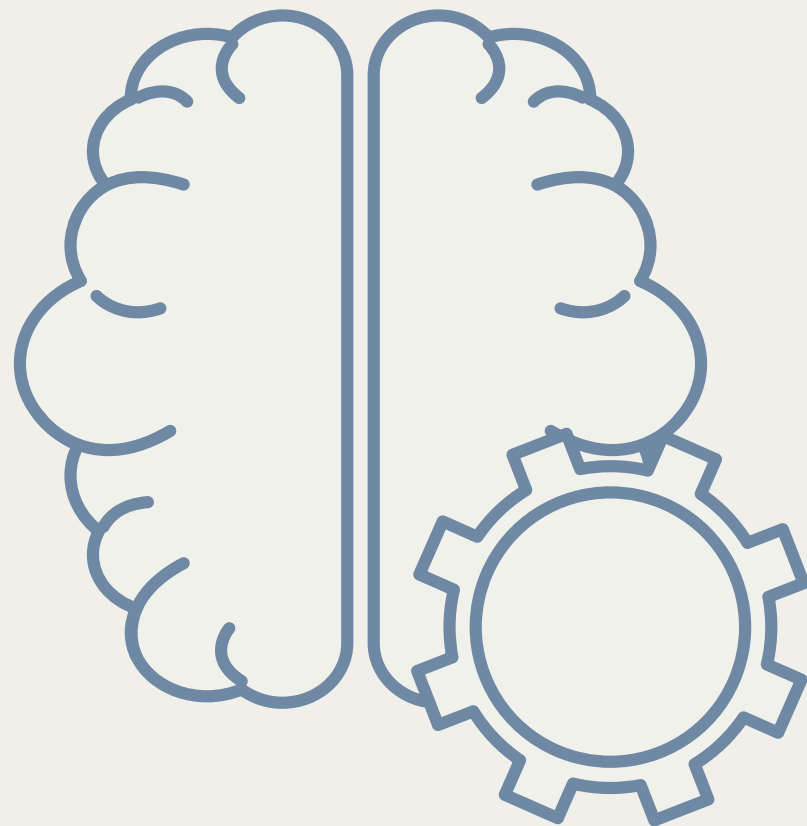


# Improving MLP-Mixer Performance on Small Image Datasets Using Locality-Aware Modifications

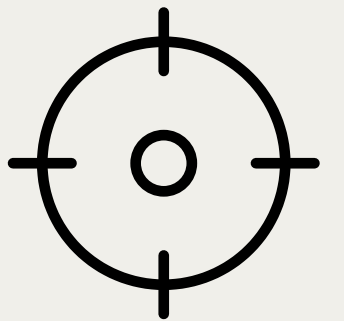
שיפור ביצועים על מאגרי תמונות קטנים באמצעות  
הוספת הטיה מרחבית

---

מגישים: גורג קנאזע, עמאר מנאע, אדם עיסא



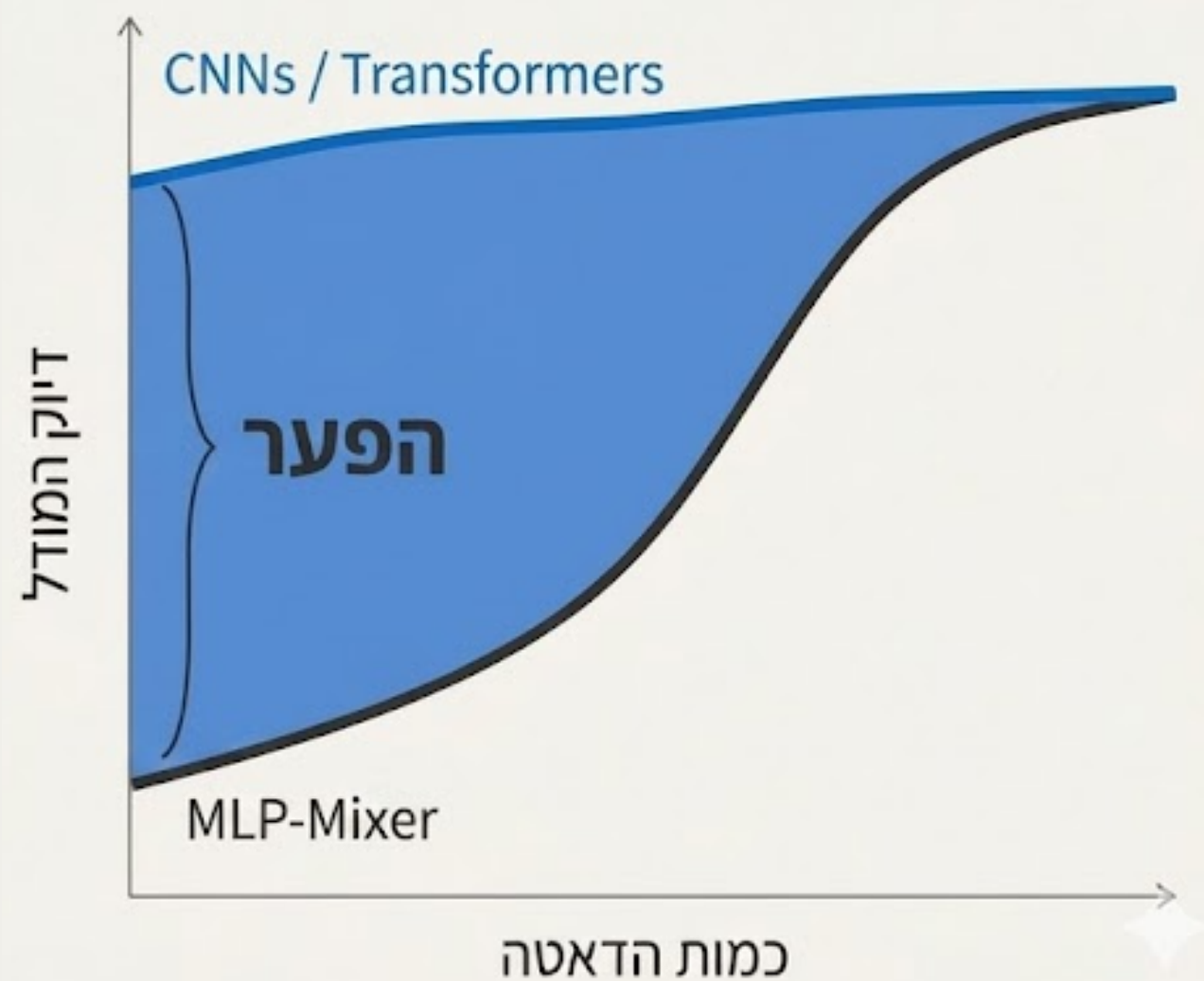
**מטרת הפרויקט:** שיפור הדיוק של ארכיטקטורת MLP-Mixer על מאגרי מידע קטנים  
(CIFAR-10/100)



**הגישה המרכזית:** החדרת הטיה אינדוקטיבית (inductive bias) מרחבית מינימלית באמצעות שכבת קונבולוציה יעילה (depth & point wise convolution) שלא מפירה את המוטיבציה של מבנה הרשת ומבלי להוסיף כמות גדולה של פרמטרים נוספים.

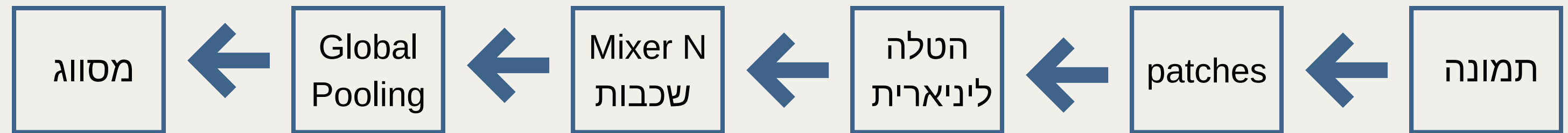


# המוטיבציה: הפער בין ביצועים על דאטה גדול לקטן



- ההבטחה של MLP-Mixer: ארכיטקטורה חדשנית שמתחרה ב-CNNs וב-Transformer ללא שימוש בקונבולוציות או attention.
- המציאות: המודל משיג ביצועים מעולים רק כאשר הוא מאומן על מאגרי מידע עצומים (כמו JFT-300M).
- הבעיה: על מאגרי מידע קטנים (כמו CIFAR), ביצועיו נמוכים משמעותית ממודלים קלאסיים.
- ההזדמנות: קיים פער ברור המזמין שיפורים ממוקדים כדי להפוך את המודל לרלוונטי גם בסביבות דלות-מידע.

# כיצד MLP-Mixer עובד?



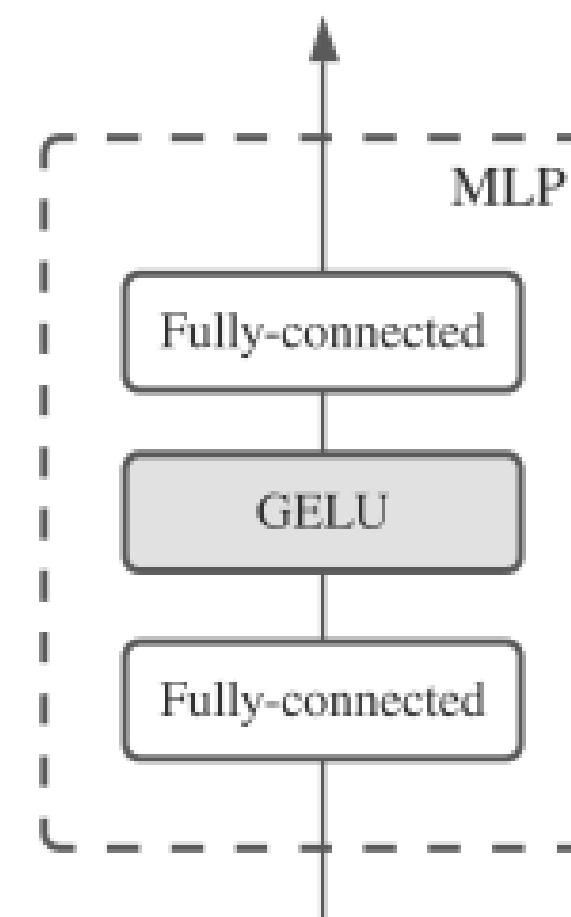
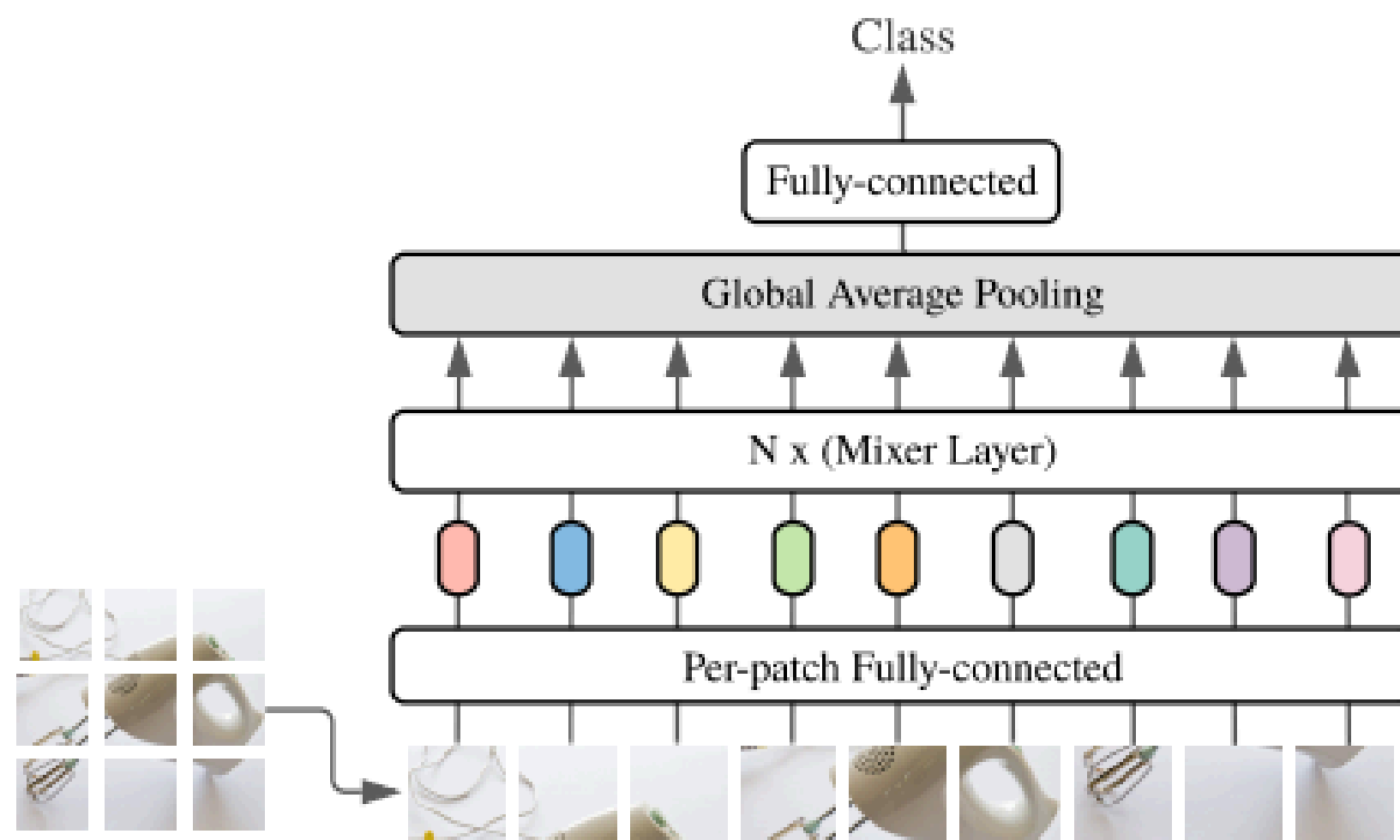
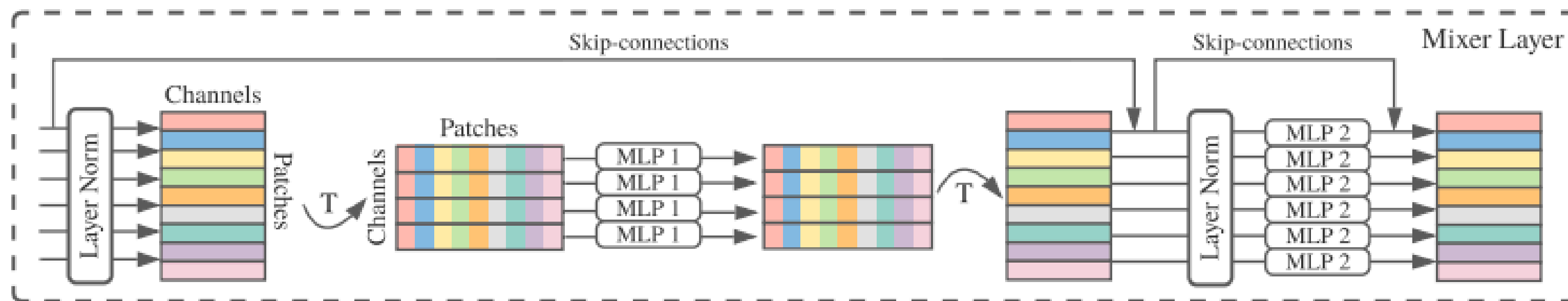
## 1. Channel-Mixing MLPs

- רשת MLP שפועלת על כל טלאי בנפרד.
- מערבבת תכונות/ערוצים בתוך אותו טלאי.
- אינטואיציה: כמו "עיבוד תכונות" מקומי בתוך הטלאי.

## 2. Token-Mixing MLPs

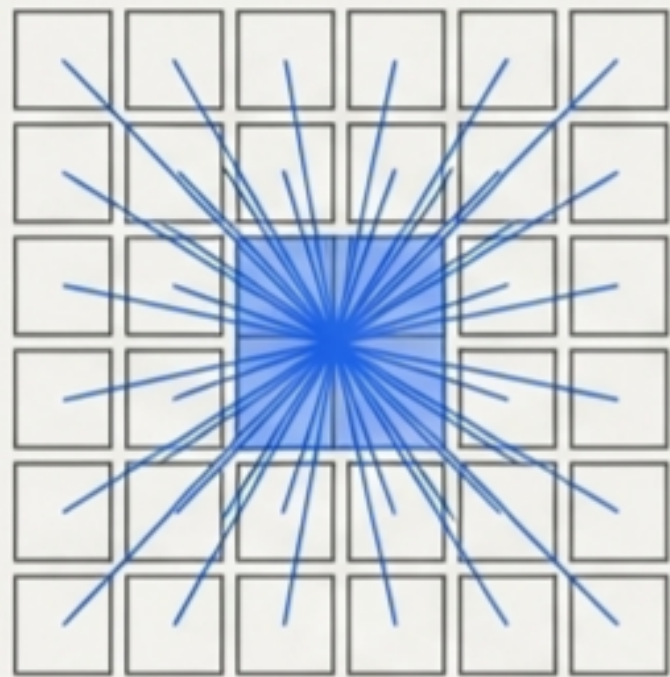
- רשת MLP שמערבבת מידע בין הטלאים בתמונה.
- מאפשרת תקשורת מרחבית בין אזורים שונים.
- נקודה חשובה: הערבוב הוא גלובלי, בלי "חוק שכנות" מובנה.

# ארכיטקטורה מלאה של הרשת



# מדוע היעדר לוקאליות פוגע בביצועים על מידע מועט?

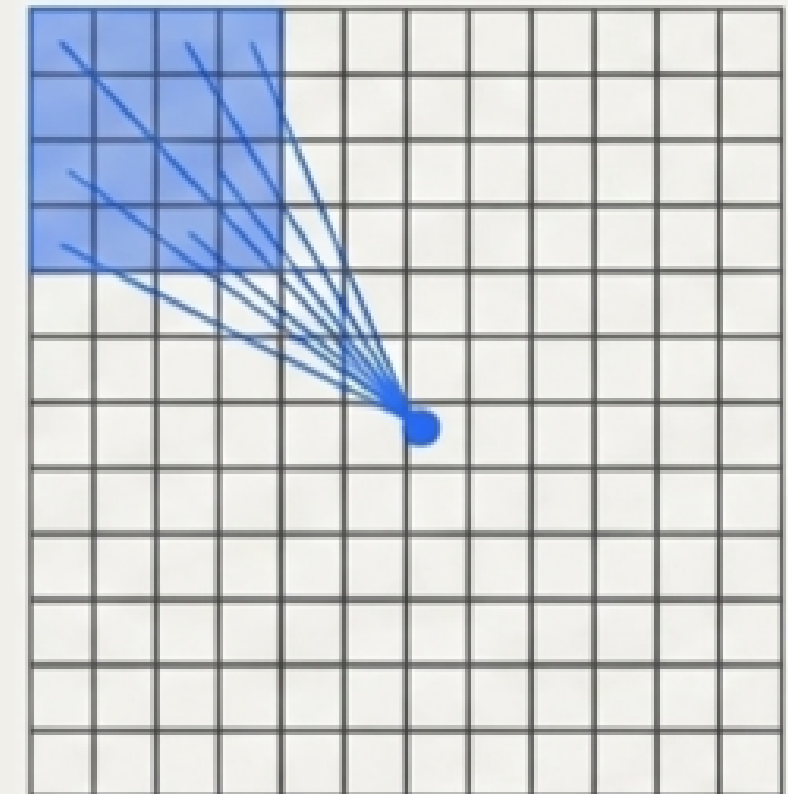
MLP-Mixer - למידה מאפס (Learning from Scratch)



- ה-Mixer "רואה" את כל התמונה בבת אחת, אך אינו יודע אילו טלאים שכנים.
- הוא צריך ללמוד את מושג ה"שכנות" והמבנה המרחבי מתוך כמות עצומה של דוגמאות.

בלי מספיק דאטה, ה-Mixer מתקשה ללמוד את הקשרים המרחביים ש-CNN מקבל ב"מתנה"

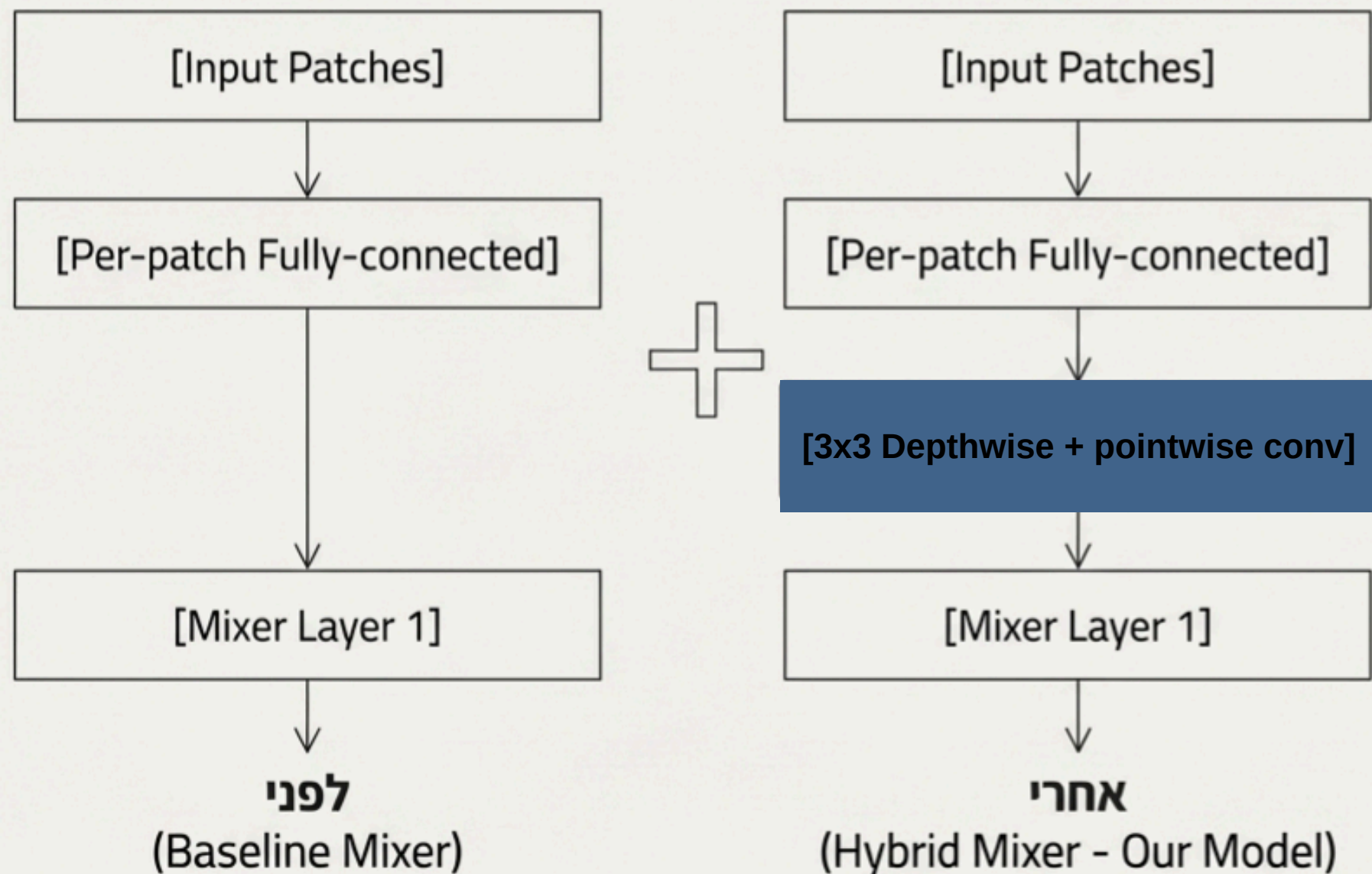
CNN - הטיה מובנית (Built-in Bias)



- CNNs מתוכננים מראש עם ההנחה שפיקסלים קרובים קשורים זה לזה.
- הפילטר לומד תבניות מקומיות (קצוות, טקסטורות) באופן טבעי ויעיל.
- הנחה זו (הטיה אינדוקטיבית) חוסכת למודל את הצורך "ללמוד" את חשיבות המבנה המרחבי מאפס.

# הרעיון שלנו : הזרקת לוקאליות בשינוי מינימלי

Before vs. After



**השינוי:**

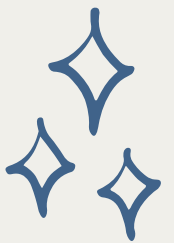
אנו מוסיפים שכבת קונבולוציה Depthwise בגודל  $3 \times 3$  ואז מפעילים קונבולוציה Pointwise בגודל  $1 \times 1$ , מיד אחרי הטמעת הטלאים (patch embedding).

**ההיגיון:**

הפעלת קונבולוציה עומקית (Depthwise) ולאחריה קונבולוציית נקודתית (Pointwise) מאפשרת תחילה לכל טלאי לתקשר עם שכניו המרחביים הקרובים, ובשלב הבא לשלב מידע בין ערוצים, וכך ליצור ייצוג ראשוני בעל מודעות מרחבית וסמנטית לפני הכניסה לבלוקי ה-Mixer הגלובליים.

# מטרות הפרויקט והשערת המחקר

מטרה מרכזית לשפר את רמת הדיוק (accuracy) ויכולת ההכללה (generalization) של ארכיטקטורת MLP-Mixer על מאגרי מידע קטנים (כדוגמת CIFAR-10/100).

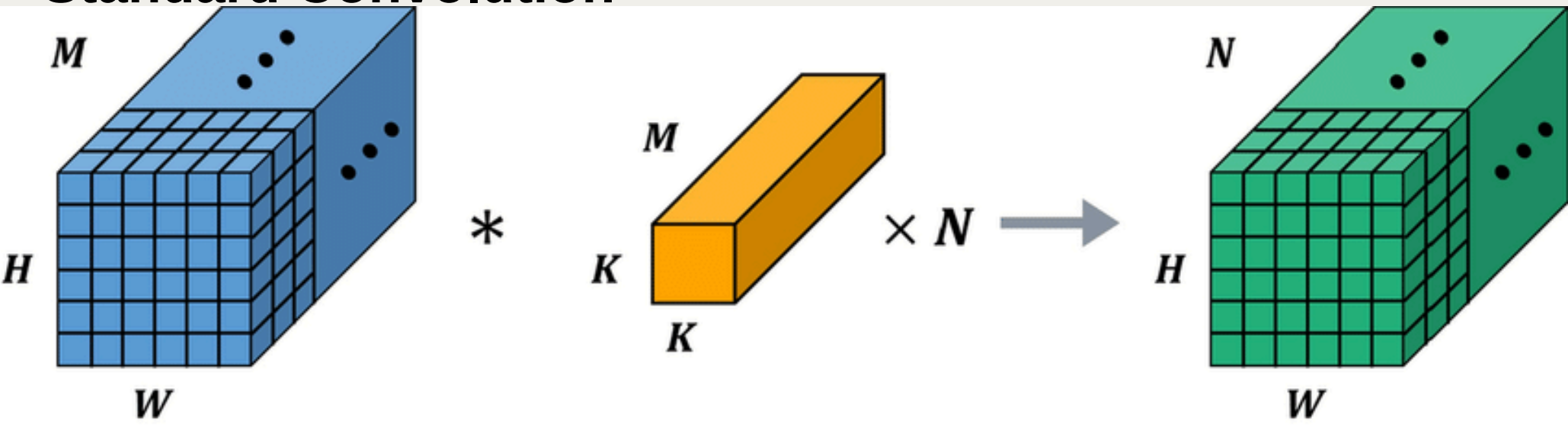


**השערת המחקר :** תוספת של שכבת קונבולוציה קלת-משקל (lightweight) תספק למודל ה-Mixer את ההטיה המרחבית החסרה לו, ותוביל להפחתת התאמת-יתר (overfitting) ולעלייה בביצועי הכללה במצב של מיעוט דאטה.





Standard Convolution



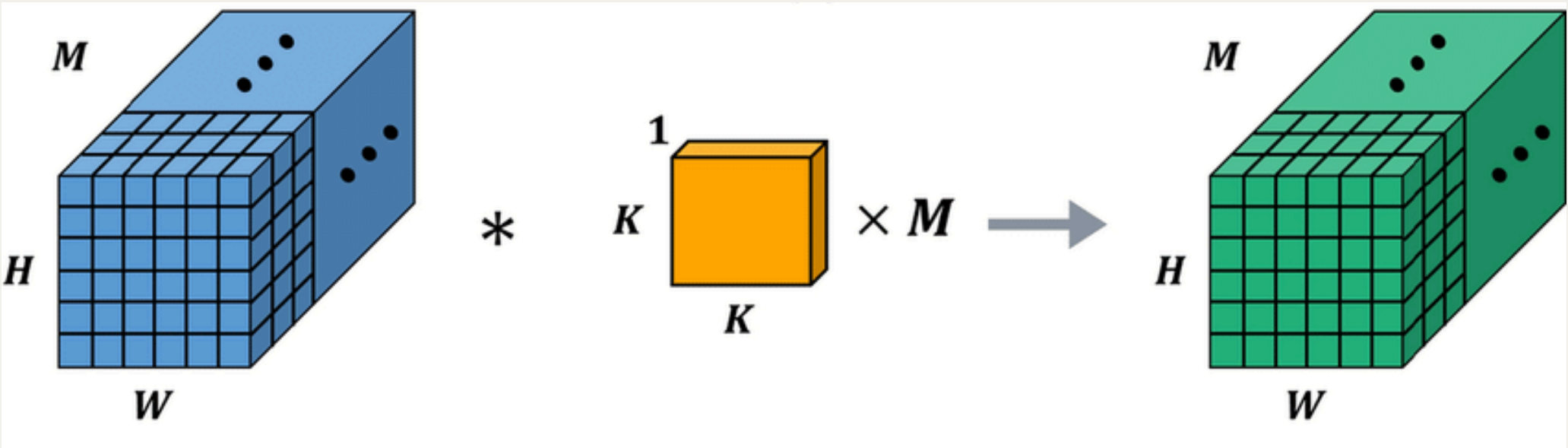
# Standard Vs Depthwise Vs Pointwise Convolution

# of params:  $\text{Kernel\_size} * \text{Kernel\_size} * \text{in-channels} * \text{out-channels}$

$3*3*128*128=147,456$

Depthwise Convolution

# of params:  
 $\text{Kernel\_size} * \text{Kernel\_size} * \text{channels}$

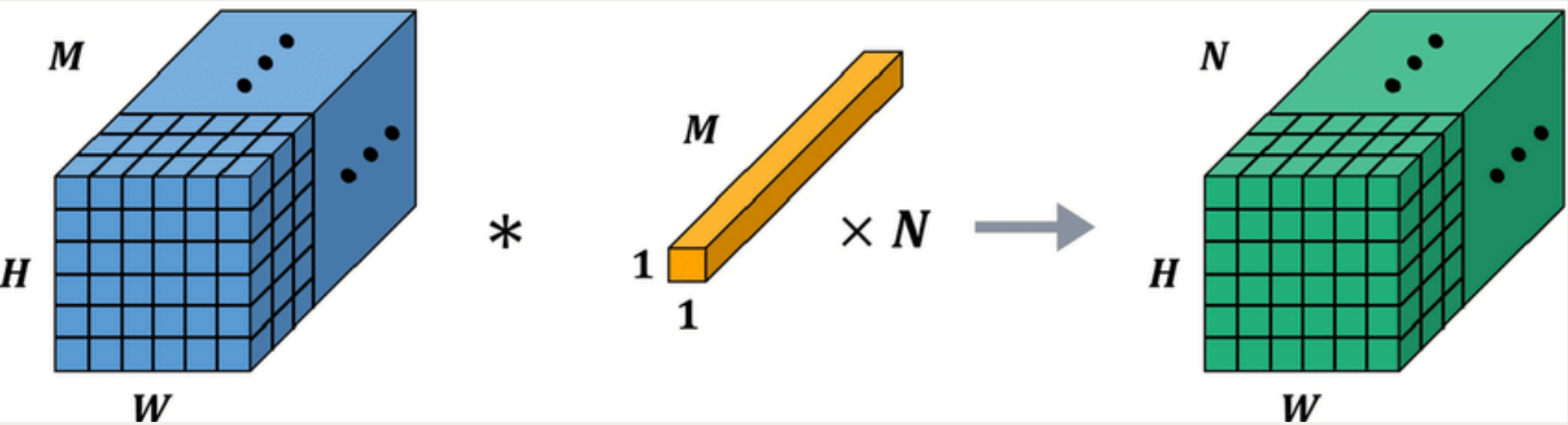


$3*3*128 = 1,152$

$128*128 = 16,384$

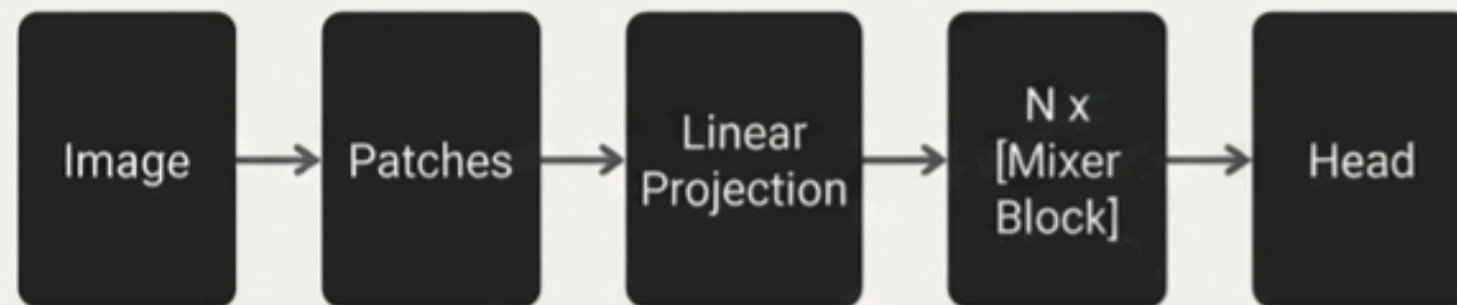
Pointwise Convolution

# of params:  
 $\text{channels} * \text{channels}$



# השוואת ארכטקטורות: Baseline VS Hybrid

## MLP-Mixer (Baseline)



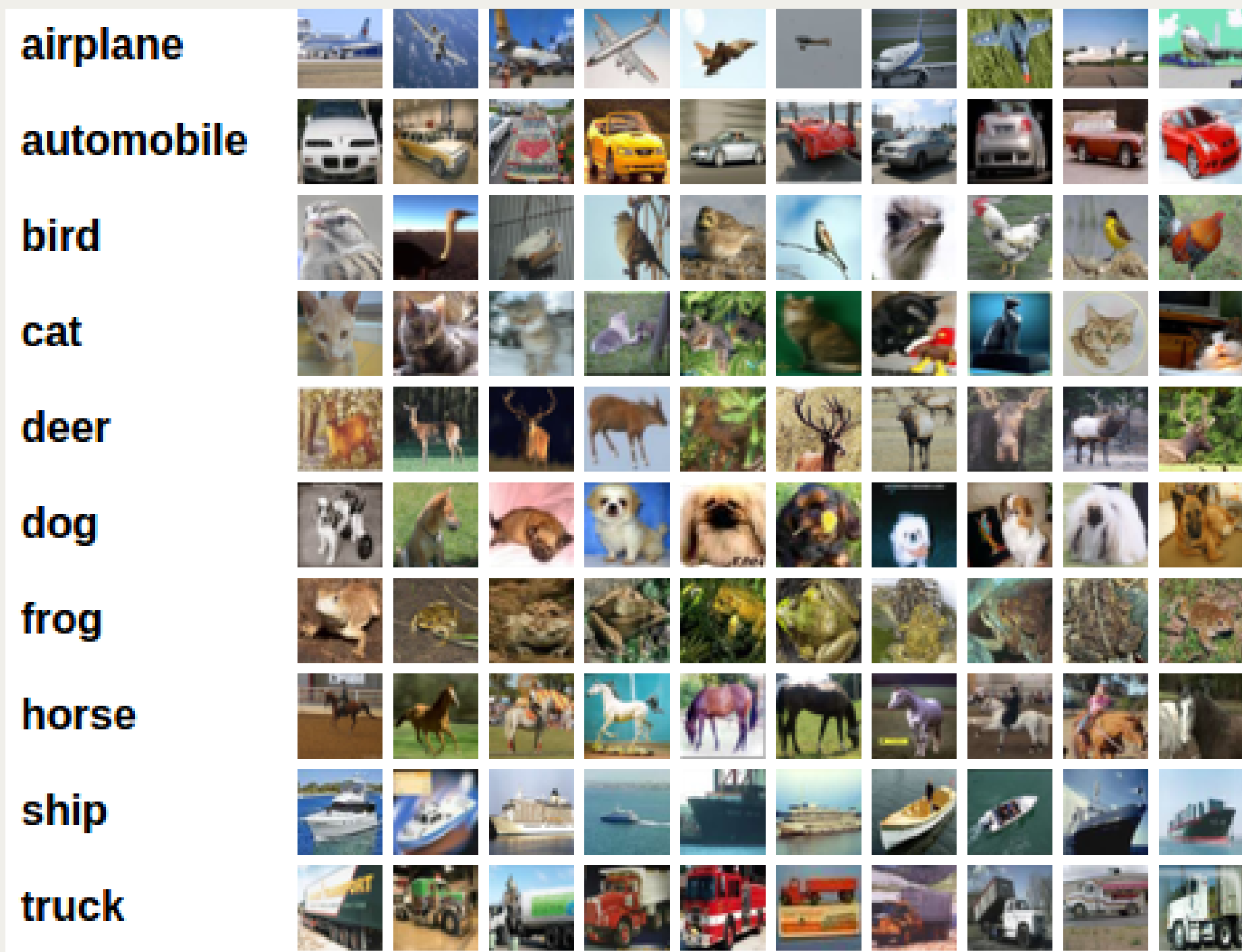
שמירה על הצורה (Shape):  
הקונבולוציה אינה משנה את  
מספר הטוקנים או המימד  
 שלהם, ומשתלבת באופן  
שקוף.

## Hybrid Mixer (Our Model)



שינוי יחיד וממוקד: התוספת  
מתרחשת לפני שכבות ה-  
Mixer, ומכינה עבורן ייצוגים  
'מודעי-מרחב'.

# מאגר הנתונים



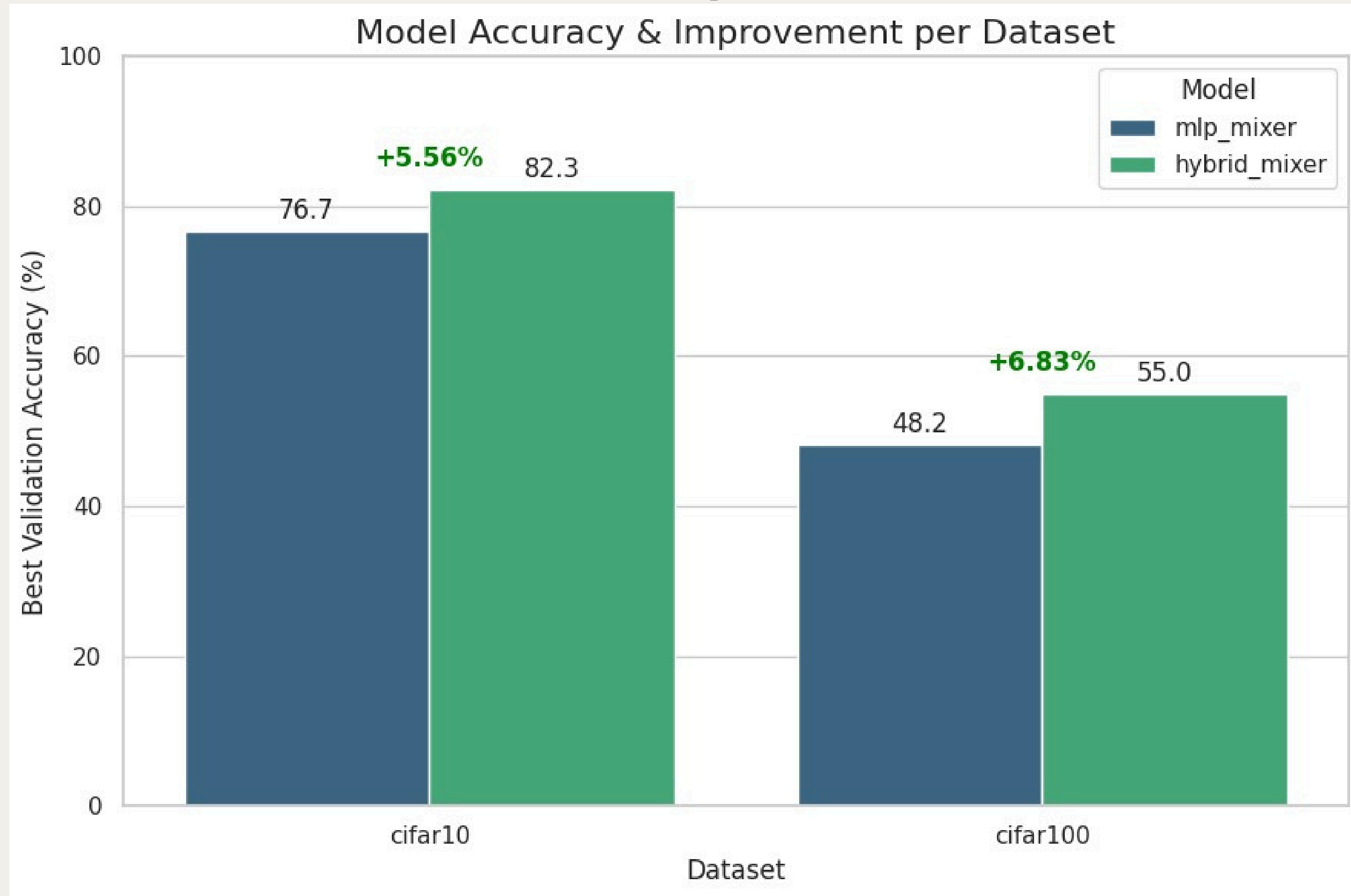
## ראשי: CIFAR-10

- מאגר סטנדרטי להשוואת מודלים (benchmark), המכיל 50,000 תמונות אימון ב-10 קטגוריות.
- גודלו המוגבל אידיאלי לבחינת ההשערה שלנו.

## משני: CIFAR-100

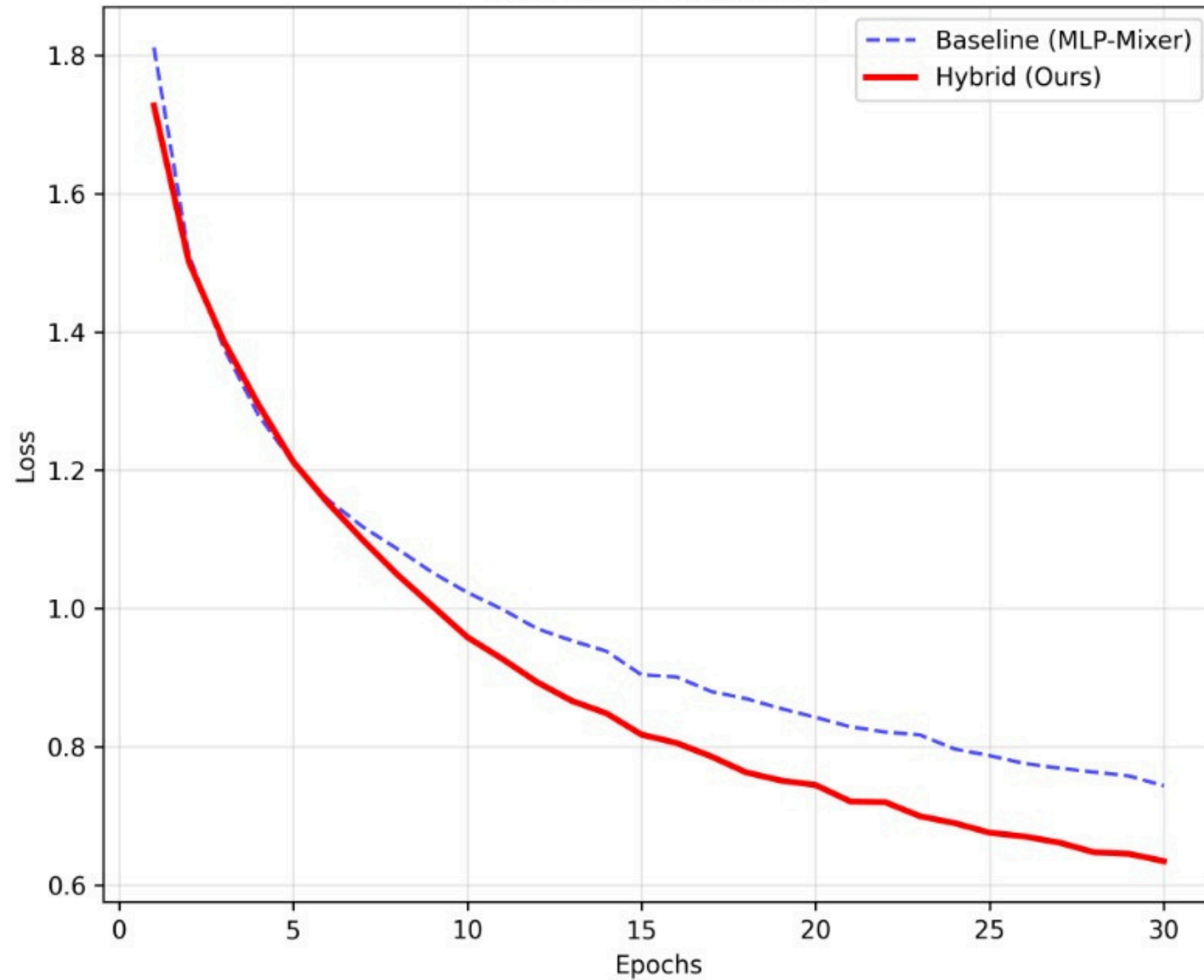
לבחינת יכולת ההכללה על בעיה מורכבת יותר.

# תוצאות: שיפור משמעותי בדיוק הסיווג

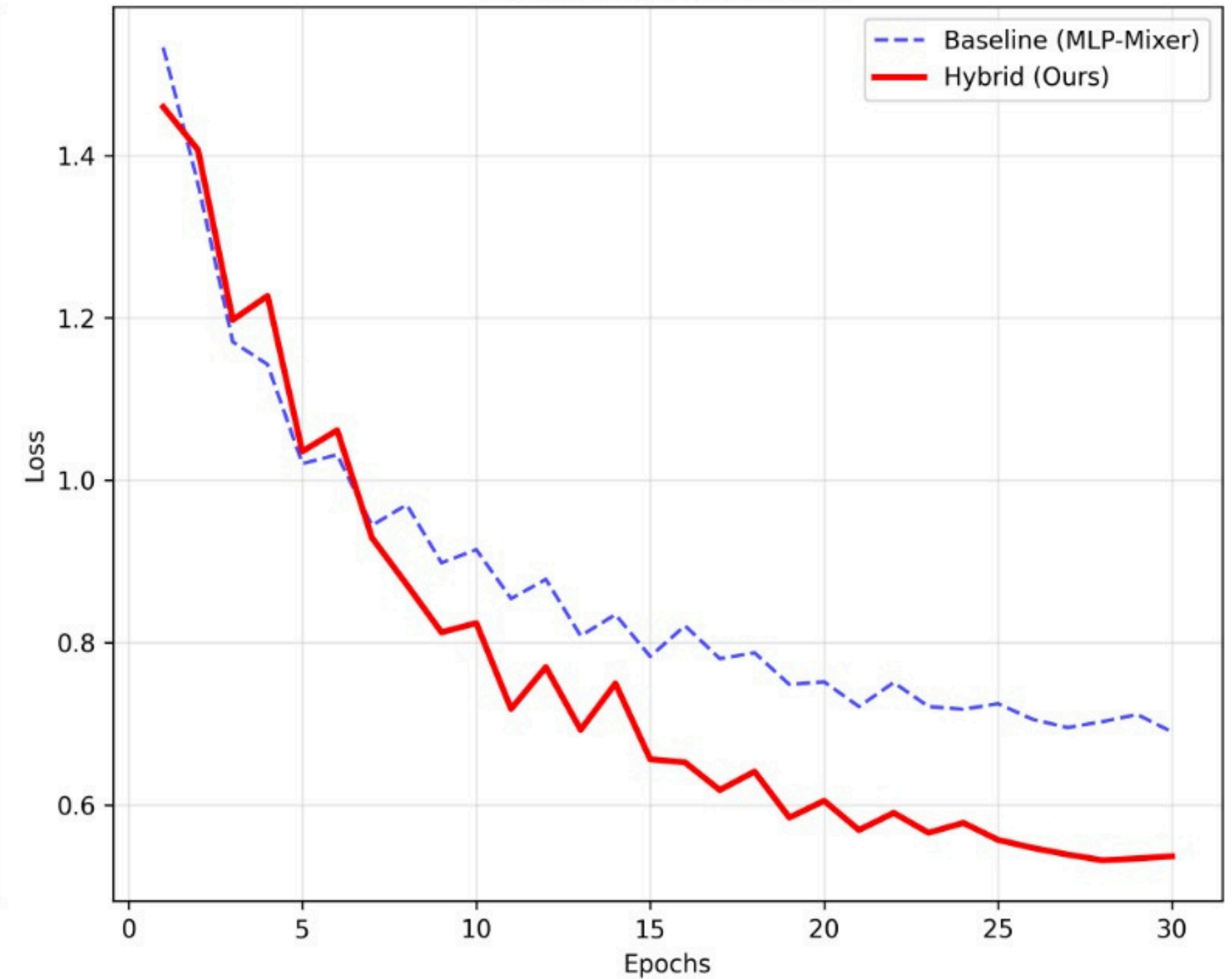


# השוואת ביצועים של המודלים (cifar 10)

Training Loss (CIFAR10)

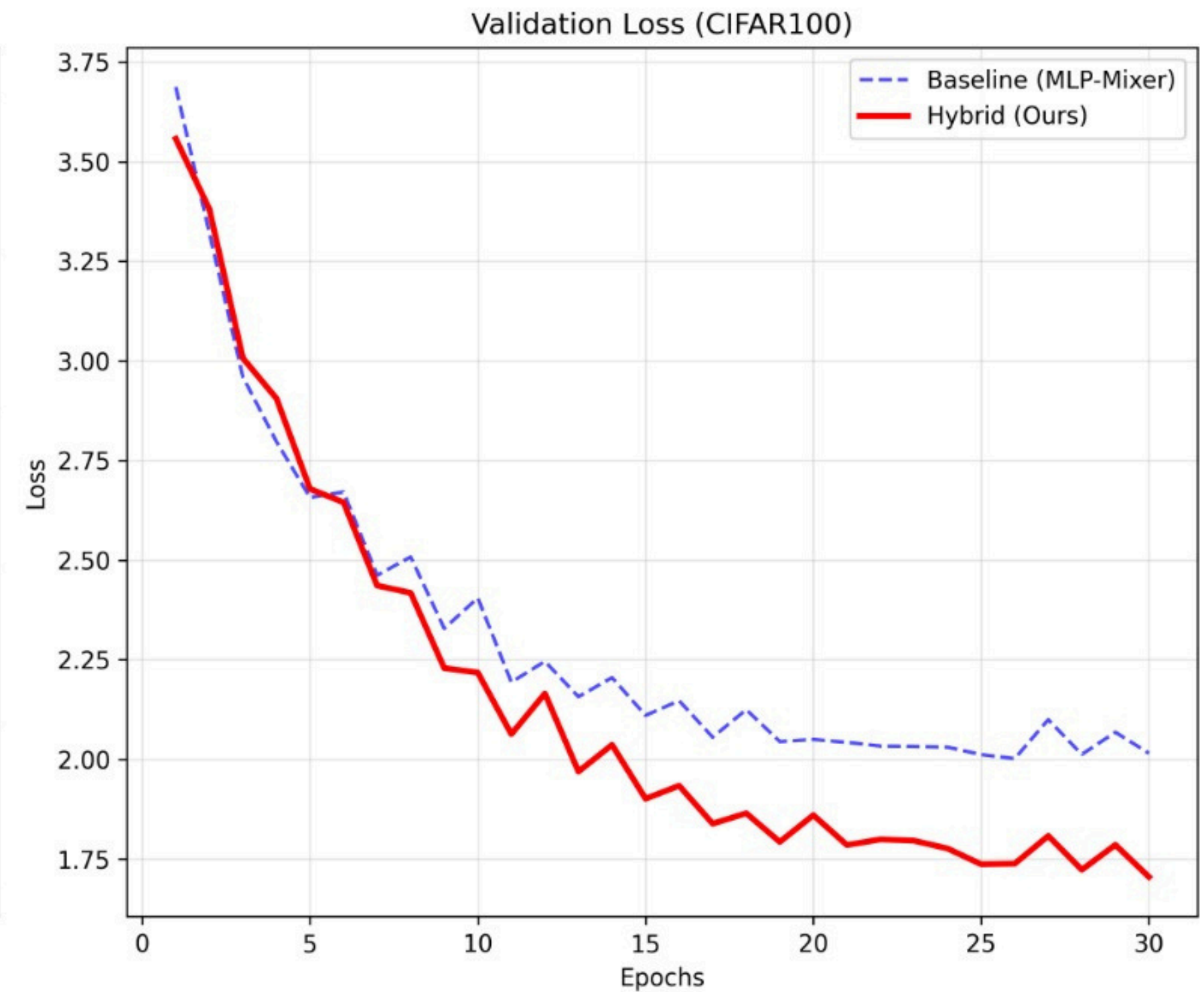
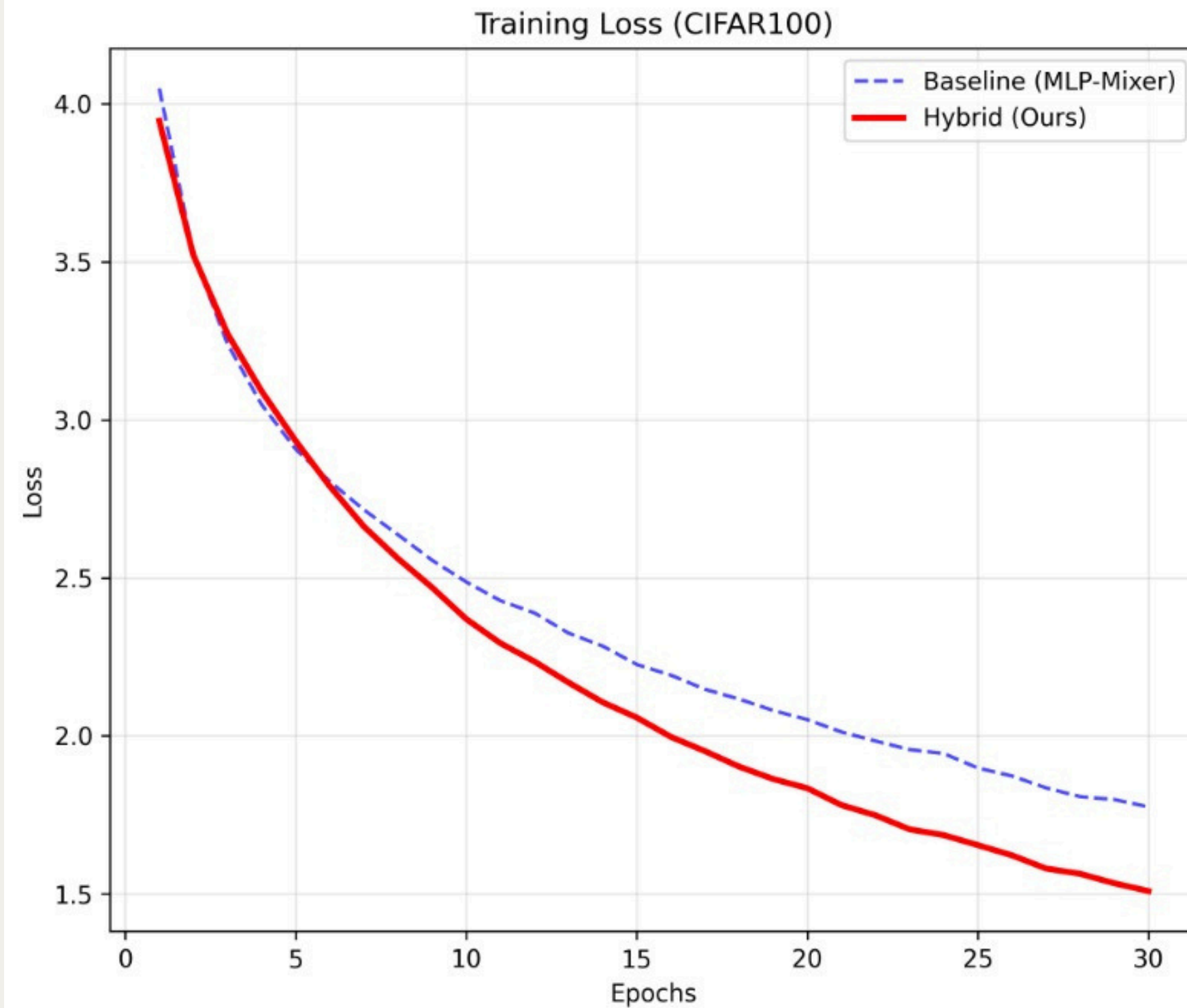


Validation Loss (CIFAR10)





# השוואת ביצועים של המודלים (cifar 100)



# מגבליות

נבדקה רק תוספת של  $3 \times 3$  DW Conv.  
ייתכנו מנגנוני הטיה אחרים  
(למשל  $5 \times 5$ , או Pooling).

הניסויים נערכו רק על מאגרי מידע  
קטנים (CIFAR). לא ידועה ההשפעה על  
מאגרים גדולים יותר כמו ImageNet.

