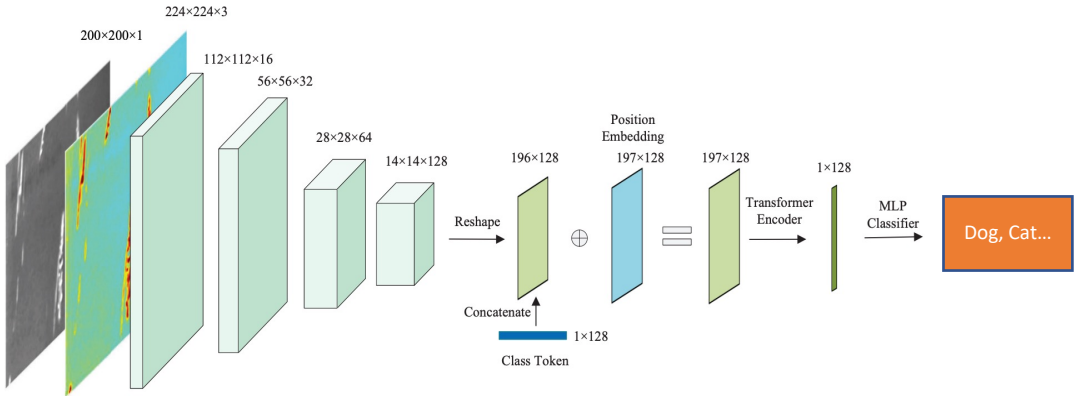
Robust Vision-

Making CNN and ViT Good Friends with Pre-Trained Vision Models

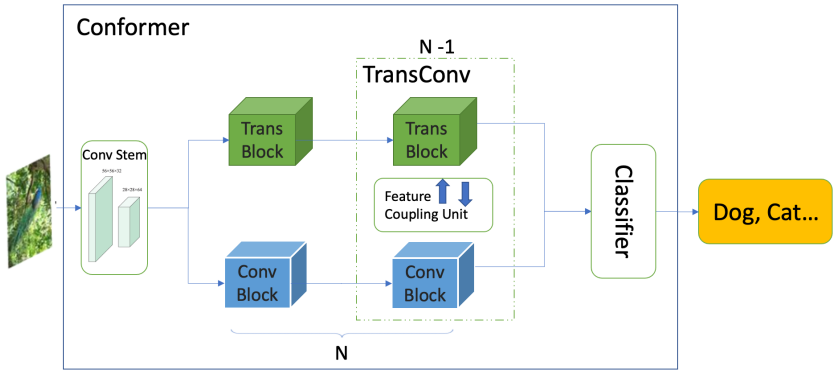
Haichao Wei, Ethan Cheng, Chunming Peng

Background / Introduction

- *There has been a competition between Convolutional Neural Networks (**CNN**) and Vision Transformers (**ViT**) in computer vision field.*
- *While **CNNs** are good at extracting local features due to its inductive bias leading to better generalization, **ViT**-like models can capture long distance feature interactions and are good at learning global representations, but ViT paper note that Vision Transformer has much less image-specific inductive bias than CNNs, and therefore is data hungry.*
- ***Is there a way of having best of both worlds in one solution? Hybrid network architecture?***
- *Recent Works of hybrid architecture: ViT [12] did experiments using several Serial hybrid ViT (CNN → Transformer). Conformer [26] proposed the first dual structure which has initial CNN based stem modules followed by dual structure of stacked CNN blocks and stacked Transformer blocks.*
- *However, the model of this hybrid approach is too complex, therefore makes it even harder to optimize and again data hungry.*



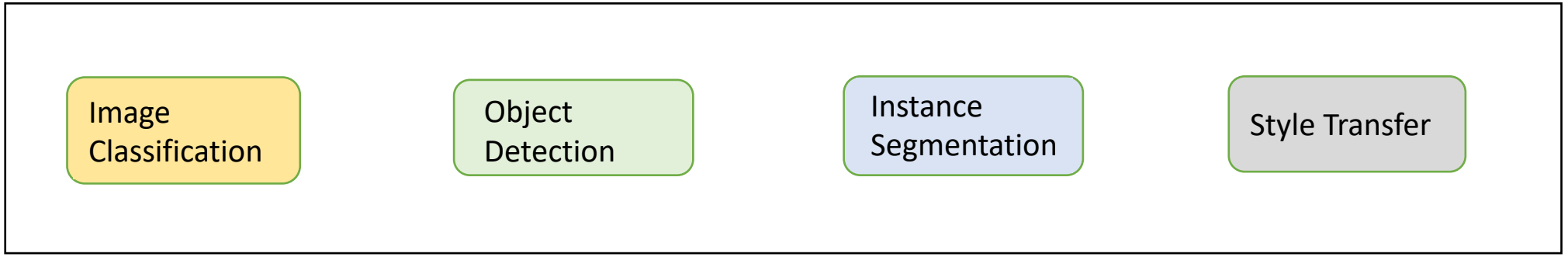
hybrid ViT (CNN → Transformer [source: hybrid arch.-Shunfeng Li etc.]



dual structure [source: conformer– Zhiliang Peng etc.]

Problem statement

- *We want to explore a **new and simpler hybrid network architecture** that can have **best of both worlds in one** solution and still **easy to optimize with much less data.***
- ***Pretrained Vision Model** sounds like a solution, as in both **CNN** and **ViT** , the pretrained model shows superior performance in downstream tasks and makes it easy to optimize with much less data.*
- ***A new architecture:** CNN + ViT + Pretrained Vision Model*
  - *Can we invent a **hybrid network architecture** that combines **CNN** and **ViT** together and easily leverages **Pretrained vision model?***
- ***Evaluation**:*
  - *The key value of this new architecture is it should work reasonably well for a **wide range of computer vision tasks** with much less data and much less training time, so we want to evaluate it with:*



Dataset

- *To evaluate our new architecture, we will use ImageNet for image classification, MSCOCO for Object Detection, Instance Segmentation, Style Transfer*

Image Classification

IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

The ImageNet Large Scale Visual Recognition Challenge. (Source: Xavier Giro-o-Nieto)

Sample data

Object Detection

Instance Segmentation

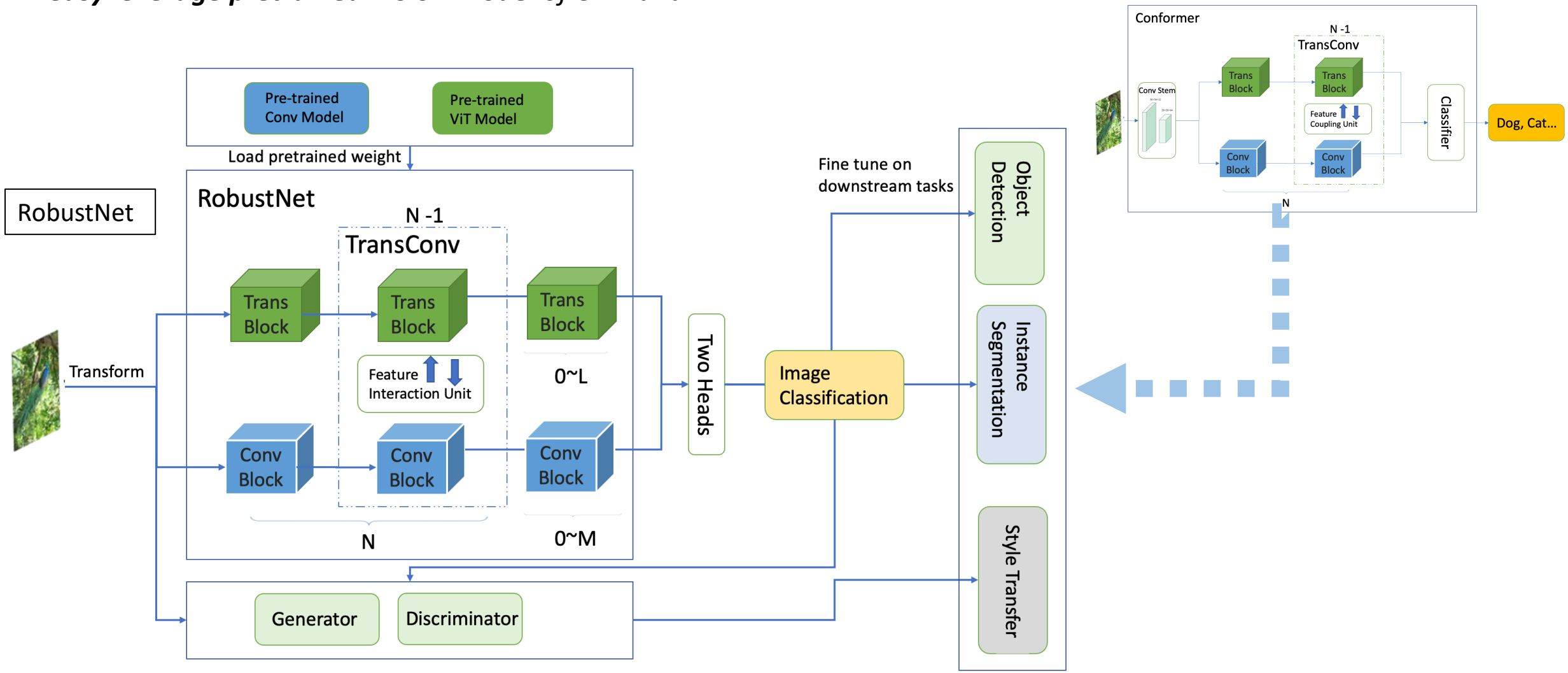
Style Transfer

MS COCO (http://cocodataset.org/)

Sample data

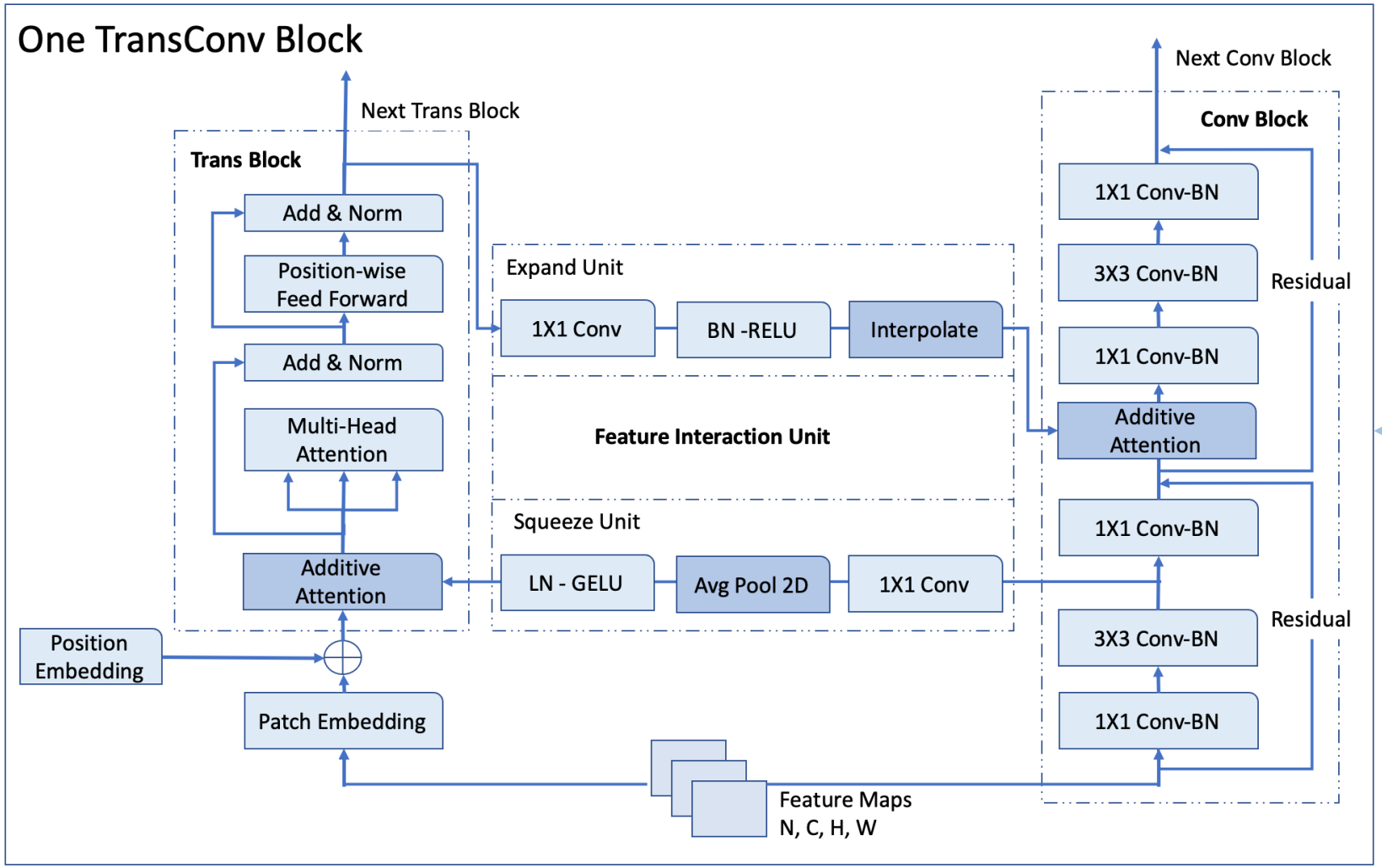
Method

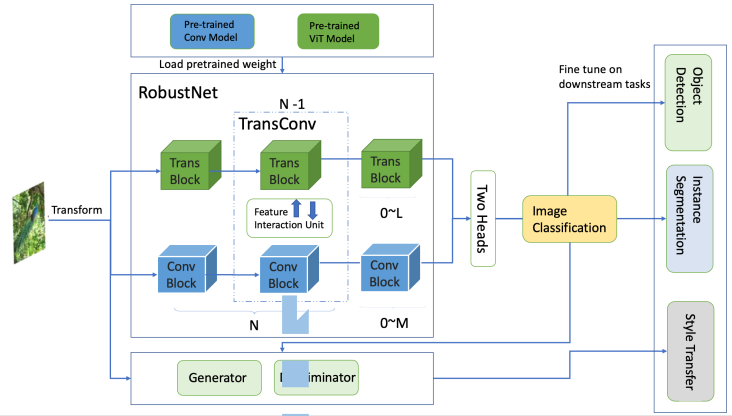
- *We propose a **new and simpler dual network architecture** that can have **best of both worlds in one** solution and can **easy leverage pretrained Vision Model** of **CNN** and **ViT**.*



Method

- *We propose a **new and simpler dual network architecture** that can have **best of both worlds in one** solution and can **easy leverage pretrained Vision Model** of **CNN** and **ViT**.*





Experiments & Analysis

- ***Experiment setting**:*
  - *We trained the model with image classification task and achieved state of art performance in ImageNet 1000, and then use the trained model as a backbone for object detection, and style transformation on MSCOCO 2017.*
  - *We have trained three variants on ImageNet and evaluated with the proposed RobustNet:* ***RN-small-patch16****(ViT- Base + Conformer-S),* ***RN-large-patch16****(ViT-Large + Conformer-B),* ***RN-base-patch14****(ViT-Huge + Conformer-B)*
- ***Experiment result:***
  - *Our results show that this proposed architecture can reach state-of-the-art results in Image Classification on Image-Net* ***with just 20 epochs, outperforming the original Conformer by 2.4%(****86.4% vs 83.6****) on ImageNet** *under comparable # of parameters and architecture, achieved* ***similar results on MSCOCO for object detection with Conformer with just 12 epochs****. We also show that this approach works across different tasks Style Transformation via fine tuning.*

Model	Epochs	Top-1(%)
ResNet-50 [17]	-	76.2
ResNet-101 [17]	-	77.4
ViT-B [12]	-	77.9
ViT-L [12]	-	76.5
Conformer-S [26]	300	83.4
Conformer-B [26]	300	84.1
MAE-ViT-Base [16]	-	83.6
MAE-ViT-Large [16]	-	85.9
MAE-ViT-Huge [16]	-	86.9
RN-small-patch16	20	83.6
RN-large-patch16	20	85.4
RN-base-patch14	20	**86.5**

Table 2. Top-1 accuracy for image classification on the ImageNet validation set.

Model	Epochs	AP(%)
ResNet-50 [17]	-	38.2
ResNet-101 [17]	-	40.0
Conformer-S [26]	20	43.6
Conformer-B [26]	20	44.9
RN-small-patch16	12	**44.6**

Table 3. Performance for object detection on the MSCOCO mini-val set. Other Results are reported by the mmdetection library or Conformer Paper

Losses	C+T	ArtFlow	MCC	AAMS	AdaIN
$L_{content}$	2.17	2.13	2.38	2.44	2.34
L_{style}	2.42	3.08	1.56	3.18	1.91

Table 4. QUANTITATIVE COMPARISONS. WE COMPUTE THE AVERAGE CONTENT AND STYLE LOSS VALUES OF RESULTS BY DIFFERENT METHODS TO MEASURE HOW WELL THE INPUT CONTENT AND STYLE ARE PRESERVED.

FiD per embedding	C+T	MSG	StyleGAN	PGGAN
CELEBAHQ	0.0136	0.008	0.009	0.012
FFHQ	0.01475	0.009	0.010	-
LSUN-BEDROOM	0.0533	-	0.012	0.037
LSUN-CHURCH	0.04117	0.030	0.067	0.030

Table 5. QUANTITATIVE COMPARISONS: FID VALUES COMPUTED WITH DIFFERENT EMBEDDINGS.

Experiments & **Analysis**

- ***Why converge so fast:***
  - *Due to limited computation budget, we did the analysis on smallest model variance t**RN-small-patch16**(**ViT-Base** + **Conformer-S**)*
  - *Model starts with Test Accuracy@1 ~ 79% even in the first epoch! Attributed to the pre-trained models and carefully designed structure, note that the two pretrained models of CNN and ViT are also on ImageNet.*
    - *Observation 1: both CNN (head 1) and Conv (head 2) are learning in the same pace.*
    - *Observation 2: the model performance drops to 76% in epoch 6 and then steadily improves afterwards. It could be because the FIU (feature interaction unit) starts to learn and stabilizes at epoch 6, when the rest of the network starts to learn.*

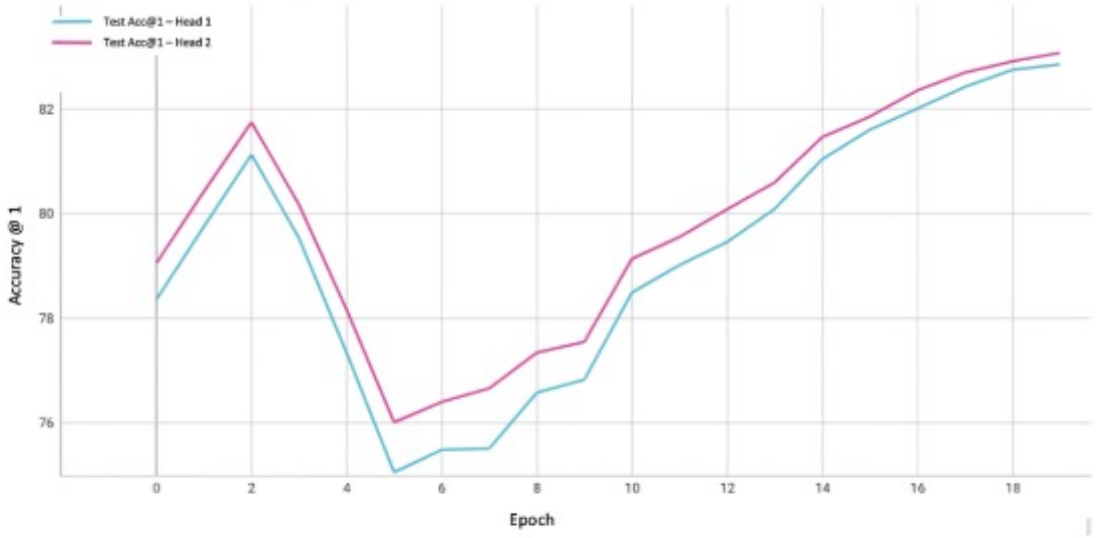


Figure 8. Why it converges so fast - Test Accuracy@1 for both branches during training

Experiments & **Analysis**

- ***Why drop – understand the learning of key component FIU(****feature interaction unit****):***
  - Before epoch 6, the weight distribution of the expand block of FIU fluctuated a lot, but after epoch 6, it is following the same distribution but steadily decays. This aligned perfectly with the test accuracy@1 curve. This indicates that the dual structure needs to learn a meaningful FIU and then steadily improves the performance. Other layers' weight histogram does not show this pattern.
  - Yes, FIU is probably the key of how the proposed model works!

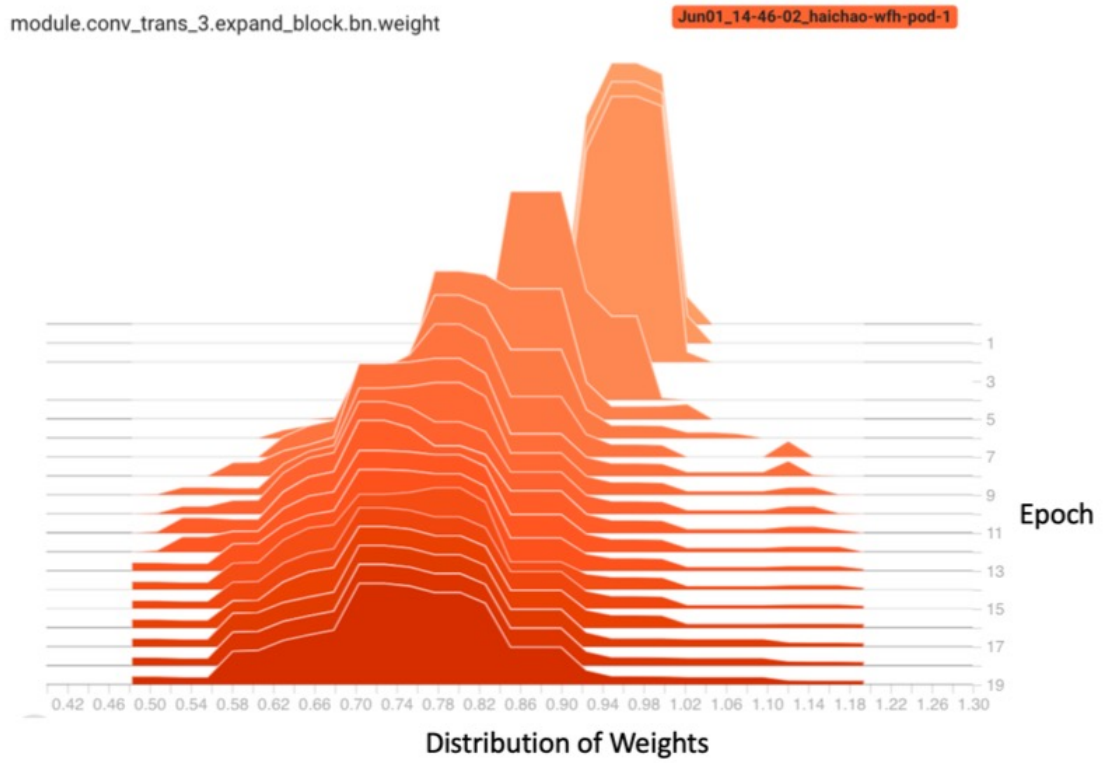


Figure 9. Understanding FIU via Weight Histogram

Experiments & **Analysis**

- ***Saliency Map:***
  - We found that dual structure as well as ViT only model does not show meaningful saliency map compared to CNN only architecture. Fig.-6 shows the saliency map of CNN-only which maps the class objects. However, Fig.-8 shows the saliency map of dual structure, and it does not show perfect match. No- tice that the saliency map is with small squared boxes. This is likely due to the ViT uses patch embedding.

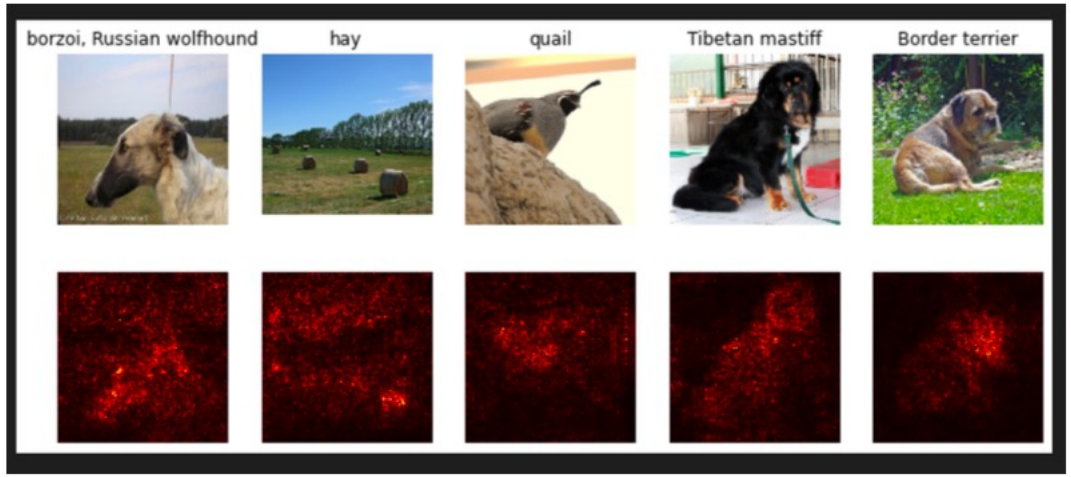


Figure 6. Salicency Map of CNN only

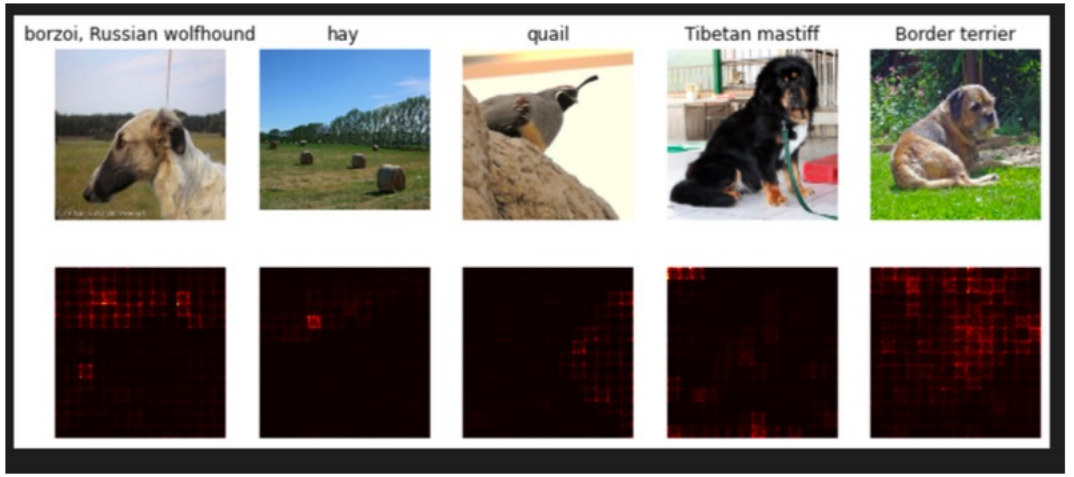
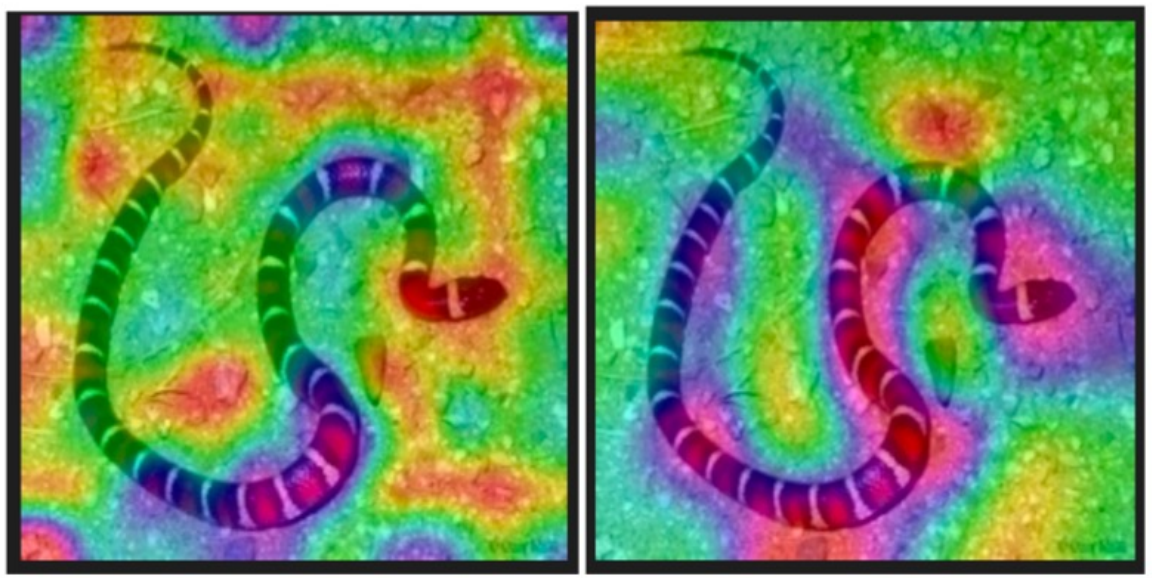


Figure 7. Salicency Map of RobustNet

Experiments & **Analysis**

- ***Class Activation Map:***
  - Conv-only branch Fig.-5 on the left as well as both Conv and Transformer branch Fig.-5 on the right . It shows that they are complementary to each other. With trained similar epochs, Using both Conv and Transformer branches shows better result!



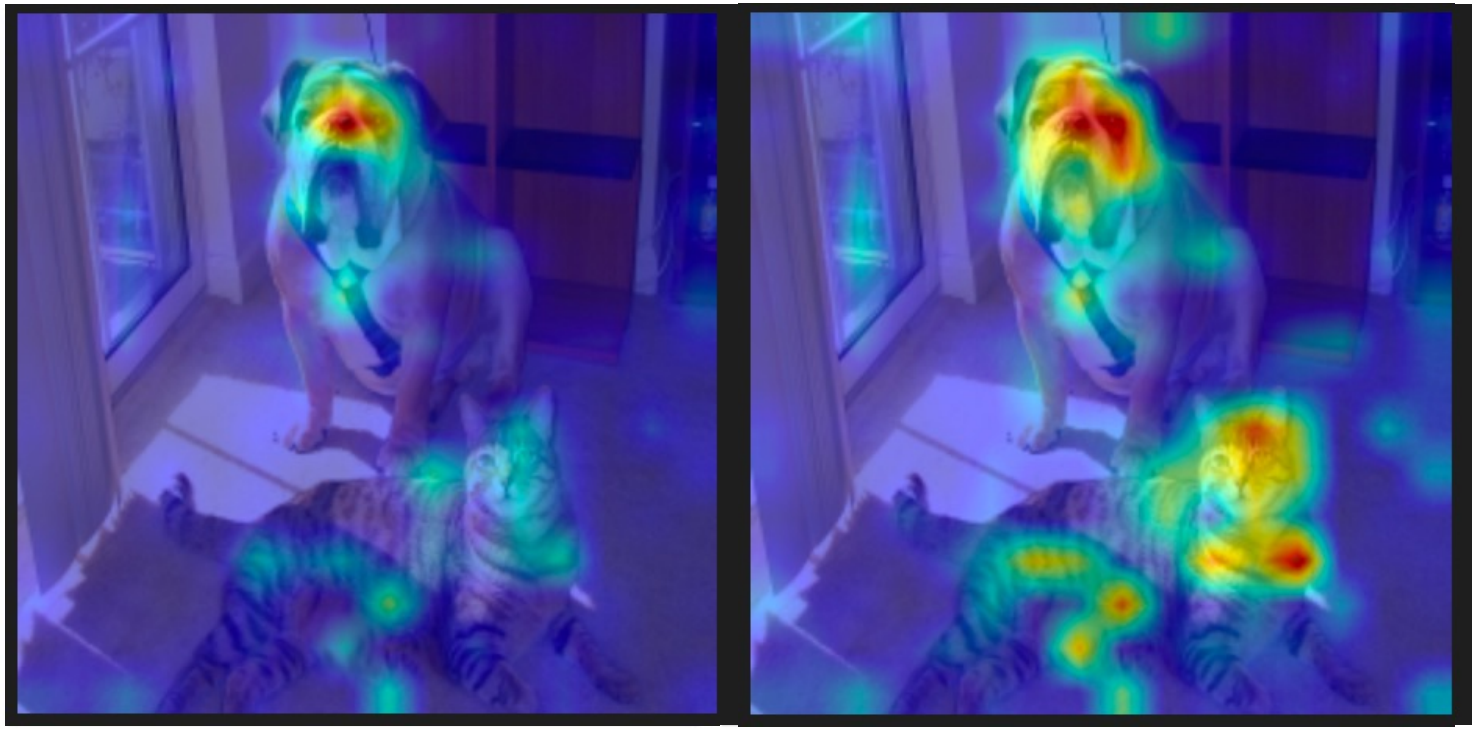
CAM of Conv Only

CAM of both Conv and Transformer Branches

Figure 5. CAM of RobustNet

Experiments & **Analysis**

- ***Local vs Global Feature Learning:***
  - We want to understand how RobustNet's dual structure learns both local and global features and pays attention to both. Surprisingly, Fig.-4 shows that Attention Map of the transformer branch pays attention to very local regions. The Class Activation Map of the Conv branch pays attention to larger region.



Attention map of dog CAM of dog

Figure 4. Local Features VS Global Features

Experiments & **Analysis**

- ***What has been learned in the feature map:***
  - It shows that Conv Layer learned local features in initial layers and abstract features in later layers.

Figure 13. The feature maps at the first convolutional layer inside the $trans\_conv$ block.

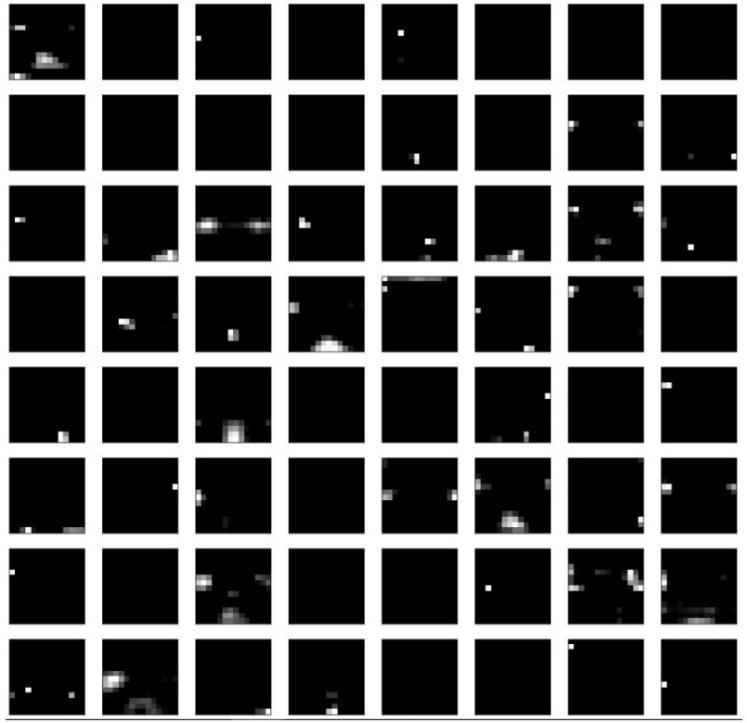


Figure 14. The feature maps at the last convolutional layer inside the $trans\_conv$ block.

Experiments & **Analysis**

- ***Qualitative Result in MSCOCO Object Detection and Style Transformation***

Figure 12. Inferenced output of MMDetector (Street objects)

Figure 10. Visualization of a sample transferred result with three style inputs.

Conclusions & Future Work

- ***Conclusions:***
  - We introduced a new dual structure that fuses CNN and ViT together, we name it RobustNet, and showed that it can perform well in different kinds of Computer Vision tasks. The key contribution is that this new architecture can lever- age state-of-the-art pre-trained models, show superior performance on a wide range of tasks with comparably less training time.
- ***Future Work:***
  - Prove this network works in Instance Segmentation, we were unable to finish training the network for the Instance Segmentation task. And apply our approach to other datasets such as Open Images to show robustness.
  - Compare this architecture with simple ensemble method to fully understand the benefit of introduced FIU(feature interaction unit) between Conv branch and Transformer branch.