Leveraging Residual Neural Network for UNet Image Denoising with Perceptual Loss

Xinrong Zhou, Jingyu Liu, Xinyu Zhang

Abstract—Image denoising is removing noise from a noisy image, so as to restore the true image. In this paper, we proposed a network that combines the structure of UNet and ResNet to do the image denoising. Our new architecture has the best PSNR in the denoising experiment. In addition, Most of denoising work has largely focused on minimizing the mean squared error (MSE). The resulting estimates have high PSNR but lacks high-frequency detail which makes them perceptually unsatisfying. To overcome this limitation, we add a perceptual loss function motivated by perceptual similarity instead of similarity in pixel space. Our joint loss function made our result has better visual effect.

Index Terms—Computational Imaging, Neural network, Computer Vision

1 Introduction

The use of images has significantly increased during the past ten years. Nowadays, because of the limit of environment, transmission channel and devices, more and more images are captured in poor conditions, therefore, distortion and loss of picture information are inevitable. The demand of clear and sharp images hastens the emergence of different image denoising techniques. However, it remains a challenging and open task. The main reason for this is that from a mathematical perspective, image denoising is an inverse problem and its solution is not unique. Computer-aided methods algorithms have been widely used in this task. Traditional model-based methods such as non-local means (NLM) [1], block-matching and 3-D filtering (BM3D) [2], weighted nuclear norm minimization (WNNM) [3] rely on image prior modeling, and their optimization algorithms are time-consuming. Nowadays, machine learning models, especially neural network frameworks are employed by many researchers on image denoising. DnCNN [4], a deep convolutional neural network for image denoising model proposed by Zhang et al. is a prominent adpation of neural network. Inspired by previous remarkable works, we intend to use the most recent developments in neural networks to improve the performance of image denoising. We are interested in combining two outstanding models: ResNet [5] and Unet [6] to explore denoising tasks and improve the performances.

2 Related Work

As one of the most significant problems in computer vision, image denoising task draws big attention of researchers. Numerous neural network techniques have been proposed for this problem. Being one of the pioneers, Jain and Seung [7] proposed a convolutional neural network(CNN) on image denoising in 2008. On that basis, Zhang et al. [4]proposed a deep convolutional neural network for image denoising (DnCNN). This model improves the denoising performance by stacking multiple blocks of convolutional layers, batch normalization, and rectified linear unit (ReLU) activations. Gurprem Singh [8] and his team members proposed a deep convolutional neural network with added benefits of residual learning for denoising. The network is composed of convolution layers and ResNet blocks along with rectified linear unit activation function, and it is capable of learning end-to-end mappings from noise-distorted images to restored cleaner versions. With a single end-to-end model, this model can tackle different levels of Gaussian noise efficiently. Another work from Javier Gurrola [9] uses a residual dense UNet neural network for image denoising. In this work, they present a residual dense neural network (RDUNet) for image denoising based on the densely connected hierarchical network. The encoding and decoding layers of the RDUNet consist of densely connected convolutional layers to reuse the feature maps and local residual learning to avoid the vanishing gradient problem and speed up the learning process. The fact that these frameworks achieve impressive results shows that it is promising to use UNet and ResNet on denoising tasks.

The common-used benchmark, PSNR [10] calculates the ratio between the maximum possible value (power) of a signal and the power of distorting noise to evaluate the denoised image. This pixel-based measurement, however, does not account for perceptual variations between the output and ground-truth images. Many recent studies([11], [12]) have proved that the perceptual loss using the feature comparison method is more in line with real visual perception, and can restore clearer images and visual effects. Much better blurring results than using only the MSE loss.

3 Methods

We propose a network that combines the structure of UNet and ResNet with the joint objective function including perceptual loss.

- *All authors are with the Department of Computer Science, University of Toronto, Toronto, ON.*
- *E-mail:xinrongzhou@cs.toronto.edu, xinyuzhang@cs.toronto.edu jingyu@cs.toronto.edu*

3.1 UNet

In 2015, Ronneberger,et al. [6] proposed the UNet framework for Biomedical Image Segmentation. There are 4 down-sampling blocks and 4 up-sampling blocks. The UNet extract the main feature of the image during down-sampling and allow the network to propagate context information to higher resolution layers during up-sampling. In order to localize, high resolution features from the contracting path are combined with the upsampled output through SkipConnection. No fully connected layers are included in the framework. By mirrioring the input image, this strategy can predict the missing context in the noised image.

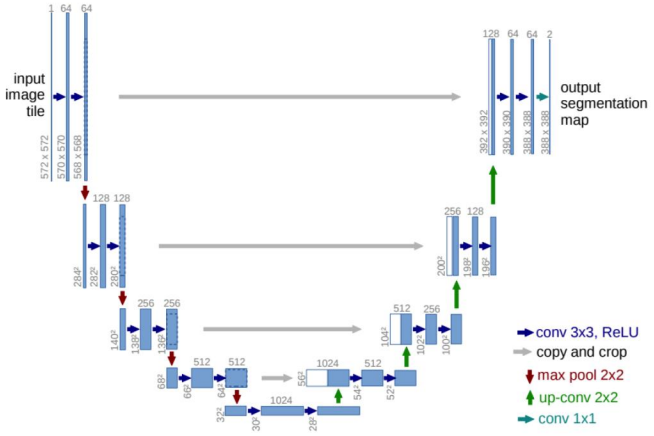


Fig. 1: UNet architecture

3.2 ResNet

Deep convolutional network seems a great tool for imaging problems, but when the network is very deep, this may result in vanishing or exploding gradients; To overcome these problems, ResNet [5] was proposed in 2016. Each block was given by adding residual learning operation in ResNet to improve the performance of image recognition, which leads to ResNet winning the mageNet LSVR in 2015. Figure 2 depicts the concept of residual learning.

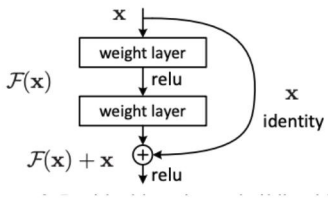


Fig. 2: ResNet framework

3.3 ResUNet

ResNet with a series of stacked residual blocks is powerful enough to extract features and strengthen the feature propagation during training and testing. Meanwhile, UNet with a symmetrical structure performs excellently for biomedical images. With the intention of combining the benefits of these two approaches, we propose a multi-stage architecture of Deep ResUNet, Figure 3 illustrates our network

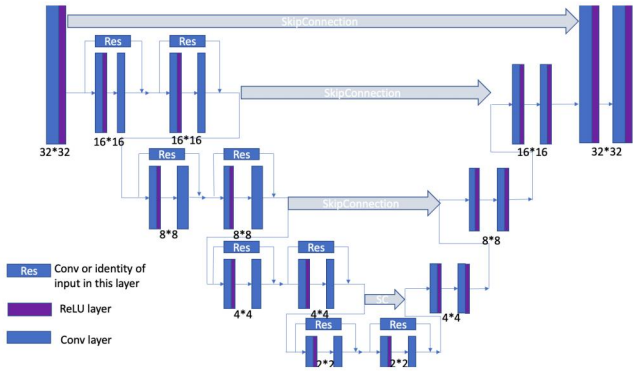


Fig. 3: ResUNet framework

architecture for image denoising. We adapted the overall structure of UNet. Generally, our network can be divided into two parts: Downsampling and Upsampling. The first part (downsampling part) composed of 4 downsampling blocks is designed to extract the features, also known as the"contracting" path in UNet. For each downsampling block, we have 2 Res-block. Each Res-block consists of a convolutional layer, followed by a ReLU layer and then another convolutional layer. The second part (upsampling part, or "expansive " path in UNet) is utilized to generate the denoised image using the extracted features at different stages of the encoding part. Between each downsampling block and upsampling block, a skip connection is operated. Unlike the conventional UNet, which implements by concatenating, we do the add operation. The proposed Deep ResUNet inherits both the benefits of ResNet and UNet.

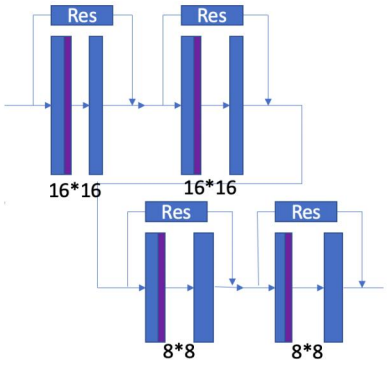


Fig. 4: Downsampling Block

Figure 4 illustrates the details of an example downsampling block. The block starts with a conv-ReLU layer and adds another convolution layer, then plug in the original input to complete a Res-block. And we repeat the same steps once again. Two Res-blocks made up a downsampling block. Then moving down to the next downsampling block, the size was reduced to half.

3.4 Objective funtion

Previous learning-based image restoration tasks use mean squared error (MSE) as the objective function to achieve a higher signal-to-noise ratio. However, this method of pixel-by-pixel comparison is found to be prone to loss

of detail information, resulting in blurred results [13]. Whereas, by comparing the image feature differences, perceptual loss can rebuild more details therefore, provides clearer result. Combining the advantages of both, we proposes a new joint loss by using λ as the coefficient to balance different loss terms:

$$\mathcal{L}_{loss} = \lambda \mathcal{L}_{MSE} + \mathcal{L}_{per}$$

Denote the original image as x, the denoised image as x', the ground truth image as y, the width of image as W and the height of image as H. We will use MSE loss:

$$\mathcal{L}_{MSE} = \frac{\|y - x'\|^2}{W * H}$$

or cross-entropy loss function to train this model. And for the perceptual loss part, the denoise images from UNet-ResNet model and the ground truth images will be both sent to a pre-trained 16 layer VGG network [14]. We will grab the (x') and (y) from one of the convolutional layer to calculate the perceptual loss:

$$\mathcal{L}_{per} = \frac{\|\phi(y) - \phi(x')\|^2}{W * H}$$

4 Experimental Results

4.1 Dataset

The dataset we use to train and test our networks is Berkeley database (BSDS300) [15]. There are 200 images in training samples and 100 images in testing samples. When we train our network to denoise the images, we clip the images to many 32*32 patches and add noise separately on them in order to augment the dataset amount and enhance the image details of learning. For our denoising experiments, we add sigma =0.01, 0.02, 0.05 and 0.1 Gaussian noise on the training dataset. For testing our models, we test on the same noise level as the training set. But we only train one single noise level sigma = 0.1 for ResUnet with perceptual loss due to the time and GPU limitation. We may add more experiments on this in future work.

Since we also curious about the utility of our ResUNet model with/without the new loss function on deblur-denoise task (like The task 3 in our Homework 5), we still use the BSDS300 dataset and also clip the images to 32*32 small patches. We extract the blur kernal from MNIST and add it on the dataset, then add sigma = 0.01, 0.02, 0.05, 0.1 Gaussian noise on the training dataset. For testing models, we test on the same noise level as the training set. Again we only train one single noise level sigma = 0.1 for ResUnet with perceptual loss due to the time and GPU limitation.

4.2 Experiments on Denoise Task

We train three different denoise networks: UNet with MSE loss, ResUNet with MSE loss and ResUNet with joint loss (MSE loss and perceptual loss) on the same training samples with different noise level and learning rate = 1e-3. For the third model, the perceptual feature is generated by a pretrained VGG-16 model. We feed the denoised results from ResUNet to the pretrained VGG-16 model and sum the 3th, 8th and 15th layers' MSE loss to calculate the perceptual loss. However, we observed that the perceptual loss and the MSE loss have different scales: the MSE loss is too much smaller than the perceptual loss. To balance the effects of these two losses on our model, we decided to multiply a coefficient value on MSE loss. We tried 1, 10 and 100 three different coefficient values and it proves that 100 has the best performances. Therefore, the loss function for the third model is

$$\mathcal{L}_{loss} = 100 * \mathcal{L}_{MSE} + \mathcal{L}_{per}$$

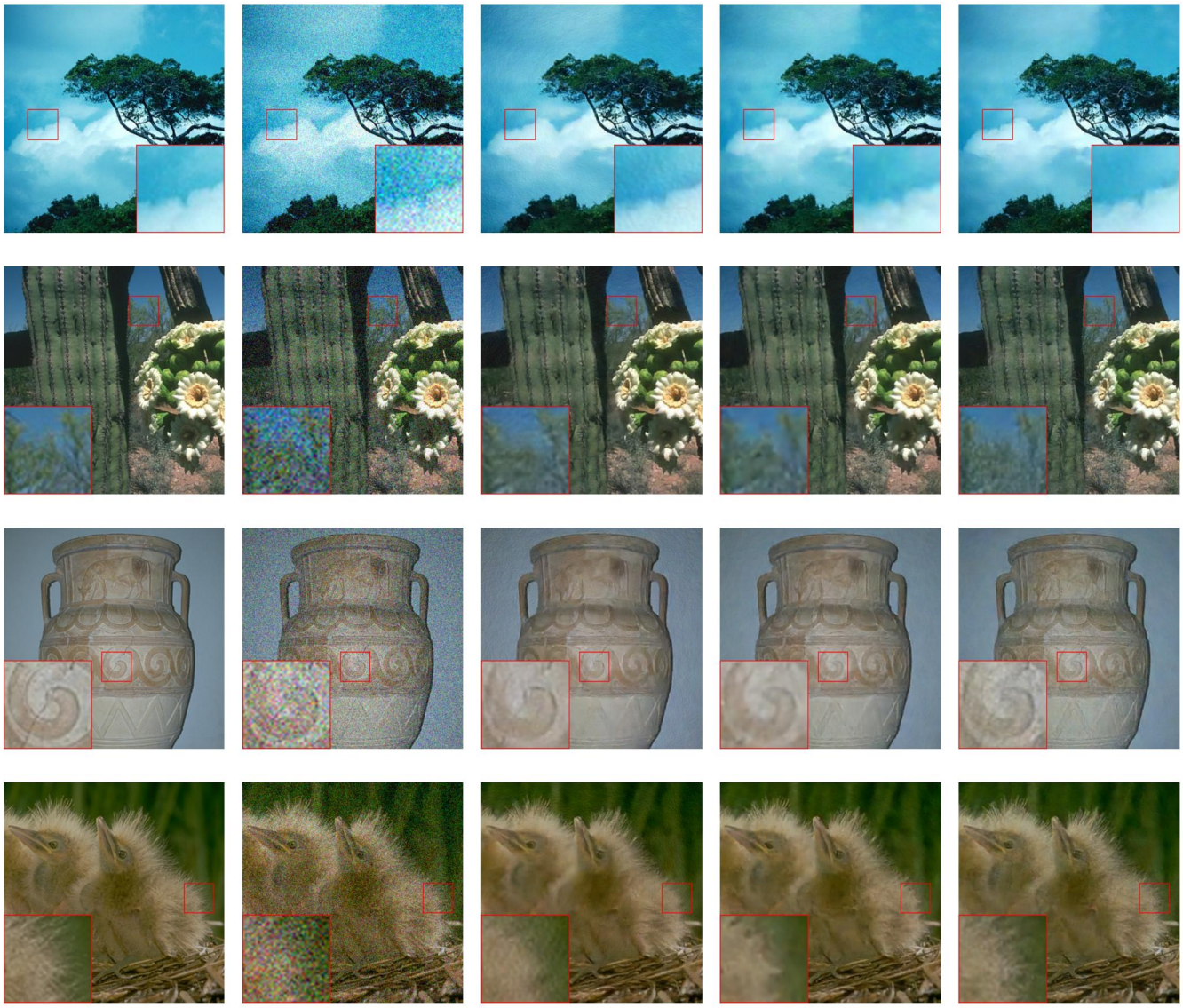
We compare the PSNR in Table 1 and the qualities of denoised images in Figure 5 generated from these three models. We also compare how the PSNR changes with the noise level in Figure 6.

Figure 5 visually compares the denoised images from these three models. In this work, We used the same test dataset with sigma = 0.1 to evaluate. We can see the UNet can dramatically reduce most of the noise but it still has some not smooth parts with lots of blurs and artifacts. ResUNet can further remove the noise and recover more details. In the first row, we can see the sky and cloud generated from ResUNet are more clear and have a stronger color comparison than the result from UNet. The artifacts in the sky disappeared and it has a better definition of the boundary between the cloud and the background. In the third row, from the results obtained by ResUNet, we can also figure that the texture on the vase is clearer than the result of UNet. It illustrates more details and more realistic light reflection. Our PSNR supports our findings as well. The PSNR of ResUNet results is around 30.11db, while the PSNR of UNet results is around 29.32db. The PSNR increases by 0.8. All the evidence can prove that our Res-blocks have a great effect on the image denoising performance.

Figure 6 displays how the PNSR varies with the noise level. Both Unet and ResUnet have excellent performance (PSNR over 43db) on lower noise level and the performance gets worse as noise level goes up. However we notice that ResUnet has more stable performance than Unet in higher noise level.

Compared the results generated by ResUNet with perceptual loss with others, we can observe this method can not only significantly restore more details but also strongly strengthen the boundaries and edges. In the third row, the results of ResUNet with perceptual loss can even indicate more details of textures on the vase than the result of ResUNet. All boundaries of the textures are clear and easy to see. In the fourth row, similar to the previous case, most of the fluff on the chick can be clearly visible. We argue that these improvements are contributed by perceptual loss. Based on the idea of being closer to perceptual similarity, perceptual loss is proposed to enhance the visual quality by minimizing the error in a feature space instead of pixel space. Each layer in the VGG-16 learns different small features of images so adding the loss from different layers to our loss function is actually extracting more information from the image.

However, the PSNR on the results generated by ResUNet with perceptual loss is around 29.50dB, which is slightly lower than the PSNR (30.11dB) on the results generated without the perceptual loss. The main reason for this is that perceptual loss can make the denoised images have sharper features so that the remainder of the pixels are



(a) original (b) noisy (c) Unet (d) ResUnet (e) ResUnet with $\mathcal{L}_{per}$

Fig. 5: Illustration of performance of different denoising approach. From left to right: original image, noisy image, UNet denoising, ResUNet denoising, ResUNet denoising with perceptual loss

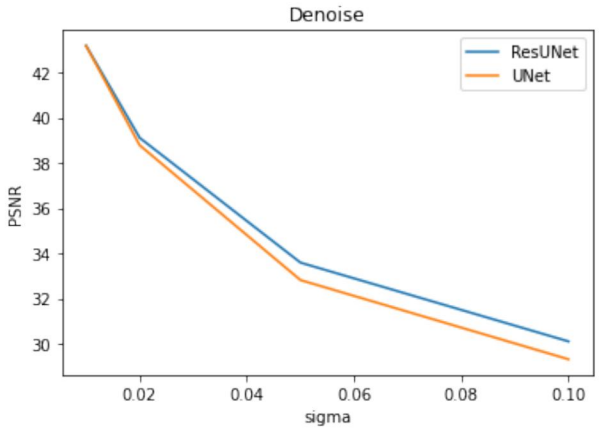
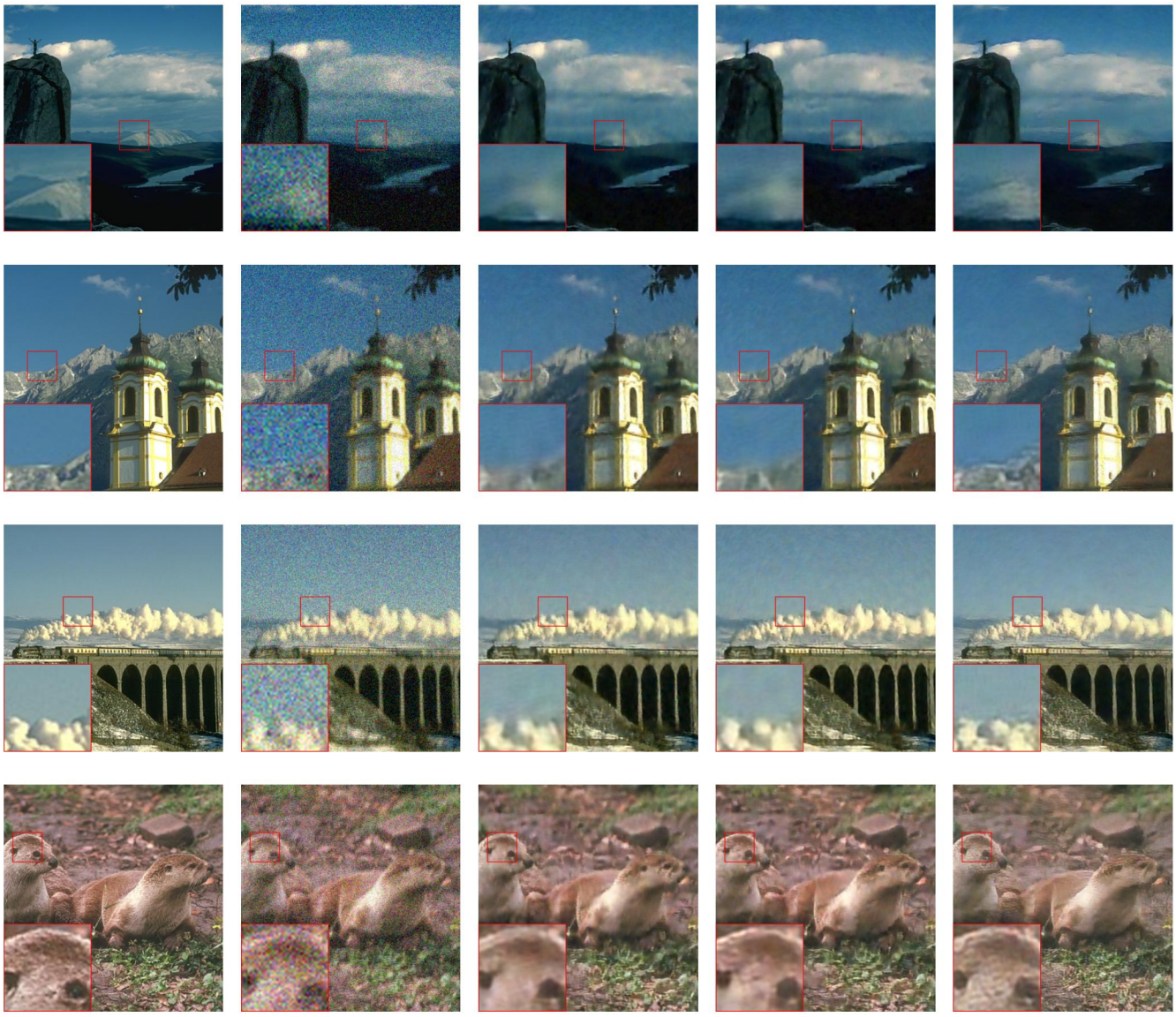


Fig. 6: Noise Level vs. PSNR in denoise task

TABLE 1: Compare PSNR of denoised images from three models

sigma	UNet	ResUNet	ResUNet with $\mathcal{L}_{per}$
0.01	43.1701	43.1752	-
0.02	38.7747	39.1049	-
0.05	32.8163	33.5873	-
0.1	29.3147	30.1055	29.5025

generally less important from a human viewer's perspective of image quality. On the other hand, from the human viewer's perspective of images, people believe that images with clearer boundaries and sharper features have higher quality, even though the results don't exactly match the ground truth. Another reason for the decreased PSNR is adding the perceptual loss changes the image color slightly. The denoised images look a little bit lighter than the ground



(a) original (b) noisy (c) Unet (d) ResUnet (e) ResUnet with $\mathcal{L}_{per}$

Fig. 7: Illustration of performance of different deblurring and denoising approach. From left to right: original image, noisy image, UNet denoising, ResUNet denoising, ResUNet denoising with perceptual loss

truth. This may be due to the loss of which layer we added from our VGG-16. In future work, we can try to add the loss from other layers to adjust the image color. Besides, we also consider that the patch size of our training samples is too small (32*32), which may result in the deeper layers of our VGG-16 fail to learn useful information.

In this image denoising task, our purposed model ResUNet is able to achieve the best PSNR and it is more resistant to higher noise level than Unet. In addition, to a certain extent, the work we explored with perceptual loss indicates that it is helpful to the restoration of the images, and it also gives us a new understanding of evaluating the image quality.

4.3 Experiments on Deblur and Denoise Task

Similar to the previous task, We train three different deblur-denoise networks: UNet with MSE loss, ResUNet with MSE loss and ResUNet with MSE loss and perceptual loss on the same training samples with different noise levels with learning rate 5e-4. This time, we also clipped the training samples and trained on the blurred and noisy images.

We compare the PSNR results in Table 2 and the qualities of deblurred-denoised images in Figure 7 generated from these three models. Again, We also compare how the PSNR changes with the noise level in Figure 8.

In this work, we evaluated all the models on the same test dataset with the same blur kernel and noise level as the training set respectively. From these images, we can find that UNet is capable of removing most of the noise but the retrieved images still have lots of distortions, and objects are pretty blurry. The boundaries of objects are too vague to recognize. However, our purposed model ResUNet can recover better the background of the images. The results

generated by ResUNet display that it can further eliminate distortions and smooth images. In the second row, we can see the sky background from ResUNet is clearer and purer than the result from UNet. Besides, we also notice that ResUNet can enhance details and sharper the edges of objects. In the fourth row, the marmots faces recovered by ResUNet have more black patterns than the ones recovered by UNet. The stone in the right top indicates very sharp edges and corners, and the marmots also have very clear outlines, which make them stand out from the background very well.

The results generated by ResUNet only have minor outperformance than UNet. There are not significant visual differences between ResUNet results and UNet results. The PSNR on ResUNet deblurred-denoised images is 35.86dB, which is only 0.2 higher than The PSNR on Unet deblurred-denoised images. We believe that ResUNet does have crucial impact on image denoising, but its performance may be weakened due to the blur that we added. In the future work, we will consider trying different combinations of blur level and noise level to investigate the utility of our purposed model on deblur task.

In the Figure 8, we can observe that both of the Unet and ResUnet have good deblur-denoise performances in lower noise lever and the performances get worse with noise level increases. But the ResUnet is always slightly better than Unet.

Comparing the results retrieved by ResUNet with perceptual loss with others, we discern that it can further smooth the hazy and ambiguous areas especially the single color blocks, and it can generate stronger edges and textures of objects. In the first row, the mountain and its boundary in the image restored with perceptual loss is remarkably clearer than the one without the perceptual loss. In the third row, in the image generated with perceptual loss, we can observe more details and small shallow parts in the smoke. It not only has sharper light and shadow borders but also contains a more prominent colour contrast to distinguish the smoke from the sky. What's more, we noticed that the images recovered with perceptual loss is more realistic than the ones recovered without perceptual loss. In the fourth row, the image retrieved with perceptual loss emphasizes the furs of the marmots and the shape of the marmots, which makes people feel the fuzzy of the marmots. It also weakens the existence of the background to highlight the presence of the marmots.

The PSNR of on the deblured-denoised images generated by ResUNet with perceptual loss is 25.46dB. Even though it is 0.4 lower by the PSNR on the restored images without perceptual loss, it still have impressed advantages on deblur and denoise task. The reasons about the decreased PSNR has been discussed in the previous part, such as the deblurred-denoised images look good in human vision but not exactly match the ground truth and the image color changed a little bit.

In this image deblur-denoise task, our purposed model ResUNet is able to achieve the best PSNR even if there is no huge difference among these three models. Our results show that ResUNet works well on image denoising but is sensitive to blur. Adding the perceptual loss is helpful with the deblurring job and enhances the vividness of image but we still need more experiments to improve it.

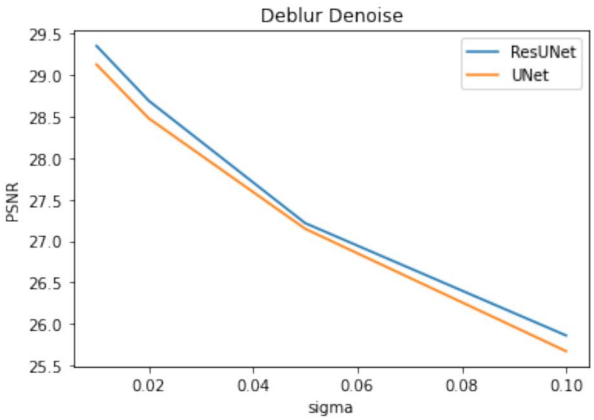


Fig. 8: Noise Level vs. PSNR in deblur-denoise task

TABLE 2: Compare PSNR of deblurred and denoised images from three models

sigma	UNet	ResUNet	ResUNet with $\mathcal{L}_{per}$
0.01	29.1276	29.3524	-
0.02	28.4789	28.6914	-
0.05	27.1468	27.2149	-
0.1	25.6711	25.8606	25.4638

5 Conclusion

Upon the task of image denoising, we came up with this new method that plug the residual neural network into UNet. Comparing to the original U-Net architecture, the performances of our porposed model ResUnet on denoising tasks are obviously superior. By using the updated loss function with perceptual loss, more distinct edges, more vivid colors and more abundant details are able to be achieved. Besides, regarding to the experiment of deblurring-denoising, we discuss the strength and limitation of the proposed architecture. Further work on using neural network to denoise images are worth investigating.

References

[1] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 2. Ieee, 2005, pp. 60–65.

[2] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[3] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2862–2869.

[4] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[7] V. Jain and S. Seung, "Natural image denoising with convolutional networks," *Advances in neural information processing systems*, vol. 21, 2008.
[8] G. Singh, A. Mittal, and N. Aggarwal, "Resdnn: deep residual learning for natural image denoising," *IET Image Processing*, vol. 14, no. 11, pp. 2425–2434, 2020.
[9] J. Gurrola-Ramos, O. Dalmau, and T. E. Alarcón, "A residual dense u-net neural network for image denoising," *IEEE Access*, vol. 9, pp. 31 742–31 754, 2021.
[10] D. H. Johnson, "Signal-to-noise ratio," *Scholarpedia*, vol. 1, no. 12, p. 2088, 2006.
[11] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.
[12] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.
[13] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.
[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[15] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, 2001, pp. 416–423 vol.2.