונים עבור סצנה בגודל 128^3 על Nvidia V100 GPU.
באימון - NG איטי למול AG בפקטור 1.2, למרות שעדיין מהיר מאלגוריתם ההשוואה NeuS. בחיזוי אין שימוש בנגזרות הנומריות ולכן הזמנים זהים. שני המימושים מהירים בחיזוי מ-NeuS עקב ה-MLP הקטן יותר שמאפשר הקידוד הפרמטרי.

	Training time (s)	Inference time (s)
NeuS [14]	0.16	0.19
NG (Ours)	0.12	0.08
AG	0.10	0.08

טבלה 11: זמני אימון ושיחזור על NVidia V100 וסצנה ברזולוציה של 1283. מתוך [7.1]

סיכום

Neuralangelo מראה שבעזרת שימוש בנגזרות נומריות במקום אנליטיות ברמות פירוט גבוהות, התבססות על קידוד פרמטרי ומשטר אימון הדרגתי מרמות פירוט נמוכות לגבוהות, ניתן להגיע לשחזור גאומטרי של סצנות גדולות בפירוט מדוייק להפליא מתוך תמונות RGB ומנחי מצלמה בלבד.

השימוש בקידוד הפרמטרי מבוסס טבלאות גיבוב מרובות רמות פירוט, שהוצג ב-Instant-NGP, מאפשר הגדלה של הסצנה המשוחזרת עקב הקטנת גודל ה-MLP.
היכולת להרחבה דינמית של מרחב הייצוג הפרמטרי מאפשר החלפה של ייצוג הגאומטריה במודל ה NeRF מפונקציית צפיפות מרחבית לפונקצית SDF.
ייצוג הגיאומטריה כ-SDF מאפשר שחזור גאומטריה מדוייק יותר, במחיר התייחסות פחותה לשחזור תנאי סביבה וראות - כגון אובך, ערפל או שליטה על תנאי תאורה.
זהו וויתור סביר מאוד מאחר ובהרבה נקודות עבודה של שחזור סצנות גדולות זו אינה יכולת שהשחזור נדרש אליו, בזמן שאיכות ייצוג הגיאומטריה הוא פרמטר המשפיע באופן משמעותי מאוד על איכות התוצר הנדרש, שכמעט בכל נקודת עבודה יוגדר כחשוב יותר.

מאחר ושחזור סצנות גדולות לא היה מטרה מרכזית מוצהרת, אין התייחסות בעבודה לסצנות שמעבר לגודל הנתמך במודל יחיד, ובדרכים לאחד מספר מודלי neuralangelo או איך לחלק סצנה למספר מודלים כאלה. שילוב של עבודה זו עם פתרונות כפי שהוצגו ב-Mega-NeRF או Block-NeRF נראים מבטיחים מאוד.

החלפת ייצוג הגיאומטריה ל- SDF מאפשרת שחזור גיאומטרי מדוייק יותר מאשר יצוג על בסיס צפיפות מרחבית, שדורשת שימוש בערכי סף כדי לקבוע איפה עובר המשטח המשוחזר, אך עלולה להשפיע על יכולת המודל לשקף תנאי סביבה אטמוספריים כגון אובך, ערפל, או דימדומים - מקרים שהודגמו באופן יפה למשל ב-Block-NeRF, ולא הוצגה להם התייחסות במאמר. קיימת אפשרות ששחזור טוב יותר או מדוייק יותר יהפוך לשחזור "טוב מדי" אם הצרכן הוא משתמש אנושי ולא אלגוריתמיקה אחרת. איור 48 מציג שחזור שעשוי להראות למשתמש אנושי פחות "טבעי" מאיור 21, למרות שמבחינת ציון PSNR איכות השחזור ב-Neuralangelo טובה יותר. נקודת העבודה והצורך הם שיכריעו.

מידע נוסף

אתר הפרויקט נמצא ב:https://research.nvidia.com/labs/dir/neuralangelo/, שם ניתן למצוא את המאמר והתוסף למאמר, סרטי וידאו, יכולת השוואה בין שחזור מבט חדש, שחזור גאומטריה ומפת נורמלים למשטח, קוד והסברים שונים.

5. סיכום, דיון ומסקנות

שחזור תלת מימד של סצנות גדולות, ובעיקר של סצנות גדולות עירוניות, מאיסופי תמונות ומיקומי מצלמה בלבד, היא בעיה ווותיקה בעולם הראייה הממוחשבת. אובדן המימד שבמעבר מהעולם התלת מימדי לדו מימד של ייצוג התמונה יכול -בעת השחזור - להיות מתורגם נומרית לפיתרונות שונים, שמסבירים באופן שונה את מבנה הסצנה. מבנה סצנה אורבנית, שבה הרבה הסתרות, משטחים חלקים גדולים שקשה להתאים למבטים אחרים, משטחים מבריקים, מחזירים או שקופים, מקשה אף יותר, ובסצנות גדולות הכיסוי היחסי של כל תמונה בודדת ביחס לכלל הסצנה הוא קטן, וקנה המידה של תמונות שנאספות מרחבי הסצנה עשויים להיות שונים מהותית.

אופי הפיתרון המועדף על משתמש זה או אחר עלול להיות מושפע באופן מהותי ממספר גורמים, בינהם: מתאר הייחוס של נקודת העבודה, אופי הקלט הזמין לו ואופן איסופו, מנגנון האיסוף ואפילו איכות המדידים הנלווים הזמינים לו.

הפתרונות שנסקרו מתייחסים לתת-סט כזה או אחר של מרחב הבעיה, ומציעים פתרונות שתואמים למתאר הייחוס שבחרו בו:

- Bungee-NeRF במוצהר לא מתייחסים לשטח הכיסוי הקרקעי בהגדרת סצנה גדולה, אלא מתמקדים בהבדלים בטווחים שבין מיקומי המצלמות האוספות - ההבדל בין גובה האיסוף בתמונת לווין לתמונת רחפן הוא פוטנציאלית עשרות קילומטרים. האלגוריתם משתמש במשטר אימון הדרגתי מהגס (טווח רחוק) לעדין (טווח קרוב) ומנצל מבנה רשת עם בלוקים שיוריים שנבנה בהדרגה בהתאם לטווח. בלוקים אלה גם מאפשרים לרנדר בעת חיזוי אזורים שונים בתמונה ברזולוציות שונות (בדומה למבנה LOD בשחזור קלסי) בלי לפגוע ברזולוציה הקרקעית המתקבלת. זהו גם הפיתרון היחיד שלא מתייחס לבעיות מראה, תאורה ותנאי סביבה (שוב במכוון ובמוצהר). הטיפול בכיסוי השטח הקרקעי הנרחב הם מותירים לעבודות אחרות או מאוחרות יותר, אבל כל מענה לנקודות עבודה שמשלב איסופים משכבות שונות משמעותית - לדוגמא איסוף רחפן ואיסוף אווירי מבוסס מטוס / כטב"ם או לווין - יוכל לנצל תובנות מעבודה זו.
- Block-NeRF מציגים פיתרון מאוד מקיף לשחזור סצנה אורבנית אמריקאית סטראוטיפית, המנצל את אופי הבניה הסדור ואת אופי האיסוף הקרקעי לבניית מערכת מאוד עמידה. הסצנה מחולקת לתת מודלים לפי הטופולוגיה של הצירים שעליהם נוסע רכב האיסוף, ובכך מנתקים בין גודל הסצנה הכללית לבין המגבלות על גודל סצנה בודדת, זמן ותנאי האיסוף, וזמן רינדור תמונה בודדת. החפיפה בין הבלוקים גדולה - 50% - ונועדה לאפשר התאמת נראות בין בלוקים סמוכים על מנת לאפשר מעבר חלק בין בלוקים, כולל שליטה בקידוד המראה עבור שעות היום. באופן כזה החיזוי מסוגל לאחד מבטים ממספר בלוקים באזורי ההשקה (1-3 בלוקים) לצרכי תמונה בודדת.
  האיסוף הקרקעי מרכב מאפשר שימוש באודומטריה איכותית יותר ובמערך מצלמות פתור הדדית[6], שאוסף בו זמנית (להבדיל מאיסופים אוויריים מסוגים שונים), מה שמאפשר מיקום מצלמות מדוייק יותר. זהו גם הפיתרון היחיד המוצג בעבודה שמשלב טיוב מיקומי מצלמות נלמד (Bundle adjustment), ולא פיתרון SFM כלשהו (ברוב המוחלט - colmap SfM) מקדים על הקלט. סביר להניח שאיכות האודומטריה המשופרת היא זאת שמאפשרת את הטיוב הנלמד, שכן פתרונות אחרים ציינו שבחנו שילוב טיוב מקום נלמד, אבל קיבלו תוצאות נחותות יחסית ל-SFM קלסי. Block-NeRF הוא הפיתרון היחיד שמתייחס בפירוט לבעיית זמן האיסוף הארוך שדורשת סצנה גדולה, וההתמודדות עם קלטים שונים שנאספו בהפרשי זמן משמעותיים. הפיתרון שהם מציגים מאלף בהיקפו.
- Mega-NeRF גם הם משתמשים בגישה של חלוקת מרחב הסצנה, אבל מאחר והאיסוף שלהם מבוסס רחפן ותרחיש הייחוס הוא איזור אסון, החלוקה גנרית יותר ואינה מתחשבת בעבירות או טופוגרפיית תא השטח המשוחזר. הסצנה מחולקת באופן ראשוני לחזית הסצנה - אליפסויד חוסם המכיל את כל מיקומי תמונות הקלט - והרקע, שמודל נפרד מייצג. את חזית הסצנה מחלקים בנוסף לגריד של תת מודלים עם חפיפה מינימלית (כ-10%) שמאומנים עצמאית. תמונות הקלט מחולקות בין תתי המודלים באופן נאיבי לפי עקבה קרקעית, ולאחר אימון

---

6 מצלמות פתורות הדדית הן מצלמות שעברו כיול משותף, ומיקומם המרחבי אחת יחסית לשניה (ולרוב גם לרכיבי האודומטריה כגון GPS ומדידים אינרציאליים) חושב במדוייק וטוייב. מיקום יחסי מדוייק בין מצלמות אוספות מאפשר להדק את הגדרת המשקולות בעקיבה בין תמונות שונות במהלך האיסוף ומקל על טיוב מיקומי המצלמות לאורך האיסוף.

ראשוני וקבלת מבנה גאומטרי גס מחולקות מחדש ברמת הפיקסל - חלקי תמונה שונים יכולים להיות משוייכים לתת מודלים שונים. שילוב של עבודה זו עם instant-NGP (שפורסם די במקביל) והאצת הביצועים המוצעת גם באימון וגם בחיזוי מתבקשת.

- Urban Radiance Fields משתמש בקלט שונה מכל שאר הפתרונות המוצעים - איסופי trakker רגליים שאספו תמונות fisheye ולידר, ביחד עם נתוני מיקום מטוייבים ע"ב SFM.
כמו Bungee-NeRF, גם URF לא מתמודד עם סצנות גדולות יותר ממה שניתן להכיל במודל בודד, אלא שפה גודל הסצנה מתבטא ביחס בין גודל הסצנה המכוסה לכמות התמונות - URF מסוגל לאמן מודל ולשחזר מבטים איכותיים וגאומטריה מדוייקת מכמות תמונות קלט קטנה מאוד יחסית למודלי NeRF אחרים, עקב ניצול המידע הנוסף מהלידר. אם Block-NeRF השתמשו ב-60-100k תמונות קלט לבלוק, URF מדברים על עשרות עד מאות בודדות של תמונות קלט, אמנם על שטח קטן מ"בלוק" בודד. זהו גם הפיתרון היחיד בסקירה זו שמתייחס לקידוד נראות אבל לא משתמש בקידוד תנאי סביבה שהוצע ב- NeRF in the wild (NeRF-W), אלא מציע קידוד משלו.
- Instant Neural Graphics Primitives לא מתייחס ישירות לסצנות גדולות, אלא מציע קידוד פרמטרי נלמד תלוי טבלאות גיבוב רב מימדיות, ושיפור ביצועים משמעותי מאוד - שעות אימון הופכות לשניות עד דקות, וזמן שיחזור של דקות הופך לקצב אינטראקטיבי מלא - כ- 30 Hz. כל הפתרונות שהוצגו עד כה ופורסמו בערך במקביל ל-INGP עשויים לשפר ביצועים משמעותית רק מהחלפת מימוש ה-Mip-NeRF הקיים בהם למימוש INGP-NeRF. בנוסף, הקטנת ה-MLP שנובעת מהוצאת הקידוד הפרמטרי לטבלאות גיבוב מאפשרת שחזור סצנה יותר גדולה במודל בודד, והרזולוציה הדינמית שטבלאות הגיבוב מאפשרות מספקת תמיכה יותר טובה בתמונות בקנה מידה משתנה. מעשית, נתוני ה-PSNR ו-F1-score שמוצגים טובים מכל התוצאות שהציגו הפתרונות הקודמים. איחוד עם פיתרון כמו Block-NeRF או Mega-NeRF יאפשר שחזור מהיר, מדוייק וניתן למיקבול של סצנה גדולה כרצוננו.
- Neurolangelo מתבססים על INGP ועל השימוש בקידוד פרמטרי מבוסס טבלאות גיבוב, כדי לאמן NeRF שבו MLP הצפיפות המרחבית (spacial density) מוחלף ב-MLP המייצג SDF. על ידי שימוש בגרדיאנטים נומריים ולא אנליטיים בחישוב הנגזרות ממעלות גבוהות, ועל ידי שימוש במשטר אימונים הדרגתי מרמת פירוט נמוכה לגבוהה, Neuralangelo מציגים איכויות שיחזור גאומטריה ושיחזור חוזי מנקודות מבט חדשות טובות מכל מה שהוצג קודם לכן.

מספר אלמנטים עולים כמשותפים לכל או לרוב הפתרונות:

<u>טיוב מיקום</u>

כל הפתרונות התייחסו לצורך בטיוב מיקום מצלמות הקלט באיסופים "in the wild" (כלומר לא סינטטיים). Bungee-NeRF, Mega-NeRF, Instant-NGP, Neuralangelo השתמשו ב-SFM קלסי כדי להפיק מיקומי מצלמה מתוקנים, Block-NeRF שילבו טיוב מיקום נילמד מבוסס BaRF (Bundle Adjustment For NeRF), ו-URF ציינו שמאגר המידע trekker שבו השתמשו כלל נתוני מצלמה שעברו SFM אוטומטי בזמן האיסוף, ושבמקרים שבהם ה-SFM האוטומטי כשל היו בעיות בשחזור.

<u>טיפול באובייקטים נעים</u>

כל הפתרונות למעט Bungee-NeRF - שמקור הקלט שלו איננו וידאו - שילבו הרצה מקדימה של סגמנטציה סמנטית על הקלט, כדי למסך סיווגים שנפוצים כאובייקטים נעים בסצנה - רכבים, אנשים וכו'. מעניין לציין שלא כולם סיננו את אותם אובייקטים: URF סיננו אנשים בלבד, Mega-NeRF סיננו אנשים וכלי רכב, ו-Block-NeRF סיננו מספר מחלקות של אובייקטים נעים, אך לא פירטו. גם מנסיון אישי וגם מתוך חלק מהתוספים שפורסמו למאמרים הנסקרים - אובייקטים נעים גורמים לארטיפקטים משמעותיים בשחזור. נקודה מעניינת היא שפיקסלים בתמונות הקלט שסומנו כחלק מפרטים המשתייכים למחלקות של אובייקטים נעים מוסכו וכלל לא השתתפו באימון, גם אם הפרט הספציפי לא היה בתנועה - כגון רכב חונה. התוצאה היא שחזור "סביבה אורבנית" נקייה לגמרי מסממנים אורבניים כגון רכבים, אופניים או הולכי רגל. כדאי לשים לב שכמות הקלט הנדרש כדי לוודא החסרה של אובייקטים נעים מחד וכיסוי מלא של הסצנה הסטטית מאידך עולה כמובן משמעותית על כמות הקלט ללא הקפדה זו.

<u>קידוד מראה ותנאי סביבה</u>

למעט Bungee-NeRF - שהשתמשו בקלט סינטטי ולכן התעלמו במוצהר ובמודע מהבעיה - כל הפתרונות

התייחסו לצורך בקידוד תנאי סביבה ומראה בסצנות חיצוניות נרחבות. בין אם המקור הוא תיקונים אוטומטיים כגון AGC במצלמה, או איסופים שונים שנדגמו בשעות ובחודשים שונים - כל הפתרונות המוצגים נדרשו לשלב קידוד מראה. למעט URF, כל שאר הפתרונות השתמשו בקידוד המראה שמוצע ב NeRF-W. URF העדיפו לנצל מטא-מידע שהיה זמין להם על מצלמת האיסוף, ולהשתמש בקידוד מטריצי של טרנספורמציית צבע אפינית, בטענה שקידוד המראה של NeRF-W מאפשר למודל "לתרץ" גם תנאי סצנה שאינם תנאי תאורה או סביבה.

<u>משטר אימון מגס לעדין</u>

מוטיב חוזר נוסף שיש להזכיר הוא משטר אימון הדרגתי, מ"גס" ל"עדין", שהוצג בחלק מהפתרונות בצורה זו או אחרת. ב-Bungee-NeRF משטר האימון הוא מהרחוק לקרוב (מאחר ומדובר על איסופים אווירים / לווייניים), כשבהוספת כל טווח מרחיבים את המודל על ידי הוספת בלוק שיורי וראש פלט. המאמר מציג באופן מעניין את "הפעלת" התדרים הגבוהים ההדרגתית במתארי פורייה שבקידוד IPE. Mega-NeRF מעדנים את שיוך תמונות ופיקסלים מתוך תמונות לתת מודלים השונים, לאחר אימון ראשוני שלומד גאומטריה גסה של הסצנה. Neurolangelo משמישים בהדרגה שכבות ברמות הפירוט של טבלאות הגיבוב, כדי להמנע ממצב שבו שכבת רזולוציה מפורטת תצטרך "לשכוח" את מה שלמדה לפני שתוכל ללמוד מחדש פרטים עדינים.

<u>זמנים</u>

זמני האימון עד Instant-NGP עמדו על שעות עד ימים לאימון ודקות לחיזוי תמונה בודדת.
Instant-NGP ו-Neuralangelo המבוסס עליו מהווים מהפכה בעולם ה-NeRF, הן מבחינת זמנים והן מבחינת איכויות השחזור. Mega-NeRF, שחלק גדול מהמחקר בו עסק במרנדר דינמי שיצליח להגיע לקצב אינטראקטיבי, הציגו עמידה בכ-1 Hz ביעף אינטראקטיבי שמניח רציפות עם פריימים קודמים.
Instant-NGP מציגים רינדור ב-30 Hz ללא תלות בפריימים קודמים.
Neurolangelo מציינים שמימוש החיזוי שלהם לא משתמש בניהול סטטיסטיקות בעת הגרלת פיקסלים וקרניים לאימון, וצופים שיפור משמעותי לאחר שילוב ניהול כזה. זמני האימון והשחזור המוצגים נמוכים משל Instant-NGP, אך עדיין טובים משמעותית משל הפתרונון האחרים.

<u>התמודדות עם סצנות נרחבות על ידי חלוקה לתת מודלים</u>

רק שניים מהפתרונות שנסקרו מתייחסים לסצנות שגדולות משמעותית מהיכולת של NeRF בודד כלשהו להכיל. בשני המקרים הגישה שואבת השראה מחלוקה ל-tiles שנפוצה במודלים תלת מימדיים קלסיים. מחלקים את מרחב הסצנה לתת מודלים בעלי חפיפה מינימלית בכיסוי, ומאמנים אותם בנפרד. אופי תרחיש הייחוס משפיע על אופי הפיתרון, ואפשרי לדון בהבדלים שבין הפתרונות מבלי להתייחס ספציפית למימוש ה-NeRF הנבחר:
ב-Mega-NeRF,מניחים שכל הסצנה נאספה באותו יעף איסוף, או ביעפים סמוכים שבוצעו פחות או יותר באותו זמן. אין כוונה לעדכן רק חלקים נבחרים מהכיסוי בהמשך, ותנאי הסביבה, למרות שיש להתייחס אליהם, לא משתנים מהותית בין תת מודולים המאומנים במקביל. אחרי אימון ראשוני שמספק מבנה גאומטרי גס, נשקלת מחדש חלוקת הקלט בין תת המודלים, והחלוקה היא ברמת פיקסלים וחלקי תמונה, כך שתמונה יכולה להשתתף במספר תת מודלים, אבל פיקסלים שונים בה יישויכו כל אחד רק לתת מודל ספציפי. תכונה זו והחפיפה המינימלית (15%) מונעים ארטיפקטים באיזורי התפר בין תת מודלים, ומאפשרים לתמונות מרוחקות עדיין להשפיע על תת מודלים. מאחר ופיקסל משוייך לתת מודל ביחד עם קידוד המראה של התמונה, מנגנון זה ממצע את קידוד המראה בין תת המודלים באימון, שמראש אופן האיסוף מניח דימיון בינהם. בעת חיזוי נבחר כל פעם מוקד תת המודל הבודד הקרוב ביותר לנקודה הדגומה, והוא אחראי לתוצאה.

Block-NeRF מציגים את המנגנון המורכב מבין השניים, בין השאר בגלל נקודת עבודה שמניחה איסוף מתמשך, בשעות יום שונות ואף בחודשים שונים - דבר שגורר מזג אוויר שונה ותנאי סביבה שונים בתכלית. בנוסף, יש הנחה שיהיה צורך לעדכן חלקים שונים בסצנה לאורך זמן. החפיפה בין "בלוקים" משיקים היא של כ-50%, תמונות שרלוונטיות למספר בלוקים משתתפות באופן מלא בכולם, וקיימים מנגנונים להשוואת נראות - כדי להתאים בין בלוקים שנאספו בתנאי נראות שונים או כדי להתאים לצורך סימולטיבי, ולאיחוד משוקלל של מספר בלוקים במהלך החיזוי.

שילוב של מודל Neuralangelo או INGP-NeRF בתוך מנגנון ניהול סצנה גדולה כמו של Mega-NeRF (שהוא גנרי יותר מאשר המנגנון של block-NeRF) ייתן כנראה תמיכה טובה למגוון של נקודות עבודה ותרחישי ייחוס. הפיתרון ש-Block-NeRF מציעים הוא השלם ביותר מבחינת יצירת חיזוי סצנה

"טבעית" ואחידה למראה לאורך מספר בלוקים. פתרונות נוספים שהוצגו פה נותנים חומר למחשבה למספר נקודות עבודה שונות או סוגי קלט יחודיים יותר.

המשך מחקר עתידי - הן של פיתרון Mega-NeRF והן של Bungee-NeRF - עם ליבה של Instant-NGP, נראה ככיוון מבטיח מבחינת זמנים. מבחינת Mega-NERF, כפי שכבר צויין, פיתרון חלוקת המרחב שהוצג ובניית סט הקלט לכל אזור אינם תלויים במימוש הנרף הנבחר, ולכן מתאימים לעידכון שכזה.
מבחינת Bungee-NeRF פיתוח כזה ידרוש לא רק החלפת מודל הנרף במימוש המהיר יותר של Instant-NGP, אלא גם מבחינת טיפול בכיסוי שטח נרחב. יהיה מעניין לבחון בלוק גס (L1) משותף לאזור שלם, עם בלוקים מפורטים יותר (L2, L3) המתפצלים לאזורים ממוקדים יותר.
שיפור דומה של שילוב Neurolangelo ב-MegaNeRF, עשוי להביא לשיפור נוסף באיכות השחזור, למרות שלא במהירות השיחזור (לעת עתה).

6. רשימת מקורות

[1] Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." arXiv preprint arXiv:2003.08934 (2020).

[1.1] Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." arXiv e-prints (2020): arXiv-2003.

[2] Müller, Thomas, Alex Evans, Christoph Schied, and Alexander Keller. "Instant neural graphics primitives with a multiresolution hash encoding." ACM Transactions on Graphics (ToG) 41, no. 4 (2022): 1-15.

[3] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. "Mega-NeRF: Scalable Construction of Large-Scale NeRFs for Virtual Fly-Throughs." In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12912-12921. IEEE Computer Society, 2022.

[4] Tancik, Matthew, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. "Block-nerf: Scalable large scene neural view synthesis." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8248-8258. 2022.

[5] Rematas, Konstantinos, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. "Urban radiance fields." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12932-12942. 2022.

[5.1] Rematas, Konstantinos, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. "Supplementary material for Urban radiance fields supplamental." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[6] Xiangli, Yuanbo, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. "Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering." In European conference on computer vision, pp. 106-122. Cham: Springer Nature Switzerland, 2022.

[7] Li, Zhaoshuo, Thomas, Muller, Alex, Evans, Russell H, Taylor, Mathias, Unberath, Ming-Yu, Liu, and Chen-Hsuan, Lin. "Neuralangelo: High-Fidelity Neural Surface Reconstruction." . In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).2023.

[7.1] Li, Zhaoshuo, Thomas Müller, Alex Evans, Russell H. Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. "Neuralangelo: High-Fidelity Neural Surface Reconstruction - supplementary." *research.nvidia.com. 2023*

[8] Kajiya, James T., and Brian P. Von Herzen. "Ray tracing volume densities." *ACM SIGGRAPH computer graphics* 18, no. 3 (1984): 165-174.

[9] Rahaman, Nasim, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. "On the spectral bias of neural networks." In *International Conference on Machine Learning*, pp. 5301-5310. PMLR, 2019.

[10] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[11] Barron, Jonathan T., Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5855-5864. 2021.

[12] Martin-Brualla, Ricardo, Noha Radwan, Mehdi SM Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. "Nerf in the wild: Neural radiance fields for unconstrained photo collections." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7210-7219. 2021.

[13] Zhang, Kai, Gernot Riegler, Noah Snavely, and Vladlen Koltun. "Nerf++: Analyzing and improving neural radiance fields." *arXiv preprint arXiv:2010.07492* (2020).

[14] Schonberger, Johannes L., and Jan-Michael Frahm. "Structure-from-motion revisited." *In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4104-4113. 2016.*

[15] Schönberger, Johannes L., Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. "Pixelwise view selection for unstructured multi-view stereo." *In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14, pp. 501-518. Springer International Publishing, 2016.*

[16] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. "PlenOctrees for real-time rendering of neural radiance fields". In ICCV, 2021.

[17] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. "Pixel-Perfect Structure-from-Motion with Featuremetric Refinement". In ICCV, 2021.

[18] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In IEEE International Conference on Computer Vision (ICCV), 2021.

[19] David Crandall, Andrew Owens, Noah Snavely, and Daniel Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[20] Yilin Liu, Fuyou Xue, and Hui Huang. "Urbanscene3d: A large scale urban scene dataset and simulator". arXiv preprint arXiv:2107.04286, 2021.

[21] Gernot Riegler and Vladlen Koltun. "Stable view synthesis". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.

[22] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. "Plenoxels: Radiance fields without neural networks". In CVPR, 2022.

[23] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 14335–14345, 2021.

[24] Jun-Yan Zhu Kangle Deng, Andrew Liu and Deva Ramanan. "Depth-supervised nerf: Fewer views and faster training for free". ArXiv:2107.02791, 2021.

[25] Takikawa, Towaki, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. "Neural geometric level of detail: Real-time rendering with implicit 3d shapes." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11358-11367. 2021.

[26] Wang, Peng, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction." *arXiv preprint arXiv:2106.10689* 2021.

[27] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. "Tanks and temples: Benchmarking large-scale scene reconstruction". ACM Transactions on Graphics (ToG), 36(4):1–13, 2017.

[28] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. "Large scale multi-view stereopsis evaluation". In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 406–413. IEEE, 2014.

[29] Li, Ruihao, Sen Wang, and Dongbing Gu. "DeepSLAM: A robust monocular SLAM system with unsupervised deep learning." IEEE Transactions on Industrial Electronics 68, no. 4 (2020): 3577-3587.

[30] Xiao, Linhui, Jinge Wang, Xiaosong Qiu, Zheng Rong, and Xudong Zou. "Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment." Robotics and Autonomous Systems 117 (2019): 1-16.

[31] Lin, Chen-Hsuan, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. "Barf: Bundle-adjusting neural radiance fields." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5741-5751. 2021.

[32] Yen-Chen, Lin, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. "inerf: Inverting neural radiance fields for pose estimation." In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1323-1330. IEEE, 2021.

[33] Hartley, Richard, and Andrew Zisserman. Multiple view geometry in computer vision. Cambridge university press, 2003.

[34] Lehmer, Derrick H. "Mathematical models in large-scale computing units." Ann. Comput. Lab.(Harvard University) 26 (1951): 141-146.

[35] Su, Shih-Yang, Frank Yu, Michael Zollhöfer, and Helge Rhodin. "A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose." Advances in Neural Information Processing Systems 34 (2021): 12278-12291.

[36] Dissanayake, MWM Gamini, Paul Newman, Steve Clark, Hugh F. Durrant-Whyte, and Michael Csorba. "A solution to the simultaneous localization and map building (SLAM) problem." IEEE Transactions on robotics and automation 17, no. 3 (2001): 229-241.

[37] Westoby, Matthew J., James Brasington, Niel F. Glasser, Michael J. Hambrey, and Jennifer M. Reynolds. "'Structure-from-Motion'photogrammetry: A low-cost, effective tool for geoscience applications." Geomorphology 179 (2012): 300-314.

[38] Jin, Hailin, Stefano Soatto, and Anthony J. Yezzi. "Multi-view stereo reconstruction of dense shape and complex appearance." International Journal of Computer Vision 63 (2005): 175-189.

[39] Seitz, Steven M., Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. "A comparison and evaluation of multi-view stereo reconstruction algorithms." In 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), vol. 1, pp. 519-528. IEEE, 2006.

[40] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

נספח א' – מושגים:

בנספח זה נעסוק במספר מושגים ומונחים, הבאים מתחומי עיסוק אחרים מ-NeRF, הרלוונטיים לעיסוקנו בסצנות גדולות, והמקילים על הבנת חלק מההסברים בעבודה.

i. *רזולוציה קרקעית פיקסלית*

25

מושג נפוץ בפוטוגרמטריה וגיאודזיה, משמעותי (בעיקר) באיסוף אווירי. מדד של איכות הכיסוי - כלומר כמה מרחק קרקעי מכסה כל פיקסל. מייצג את היחס שבין פרמטרי המצלמה הפנימיים - בעיקר הרזולוציה הפיקסלית של המצלמה וה FOV - לבין גובה / מרחק האיסוף. FOV צר ומרחק גדול יכולים להיות שקולים ל-FOV רחב ומרחק קצר, וכך רזולוציה קרקעית נותן מדד אינפורמטיבי יותר מאשר טווח. לרוב נמדד בס"מ לפיקסל, תוך הנחה של פיקסלים ריבועיים (כלומר מימד זהה ב X,Y). באיסוף אורטוגרפי (המצלמה בנדיר[7] יחסית לקרקע) הערך אחיד, ואילו באיסוף אלכסוני (oblique) עשוי להיות הבדל משמעותי בין החלק הקרוב למצלמה לבין החלק הרחוק. במקרים כאלה נהוג לציין האם הערך המצויין הוא מקסימום או ממוצע. באיסופים קרקעיים יכול לשמש רק כמדד כללי לפירוט הממוצע של אובייקט העניין (foreground), מאחר וכל תמונה מכילה מרזולוציה מילימטרית ועד אינסוף[8], בהתאם למרחק האובייקט מהמצלמה, הסתרות ומבנה הסצנה הכללי.

ii. *עקבה קרקעית*

מונח מתחום הגיאודזיה / פוטוגרמטריה. בתמונות אוויריות, מדובר במצולע הבנוי מהטלת פינות התמונה (הפרוסטום) על הקרקע (או על מישור הקרקע המשוערך). משמש לחישובי כיסוי מהירים ולהערכת חפיפה בין מצלמות המשתתפות באותה סצנה.

iii. *Level Of Detail (LOD)*

מונח שמקורו בגרפיקה תלת מימדית, בייצוג של מודלים תלת מימדיים מורכבים. משמש הרבה בתחום של רינדור אינטראקטיבי (כגון משחקי מחשב). מטווח מסויים והלאה, פירוט רב מדי (הן ברמת הגיאומטריה והן ברמת הטקסטורה) לא יעיל - וורטקסים שונים במודל ימופו לאותו פיקסל מרונדר, והמודל יעמיס על הכרטיס הגרפי עם מספר וורטקסים גדול וטקסטורות כבדות ומפורטות, מבלי לשפר את התמונה המרונדרת. כדי להמנע מכך, מגדירים מראש מספר רמות פירוט, בהן גיאומטריות מפושטות וטקסטורות מרודדות יותר ויותר, ומגדירים מטריקה ידועה (לרוב - טווח של המודל מנקודת המבט או גודל פיקסלי מוערך של המודל על התמונה המרונדרת) שלפיה מנוע הרינדור יכול לבחור את ה- LOD המתאים בכל פריים. בטווח רחוק יבחר ברמת פירוט נמוכה ויחסוך משאבי מערכת מבלי לפגוע בתוצר המוגמר, ובטווח קרוב יצייר את הרזולוציה הטובה ביותר האפשרית.

iv. *Peak Signal to Noise Ratio – PSNR*

21

מדד למדידת איכות תמונה. הוא מחושב על ידי השוואת התמונה המקורית לתמונה המשוחזרת, ומדידת הפער בין שתי התמונות. הפער מחושב כ-MSE (Mean Squared Error), וה-PSNR מחושב על ידי לקיחת לוג של 1/MSE.
PSNR הוא מדד ללא יחידות (היות ו MSE הוא מדד ללא יחידות), והוא מוגדר בדרך כלל בדציבלים [dB]. ערך PSNR גבוה יותר פירושו איכות תמונה טובה יותר. לשם קבלת תחושה - ערך PSNR של 30 dB נחשב לאיכות תמונה טובה מאוד, ערך של 20 dB נחשב לאיכות תמונה טובה, וערך של 10 dB נחשב לאיכות תמונה בינונית.
שתי תמונות המכילות את אותו תוכן אבל ברזולוציות שונות יחזירו אותו ערך PSNR.
PSNR הוא מדד פופולרי לשימוש בתחום דחיסת תמונה, מכיוון שהוא יכול לשמש כדי למדוד את איכות התמונה המשוחזרת לאחר דחיסה / פריסה. ב-NeRF משמש למדידת איכות התמונה המשוחזרת מנקודת מבט נתונה מול תמונת מקור שלא השתתפה באימון.

---

7 Nadir – קו ראייה היישר למטה, במאונך למישור הקרקע האידיאלי. קו הופכי לזנית (Zenith). ראו https://en.wikipedia.org/wiki/Nadir .

8 נקודת ההעלמות – vanishing point היא הנקודה בגאומטריה פרוייקטיבית שבה נפגשים קווים מקבילים. בתמונה קרקעית המסתכלת לכיוון האופק יש פיקסל ספציפי שבו נמצאת נקודת ההעלמות, והרזולוציה הקרקעית בו היא אינסופית. ראו https://en.wikipedia.org/wiki/Vanishing_point להסבר קצר או [33] להסבר מעמיק.

v. *LIDAR (חיישן)*

41

הוא חיישן (LIght Detection and Ranging- ידוע גם כ- Laser Imaging Detection And Ranging) אקטיבי (קורן), לרוב באינפרה אדום (סנסורים קרקעיים לרוב ב -NIR, סנסורים מוטסים לרוב בתדר נמוך יותר) המחזירים טווח על פי זמן מעוף (time of flight). לרוב מדובר בחיישן שורה סורק (או גזרה או 360°), כך שפוטנציאלית לכל שורה יש מיקום מקור נפרד -בהתאם למהירות תנועת הפלטפורמה הנושאת. בשילוב GPS+IMU יכול לגזור סדרה של מיקומים תלת מימדיים אבסולוטיים שבהם המערכת זיהתה מכשול.

vi. *מרחק שמפר - chamfer distance*

55,44

מטריקה המשמשת להשוואה של ענני נקודות. זהו סכום המרחק בין כל נקודה בענן אחד לנקודה הקרובה ביותר בענן השני. ככל שהערך נמוך ענני הנקודות דומים יותר. נמוך יותר עדיף.

vii. *F_1-Score*

מדד דיוק (למשל של מסווג) המוגדר כ: $F = 2*\frac{(אחזור \cdot דיוק)}{(אחזור + דיוק)}$ כאשר דיוק (precision) הוא כמה תוצאות חיוביות הן אכן נכונות (סווגו נכון) , ואילו איחזור (recall) מוגדר כיחס הערכים החיוביים ששוערכו ככאלה (כמה מתוך התוצאות הנכונות סווגו ככאלה). דיוק מושלם יחזיר את הערך 1.0. מדד F_1 ממשקל באופן שווה דיוק ואחזור, אך לפעמים זאת לא התוצאה הרצויה.

מדד כללי יותר הוא $f_\beta = (1+\beta^2)\frac{אחזור \cdot דיוק}{(\beta^2 \cdot דיוק) + אחזור}$ שבו ניתן למשקל את השפעת הדיוק לעומת האחזור.
β הוא פקטור חיובי ממשי הממשקל את השפעת אחד החלקים במשוואה לעומת השני -השפעת הדיוק ממושקלת פי β^2 מאשר האיחזור. F_1-Score הוא מקרה פרטי של F_β-Score עם השפעה שווה בין הדיוק והאחזור ($\beta=1$). בהתאם לנקודת העבודה הרצויה עשויים לשים דגש על רכיב אחד או השני - לדוגמא בתצפית ואבטחה עשויים לבחור מערכת שתדגיש את האחזור לעומת הדיוק - מפעיל מצלמת אבטחה עשוי להעדיף שהמצלמה תציף את כל הזיהויים האפשריים, גם אם חלקם שגויים (false positive), ותאפשר שיקול דעת, בעוד במקרים אחרים ההפך הוא הרצוי - מערכות המגיבות אוטומטית עלולות להעדיף את הדיוק על האחזור - למשל כאשר הנזק מתגובה לזיהוי שגוי משמעותי, או כשיש יתירות מערכתית בנוגע לאחזור.

viii. *Signed Distance Function - SDF*

50,46

SDF הינה פונקציה המייצגת צורה תלת מימדית, כך שסט האפס שלה (zero-set) מייצג את פני השטח של הצורה. בכל מקום שאיננו על פני האובייקט הפונקציה מחזירה מרחק (בעל סימן) של הנקודה הנתונה מהאובייקט או המשטח התלת מימדי (3D surface). ערך שלילי פירושו שהנקודה מעל המשטח או מחוץ לאובייקט, בעוד ערך חיובי פרושו מתחת למשטח או בתוך אובייקט - בהתאם לשימוש. SDF-ים משמשים בסימולציות, חישובי מסלול, משחקי מחשב וכלי מידול תלת מימדי.

ix. *Intersection Over Union - IoU*

50

IoU הוא מדד התאמה בין שטחים או נפחים. בהקשר לנפחים וצורות IoU מוגדר להיות היחס בין נפח פנים החיתוך לבין נפח איחוד הצורות המושוות. בדו מימד ההגדרה דומה.
$0.0 \leqslant IoU \leqslant 1.0$ תמיד, כש-1.0 פרושו התאמה מלאה.

Abstract

Reconstructing Large-Scale Urban Scenes Using Neural Radiance Fields

This research focuses on the 3D reconstruction of large-scale urban scenes leveraging Neural Radiance Fields (NeRF). 3D reconstruction, both visually and geometrically, from 2D imagery is a complex problem, and large-scale outdoor scenes introduce additional challenges, such as high geometric complexity, varying lighting conditions, and massive input data. For years, this field was dominated by classical algorithms like Multi-View Stereo (MVS) and Simultaneous Localization and Mapping (SLAM).

In recent years, deep learning algorithms, primarily NeRF and its variants, have made significant inroads. NeRFs are deep learning-based networks that enable an implicit and differentiable representation of a 3D scene's structure–both geometrically and color-wise–allowing for novel view synthesis and depth map generation. Despite the abundance of research on NeRF, few studies have focused on large-scale scenes, especially large urban environments. Even fewer explicitly address scenes that cannot be represented by a single NeRF.

This work reviews the theoretical and algorithmic foundations of NeRF and Mip-NeRF, essential for understanding the presented solutions. It also explores the unique characteristics and challenges of large-scale scenes and the leading solutions addressing them.

Investigated methods include Bungee-NeRF, Block-NeRF, Mega-NeRF, URF, Instant-NGP, and Neuralangelo. The paper presents the various algorithmic solutions, discusses the specific challenges each solution addresses, examines the problem set each solution aims to solve, and proposes potential combinations of these methods. Additionally, the paper delves into fundamental questions related to large-scale scene reconstruction, such as defining a large scene, dividing a scene into subparts, data acquisition, and handling visual complexity.

As a foundation for the discussion, the paper also provides a brief theoretical background on classical methods and tools for addressing 3D reconstruction. These methods are relevant to the discussion, whether as preprocessing tools, inspiration for parts of the presented algorithms, or a way to explain alternative processes presented in the reviewed works.

The results presented in the paper demonstrate the significant progress achieved in reconstructing large-scale urban scenes using NeRF. From the vast body of work under "NeRF algorithms," this paper centralizes the solutions addressing the sub-problem of large-scale scenes, presents the advantages and disadvantages of each solution–both relative to others and independently–and suggests future research directions.

Table of Contents

The Open University of Israel
Department of Mathematics and Computer sciences

Large Scene NeRFs:
Theory and current applications

Final Paper submitted as partial fulfillment of the requirements
towards an M.Sc degree in Computer Sciences
The Open Univerity of Israel
Department of Mathematics and Computer Sciences

By
Aviv Sever

Prepared under the supervision of Dr. Mireille Avigal

07/2024