Deep Convolutional Networks for PET Super-Resolution

Kaan Özaltan[a], Emir Türkölmez[a], I. Jacques Namer[b], A. Ercüment Çiçek[a], and Selim Aksoy[a]

[a]Dept. of Computer Engineering, Bilkent University, Ankara, Turkey
[b]Dept. of Nuclear Medicine and Molecular Imaging, Strasbourg University, Strasbourg, France

ABSTRACT

Positron emission tomography (PET) provides valuable functional information that is widely used in clinical domains such as oncology and neurology. However, the structural quality of PET images may not be sufficient to effectively evaluate small regions of interest. Image super-resolution techniques aim to recover a high-resolution image from an input low-resolution version. We study adaptations of deep convolutional neural network architectures for improving the spatial resolution of PET images. The proposed super-resolution model involves a deep architecture that uses convolutional blocks together with various residual connections for more effective and efficient training. We use the supervised setting where the downscaled versions of the original PET images are given as the low-resolution input to the deep networks and the original images are used as the high-resolution target data to be recovered. Experiments show that the proposed model performs better than a multi-scale convolutional architecture according to both quantitative performance metrics and visual qualitative evaluation.

Keywords: Positron emission tomography, image super-resolution, convolutional neural networks

1. INTRODUCTION

Positron emission tomography (PET) is a medical imaging technique for mapping the positron emissions of injected radiopharmaceuticals into a 3D image where biological function can be accurately represented at molecular level.[1] As a result, PET has been an important modality in clinical domains such as oncology and neurology for the interpretation of various disorders such as dementias, epilepsy, infection, and brain tumors. However, the structural quality of PET is limited, unlike magnetic resonance (MR) images that convey high amounts of structural information. For instance, the widely used ^{18}Fluor-fluorodeoxyglucose PET may not have sufficient structural quality to effectively evaluate brain tumors[2] particularly when the target regions of interest are small.[3]

Image super-resolution (SR) techniques aim to recover a high-resolution (HR) image from an input low-resolution (LR) version.[4] The mapping for this ill-posed problem is learnt by using some form of prior that constrains the solution space. Deep learning-based SR models have recently been popular as they can directly learn the mapping from the LR to HR images without any explicit assumptions. These architectures typically make use of convolutional layers for an end-to-end learning of the mapping function.

This paper studies adaptations of deep convolutional neural network (CNN) architectures for improving the spatial resolution of PET images. The proposed architecture uses convolutional blocks together with various residual connections for more effective and efficient training. In the rest of the paper, we discuss several related work, describe the network architectures and the methodology for training these architectures in the supervised setting, and present experimental results with quantitative and qualitative evaluation.

2. RELATED WORK

Dong et al.[4] propose a single-image SR technique, namely super-resolution CNN (SRCNN), which is a CNN-based SR architecture that learns direct mappings between LR and HR images. Older methods such as sparse-coding[5] rely on learning mapping functions between example image pairs. SRCNN proposes a similar method but makes the assumption that such mappings are implicitly-learned convolutional layers rather than explicitly-learned dictionaries, and inherently performs patch extraction and aggregation which are additionally done in such

Send correspondence to S.A. and E.C.: E-mail: saksoy@cs.bilkent.edu.tr, cicek@cs.bilkent.edu.tr

previous methods. In the SRCNN architecture, 3 convolutional layers that respectively perform patch extraction and representation, non-linear mapping, and reconstruction are used.

Kim et al.[6] propose a CNN-based SR architecture inspired by VGG-net,[7] which is coined very deep SR (VDSR). This architecture, which consists of 20 layers in its final form, aims to build on top of and improve SRCNN. Three main issues are addressed: context, convergence, and scale factor. To improve the reliance of SRCNN on small regions of the input image, the very deep nature of VDSR is utilized. Through these layers, the model achieves a large receptive field and can exploit the global context in contrast to SRCNN. The problem of slow convergence during training is solved by using residual learning, along with higher learning rates compared to SRCNN. Finally, to make the VDSR scale independent and omit the need for training a new model for each SR factor that the user may require, resizing through interpolation is used.

Lim et al.[8] propose an enhanced deep SR network (EDSR) that improves the performance of similar methods by removing unnecessary modules from conventional residual blocks and increasing the model size and parameter count. Architecture optimality is asserted to be the main problem of current deep neural network models, where small changes in network architectures are said to affect the SR image quality. The scale problem also exists, where most models need to be trained for various SR factors without utilizing common relationships between them. These problems are solved by introducing a learnable upscaling layer to the model, along with removing unnecessary normalization layers.

Song et al.[3] propose an SR technique built on top of the SRCNN and VDSR architectures that is specific to PET images. The essential contribution is that HR MR images are incorporated into the learning process in order to increase the SR image quality. The architecture of the model takes in some or all of the following as inputs: LR PET, HR MR, radial location, and axial location. In addition to the LR PET, HR MR is used due to its high structural quality. Moreover, the aforementioned spatial locations of the voxels inside each patch are also included for improving the training by incorporating information regarding the cylindrical symmetry of PET scanners.

3. METHODOLOGY

3.1 Model architectures

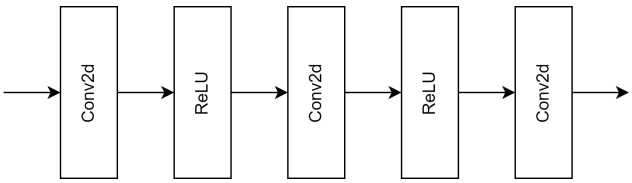
We implement two CNN-based architectures: a modification of SRCNN and a deep architecture, referred to as DeepSR from now on, combining the benefits of VDSR and EDSR while also complying with the computational constraints at hand. Both architectures take an LR image as input and try to learn the mapping from this LR image to a target HR image in a supervised setting. Before passing the LR image into either of these architectures, bicubic interpolation is applied in order to obtain an upsampled LR image at the same pixel resolution as the target. This allows the calculation of per-pixel losses while also making the model independent to scale differences in the input.

3.1.1 Multi-scale SRCNN

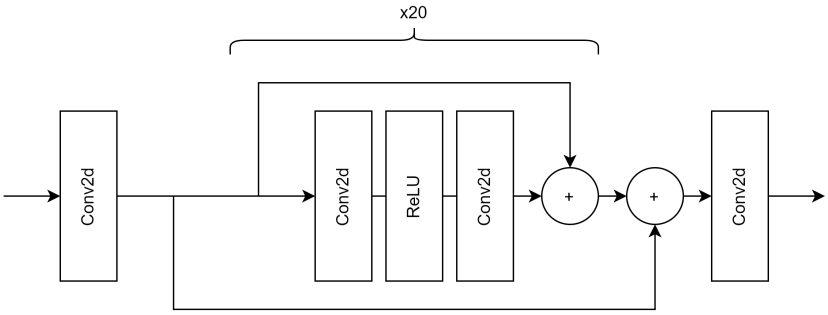
The multi-scale SRCNN architecture consists of 3 convolutional layers with rectified linear unit (ReLU) activation. The first convolutional layer has 64 filters with size 9×9 and a padding size of 4. The second convolutional layer has 32 filters of size $64 \times 1 \times 1$ with no padding. The third convolutional layer has one filter with size $32 \times 5 \times 5$ and a padding size of 2. The kernel size and padding values are carefully selected to make sure that the input dimensions are preserved at the output. The main difference of this architecture from the original SRCNN is that it works with all super-resolution scales thanks to the application of bicubic interpolation on the input image. The overall architecture is illustrated in Figure 1(a).

3.1.2 DeepSR

The proposed DeepSR architecture derives from both VDSR and EDSR, combining their useful features. Also, a reduced parameter count is used due to computational constraints, along with making the network shallower. This, however, does not significantly take away from the overall model performance. The input convolutional layer has 64 filters of size 3×3. Then, 20 convolutional blocks with two convolutions and a ReLU in-between follow. These convolutions have 64 filters of size $64 \times 3 \times 3$ each. Finally, there is an output convolution that



(a) SRCNN



(b) DeepSR

Figure 1. Illustration of multi-scale SRCNN and DeepSR architectures.

has one filter of size $64 \times 3 \times 3$. All convolutions have a padding value of 1, in order to ensure input dimension preservation. The architecture is visualized in Figure 1(b).

An important detail is the use of residual connections both from the input to the output and around each of the 20 convolutional blocks. The former type of residual connection allows prediction of only the details in the difference between the input and target images and leads to more effective training because the input and output images are often very similar.[6] The latter type of residual connections enables an efficient model by simplification of the layers in more complex residual networks that are typically used for solving higher-level computer vision problems.[8]

3.2 Dataset and preprocessing

The dataset used in this study includes PET scans from 33 patients. The images are acquired with the General Electric Healthcare Signa PET/MR scanner at the Hautepierre Hospital — University Hospitals of Strasbourg (Hôpital de Hautepierre — Hôpitaux Universitaires de Strasbourg). Each scan consists of 89 axial slices, capturing detailed sections from the skull down to the lower jaw. Each slice is a single-channel grayscale image with a bit depth of 16. The pixel resolution of each slice is $4mm \times 4mm$ through a combination of OSEM, TOF, and PSF reconstruction for 13 iterations and 8 subsets.

The dataset is split into training (20 patients), validation (4 patients), and test (9 patients) folds. For each patient, a total of 40 slices either were identified as noisy or did not correspond to parts of the brain so only 49 out of the 89 slices are used to maintain focus on the brain in the validation and test folds. However, during the training phase, all slices are used to include variability to prevent the model from simply memorizing the brain images that often bear a high resemblance to one another. This decision aims to encourage the model to face a controlled level of noise and genuinely learn and understand the image patterns, enhancing its ability to handle a diverse range of brain images. Overall, the training set includes 1780 slices, the validation set includes 196 slices, and the test set includes 441 slices. Each slice is processed as an independent 2D image. The slices obtained from the same patient remain in the same fold.

During both training and evaluation, the original PET images are used as the target HR images. The input LR images are simulated by downscaling the HR images with different factors such as $8\times$, $12\times$, and $16\times$. For $8\times$ downscaling, this corresponds to reducing the resolution to 48×48 pixels from the original 384×384 pixels. Since the deep networks described in the previous section require input images with the same size as the target, the input images are upscaled back to the original pixel resolution using bicubic interpolation. This two-step

Table 1. Quantitative quality metrics (average) computed from the input LR images as well as the output SR images of the SRCNN and DeepSR models for 8× downscaled simulation.

Image	PSNR	SSIM	LPIPS
Input LR (8× + bicubic)	30.53	0.889	0.152
Output SR from multi-scale SRCNN	32.66	0.926	0.119
Output SR from DeepSR	33.63	0.936	0.108

Table 2. Quantitative quality metrics (average) computed from the input LR images as well as the output SR images of the SRCNN and DeepSR models for 12× downscaled simulation.

Image	PSNR	SSIM	LPIPS
Input LR (12× + bicubic)	25.97	0.796	0.271
Output SR from multi-scale SRCNN	26.71	0.822	0.246
Output SR from DeepSR	27.46	0.861	0.182

Table 3. Quantitative quality metrics (average) computed from the input LR images as well as the output SR images of the SRCNN and DeepSR models for 16× downscaled simulation.

Image	PSNR	SSIM	LPIPS
Input LR (16× + bicubic)	22.799	0.699	0.356
Output SR from multi-scale SRCNN	23.433	0.714	0.343
Output SR from DeepSR	24.118	0.774	0.266

process, downscaling followed by upscaling, results in an image that visually resembles the original but with a lower quality, as the intricate details lost during downscaling cannot be fully recovered during the upscaling phase. We also considered 2× and 4× downscaling but training at these levels were ineffective because there was no significant difference between the HR (original PET) and LR (after bicubic interpolation) images.

All PET image slices are processed independently. Both the multi-scale SRCNN and the DeepSR models are trained on a single NVIDIA Quadro P6000 GPU. During the training of both models, the mean squared error (MSE) loss that is calculated between the target HR and the output SR images is used along with the Adam optimizer. Multi-scale SRCNN is trained with a learning rate of 10^{-3} for faster convergence while DeepSR is trained with a learning rate of 10^{-4} since convergence speed is already assisted with residual connections. Both models are trained over 50 epochs for fair comparison. We also considered the L1 loss but MSE had better experimental performance.

4. EXPERIMENTS

4.1 Quantitative evaluation

The performances of the models are measured by using three quality metrics: peak signal-to-noise ratio (PSNR),[6] structural similarity index measure (SSIM),[9] and learned perceptual image patch similarity (LPIPS).[10] PSNR can take any non-negative real value, whereas SSIM and LPIPS range from 0 to 1. While higher PSNR and SSIM scores are better, a lower LPIPS score shows better performance. PSNR measures the pixel-wise ratio of maximum intensity to disturbance, SSIM measures perceived changes in the image structure, and LPIPS measures the perceptual distance through a feature extractor (in our case, VGG[7]). Therefore, PSNR focuses on lower-level features, while SSIM and LPIPS focus on higher-level ones.

The quantitative results obtained using 441 test slices from 9 patients are given in Tables 1–3. All metrics show better performance when the downscaling factor is 8× compared to 12× and 16× settings. This is expected because while recovering an image of size 48×48 pixels after 8× downscaling is already a difficult task, 32×32 pixels after 12× downscaling and 24×24 pixels after 16× downscaling result in significant loss of details. Within the same downscaling setting, the metrics show that both of the CNN-based models are effectively performing the SR task compared to the input image obtained by bicubic interpolation. Overall, it is also observed that the proposed DeepSR model obtains the best scores compared to the SRCNN network for all downscaling settings.

4.2 Qualitative evaluation

Figures 2–4 present the SR results for example test slices from three different patients. As we compare the SR results of the two models, we see that DeepSR yields more clear reconstructions compared to multi-scale SRCNN due to its deeper and more effective residual architecture. These observations are also consistent with the quantitative results where DeepSR achieves higher PSNR, higher SSIM, and lower LPIPS scores compared to SRCNN for all downscaling factors. Regarding the results for different downscaling factors, reconstructions with the 8× model is visually closer to the original PET images whereas the results for 12× and 16× are still blurry because those downscalings result in unrecoverable loss of details.

All of the previous experiments aim reconstruction, which corresponds to downscaling an input PET image and trying to recover it back through a deep network model. As an additional scenario, we perform enhancement, where the original PET image is given as input to an SR model to perform super-resolution inference to achieve a higher resolution that did not exist in the original dataset. Figure 5 shows the results where the original HR test slices are given as input to the SR networks trained in the 8× downscaling setting. It is observed that the SR output has better contrast compared to the original PET image. This figure also shows the corresponding MR slices that we plan to incorporate into the super-resolution framework in future work.

5. CONCLUSIONS

In summary, the PET SR task was performed using two CNN-based architectures. Both models gave meaningful results as validated with quantitative image quality metrics and in qualitative comparisons. Our current work for further improving the PET SR pipeline includes the incorporation of HR MR images as complementary channels of input in addition to the LR PET slices to improve the structural quality of the output SR images.

REFERENCES

[1] Lameka, K., Farwell, M. D., and Ichise, M., "Positron emission tomography," *Handbook of Clinical Neurology* **135**, 209–227 (2016).

[2] Somme, F., Bender, L., Namer, I. J., Noël, G., and Bund, C., "Usefulness of 18F-FDOPA PET for the management of primary brain tumors: a systematic review of the literature," *Cancer Imaging* **20**(1), 1–13 (2020).

[3] Song, T.-A., Chowdhury, S. R., Yang, F., and Dutta, J., "Super-resolution PET imaging using convolutional neural networks," *IEEE Transactions on Computational Imaging* **6**, 518–528 (2020).

[4] Dong, C., Loy, C. C., He, K., and Tang, X., "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(2), 295–307 (2016).

[5] Yang, J., Wright, J., Huang, T. S., and Ma, Y., "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing* **19**(11), 2861–2873 (2010).

[6] Kim, J., Lee, J. K., and Lee, K. M., "Accurate image super-resolution using very deep convolutional networks," in [*IEEE Conference on Computer Vision and Pattern Recognition*], 1646–1654 (2016).

[7] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," in [*International Conference on Learning Representations*], (2015).

[8] Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K., "Enhanced deep residual networks for single image super-resolution," in [*IEEE Conference on Computer Vision and Pattern Recognition Workshops*], 136–144 (2017).

[9] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: from error measurement to structural similarity," *IEEE Transactions on Image Processing* **13**(1) (2004).

[10] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O., "The unreasonable effectiveness of deep features as a perceptual metric," in [*IEEE Conference on Computer Vision and Pattern Recognition*], 586–595 (2018).

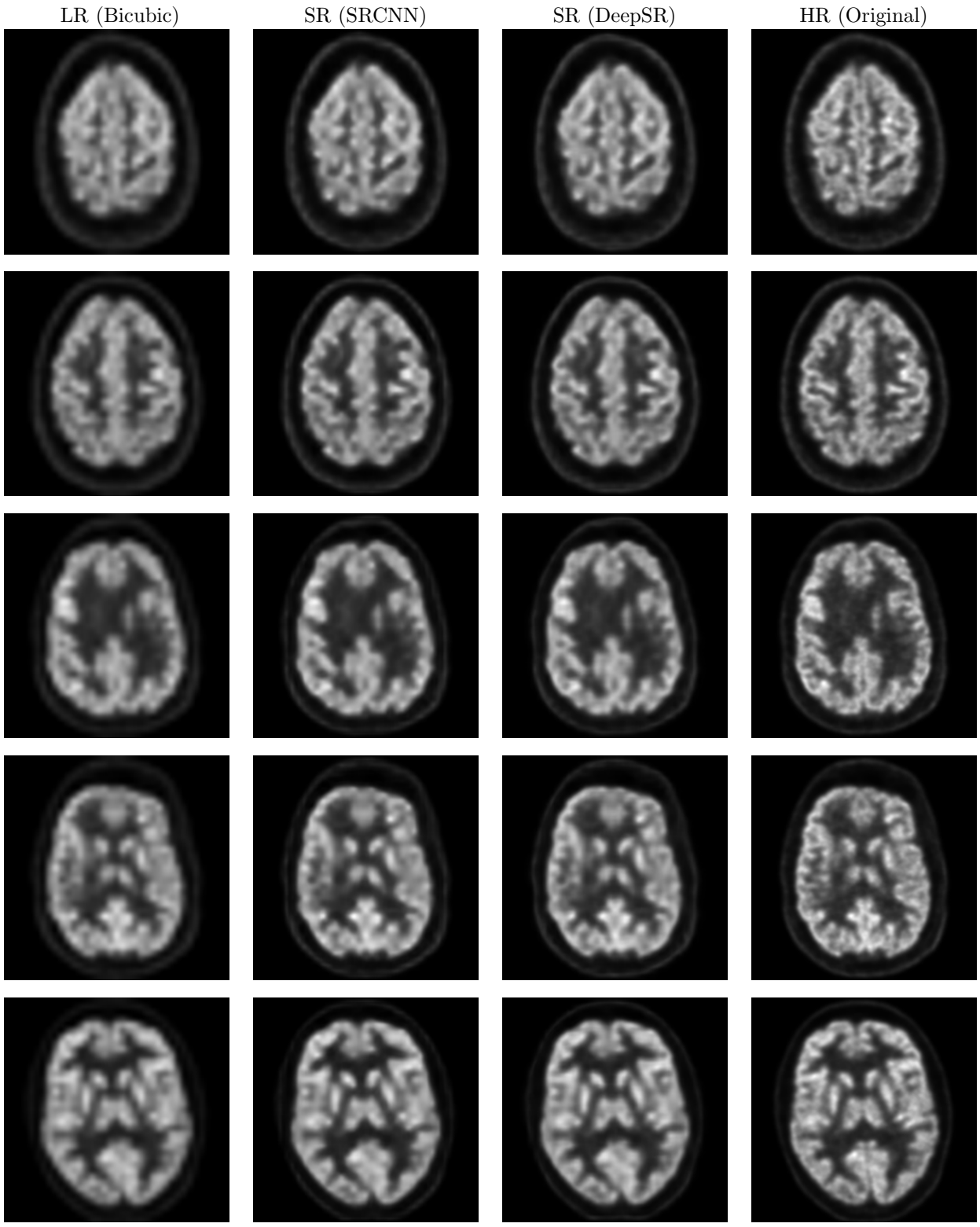


Figure 2. Illustration of the network outputs for 8× downscaled images. Each row corresponds to a different sample from the test dataset. From left to right: input LR image processed with bicubic interpolation, SR output from SRCNN, SR output from DeepSR, target HR image.

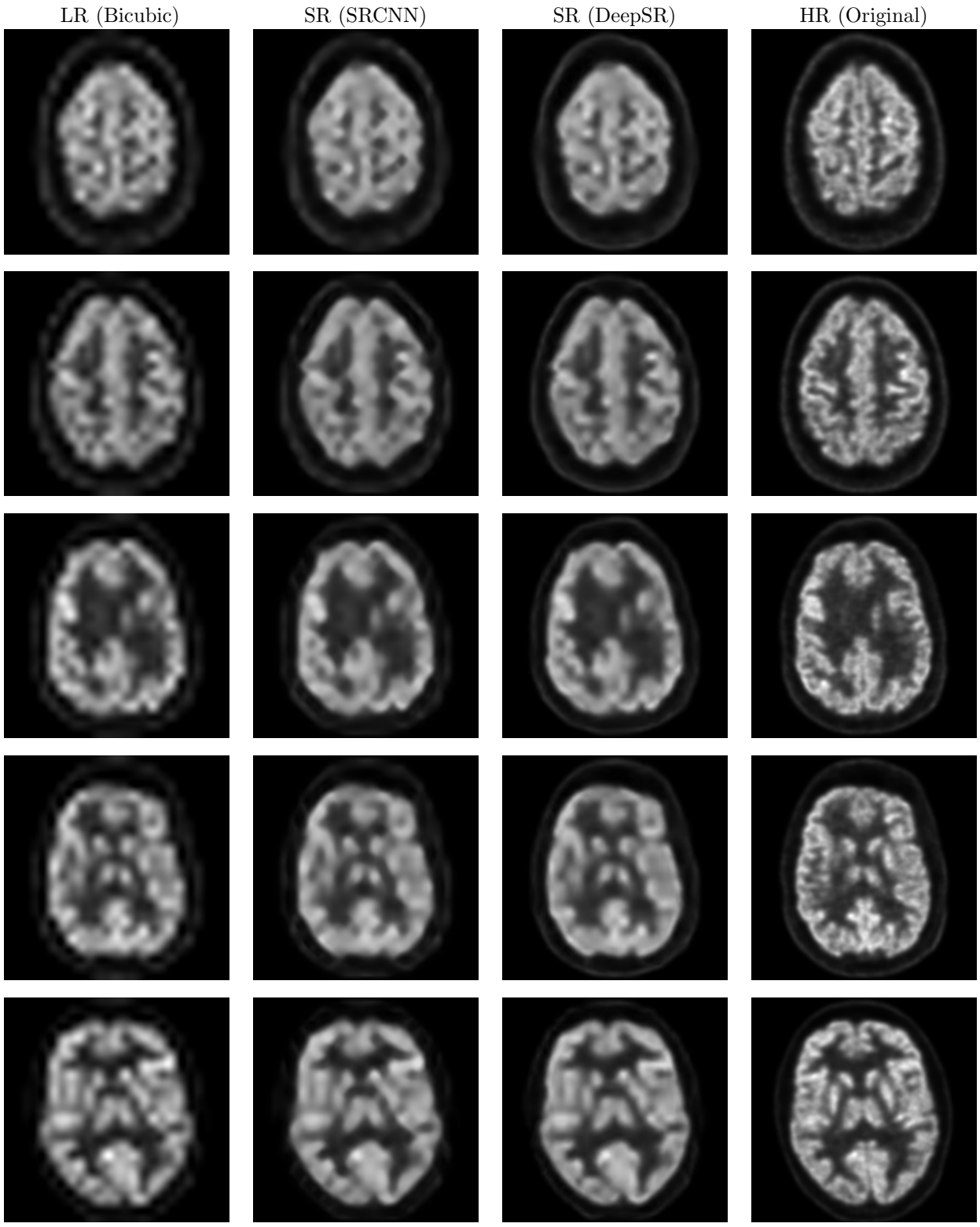


Figure 3. Illustration of the network outputs for 12× downscaled images. Each row corresponds to a different sample from the test dataset. From left to right: input LR image processed with bicubic interpolation, SR output from SRCNN, SR output from DeepSR, target HR image.

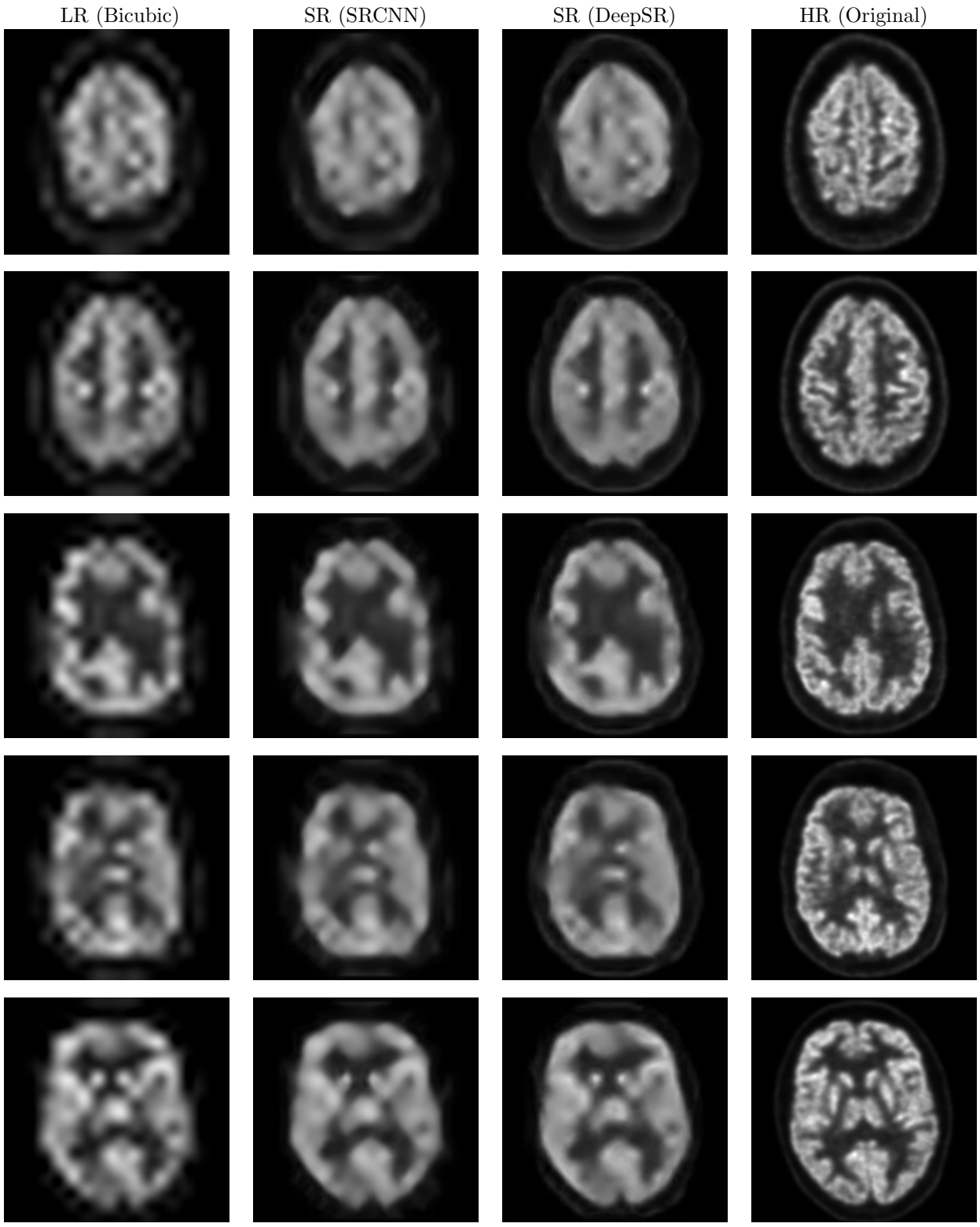


Figure 4. Illustration of the network outputs for 16× downscaled images. Each row corresponds to a different sample from the test dataset. From left to right: input LR image processed with bicubic interpolation, SR output from SRCNN, SR output from DeepSR, target HR image.

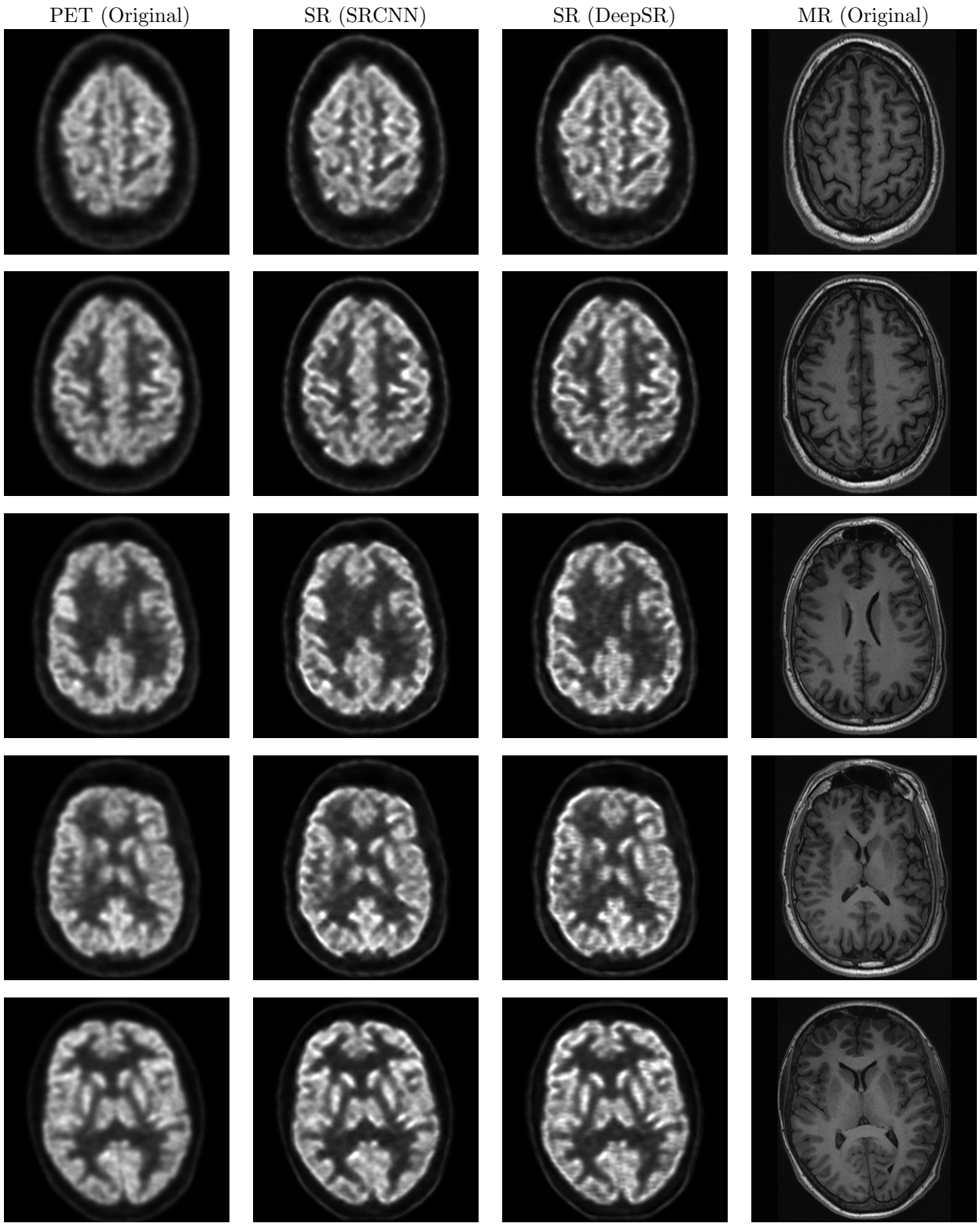


Figure 5. Illustration of the 8×-trained network outputs for the original PET images. Each row corresponds to a different sample from the test dataset. From left to right: input HR PET image, SR output from SRCNN, SR output from DeepSR, MR slice corresponding to the input PET image.