# Finetuning Dense Prediction Vision Transformers for Image Restoration

Shirley Wang
Vector Institute, University of Toronto

## Introduction

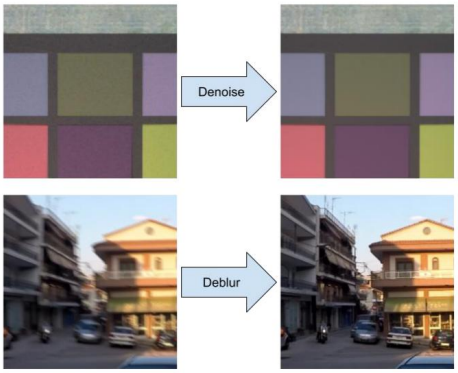
**Background:** "Dense Prediction" Vision Transformers are currently the standard for image segmentation.

**Motivation:** Dense prediction vision transformers have a good understanding of the content of images and predict a value for every pixel. They have an architecture that should be capable of image restoration, but how well will they actually perform?

**Summary/Contribution:** We explore finetuning dense prediction vision transformers on denoising and deblurring with minimal changes to their architecture.

**Models:**
- Encoder: Swin / Decoder: UperNet
- Encoder: Swin / Decoder: Mask2Former
- Encoder: ViT-Adapter / Decoder: Mask2Former

**Datasets:**
- SIDD-Medium: Image denoising
- GoPro: Image motion deblurring




## Related Work

**Dense Prediction Vision Transformers**
- Swin [1] is a hierarchical transformer with shifted windows, and uses UperNet as its decoder, which fuses features from a feature pyramid for predicting values for every pixel. It's a highly competitive computer vision model on dense prediction tasks.
- ViT-Adapter [2] is a state of the art segmentation model based on the original ViT, which uses adapter modules to inject additional information for dense prediction, with BEiT pretraining. It also uses Mask2Former as its decoder, which constrains cross-attention to predicted mask regions.

**Image Restoration Models**

Models for image restoration are usually specially crafted for the task of image restoration.
- NAFNet [3] achieves state of the art results on image restoration with a simple design and no nonlinear activation functions, exceeding previous models with a less computationally expensive structure.
- Restormer [4] is a competitive image restoration model which uses special attention modules and gating mechanisms for a more efficient transformer.

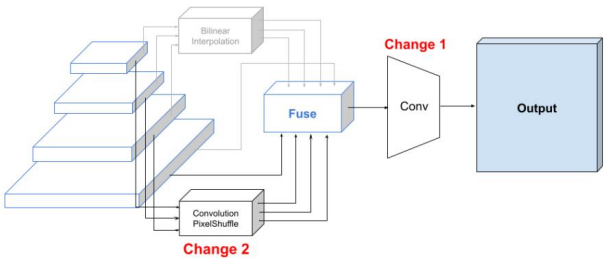
## References

[1] Z. Liu et al., 'Swin Transformer: Hierarchical Vision Transformer using Shifted Windows', in ICCV, 2021.
[2] Z. Chen et al., 'Vision Transformer Adapter for Dense Predictions', arXiv preprint arXiv:2205. 08534, 2022.
[3] L. Chen et al., 'Simple Baselines for Image Restoration', arXiv preprint arXiv:2204. 04676, 2022.
[4] S. W. Zamir et al., 'Restormer: Efficient Transformer for High-Resolution Image Restoration', in CVPR, 2021.

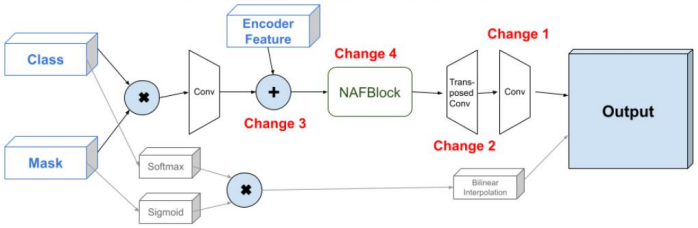
## New Technique

**Architecture Changes:** Entirely within the decoder. Changes are incremental (e.g.: change 2 also includes change 1)
1. Change the final output channel from number of classes to three (RGB)
2. Change upsampling within the decoder from bilinear interpolation to convolutions
3. Add skip connection
4. Add a NAFBlock before the final upsampling

**UperNet Architecture Modifications**




**Mask2Former Architecture Modifications**




## Experimental Results

| Model | | GoPro PSNR | GoPro SSIM | SIDD PSNR | SIDD SSIM |
|---|---|---|---|---|---|
| Swin-UperNet | Output Channels | 28.05 | 0.8581 | 27.79 | 0.5346 |
| | Upsampling | 29.03 | 0.8734 | 28.04 | 0.5412 |
| Swin-Mask2Former | Output Channels | 28.44 | 0.8675 | 27.89 | 0.5374 |
| | Skip Connections | 30.22 | 0.9013 | 30.63 | 0.6680 |
| | NAFBlock | 30.43 | 0.9059 | 31.72* | 0.7027* |
| ViTAdapter-Mask2Former | Output Channels | 29.50 | 0.8808 | 27.82 | 0.5356 |
| | NAFBlock | **31.83*** | 0.9224* | 29.06 | 0.5880 |
| NAFNet | | 30.37 | **0.9404** | **42.01** | **0.9706** |
| Restormer | | 29.93 | 0.9337 | 40.20 | 0.9608 |
| HINet | | 30.31 | 0.9350 | 41.27 | 0.9677 |

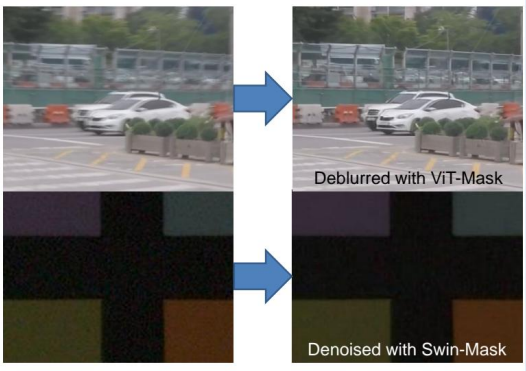
ViT-Adapter with Mask2Former achieves competitive results on deblurring.

Only Swin with Mask2Former succeeds at denoising, and it still falls short of state of the art models.

Changes 2 and 3 provides a large boost in performance. Change 4 only provides a small boost.

Instance segmentation models are capable of recognizing objects while also recognizing that they are blurry. Their outputs to noisy images are overall very similar. This suggests why segmentation models do not finetune well to denoising.

These results suggest segmentation models as a potential future basis for deblurring.




Swin-UperNet Segmentation Masks