IMAGE-TO-IMAGE MLP-MIXER FOR IMAGE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks are highly effective tools for image reconstruction problems such as denoising and compressive sensing. To date, neural networks for image reconstruction are almost exclusively convolutional. The most popular architecture is the U-net, a convolutional network with a multi-resolution architecture. In this work, we show that a simple network based on the multi-layer perceptron (MLP)-mixer enables state-of-the art image reconstruction performance without convolutions and without a multi-resolution architecture, provided that the training set and the size of the network are moderately large. Similar to the original MLP-mixer, the image-to-image MLP-mixer is based exclusively on MLPs operating on linearly-transformed image patches. Contrary to the MLP-mixer, we incorporate structure by retaining the relative positions of the image patches. This imposes an inductive bias towards natural images which enables the image-to-image MLP-mixer to learn to denoise images based on fewer examples than the original MLP-mixer. The image-to-image MLP-mixer requires fewer parameters to achieve the same denoising performance than the U-net and its parameters scale linearly in the image resolution instead of quadratically as for the original MLP-mixer. If trained on a moderate amount of examples for denoising, the image-to-image MLP-mixer outperforms the U-net by a slight margin. It also outperforms the vision transformer tailored for image reconstruction and classical un-trained methods such as BM3D, making it a very effective tool for image reconstruction problems.

1 INTRODUCTION

Deep neural networks have emerged as highly successful tools for image and signal reconstruction, restoration, and manipulation. They achieve state-of-the-art image quality on tasks like denoising, super-resolution, image reconstruction from few and noisy measurements, and image generation.

Current state-of-the-art image reconstruction networks are convolutional. Convolutional neural networks (CNNs) achieve better denoising image quality than classical methods such as BM3D (Zhang et al., 2017; Brooks et al., 2019). They also perform excellent on many other imaging problems including computed tomography (McCann et al., 2017) and accelerated magnetic resonance imaging (MRI) (Zbontar et al., 2018). For example, all top-performing methods at the FastMRI competition, a challenge for accelerated magnetic resonance imaging (Zbontar et al., 2018; Knoll et al., 2020), are CNNs.

For the related problem of *image classification* CNNs are also state-of-the-art. However, recent work has shown that new non-convolutional networks can perform comparable when trained on huge datasets. For instance, the vision transformer (Dosovitskiy et al., 2021) is an attention-based architecture without convolutions that achieves excellent classification accuracy when pre-trained on very large datasets. Most recently, networks solely based on multi-layer perceptrons (MLPs) were proposed, including the MLP-mixer (Tolstikhin et al., 2021; Liu et al., 2021a; Chen et al., 2021). Trained on a huge dataset, the MLP-mixer performs almost as well as the best convolutional architectures while having lower computational costs at inference.

Non-convolutional architectures such as the ViT and MLP-mixer impose a lower inductive bias than CNNs. This inductive bias enables CNNs to perform well when little to moderate amounts of training data are available, but might limit performance if abundant data is available.

Motivated by this development, and by the simplicity of the MLP-mixer, we propose and study a variant of the MLP-mixer for image reconstruction tasks, with the premise that such a network can give better image quality than convolutional networks if trained on sufficient data.

The architecture of the image-to-image MLP-mixer is depicted in Figure 1. The image-to-image MLP-mixer differs to the MLP mixer in that it retains the relative positions of the patches, which leads to significantly better performance for image reconstruction tasks.

Our results show the image-to-image mixer can outperform a state-of-the-art image reconstruction architecture, the U-net Ronneberger et al. (2015), by a small margin. We show that the gap in performance between the image-to-image mixer and a U-net increases with the number of training images and the model size (see Figures 2 and 3). We also show that, even in the regime of relatively few training images, the image-to-image MLP-mixer slightly outperforms a U-net of similar size in image quality, both for denoising images perturbed with Gaussian noise, denoising images perturbed by real-world camera noise, and for compressed sensing reconstruction in magnetic resonance imaging. Phrased differently, to achieve the same denoising performance, the image-to-image MLP-mixer requires fewer parameters (see Figure 2). The MLP-mixer also outperforms a vision transformer tailored to image-to-image tasks, and BM3D, a classical un-trained denoising algorithm at denoising.

2 IMAGE-TO-IMAGE MLP-MIXER NETWORK ARCHITECTURE

In this section, we introduce an image-to-image MLP-mixer architecture that builds on the original MLP-mixer (Tolstikhin et al., 2021). The image-to-image MLP-mixer operates on linearly transformed image patches, just like the MLP-mixer, as illustrated in Figure 1. However, contrary to the MLP-mixer, the image-to-image mixer imposes some structure by retaining the spacial order of image patches, which turns out to be critical for image reconstruction performance.

We start by splitting the image into non-overlapping patches of size $P \times P \times 3$ (our default choice is $P = 4$). Each patch is viewed as a vector of dimension $3P^2$ that is linearly transformed with the same trainable matrix to a space of arbitrary embedding dimension C. This patch embedding step thus transforms an image of dimension $H \times W \times 3$ (or $H \times W \times 1$ for greyscale images) to a volume of dimension $H/P \times W/P \times C$. The patch embedding step retains the relative positions of the patches in the image. The MLP-mixer and the vision transformer (Tolstikhin et al., 2021; Dosovitskiy et al., 2021) also split an image into patches and linearly project the patches, and so do several other architectures for example the swin transformer (Liu et al., 2021b).

We then apply an MLP-mixer layer inspired by the original MLP-mixer module. This MLP-mixer layer mixes the tensor in height dimension, then in width dimension, and finally in channel dimension. Mixing in channel dimension means viewing the tensor of dimension $H/P \times W/P \times C$ as a collection of $H/P \cdot W/P$ vectors of dimension C and passing each of them through the same MLP consisting of a linear layer, followed by a GeLU-non-linearity and then another linear layer. The hidden layer dimension is the input dimension of the respective vector multiplied by a factor of f. We also add skip connections and layer norms to help with the optimization. A mixer layer does not alter the dimensions of the input volume.

After N many such mixer layers, the volume is transformed back to an image via a patch expansion step. The patch expansion step transforms the volume consisting of flattened patches, each of dimension C, back to an image of dimension $H/P \times W/P \times 3$ as follows: First, we linearly transform each patch of dimension C to a patch of dimension CP^2 using a shared linear transformation. This maps the volume of shape $H/P \times W/P \times C$ to a volume of shape $H/P \times W/P \times CP^2$. Second, we reshape the volume to a volume of shape $H \times W \times C$, and finally transform this volume to an image of shape $H \times W \times 3$ by linearly combining the layers (which can be implemented with a 1×1 convolution). A similar patch expansion step has been used by the Swin U-net Transformer (Cao et al., 2021).

The main difference between our image-to-image MLP-mixer architecture and the original MLP-mixer is that we transform the image to a 3D tensor instead of a 2D tensor, and the mixer layer is modified to act on a 3D volume. This modification retains the relative location of the patches in the 3D volume which induces an inductive bias enabling the image-to-image MLP-mixer to perform

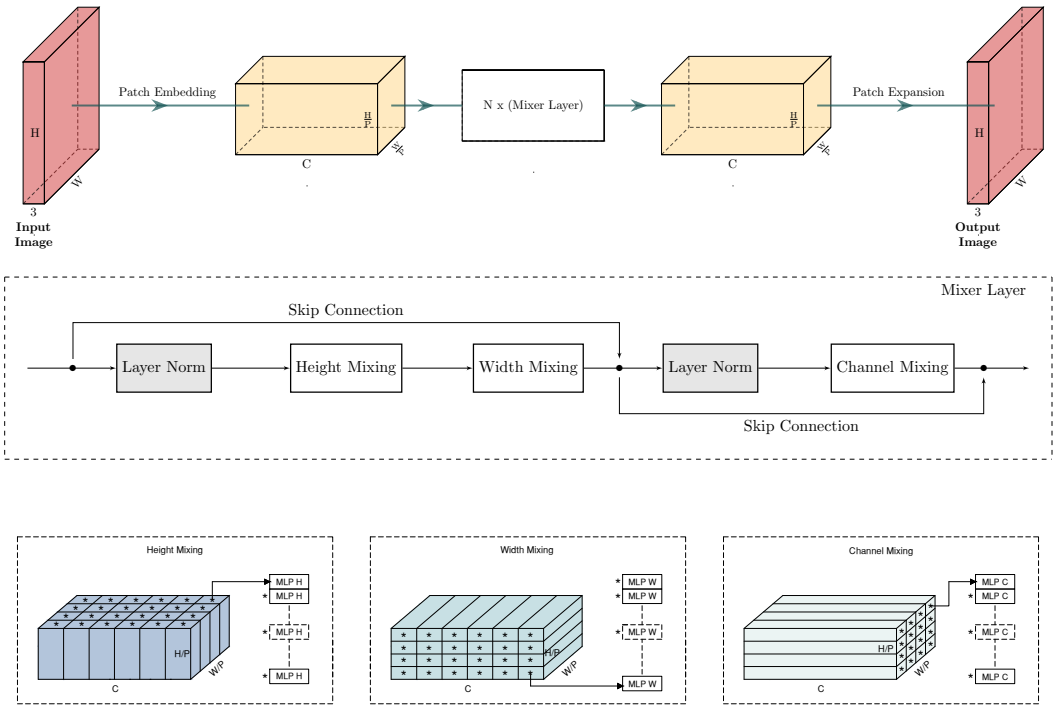

Figure 1: An illustration of the image-to-image MLP-mixer: First, the image is divided into patches which are transformed with a linear layer to a volume of size $W/P \times H/P \times C$. This volume is transformed with 3D mixer-blocks (16 in our standard architecture), and finally projected back with a linear transformation to an image. There are many potential choices for the mixer block, our default one mixes across channels, width, and height separately. For example, mixing in channel dimension means applying the same MLP to each of the $H/P \cdot W/P$-many vectors in channel direction. The mixer block we propose retains the relative order of the patches.

very well when trained on relatively few images. As we show later in Section 3.4, the inductive bias is less than that of a convolutional network, but more than the original MLP-mixer.

A further difference of the image-to-image-Mixer over the original MLP-mixer is the scaling of the number of parameters: The trainable parameters of the token mixing in the original mixer are $\mathcal{O}\left(H^2W^2\right)$, while the height- and width mixing of the image-to-image MLP-mixer are $\mathcal{O}\left(H^2 + W^2\right)$. The linear scaling in image resolution of the image-to-image MLP-mixer keeps the total number of trainable parameters low and the architecture memory efficient.

3 Experiments

We evaluate the performance of the image-to-image mixer for a variety of image reconstruction problems. We focus on image denoising as it is considered to be a fundamental image reconstruction problem, for its practical importance, and since a good denoiser typically serves as a building block for other tasks such as recovering images from few and noisy measurements. For example, a state-of-the-art approach for reconstructing an image from few and noisy linear measurements is a so-called variational network which uses a denoiser as a building block (Sriram et al., 2020). Reconstructing an image from few and noisy linear measurements is an important inverse problem that arises in accelerated magnetic resonance imaging and sparse-view computed tomography.

Baseline methods: We compare the denoising performance of the image-to-image MLP mixer to three baselines: BM3D (Dabov et al., 2007), a standard and well performing denoising algorithm that does not rely on any training data. The U-net (Ronneberger et al., 2015), a standard image-

to-image convolutional network that is a go-to for image reconstruction problems. The U-net performs slightly better than a standard multi-layer convolutional network for image denoising (Brooks et al., 2019) (for example better than the multi-layer convolutional network proposed by Zhang et al. (2017)). We also compare to the vision transformer (Dosovitskiy et al., 2021), which we adapted for image recovery tasks as follows. We disposed the classification token and replaced the classification head by a linear layer that maps each element of the transformer output to a corresponding image patch.

All networks (the image-to-image MLP-mixer, U-net, and ViT) are all trained in the same fashion as described next.

3.1 Gaussian Denoising

We first consider the problem of removing Gaussian noise from ImageNet color images (Deng et al., 2009). We constructed a dataset as follows: We collected images of different classes from ImageNet and center-cropped them to a size of $256 \times 256 \times 3$. We then added zero-mean Gaussian noise of standard deviation $\sigma = 30$ to each image channel independently, resulting in a data set consisting of pairs of noisy image $\mathbf{y}_i = \mathbf{x}_i + \mathbf{z}_i$ and corresponding clean image $\mathbf{x}_i$. Here, $\mathbf{z}_i$ is the Gaussian noise. The noisy images have a peak signal-to-noise-ratio (PSNR) of 19 dB.

We trained the image-to-image MLP-mixer $f_{\boldsymbol{\theta}}$ with trainable parameters $\boldsymbol{\theta}$ (and the baseline architectures) to map the noisy image to the noise by minimizing the loss function

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \|\mathbf{y}_i - f_{\boldsymbol{\theta}}(\mathbf{y}_i) - \mathbf{x}_i\|_2^2.$$

Here, n is the total number of training images. At inference, we are given a noisy image $\mathbf{y}$ and estimate a clean image by subtracting the estimated residual from the noisy observation as $\hat{\mathbf{x}} = \mathbf{y} - f_{\boldsymbol{\theta}}(\mathbf{y})$. This is referred to as residual learning (Zhang et al., 2017), because the network learns to predict the residual. Training the network directly to map a noisy image to a clean image also works, but performs worse than residual learning for all architectures considered here.

We split the data set into train and test sets and ensured that images from the same ImageNet class do not exist in both sets simultaneously. This guarantees that the network is not just learning to denoise a specific class only.

In Figure 2, we depict the denoising performance of the different architectures as a function of the number of training examples, ranging from 1000 to 100.000 training images, with constant model size, and as a function of the number of parameters, with constant training set size. The plots show that even in the regime of small training data (**left:** 4000 images) and small model size (**middle:** 3 million parameters), the image-to-image mixer can outperform the Unet. Thus, the image-to-image MLP-Mixer is more parameter effective in that it outperforms the U-net with fewer parameters, i.e. a 3M version of the image-to-image mixer performs slightly better than a 12M version of the Unet.

Most importantly, Figure 2 shows that the image-to-image mixer scales better than the U-net when both the dataset size and the size of the models grow. Moreover, the right panel in Figure 2 shows that for large models (24M) the gap in performance between the image-to-image mixer and the U-net increases: The U-net shows a relatively smaller accuracy improvement when increasing the training set size from 10k to 100k. Thus, we expect even larger improvements when moving to even larger datasets.

In the experiment, the model parameters of the image-to-image mixer are varied by changing the number of channels and the hidden dimension of the MLPs. The exact hyperparameter configurations are in Table 3 in the appendix. For the U-net, we increased the model size by increasing the number of channels, and for ViT we increased the model size by increasing both its depth and width.

3.2 Denoising performance on real-world camera noise

We next evaluate the performance of the image-to-image MLP-mixer on real-world camera noise, which is often not well approximated by Gaussian noise. We evaluated the architectures on the Smartphone Image Denoising Dataset (SIDD) (Abdelhamed et al., 2018). The data set consists of high-resolution images from 10 scenes obtained under different lighting conditions with five

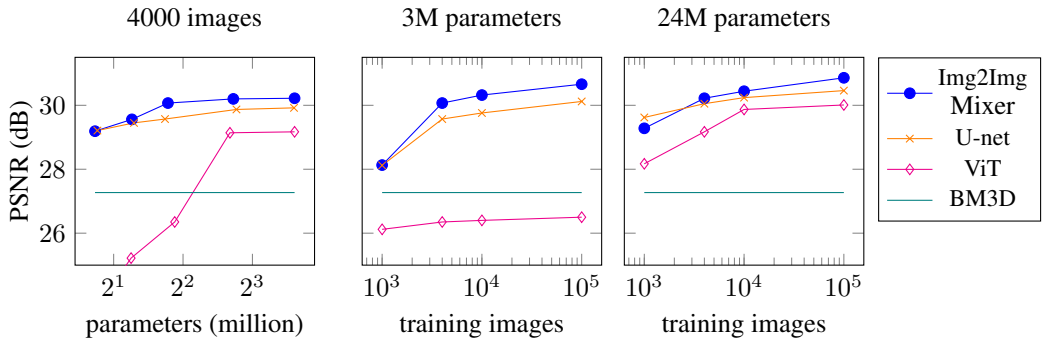

Figure 2: **Left:** Gaussian denoising performance for image-to-image MLP-mixer networks of different sizes trained on 4000 noisy images of 19 dB PSNR. The mixer is more parameter efficient in that it outperforms the U-net and ViT for almost all sizes considered. **Middle:** Gaussian denoising performance as a function of the number of training images for a network of size 3.4 million parameters. For a very small number of training images, the U-net performs on paar with the image-to-image MLP-mixer, but in the regime where a moderate number of training examples is available, the mixer slightly outperforms the U-net. **Right:** Same as the middle plot, but with a larger network size. It can be seen, that the Unet starts reaching a plateau as the training images increase, but the image-to-image mixer continues to exhibit an improvement in performance.

representative smartphone cameras. We center-cropped a 2048×2048 patch from each image and divided that into non-overlapping images of size 256×256. We used 4700 images from 8 of the scenes for training and 700 images from the remaining 2 scenes for testing.

The noisy images have a PSNR of 24 dB, 5 dB higher than the noisy ImageNet images, suggesting that the SIDD images have a much lower noise level than the noisy imagenet images we denoised in the previous section. The image-to-image MLP-Mixer achieved a denoising PSNR of 33.66 dB whereas the U-net and the ViT obtained slightly lower values of 33.13 dB and 32.87 dB, respectively. BM3D got 28.87 dB. BM3D has one hyperparameter for the noise level, this we set to the noise variances estimated from the ground truth, which resulted in the best performance for BM3D.

This experiment shows that the image-to-image MLP-mixer can obtain state-of-the-art denoising performance on real-world noise as well.

3.3 Compressive sensing

Next, we evaluate the image-to-image-mixer on the task of recovering an image $\mathbf{x} \in \mathbb{R}^n$ from few linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$, with $m < n$, is a wide and known measurement matrix. This compressive sensing problem arises in sparse-view tomography and accelerated magnetic resonance imaging. Our results show that, similar as in the previous section, that the performance of the image-to-image MLP-mixer scales well with number of training images and size of the network.

A standard approach to address the compressive sensing problem with a neural network is to first compute a coarse reconstruction via least-squares as $\mathbf{A}^\dagger \mathbf{y}$ and then train a neural network to map the coarse least-squares reconstruction to a clean reconstruction by minimizing the loss $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(1 - \mathrm{SSIM}(f_{\boldsymbol{\theta}}(\mathbf{A}^\dagger \mathbf{y}_i), \mathbf{x}_i)\right)$, where $f_{\boldsymbol{\theta}}$ is a neural network with parameters $\boldsymbol{\theta}$ mapping an image to an image. Here, SSIM is the structural similarity index metric (Zhou Wang et al., 2004), a metric indicating the visual similarity between two images, larger is better, and a value of 1 indicates that the two images are equivalent. This approach has been pioneered by Jin et al. (2017) for computational tomography and serves as a baseline for a competition for accelerated MRI, called FastMRI (Zbontar et al., 2018).

We evaluate the image-to-image mixer, the U-net, and the ViT on a four-times accelerated MRI knee-reconstruction problem (i.e., $m = n/4$). We trained the networks of equal size of about 8 million parameters on the FastMRI knee training dataset containing 2k, 10k, 17k, and 35k training images and evaluated their performance on the FastMRI knee validation set. Figure 3 depicts the

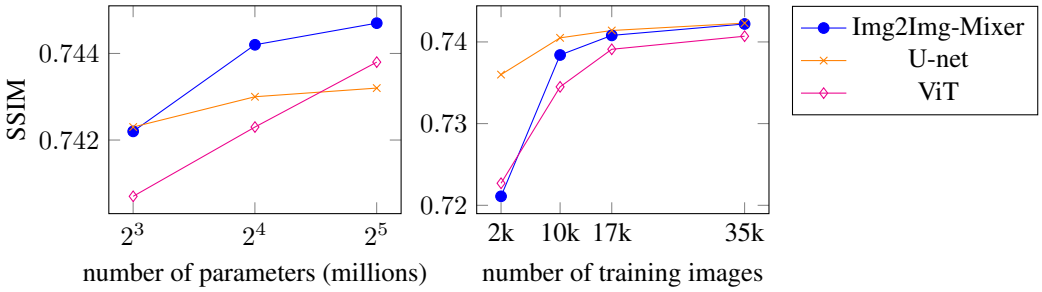

Figure 3: **Left:** Compressed sensing performance for different models sizes when trained on the entire fastMRI knee dataset (35k data). **Right:** Compressed sensing performance as a function of the number of training images for networks of size about 8 million parameters. For a small number of training images, the U-net reaches slightly higher SSIM than the image-to-image MLP-mixer, but this performance gap quickly diminishes as training data grows.

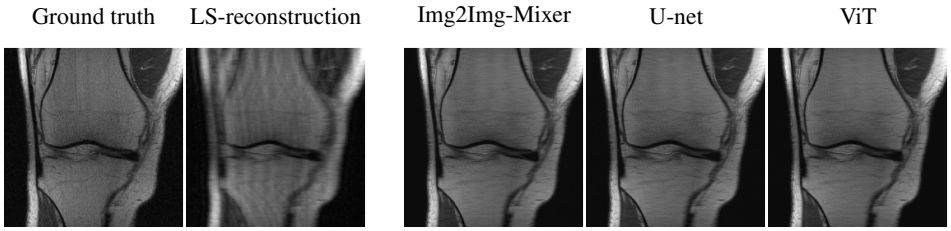

Figure 4: We trained the different architectures to map a coarse least-squares (LS)-reconstruction to the ground truth. All networks yield very similar reconstruction performance, three example images are given, for training on 35k images.

reconstruction performance as a function of the number of training examples. Example reconstructions are given in Figure 4. In this experiment, all three architectures yield similar performance. The same trends as in the denoising experiment in Figure 2 hold true: that the performance of the image-to-image MLP-mixer scales well with number of training images and size of the network, and surpases that of U-net if sufficiently large and trained on sufficiently many images.

The results show that the image-to-image mixer yields competitive performance beyond plain denoising tasks. Together, our denoising and compressive sensing results demonstrate that convolutions and a multi-resolution architecture are not necessary for state-of-the-art imaging performance.

3.4 Measuring the inductive bias of the Image-to-Image Mixer

In this section we measure the inductive bias of the the image-to-image architectures considered here. We find that low inductive bias correlates with more significant performance improvements as both the model size and size of the dataset are increased: Both the image-to-image MLP-mixer and the ViT have a lower inductive bias (as shown in this section) show a larger increase in performance as the number of parameters and dataset are increased (see Figure 2).

Convolutional neural networks have an inductive towards natural images in that they are well suited to generate natural images. The inductive bias of convolutional neural networks is so strong that a convolutional neural network can perform image reconstruction without any training. This has first been shown for a U-net in the deep image prior paper: Ulyanov et al. (2018) has shown that a randomly initialized, un-trained U-net fits a natural image with significantly fewer gradient descent iterations than it fits noise. This effect can be reproduced with a very simple convolutional network, without any skip connections and without an encoder-decoder structure (Heckel & Hand, 2019). This inductive bias has been theoretically explained by wide convolutional networks trained with gradient descent fitting the lower frequencies of a signal before fitting the higher frequencies (Heckel & Soltanolkotabi, 2020). Since a natural image has much of its energy concentrated

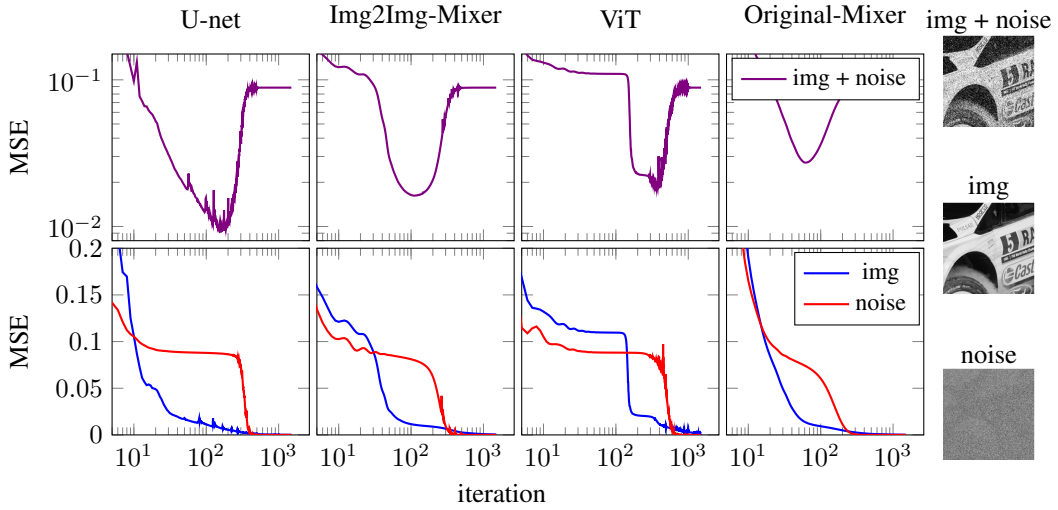

Figure 5: Measuring the inductive bias of architectures. We fitted the {Img2Img-Mixer,U-net,ViT, Original-Mixer} to map a random input to (a) an image, (b) Gaussian noise, and (c) the image plus noise. MSE denotes Mean Square Error of fitting the network output with respect to the clean image for img and img + noise, and with respect to the noise for noise. All networks have significantly more parameters than pixels and can fit both noise and a natural image perfectly, but have an inductive bias, in that they fit the image faster than the noise.

on low-frequency components, a natural images is fitted faster than Gaussian noise which has in expectation the same energy on all components.

Motivated by this observation, we measured the inductive bias of the image-to-image mixer, the original MLP-mixer, the U-net, and the ViT, by fitting the respective randomly initialized networks to i) a natural image, ii) Gaussian noise, and iii) the natural image plus the Gaussian noise. The three signals are displayed in Figure 5 along with the training curves obtained by minimizing the loss $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\text{signal} - f_{\boldsymbol{\theta}}(\mathbf{y})\|_2$ with Gradient descent. Here, signal is the respective signal (i.e., img, noise, and img+noise), and $f_{\boldsymbol{\theta}}(\mathbf{y})$ is the respective network, initialized randomly, and fed with a random input $\mathbf{y}$.

Figure 5 shows that the U-net has a larger inductive bias than the image-to-image mixer, which in turn has a larger inductive bias than the ViT and the original MLP-mixer: without any training, U-net achieves a denoising performance of 20.2 dB, the image-to-image mixer of 18.1 dB, the ViT of 17.3 dB and the original MLP-mixer of 16.3 dB. Interestingly, all four networks have an inductive bias in that they fit a natural image significantly faster than noise.

Figure 6 in the appendix illustrates the type of inductive bias: The convolutional U-net has an inductive bias towards smooth signals, the image-to-image mixer towards fitting vertical and horizontal lines first.

3.5 Ablation studies

In this section we discuss a few variants of the image-to-image mixer in order to understand which elements of the network are critical for its image reconstruction performance.

Linear MLP-mixer layers. The perhaps most interesting variation of the image-to-image mixer works with linear mixer layers. Recall that the height mixing, width mixing, and channel mixing blocks all consist of a linear transformation, non-linearity, followed by another linear transformation. We studied a mixer version where instead of three such mixer layers we only have one mixer layer, which has a first linear layer mixing in height dimension, a second linear layer mixing in width dimension, a third linear layer mixing in channel dimension, followed by a non-linearity, and finally a linear layer in channel dimension. This mixer architecture is extremely simple and, perhaps surprisingly, performs almost as well as the default image-to-image mixer architecture introduced

earlier. Specifically, we designed a 3.5 M version with linear MLP-mixer layers and trained it on 4000 ImageNet images. It achieved 29.92 dB, only 0.15 dB less than the default Img2Img-Mixer of similar model size.

Incorporating multi-resolution. The most successful architectures to date for image reconstruction and dense predictions incorporate a notion of multi-scale. For example, the U-net (Ronneberger et al., 2015) transforms an image by first decreasing the spacial dimensions and increasing number of channels, and second increasing the spacial dimensions while decreasing the number of channels. Even image-to-image transformers (or attention-based networks) incorporate such multi-resolution structure (Liu et al., 2021b) successfully. For convolutional architectures, incorporating such multi-scale architecture improves performance, in that the U-net outperforms a standard multi-layer convolutional network at denoising (Brooks et al., 2019).

We incorporated such multi-resolution structures by implementing patch merging as in the Swin Transformer (Liu et al., 2021b) and patch expanding as in the Swin U-net Transformer (Cao et al., 2021). Patch merging can be seen as an encoding step, where the spatial dimensions (height and width) are reduced by a factor of two and the channel dimension is increased by a factor of two. Patch expanding acts as the decoder by reversing the merging operation, i.e., it increases the spatial dimensions and decreases the number of channels. The merging and expanding steps are implemented by linear transformations and reshaping as in the patch combining step. Figure 1 in the paper (Cao et al., 2021) visualizes the similar Swin Transformer architecture, but instead of swin transformer blocks we used the mixer layers.

Our results show that incorporating multi-resolution structure does not improve performance and instead marginally decreases performance. We considered a multi-resolution image-to-image mixer with 6.45 M parameters, and compared it to our proposed image-to-image mixer, the U-net, and ViT, on the Gaussian denoising experiment described in Section 3.1. The multi-resolution mixer achieved 28.77 dB, less than the U-net (29.87 dB), the Img2Img-Mixer (30.20 dB) and the ViT (29.14 dB). We also evaluated the multi-resolution architecture on the SIDD images, where it achieved 33.48 dB, slightly better than the U-net (33.13 dB) and the ViT (32.87 dB), but still slightly worse than the Img2Img-Mixer (33.66 dB).

Comparison to the original MLP-mixer. The original MLP mixer was designed for classification. We modified it to perform image reconstruction as follows. We omitted the global average pooling and fully connected layer at the end, and used a projection matrix to linearly transform the hidden dimension C back to dimension $3P^2$. That results in a volume of dimension $S \times D$, where $S = HW/P^2$ and $D = 3P^2$. Each row of dimension D of the volume represents a flattened image patch. By unflattening each row in the table, i.e., by reshaping each row to dimensions $P \times P \times 3$, we end up with an image of dimension $H \times W \times 3$.

We used a 6.87 M version of the modified original mixer and trained it on 4000 ImageNet images on the Gaussian denoising experiment (Section 3.1). The network achieved 27.36 dB, significantly worse than the image-to-image mixer (30.20 dB), Unet (29.87 dB) and ViT (29.14 dB). This shows that it is critical for performance to retain the position of the image patches during the mixing operations.

Effect of patch size. Our default Img2Img-Mixer has a patch size of $P = 4$ as in the Swin transformer (Liu et al., 2021b) and the Swin U-net transformer (Cao et al., 2021). Here, we evaluate versions of the network with varying patch sizes on the Gaussian denoising experiment described in Section 3.1. We varied the patch sizes and changed the other hyperparameters to keep the model size similar. Table 1 depicts the PSNR values for two model sizes. The results show that having a smaller patch size than 4 actually leads to a marginally better performance. However, a smaller patch size also leads to a higher computational cost, which is why $P = 4$ is a good tradeoff. The exact hyperparameter configurations are in Table 4 in the appendix.

Throughput We measured the throughput of the networks by calculating the average speed of a forward pass on the GPU at inference. Since the networks benefit from different batch sizes, we report the results in Table 2 for the best performing batch size and also for batch size = 1, which is most relevant if we process one image at inference, which is common for example in the MRI application we discussed. All networks have a size of about 3.4 M parameters. For batch size 1,

Size | P	1	2	4	8
3.4M	30.19	30.18	30.07	23.85
6.8M	30.26	30.26	30.20	24.01

Table 1: Denoising PSNR (dB) for varying patch sizes P. Smaller patch sizes than 4 work better, but only marginally and at the expense of higher computational cost.

the Img2Img-Mixe is faster than DnCNN (Zhang et al., 2017), a popular network for denoising, and similar than ViT, but slower than the Unet.

Batch Size	ViT	Unet	Img2Img-Mixer	DnCNN
1	116	287	89	28
best	690	524	98	30

Table 2: Throughput of the different networks measured at inference as number of images per second per GPU.

4 Related literature

Our work builds on the recently introduced MLP-mixer (Tolstikhin et al., 2021). While there are a number of works that also build on the MLP-mixer, to the best of our knowledge, this is the first work exploring a structured MLP-based architecture for image reconstruction tasks.

There are several recent works that build on the MLP-mixer for classification tasks: Chen et al. (2021) proposed to mix the spatial dimensions in a cyclic way resulting in an architecture that performs well on detection and segmentation tasks. The ResMLP network (Touvron et al., 2021) replaces the self-attention layers of a ViT by an MLP, yielding competitive image classification performance. RaftMLP (Tatsunami & Taki, 2021) modifies the MLP mixer for classification by mixing the spacial dimensions in a similar way as we do, and achieve a more parameter efficient model for classification. Cazenavette & De Guevara (2021) proposed an image-to-image GAN that utilizes MLP-mixer blocks followed by convolutional layers in the decoder part. Liu et al. (2021a) proposed to substitute attention with MLPs paired with gating, demonstrating that attention is not critical for ViTs to perform well.

We finally note that even a completely unstrutured MLP can perform well for denoising small image patches. Specifically Burger et al. (2012) trained an MLP to denoise image patches of size 17×17 and achieved performance competitive with BM3D.

5 Conclusion

We introduced and evaluated a simple architecture based on the MLP-mixer (Tolstikhin et al., 2021) for image-to-image reconstruction tasks. Image reconstruction tasks are currently dominated by convolutional networks that incorporate a multi-resolution structure such as the U-net. Our work shows that an architecture based on MLPs and without a multi-resolution structure gives even slightly better performance at both small, moderate, and large network sizes for denoising. The image-to-image architecture incorporates structure by retaining the relative positions of the image patches. This inductive bias is manifested by the image-to-image mixer enabling to denoise an image even without training data, but admittedly gives relatively poor image quality without training data. If trained on a moderate amount of images, the image-to-image MLP mixer slightly outperforms both the U-net as well as the ViT at synthetic denoising tasks and at real-world denoising tasks. For compressive sensing, we found all three architectures to perform very similarly. Our work shows that training on millions of images is not essential for non convolutional networks to compete with CNNs. Even in the regime of moderately sized training sets, CNNs can be outperformed. The necessity of massive datasets has been a limiting factor for further research in non-convolutional networks. We therefore hope that our work serves as a starting point for ending the dominance of CNNs in image processing tasks.

6 REPRODUCIBILITY STATEMENT

The code to reproduce the results in this paper is available at this URL: `https://www.dropbox.com/sh/0jb0368lp4ofw74/AACweyZExAsekWJCnD33wJr5a?dl=0`. The experiments were carried out on a server with four RTX6000 GPUs, most experiments reported here run on a single GPU for less than a day.

REFERENCES

Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A High-Quality Denoising Dataset for Smartphone Cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1692–1700, 2018.

Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing Images for Learned Raw Denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11036–11045, 2019.

H. C. Burger, C. J. Schuler, and S. Harmeling. Image denoising: Can plain neural networks compete with BM3D? In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2392–2399, 2012.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv:2105.05537 [cs, eess]*, 2021.

George Cazenavette and Manuel Ladron De Guevara. MixerGAN: An MLP-Based Architecture for Unpaired Image-to-Image Translation. *arXiv:2105.14110 [cs]*, 2021.

Shoufa Chen, Enze Xie, Chongjian Ge, Ding Liang, and Ping Luo. CycleMLP: A MLP-like Architecture for Dense Prediction. *arXiv:2107.10224 [cs]*, 2021.

K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Reinhard Heckel and Paul Hand. Deep Decoder: Concise Image Representations from Untrained Non-convolutional Networks. In *International Conference on Learning Representations*, 2019.

Reinhard Heckel and Mahdi Soltanolkotabi. Denoising and Regularization via Exploiting the Structural Bias of Convolutional Generators. In *International Conference on Learning Representations*, 2020.

K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep Convolutional Neural Network for Inverse Problems in Imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

Florian Knoll, Tullie Murrell, Anuroop Sriram, Nafissa Yakubova, Jure Zbontar, Michael Rabbat, Aaron Defazio, Matthew J. Muckley, Daniel K. Sodickson, C. Lawrence Zitnick, and Michael P. Recht. Advancing machine learning for MR image reconstruction with an open competition: Overview of the 2019 fastMRI challenge. *Magnetic Resonance in Medicine*, 84(6):3054–3070, 2020.

Hanxiao Liu, Zihang Dai, David R. So, and Quoc V. Le. Pay Attention to MLPs. *arXiv:2105.08050 [cs]*, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv:2103.14030 [cs]*, 2021b.

Michael T. McCann, Kyong Hwan Jin, and Michael Unser. Convolutional Neural Networks for Inverse Problems in Imaging: A Review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C. Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-End Variational Networks for Accelerated MRI Reconstruction. *arXiv:2004.06688 [cs, eess]*, 2020.

Yuki Tatsunami and Masato Taki. RaftMLP: Do MLP-based Models Dream of Winning Over Computer Vision? *arXiv:2108.04384 [cs]*, 2021.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-Mixer: An all-MLP Architecture for Vision. *arXiv:2105.01601 [cs]*, 2021.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. ResMLP: Feedforward networks for image classification with data-efficient training. *arXiv:2105.03404 [cs]*, 2021.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep Image Prior. In *Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454, 2018.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Matthew J. Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, James Pinkerton, Duo Wang, Nafissa Yakubova, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI. *arXiv:1811.08839 [physics, stat]*, 2018.

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*, 26(7): 3142–3155, 2017.

Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

A SUPPLEMENTARY MATERIAL

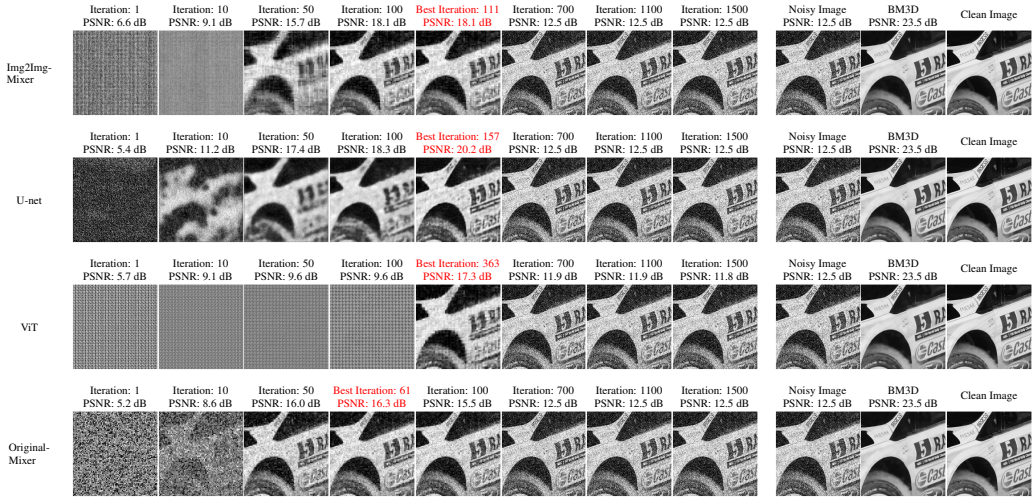

Figure 6: Visualizing the images obtained by fitting a single noisy image with different architectures. The U-net fits the low-frequency components of the image before the high-frequency one, while for the image-to-image mixer, we can see the structure imposed by mixing horizontally and vertically.

Size	P	N	C	f
1.66M	4	16	64	4
2.40M	4	16	96	4
3.44M	4	16	128	4
6.61M	4	16	128	8
12.19M	4	16	192	8

Table 3: Hyperparameter configuration for varying the network size of the Img2Img-Mixer in Figure 2.

Size	P	N	C	f
3.45M	1	12	107	1
3.46M	2	16	140	2
3.44M	4	16	128	4
3.46M	8	16	128	4

Size	P	N	C	f
6.81M	1	12	100	2
6.82M	2	16	140	4
6.61M	4	16	128	8
6.86M	8	16	140	8

Table 4: Hyperparameter configuration for varying the patch size of 2 model sizes of the Img2Img-Mixer in Table 1.