רשתות נוירונים לתמונות | תרגיל 4

יואב שפירא 312492838 | עידו קליינר 313198236

6 ביולי 2023

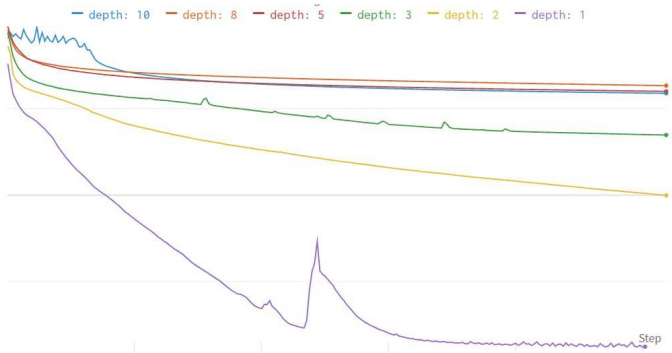
חלק 1

מימשנו את $StyleLoss$ ואת $ContentLoss$, מודולים שמקבלים שכבות של VGG כאינפוט, ומעתיקים את הארכיטקטורה של הרשת $VGG-19$ עד לשכבה הכי עמוקה שניתנה להם (מעבר לכך לא צריך). ה $ContentLoss$ מחשב את הMSE ישירות על שכבות האקטיבציה שניתנות לו כאינפוט, וה$StyleLoss$ מחשב את הMSE בין מטריצות ה$Gram$ של שכבות האקטיבציה - כמו שלמדנו בשיעור.
בנינו $Style-loss\ Model$ שמקבל תמונות רפרנס של תוכן ושל סטייל, ומאתחל מודולים $ContentLoss$ ו$StyleLoss$ - עם הרפרנסים כ$target$, בהתאם. אחר כך הוא מאתחל תמונת רעש, ומאפטם אותה לפי ה$loss$ שהוא מקבל מהשוואה בין התוכן והסטייל של המודולים $ContetntLoss$ ו$StyleLoss$, לפי משקול מסוים שניתן כהיפר פרמטר.
לאורך כל התרגיל עשינו שימוש בתמונת תוכן הבאה:

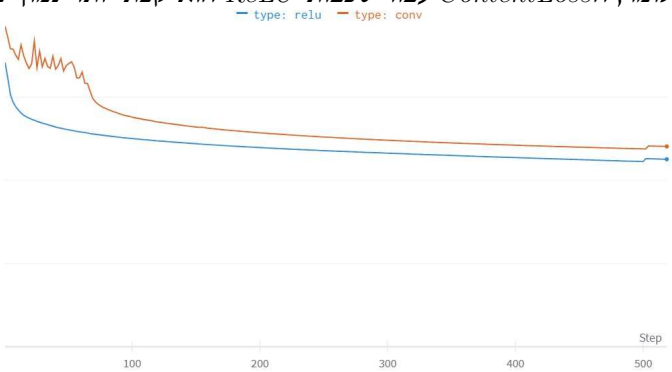
ובתמונת סטייל הבאה ($"Seated\ Nude"\ by\ Picasso$):

$Features\ Inversion$:

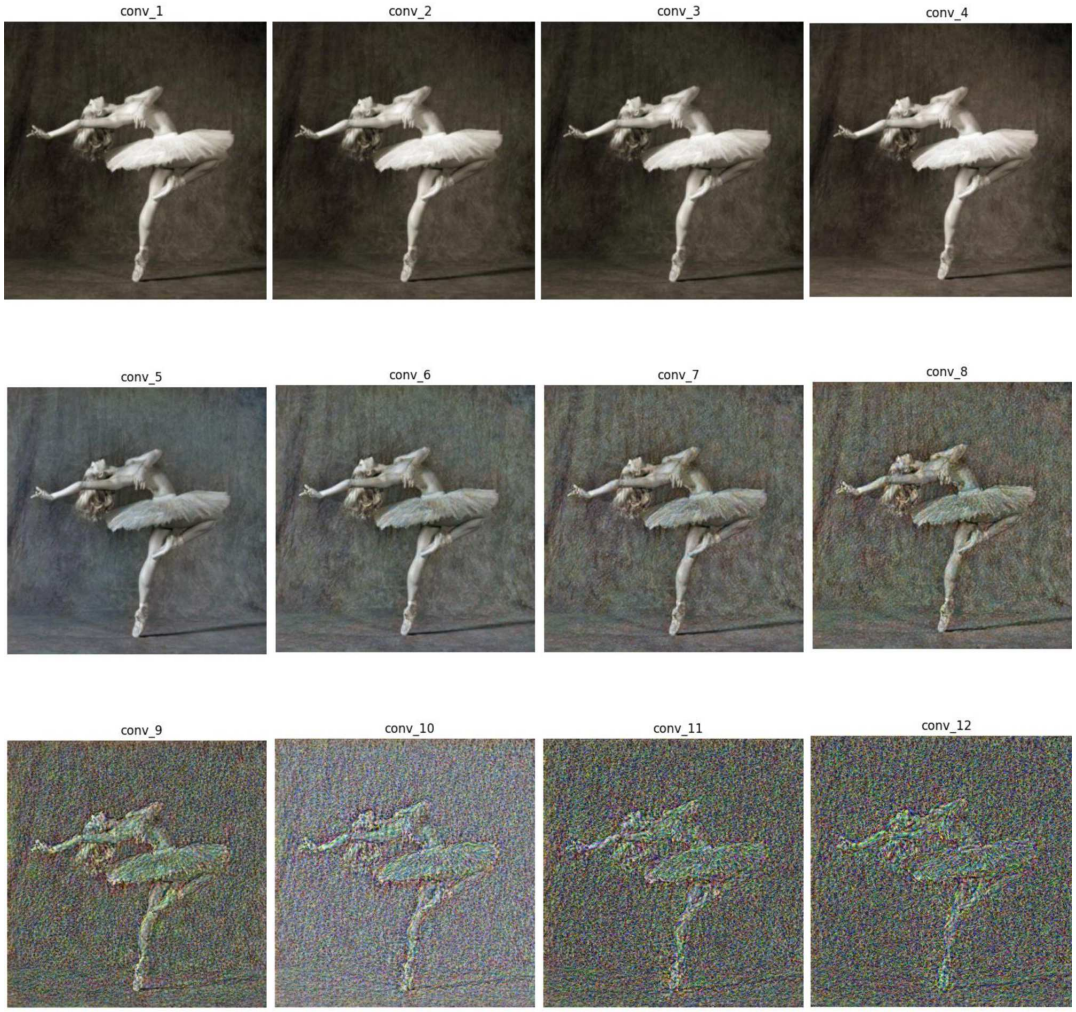
בשאלה הזו אנחנו נדרשים לשנות אל האלגוריתם כדי לחקור את האופטימיזציה של הרעש רק לפי $Content\ Loss$ משכבות שונות. בנינו מודל $Style-loss\ model$ שמאתחל רק $ContentLoss$, בכל פעם עם שכבה אחרת מתוך הארכיטקטורה של הVGG (למעשה, הוא מאתחל גם $StyleLoss\ Module$ אבל בלי שכבות לחלץ מהן את הלוס, וכך בפועל לא מתבצעת השוואה של סטייל אלא רק של תוכן). ביצענו את זה עבור שכבות בעומק שונה (מ1 עד 13) ושכבות מסוגים שונים (גם $ReLU$ וגם $conv$).
התוצאות מראות שככל שמעמיקים ברשת, התוכן נעלם: ה$ContentLoss$ הולך ונהיה גדול יותר ככל שמעמיקים ברשת, באופן כללי. זו תוצאה מאוד הגיונית, כי השכבות העליונות הן השכבות שהכי קרובות לאינפוט - התמונה המקורית שבאמת מכילה את התוכן במלואו.
להלן גרף של ה$loss$ שממחיש זאת:



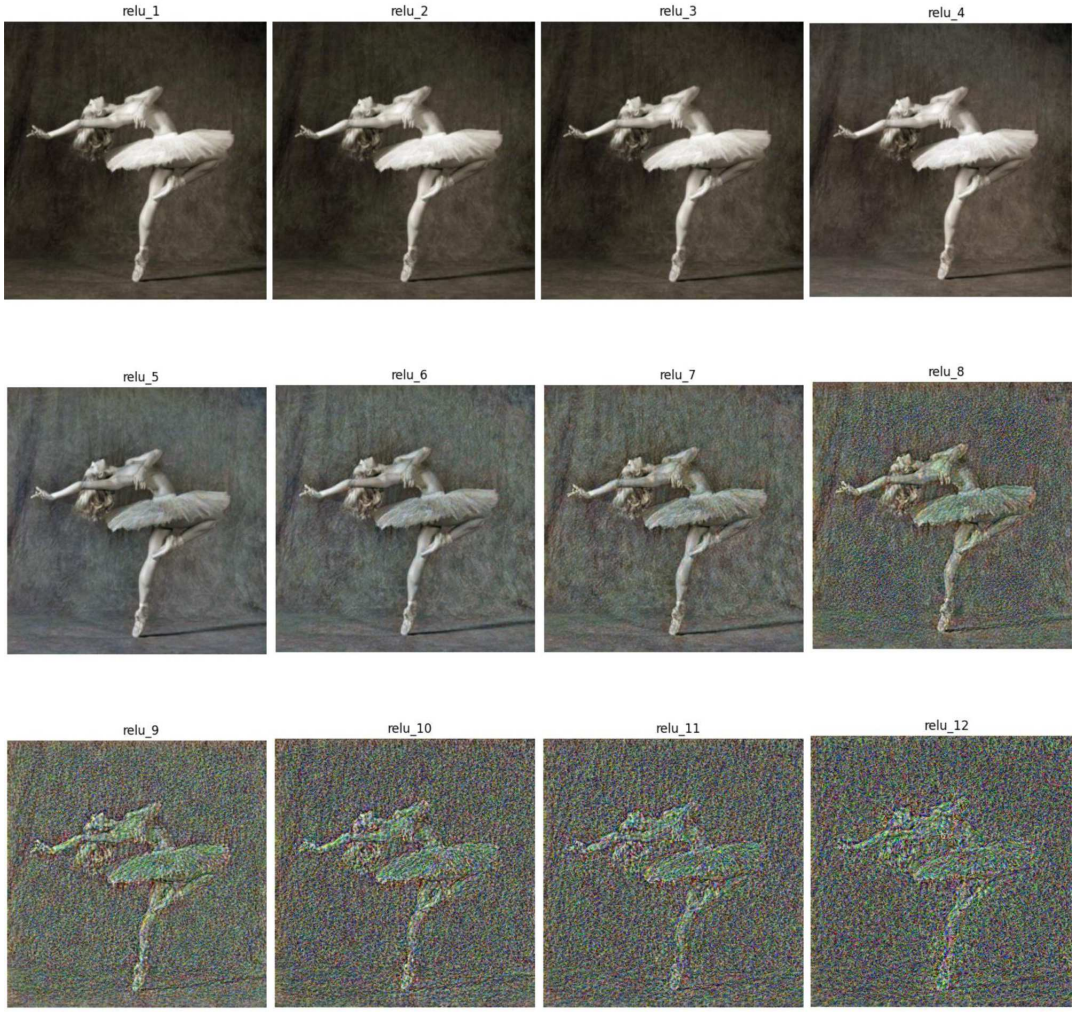
בנוסף, ראינו שהאקטיבציות מיד לאחר שכבות ה$ReLU$ קצת יותר מעידות על התוכן. כלומר, ה$ContentLoss$ עבור שכבות $ReLU$ הוא קצת יותר נמוך משכבות של $Conv$:



התוצאות גם משתקפות בצורה מאוד ברורה לעין אנושית. אלו התוצאות בעומקים 1-12 (האחרונה לא נכנסה יפה לקולאז' אבל הבנו את הרעיון...), לפי השוואה משכבות קונבולוציה:



ואלו מהשוואה לפי שכבות $ReLU$:

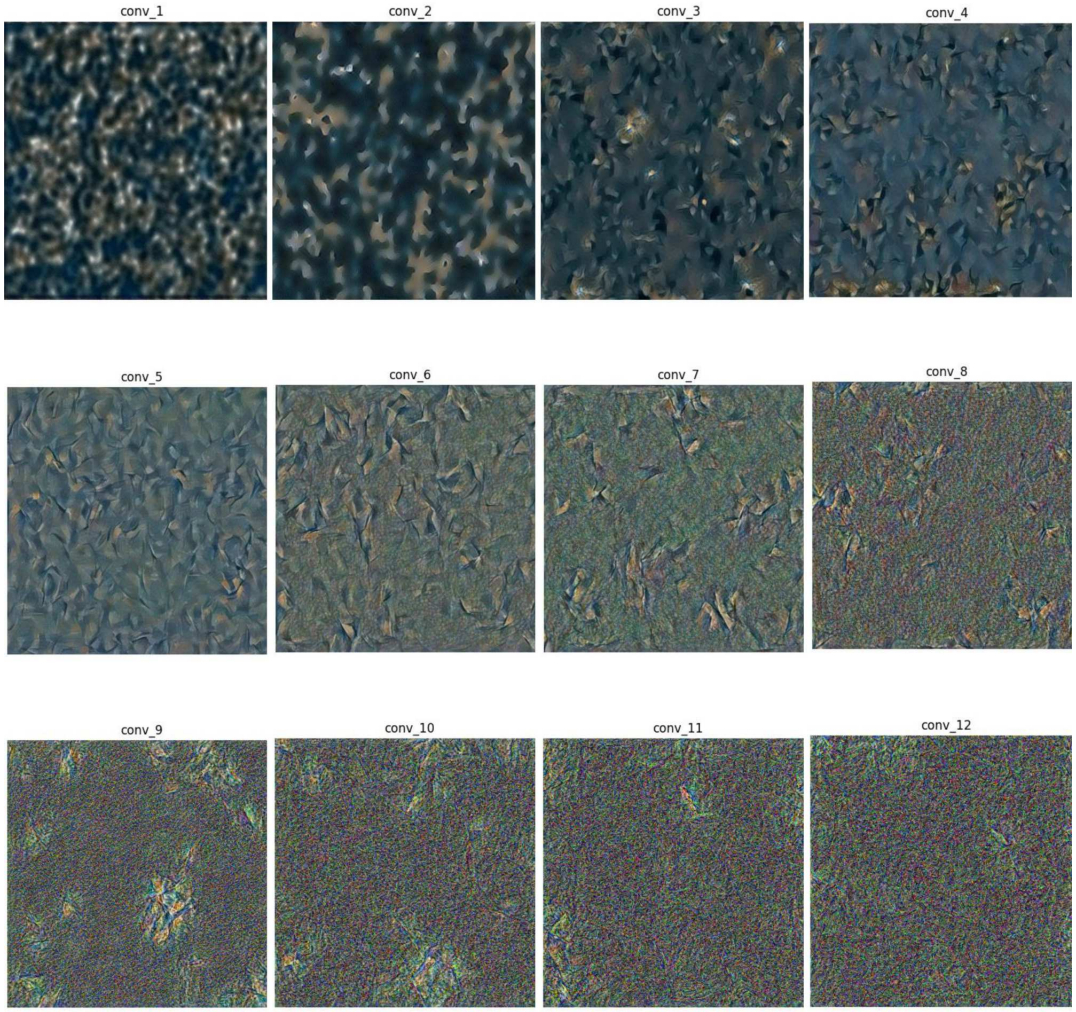


אפשר קצת להבחין שרוב התמונות המשוחזרות מהניסוי על שכבות $ReLU$ יוצאות מעט יותר בהירות. זה ניתן להסבר על ידי כך ש$ReLU$ חותכת את הערכים השליליים, כך ששחזור תמונה מיד לאחר שכבת $ReLU$ תהיה 'פחות מגוונת' מבחינת אזורים כהים יותר. ביצענו השוואה של בהירות ממוצעת ואכן ברוב השכבות זה המצב:

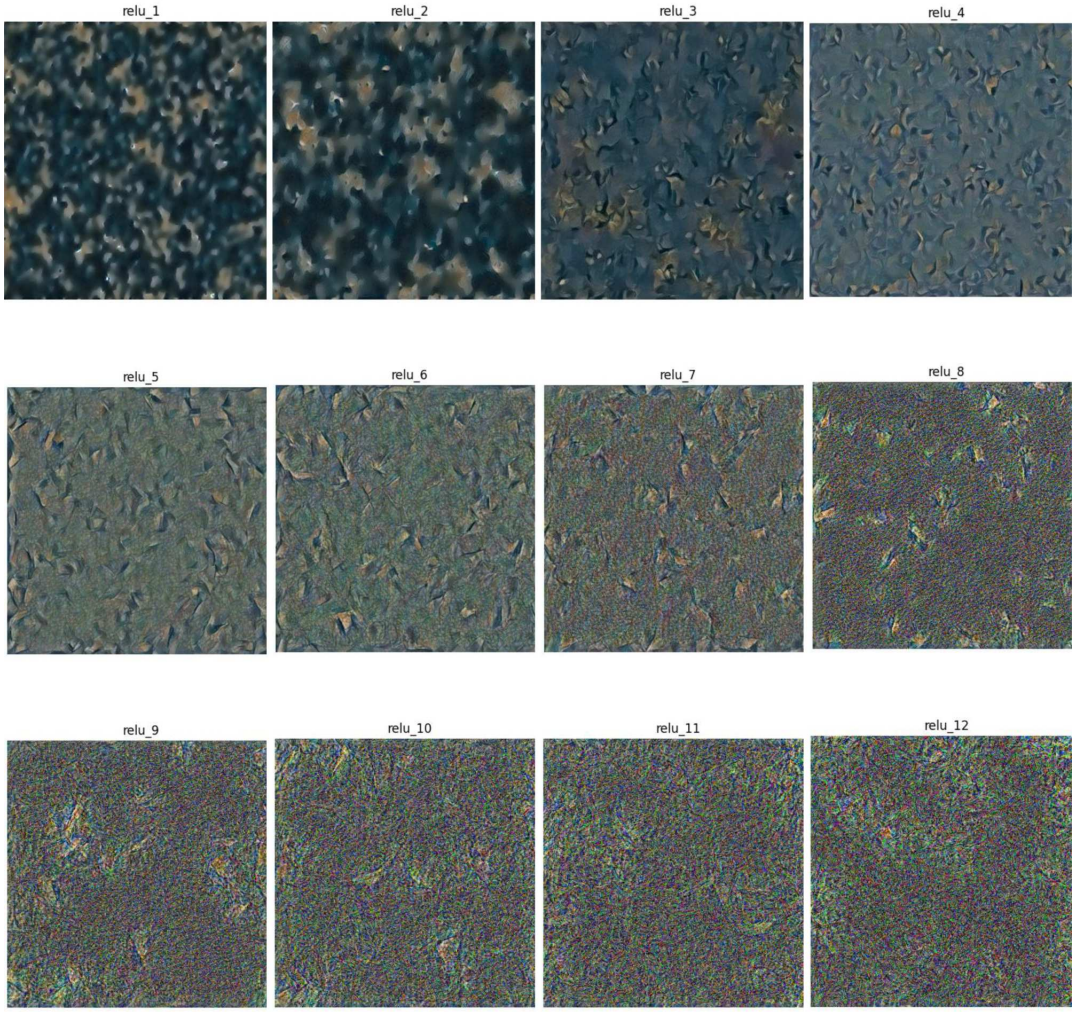
```
Higher Intensities in depth 0:  ReLU
Higher Intensities in depth 1:  ReLU
Higher Intensities in depth 2:  ReLU
Higher Intensities in depth 3:  ReLU
Higher Intensities in depth 4:  ReLU
Higher Intensities in depth 5:  Conv
Higher Intensities in depth 6:  ReLU
Higher Intensities in depth 7:  Conv
Higher Intensities in depth 8:  Conv
Higher Intensities in depth 9:  ReLU
Higher Intensities in depth 10:  ReLU
Higher Intensities in depth 11:  ReLU
Higher Intensities in depth 12:  ReLU
```

$Texture\ synthesis$:

בשאלה הזו אנחנו נדרשים לשנות את האלגוריתם כדי לייצר טקסטורות לפי תמונת רפרנס של טקסטורה. למעשה המשימה הפעם היא הפוכה מסעיף קודם - לאפטם את הרעש רק לפי ה$StyleLoss$. בדומה לקודם, מימשנו את זה על ידי אתחול של $ContentLoss$ ללא שכבות שיעשה עליהם השוואה, ועבור ה$StyleLoss$ חקרנו שכבות מכל העומק של הVGG, כל פעם ניסינו לסנתז את הטקסטורה לפי השוואה לשכבה אחת.
התוצאות מראות שיש איזשהי שכבה (חמישית במקרה שלנו) שלפי העין - בה מתקבלת טקסטורה די דומה לטקסטורה שהיא הרפרנס. עמוק יותר ממנה - הטקסטורות שיוצאות דומות כבר לרעש, וגבוה יותר ממנה - הטקסטורות מטושטשות ופחות מפורטות. ואמנם, אפשר להסתכל על השכבות הראשונות לא כמטושטשות אלא למעשה כמעידות על הפרטים הגסים יותר של הטקסטורה.
בנוסף, שמנו לב להבדלים נראים לעין בין המקרה של השוואה משכבות קונבולוציה למקרה של השוואה משכבות ה$ReLU$ - במקרה של $ReLU$ אנחנו מקבלים תמונות מורעשות **מוקדם יותר**, כלומר מתרחקים מהאינפוט מוקדם יותר. אפשר להסביר את זה מהעובדה שבארכיטקטורה של הרשת המקורית, ה$ReLU$ מגיעות אחרי קונבולוציה. התוצאות לפי השוואה משכבות קונבולוציה:



ולפי השוואה משכבות $ReLU$:



$MeanVar\ Style\ Loss$:

בשאלה הזאת נדרשנו לייצר $StyleLoss$ שמחשב את השגיאה לא לפי ה$Gram\ Matrix$ של האקטיבציה ושל ה$target$, אלא לפי מטריקה חדשה - $MeanVar$.
היינו מצפים לראות בשיטה הזאת תוצאות שמצליחות פחות לשחזר את הסטייל של התמונה: גם בגלל שה$MeanVar$ מספק פחות מידע באופן נומרי (שני משתנים לעומת מטריצת $Gram$), וגם מכיוון שהוא מספק פחות מידע מרחבי (מטריצת $Gram$ מכילה את מידת הקורלציה בין פיקסלים באופן **מקומי**). ביצענו את הניסוי עם משקולים שונים של הסטייל, ולמרות שגם כאן התוצאות די יפות - ניתן לראות את האפקט שדיברנו עליו: הטקסטורות פחות מגוונות ומשתנות במרחב מאשר במקרה של $Gram-based$.
זו התמונה המקורית שיוצאת מה $Style-Transfer$ עם חישוב סטייל לפי $Gram\ matrix$:

dancing in picasso style

ואלו תוצאות לפי חישוב $MeanVar$ ועם משקולים שונים:

dancing in picasso style - meanVar Loss style_weight100

dancing in picasso style - meanVar Loss style_weight1000

dancing in picasso style - meanVar Loss style_weight5000

dancing in picasso style - meanVar Loss style_weight10000

dancing in picasso style - meanVar Loss style_weight20000

חלק 2

$Style\ Loss{+}Reference\ vs.\ Classifier\ Guided\ Diffusion$

בשאלה הזו נדרשנו לממש אלגוריתם של $Classifier-Guidance\ Diffusion$, שמשתמש ב$Stable\ Diffusion$, מודל שמקבל טקסט והופך אותו לתמונה. הוא משתמש במודלים שיודעים לעשות $embedding$ לטקסט, ומשתמש בקידוד שהם מוציאים כ$target$ שלפיו הוא מאפטם את הלוס. הגרסא של $Classifier-Guided$ היא למעשה שילוב של הנחיה נוספת, חיצונית, אל תוך תהליך סינתוז התמונה: במקרה שלנו, השתמשנו ב$Style\ loss\ model$ שהוגדר בחלק הראשון, על מנת לחשב את ה$StyleLoss$ בכל שלב של הסנתוז, לחשב את הגרדיאנטים שלהם, ולעשות אופטימיזציה גם בעזרתם בעזרת הנוסחה שהובאה באלגוריתם בתרגיל.

בסעיף הראשון נשווה את התמונות של הבלרינה בסטייל פיקאסו מהחלק הקודם.

הטקסט שהזנו לאלגוריתם היה $a\ ballerina$, וכתמונת הרפרנס נתנו לו את התמונת סטייל של פיקאסו. שיחקנו לא מעט עם הפרמטרים ועם האלגוריתם וראינו שאם המשקל של הסטייל הוא אדפטיבי - כלומר מאיטרציה כלשהי הוא גדל - התוצאות נראות בעינינו יותר יפה. בפועל התוצאה הטובה ביותר ניתנה על ידי משקול של הסטייל 0.6, והגדלה שלו פי 1.5 החל מהאיטרציה ה50. ה$gradient\ scale = 7.5$, והרצנו 100 איטרציות. התוצאה:

a ballerina_in_style_picasso

בשלב הזה שמנו לב שהתוכן יוצא לא מספיק מפורט. הוספנו תנאי שב20 האיטרציות הראשונות **אין** התחשבות ב$StyleLoss$, כדי שהמודל ייצר תוכן יותר טוב, ורק לאחר מכן

מתחילים להתחשב בסטייל. התוצאה:

a ballerina_in_style_picasso

אפשר להבחין שיש כאן באמת יותר פרטים: הפנים מפורטות, הלבוש, שתי רגליים...
בהשוואה לתוצאות מהחלק הקודם, נתייחס אל החיקוי של הסטייל ולא אל התוכן:
המודל דיפוזיה משיג תוצאות פחות דומות לסטייל, במיוחד בתמונה השנייה (שזה הגיוני, כי בראשונה התחשבנו בסטייל לכל אורך הדרך).
הדמיון לסטייל מתבטא בעיקר בצבעים, ופחות בטקסטורה. המוקומות שבהם הטקסטורה בכל זאת התקרבה לרפרנס, הם ברקע - איפה שאין הרבה תוכן. במקומות האלה, המודל דיפוזיה השיג טקסטורה יחסית דומה למה שהמודל $MeanVar$ בחלק הקודם השיג: טקסטורה צפופה יחסית, כלומר לא משתנה הרבה בכל התמונה.

$Classifier - Guidance\ vs.\ Regular\ Diffusion$

בשאלה הזו התבקשנו להשתמש ב$Stable\ Diffusion$ כמו שהוא, בלי $Classifier - Guidance$ של ה$StyleLoss$ (כדי לכוון את המודל לסטייל במקרה הזה, אנחנו מכניסים את הסטייל המבוקש כחלק מה$text\ prompt$), ולהשוות את התוצאות לתוצאות מהסעיף הקודם.
הרצנו את ה$prompt$ הבא: "$a\ ballerina\ in\ Picasso's\ Seated\ Nude\ style$". התוצאה היא:

a ballerina in picasso's seated nude style_in_style

הסטייל כאן לא דומה בכלל לסטייל שרצינו להשיג. למעשה נראה שהמודל פשוט צייר מישהו יושב ערום עם יד של בלרינה בסטייל שאופייני יחסית לפיקאסו. גם עם היפר פרמטרים שונים ו$prompt$ שונים (למשל $a\ ballerina\ in\ cubism\ style\ Picasso$) התוצאות יצאו יחסית דומות לסטייל הזה. הסבר אפשרי הוא שהקידוד של הטקסט ($embeddings$) מכיל מידע שהמקודד למד על העולם, וכך גם על פיקאסו, אבל הוא לאו דווקא מכיל מידע שמקשר בין היצירה הספציפית $seated\ nude$ לאלמנטים חזותיים שמאפיינים אותה (מילולית). מה גם, שייתכן שהמודל מעולם לא נתקל ביצירה הזו או בשם שלה, וייתכן שכל הקשר סמנטי של פיקאסו או קוביזם מוביל את המודל לתוצאות של סטייל יחסית דומות לאחד הנ"ל. בדיוק במקרים כאלה, הרעיון של $classifier-guidance$ יכול לבוא לעזרתנו: המודל לא צריך להכיר את הסטייל (או באופן כללי את ההנחיה הנוספת) כדי לנסות להתחקות אחריו מאחר שיש לו רפרנס ישיר. כמובן שדרושה מטריקה טובה דיה כדי שההדרכה תהיה משמעותית, כמו למשל ההבדלים ב $Gram\ matrix$ במקרה שלנו.