# Review 129: [Short] Watch Your Steps: Local Image and Scene Editing by Text Instructions, 26.08.23 https://arxiv.org/abs/2211.09800.pdf

**Paper: https://arxiv.org/abs/2308.08947v1**

מודלי הדיפוזיה לגמרי השתלטו כמעט על כל המשימות של הראייה הממוחשבת. למשל עריכה של תמונות (למשל להחליף ציפור בפרפר)בהתאם לתיאור טקסטואלי כבר מזמן עושים רק באמצעות מודלי דיפוזיה חזקים כמו InstructPix2Pix או IP2P בקצרה. למרות התוצאות המדהימות עדיין יש אי התאמות בין התמונה הערוכה לבין המקורית.   
  
היום ב-#shorthebrewpapereviews נסקור מאמר שמנסה לתקן את אי דיוקים אלו בצורה די אלגנטית. בשלב הראשון המודל המוצע מאתר את מיקום הפיקסלים שאותם צריך לשנות(מסכה) ובשלב השני עורכים את התמונה רק באזורים של המסכה. כל זה נעשה באמצעות מודלי דיפוזיה באופן די אלגנטי.   
  
בשלב הראשון מרעישים את התמונה המקורית (עד רמת רעש מסוימת שהיא מהווה הייפרפרמטר חשוב מאוד) משתמשים במודל IP2P כדי לשערך את הרעש נוסף עבור ללא תופסת טקסט לעריכה ויחד איתו. כלומר במקרה הראשון אנו מפעילים מודל דיפוזיה סטנדרטי (ללא עריכה) ובמקרה השני כן עורכים את התמונה בהתאם לתיאור הטקסטואלי. לאחר מכן מחשבים את הערך המוחלט של ההפרש בין השערוכים אלו, מקצצים את החריגים (עם IQR עם מקדם 1.5). המסכה מקבלת ערך 1 (פיקסלים לעריכה) במקומות שההפרש הזה עולה על סף מסוים (הייפרפרמטר נוסף).   
  
בשלב השני מרעישים את התמונה (רמת הרעש עוד הייפרפרמטר). ואז באמצעות מסירים את הרעש עם מודל IP2P (עם תיאור טקסטואלי) באיזורים של המסכה ובכל האזורים האחרים עושים זאת עם מודל דיפוזיה רגיל (הטקסט המוסף הוא ריק). בנוסף המחברים מכלילים את הגישה שלהם ל-NeRF (ייצוג של מודלי 3D). בגדול עושים את מה שמתואר למעלה על views מכל הזווית תוך שמירה של קוהרנטיות ביניהם.