Finetuning Dense Prediction Vision Transformers for Image Restoration

Shirley Wang

Abstract—Dense prediction vision transformers are the current standard of image segmentation, which arises from the model's incredible ability to understand the contents of the images it is fed, as well as predicting a value for every pixel of the image. Image restoration is another task that involves dense prediction and requires a strong understanding of the content of images to perform accurately. This paper provides an exploration of finetuning state of the art image segmentation models on image restoration tasks with minimal changes to the architecture, to see what the fields can borrow from each other. We find that with minimal changes, the segmentation model ViT-Adapter can be finetuned to achieve competitive results on deblurring, suggesting potential directions for future research in image deblurring. We also provide potential reasons for why image segmentation models have qualities that image deblurring may find useful, and also why they fail on image denoising.

Index Terms—Image Restoration, Image Segmentation, Dense Prediction, Vision Transformers, Finetuning

1 Introduction

Ever since transformers proved their usefulness in NLP, they have been continuously adopted in a variety of fields, and computer vision is no different. Vision transformers have been seen to quickly overtake convolutions in numerous different vision tasks, and are the fan favourite model nowadays. Lots of work has gone into adapting them for more complicated imaging tasks like segmentation, which fall under the category of ”Dense Prediction” due to how a prediction is required for each pixel of the image. To perform segmentation accurately, these dense prediction vision transformers require a strong understanding of the content captured within the image. In this paper, we explore the flexibility and robustness of these segmentation models by finetuning them on a different dense prediction task: image restoration.

Image restoration covers a pretty broad range of tasks, but this paper will mainly focus on denoising and deblurring since those are broadly applicable and also the most popular. For these tasks, we assume we have a corrupted version of an image, and would like to recover the original uncorrupted version. Since dense prediction vision transformers should have a good understanding of the content of images, and are capable of creating predictions for every pixel, they already have an architecture that should be able to perform image denoising and deblurring.

This paper contributes an exploration of bridging the fields of image segmentation and image restoration by finetuning these segmentation models to perform image restoration with minimal changes to their architecture. We find that segmentation models are capable of achieving competitive results on image deblurring, but fall short for denoising. Furthermore, by making use of a few small tricks, the performance on these segmentation models improves dramatically for image restoration. These changes can be summarized as:

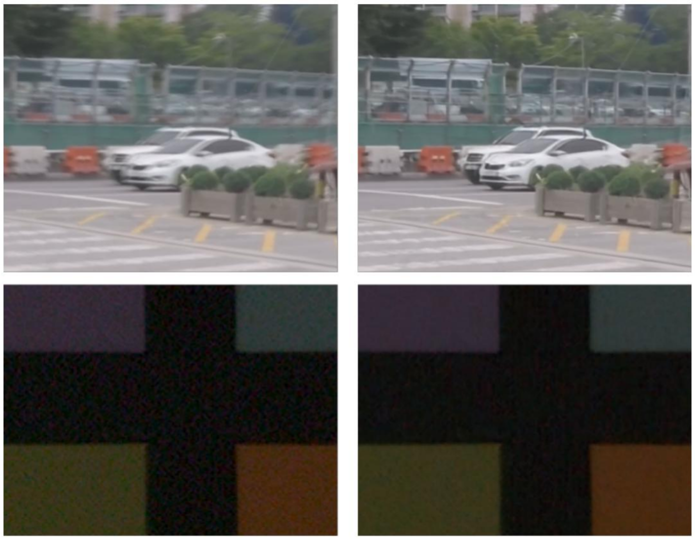

Fig. 1. Visualized results of our models on image restoration tasks. Top-Left: GoPro blurry image. Top-Right: Deblurred image with our ViT-Adapter-Mask2Former model. Bottom-Left: SIDD noisy image. Bottom-Right: Denoised image with our Swin-Mask2Former model.

1) Changing the number of output channels of a segmentation head from the number of classes it has been trained to segment over, to three (for the three RGB channels).
2) Changing the upsampling method in the segmentation decoder from basic bilinear interpolation into trained convolutions.
3) Adding in skip connections from the encoder features to the decoder features.
4) Adding an additional module from NAFNet to further process the final features.

- *Shirley Wang is with Vector Institute, and the Department of Computer Science, University of Toronto.*
  E-mail: shirleywang@cs.toronto.edu

2 RELATED WORK

2.1 Dense Prediction Vision Transformers

The original Vision Transformer (ViT) [1] was created and applied on image classification. The independent patch structure of transformers wasn't well-suited for dense prediction, so there have been many approaches to adapting the original transformer structure to something more suited for dense prediction. Pyramid Vision Trasformer (PVT) [2], Swin Transformer [3], and SegFormer [4] all take different approaches to creating features from different scales in the image, as being able to create and aggregate features from varying scales in the image is important for accurate segmentation in difficult areas. Vision Transformer Adapter (ViT-Adapter [5]) on the other hand, takes the original ViT and adds on some additional adapter modules so that after pretraining, the adapter modules inject additional information into it so that it is capable of performing well on dense prediction tasks. Mask2Former [6] makes use of masked attention, constraining cross-attention to the predicted mask regions. Pretraining the backbone has also taken a large focus in recent years, with Contrastive Learning [7] and Masked Image Modeling [8] both becoming competitive pretraining methods, and BEiT-3 [9] achieving state of the art results on ADE20k and COCO as of November 2022.

2.2 Image Restoration

Previous research that focuses on image restoration often test their models on both denoising and deblurring tasks, although there is some research into models specially for one task. Uformer [10] uses a hierarchical encoder-decoder network with transformer blocks, with nonoverlapping window-based self-attention and a multi-scale restoration module. Restormer [11] makes use of Multi-DConv Head Transposed Attention (MDTA) blocks and gating mechanisms on the linear layers. HINet [12] integrates instance normalization into the blocks of their network to achieve high quality results. "Simple Baselines for Image Restoration" [13] suggests a Nonlinear Activation Free Network (NAFNet), which uses no nonlinear activation functions, and achieves state of the art results on both image deblurring and denoising. Swin Transformer has also been explored on image restoration before [14], however that exploration only utilizes Swin Transformer blocks for deep feature extraction, and creates new architecture for shallow feature extraction and image reconstruction modules. To our knowledge, this is the first exploration of segmentation models with minimal changes applied on image restoration tasks.

Most of the competitive image restoration methods can be viewed as variants of the classic UNet [15], which stacks blocks in a U-shaped architecture with skip connections for features on the encoder and decoder halves of the same size. This motivates the exploration of using skip connections in this paper. Uformer and Restormer also make use of the transformer architecture, as can be inferred from their names.

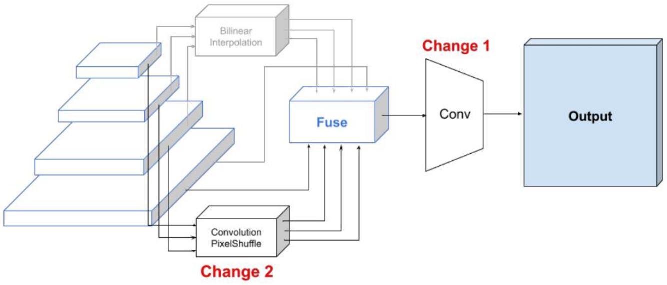

Fig. 2. UperNet decoder and our proposed architecture changes. The original goes through the top (gray) path while our proposed architecture goes through the top (black) path. Changes are noted with red where they happen.

3 METHOD

3.1 Segmentation Models

We will be exploring three established segmentation models in this paper: Swin, Mask2Former, and ViT-Adapter. These three models also tend to overlap in terms of encoder and decoder.

Swin Transformer uses an encoder made of Swin transformer blocks, and UperNet [16] as the decoder. We specifically use the base Swin model that was pretrained on ImageNet-22K, and trained on semantic segmentation on ADE20K. This version originally achieved an mIoU of 50.76 on ADE20K.

The original Mask2Former uses Swin as its backbone and its titular Mask2Former as the decoder. We specifically use the version with the base Swin model as its backbone, pretrained on ImageNet-21K, and trained on semantic segmentation on ADE20K. This version originally achieved an mIoU of 53.9 on ADE20K.

ViT-Adapter uses its ViT with Adapter modules for the backbone, and Mask2Former as the decoder. We specifically use ViT-Adapter-Large for the backbone, pretrained on ImageNet-22K using masked image modelling. For consistency, we use the same Mask2former used for the Swin-Mask2Former model, that has a feature dimension of 256 in its decoder. However, since ViT-Adapter-L does not have a corresponding Mask2Former of this size (its original Mask2Former uses a feature dimension of 1024), we reload weights for the backbone from the pretrained version and the weights for the decoder from the same mask2former weights as Swin-Mask2former. This does mean this model does not have a backbone and decoder that have been trained to work together already, but ViT-Adapter has achieved state of the art results with it's backbone, so this shouldn't affect the results of this model once finetuned.

3.2 Architecture Changes

There are four main architecture changes we explore in this paper, all contained within the decoders of the segmentation models. Along with these changes, the model is trained to predict the residual between the original and corrupt image, rather than just predicting the original image itself, as most image restoration models are also trained in this manner. All changes detailed are incremental: the change in upsampling is always implemented alongside the change in

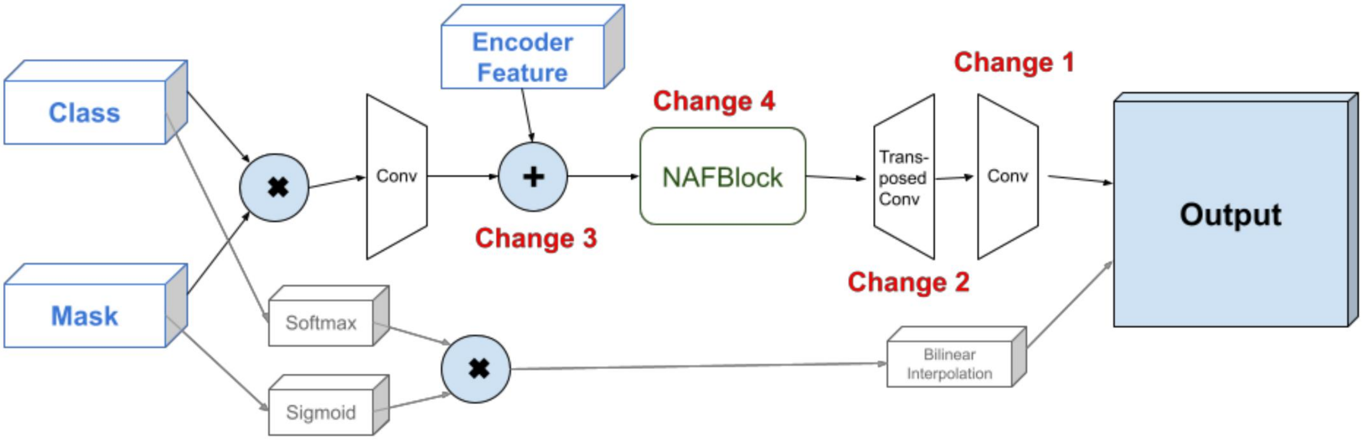

Fig. 3. Final processing portion of Mask2Former and our proposed architecture changes. The original goes through the bottom (gray) path while our proposed architecture goes through the top (black) path. Changes are noted with red where they happen.

output channels, the change in skip connections is always implemented alongside the change in upsampling, and so on. The changes to the UperNet decoder are detailed in Figure 2. The changes to the Mask2Former decoder are detailed in Figure 3.

In image segmentation, we need a good idea of exactly where the object is, and predict scores for how likely a pixel is a part of the object. However, in image restoration, since we want to predict the precise offset between the corrupt and original image, the usual nonlinear activation functions (e.g. sigmoid) that create probabilities for segmentation masks are actually undesirable here. The predictions for image restoration models need to be more precise in exact values than predictions for image segmentation masks, which is what motivates these changes.

3.2.1 Output Channels

For the decoder UperNet, the output channels of the model have been changed from the original number of classes it was trained to perform segmentation on, to three for the three RGB channels of an image. This is the bare minimum change necessary for segmentation models to be applicable for image restoration.

The decoder Mask2Former has a different structure in it's final output that requires slightly more changes. The original Mask2Former creates class scores and mask scores, and then combines them to create their final segmentation mask scores.

$$Softmax(c) \times Sigmoid(m)$$

But the Softmax and Sigmoid constrains the model outputs to be within an interval of [0, 1], which is undesirable when we are predicting the residual in image restoration where we sometimes may desire predicting a negative offset. So for Mask2Former, the change to the final output channels is instead:

$$Conv(c \times m)$$

Where we keep the general structure of how the original created their output, but remove the Softmax and Sigmoid constraints, and add a convolutional layer that takes the in the number of classes dimension the original was trained on, and outputs three channels.

3.2.2 Upsampling

Both UperNet and Mask2Former use bilinear interpolation for when they need to upsample. However, due to the precise nature of predictions in image denoising and deblurring, this may be undesirable for getting the correct per-pixel predictions. As such, the second potential change is for the bilinear interpolation used within the decoders of the models to be updated to something that uses trained convolutions instead, to hopefully obtain more precise details upon reconstruction.

UperNet makes use of bilinear interpolation to go from each feature map to the next, combining features from different scales to create its final segmentation mask predictions. Swin has already been explored on image restoration [14], and there, the authors use a convolutional layer along with pixel shuffle to upsample within their image reconstruction decoder. We adopt that practice here, replacing bilinear interpolation within UperNet with a convolutional layer followed by pixel shuffle to upsample.

Mask2Former makes use of bilinear interpolation within its pixel decoder when aggregating features, along with a final bilinear interpolation from the segmentation mask scores created from combining class and mask scores together to go back to the original size of the image. We will replace only the final bilinear interpolation with a transposed convolution for upsampling. We do not replace the interpolation within the pixel decoder for simplicity, and we use two transposed convolutions to upsample (with each upsampling by a factor of two) because through experimentation, transposed convolutions perform better than a regular convolution followed by pixel shuffle.

3.2.3 Skip Connections

Skip connections are another common feature of image restoration networks. This allows the model to not forget about finegrained details while upsampling features from deep within the network. Due to the importance of having finegrained predictions for every pixel, skip connections should be an important part of image reconstruction. We implement a final skip connection within Mask2Former after it's computed the product between class and mask scores but before upsampling.

3.2.4 NAFBlock

NAFNet currently achieves state of the art results on image restoration with a simple design consisting of its NAFBlocks (Nonlinear Activation Free Blocks) and up/downsampling. We borrow one of these NAFBlocks and add it to the decoder after the skip connection but before upsampling within Mask2Former, to see how these blocks operate in a new setting and see if adding additional layers can improve the model's performance.

4 Experiments

4.1 Datasets

We train our models on SIDD [17] for image denoising and GoPro [18] for image deblurring. For SIDD, we will specifically be using the SIDD-Medium Dataset with sRGB images, not the full SIDD dataset, due to size constraints. For GoPro, we will be using the full GoPro dataset, but taking out a subset of 200 instances for evaluation. We will be measuring model performance on PSNR and SSIM for both datasets.

4.2 Training

We use MMSegmentation [19] and the ViT-Adapter [5] code base for implementing the three models. For both datasets, we use train with an image size of 256×256. We use random resized cropping, along with random horizontal and vertical flipping. All models are trained on one GPU with 12GB memory. All models are trained with the AdamW optimizer to minimize the Dice Loss between the corrupted image plus model output, and the ground truth image.

We use the exact same training schedule that was used for training the original models, except with the learning rate halved. Models are trained for at least 1000 iterations, and early stopping occurs in the models no longer improve on the validation set. To be more precise, the Swin-Upernet model is trained with learning rate 3e-5, weight decay 0.01, batch size 16, and no optimization on the absolute position embeddings, relative position biases, and normalization. The Swin-Mask2Former model is trained with learning rate 5e-5, weight decay 0.05, batch size 8, and the learning rate is decreased by a factor of 0.1 for the backbone. The VitAdapter-Mask2Former model is trained with learning rate 1e-5, weight decay 0.05, batch size 4, and decreasing learning rate in the backbone by a factor of 0.9 for every additional layer.

We use a linearly decreasing learning rate starting from the original learning rate and ending at a learning rate a factor of 1e-3 smaller than the initial rate 4000 iterations later. However, due to early stopping during training, most models will not reach this point.

Due to size constraints, our models are unable to process the entire image at once during evaluation time. To mitigate this, we split the image up into 256×256 overlapping patches, and send all patches through the model, then recombine them back into the image of the original size, with overlapping areas averaged.

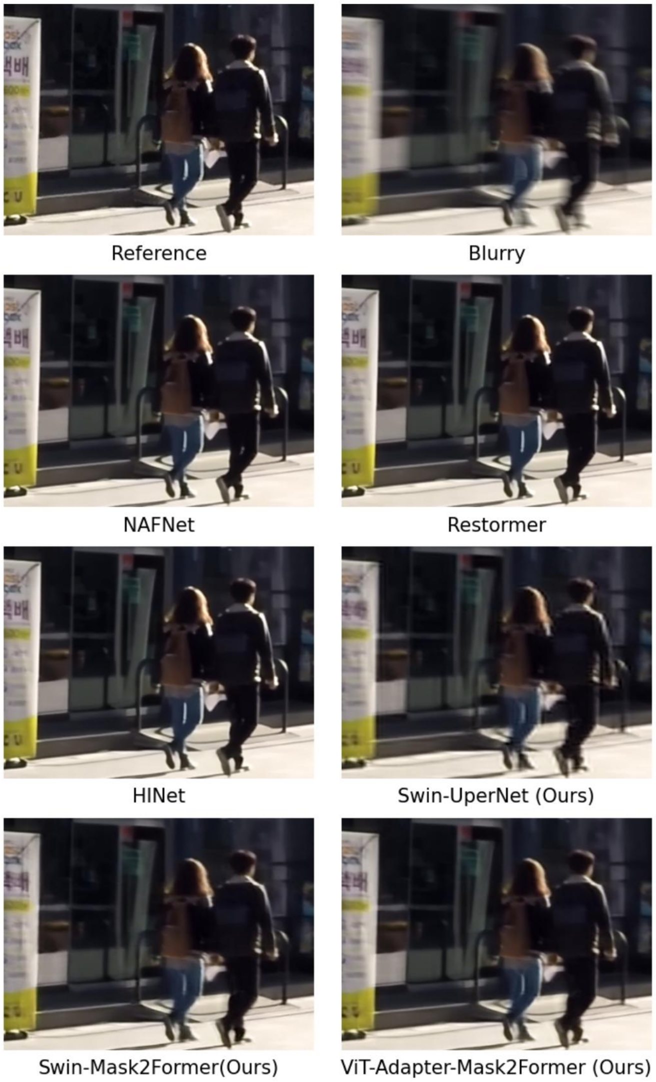

Fig. 4. Qualitative comparison of deblurring results of our models and benchmark models.

4.3 Evaluation

We compare the deblurring and denoising results of our models to state of the art methods on the subset of GoPro and SIDD we remove to use for evaluation. NAFNet, Restormer, and HINet are all competitive models on image restoration, and they are also evaluated on the same subset of SIDD and GoPro that we set aside from the training sets for evaluation. Numerical comparisons can be found in Table 1.

4.3.1 Deblurring

We compare the deblurring results of our models to state of the art methods on GoPro dataset. As we show in Table 1, our modifications to Vit-Adapter allow it to surpass the previous best method NAFNet by 1.46dB in PSNR. Qualitative results are shown in Figure 4.

4.3.2 Denoising

We compare the denoising results of our models to state of the art methods on SIDD dataset. As we show in Table 1, our modifications to the models do improve them slightly, but

TABLE 1
Evaluation Results on GoPro and SIDD. We report the PSNR and SSIM on the evaluation sets we created for our models, along with three existing image restoration models. Our best deblurring model exceeds current state of the art deblurring models on our GoPro evaluation set.

Model		GoPro		SIDD	
		PSNR	SSIM	PSNR	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

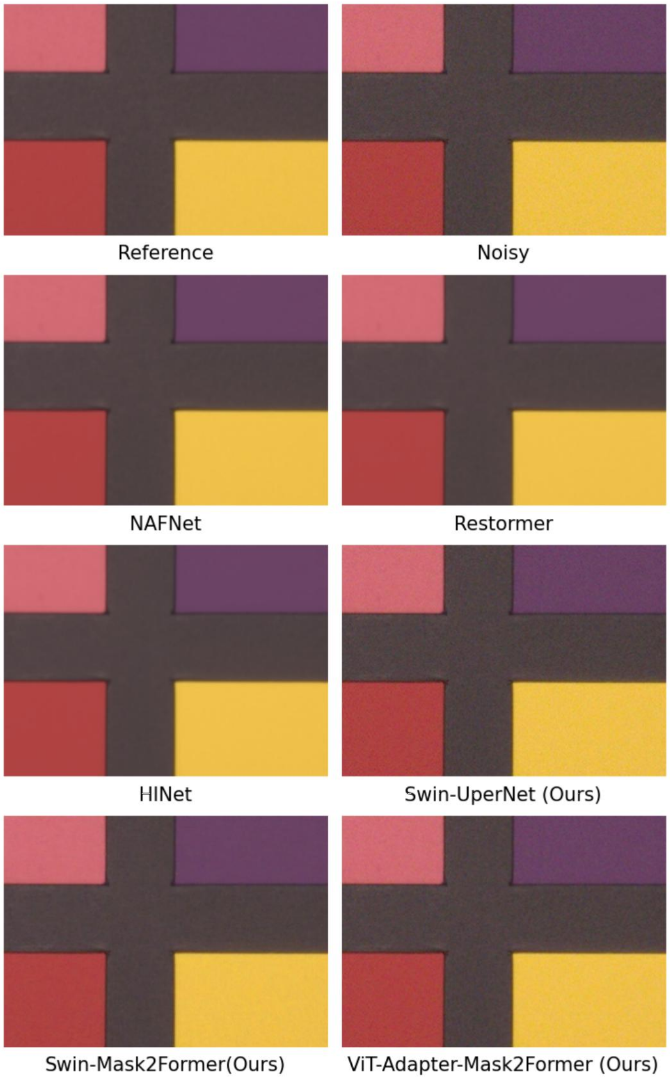
Reference
Noisy
NAFNet
Restormer
HINet
Swin-UperNet (Ours)
Swin-Mask2Former(Ours)
ViT-Adapter-Mask2Former (Ours)

Fig. 5. Qualitative comparison of denoising results of our models and benchmark models.

the results are still nowhere close to state of the art denoising methods. Our best model in this case is our modified Swin-Mask2Former. Qualitative results are shown in Figure 5.

4.4 Ablation Studies

The change made to output channels is necessary for segmentation models to perform image restoration, so there is no further analysis into it's benefits.

As we show in Table 1, the change in upsampling improves results for the Swin-UperNet model by 0.98dB PSNR and 0.0153 SSIM. The improvements for SIDD are much smaller and inconsequential. Swin-UperNet also has the lowest performance of the three, which was expected since the original Swin transformer came out much longer ago and also performs worse on segmentation than the other two models.

For Swin-Mask2Former, the change in upsampling as well as adding in an additional skip connection improves its results on deblurring by 1.78dB PSNR and 0.0388 SSIM, while also improving results on denoising by 2.74dB PSNR and 0.1306 SSIM. The improvements are larger for denoising than deblurring. This does potentially suggest that if we adopted the typical image restoration model structure with skip connections and transposed convolutions for features of every scale, instead of only at the last scale, we could potentially achieve even better performance. Additionally adding an NAFBlock makes slight improvements to its performance on deblurring, while making large improvements to its ability to denoise.

For ViTAdapter-Mask2Former, the change in upsampling, additional skip connection, and an additional NAFBlock result in a performance increase of 2.33dB PSNR and 0.0416 SSIM on deblurring, achieving competitive results on deblurring in terms of PSNR. It also results in a performance increase of 1.24dB PSNR and 0.0524 SSIM on denoising, but performance is still far from Swin-Mask2Former and state of the art models. Both this model and Swin-UperNet achieve a initial performance on SIDD, so it's also possible that adding in skip connections and an NAFBlock for UperNet would also improve it's performance there. However, since both of these models struggle greatly with denoising and are far from state of the art results, further research in this direction is probably not worth it. Since the performance is so outstanding on deblurring, further research into using this model's architecture and pretraining techniques may be useful for improving current deblurring models.

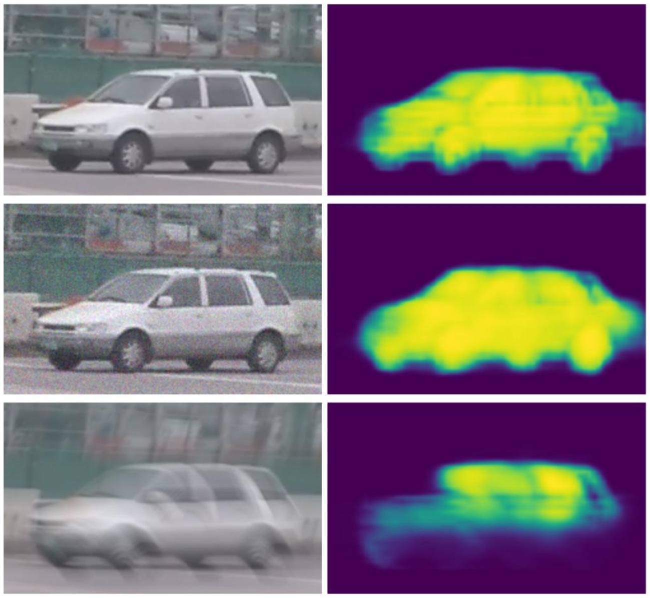

Fig. 6. Swin-Upernet final segmentation masks on the original (top) image, a noisy (middle) image, and a blurry (bottom) image for the class "Vehicle"

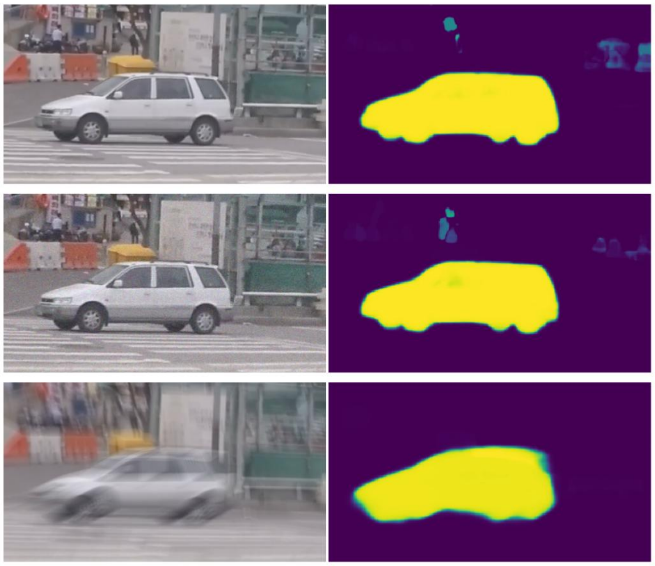

Fig. 7. Swin-Mask2Former final segmentation masks on the original (top) image, a noisy (middle) image, and a blurry (bottom) image for the class "Vehicle"

Swin-Mask2Former performs the best on denoising out of all three models, beating ViT-Adapter-Mask2Former by a wide margin. This does suggest that the Swin backbone is more suited to denoising than Vit-Adapter is. This could potentially be due to the differences in architecture between the two, where Swin is better at paying attention to tiny patch details while ViT-Adapter is better at understanding the overall content within the pictures. Another potential reason is the Swin backbone was original pretrained using the typical supervised method on ImageNet, while the ViT-Adapter backbone was pretrained using Masked Image Modelling, giving it better potentially a stronger understanding of the content and objects within the image, but as a result struggling greating with caring more about the tiny noise around objects instead of the objects itself.

4.5 Segmentation Masks

We display the segmentation mask outputs for images sent through the already trained Swin-Upernet and Swin-Mask2Former on semantic segmentation (which are also the model weights we reload training from). As seen in Figure 6 for Swin-UperNet, the model is still capable of recognizing the object and its class even when the image is blurry, although the regions it predicts that are a part of the vehicle are much less certain. The same applies for Figure 7 for Swin-Mask2Former, although as can be seen, this model is much more capable of recognizing where the vehicle is despite it being blurry. It is still much more uncertain around the edges, but it still performs much better on segmentation than on a blurry object than Swin-UperNet. This could potentially be why it performs better than Swin-UperNet on image deblurring. This also gives a motivation for further analysis into borrowing components from image segmentation for image restoration.

As can be seen in both Figures, the model outputs are not sensitive to the noise in the image, and the segmentation masks have very little change between the original and noisy images. While this is a desirable trait for image segmentation models, it is not desirable for image denoising as the model must be sensitive to the noise This does provide a potential reason for why the models performed much worse than state of the art for image denoising.

5 Conclusion

In this paper, we explore finetuning image segmentation models for image restoration instead. To be specific, we explore using the backbones Swin and ViT-Adapter, and the decoders UperNet and Mask2Former. The changes implemented are minimal and contained within the decoder, and mimic common architecture choices used in other image restoration models. Specifically, we explore changing the output channels, changing the upsampling method to one that uses trained convolutions, adding in a skip connection, and adding in an NAFBlock from NAFNet.

We find that these image segmentation models are capable of achieving competitive results on image deblurring, but do not have the structure required to perform image denoising. Our best image deblurring model (ViT-Adapter with Mask2Former) improves over our second best image deblurring model (Swin with Mask2Former) by 1.4dB PSNR, as well as achieving competitive performance with other image restoration models. Since the only change between the two models is the backbone (Swin and ViT-Adapter), we believe that either the architecture of ViT-Adapter is superior, or ViT-Adapter's use of masked image modelling for pretraining is better than the traditional supervised pretraining on ImageNet. We believe that it may be beneficial for further research in image deblurring to borrow some ideas from these state of the art image segmentation models in terms of their backbone architecture or their pretraining method.

Acknowledgments

The author would like to thank Professor David Lindell for the opportunity to work on this, Shayan Shekarforoush for providing many helpful suggestions for the direction of the exploration, and the University of Toronto for providing the resources to make this possible.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[2] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

[4] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Neural Information Processing Systems (NeurIPS)*, 2021.

[5] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," *arXiv preprint arXiv:2205.08534*, 2022. [Online]. Available: https://arxiv.org/abs/2205.08534

[6] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation." arXiv, 2022. [Online]. Available: https://arxiv.org/abs/2112.01527

[7] Y. Wei, H. Hu, Z. Xie, Z. Zhang, Y. Cao, J. Bao, D. Chen, and B. Guo, "Contrastive learning rivals masked image modeling in fine-tuning via feature distillation," *Tech Report*, 2022.

[8] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *arXiv:2111.06377*, 2021.

[9] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a foreign language: Beit pretraining for all vision and vision-language tasks," 2022. [Online]. Available: https://arxiv.org/abs/2208.10442

[10] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 17 683–17 693.

[11] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *CVPR*. arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2111.09881

[12] L. Chen, X. Lu, J. Zhang, X. Chu, and C. Chen, "Hinet: Half instance normalization network for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. arXiv, June 2021, pp. 182–192. [Online]. Available: https://arxiv.org/abs/2105.06086

[13] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," *arXiv preprint arXiv:2204.04676*, 2022. [Online]. Available: https://arxiv.org/abs/2204.04676

[14] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," *arXiv preprint arXiv:2108.10257*, 2021.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[16] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *European Conference on Computer Vision*, Springer. arXiv, 2018. [Online]. Available: https://arxiv.org/abs/1807.10221

[17] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *CVPR*, July 2017.

[19] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.

Shirley Wang is a graduate student at the University of Toronto pursuing a masters in computer science.