Targeted Gradient Descent: A Novel Method for CNN Fine-tuning and Online-learning

Junyu Chen[1,2,*], Evren Asma[3], Chung Chan[3]

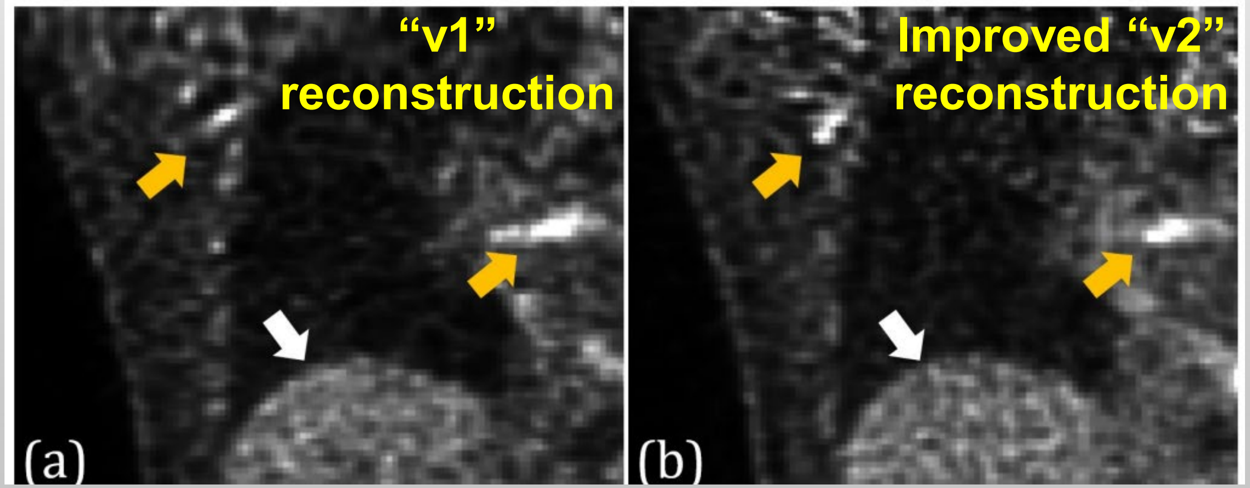
[1]Department of Radiology and Radiological Science, Johns Hopkins University, MD, USA
[2]Department of Electrical and Computer Engineering, Johns Hopkins University, MD, USA
[3]Canon Medical Research USA, INC., IL, USA

*: This work was done while the author was an intern at Canon Medical Research USA

Background

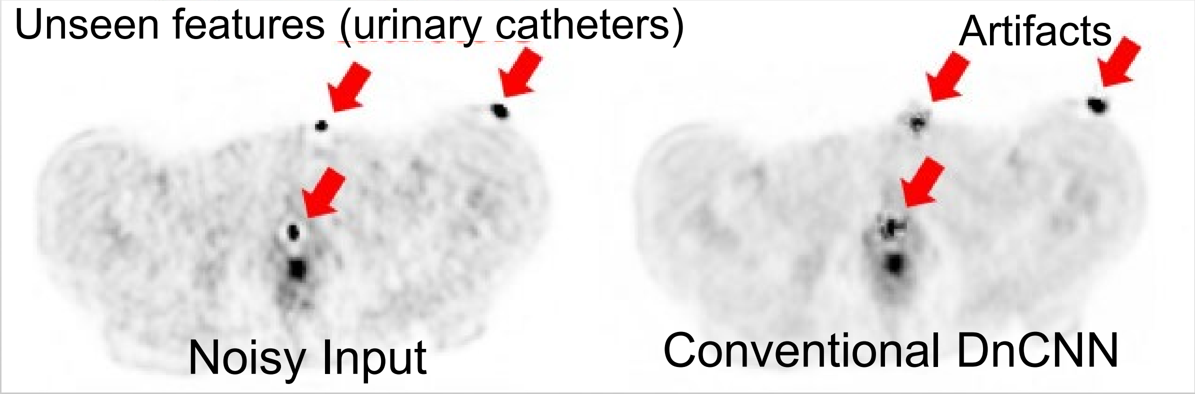
- Training a CNN typically requires a large amount of data.
- Acquiring training data is time consuming and expensive.
- **Challenge 1:**
  - When new imaging systems and or updated reconstruction algorithms are employed:
    - *Image quality and appearance will change.*
    - *Neural networks need to be retrained to adapt the changes.*
- **Challenge 2:**
  - A trained DNN often produces suboptimal predictions on unseen features.

Challenge 1:



F-18 FDG PET images of the right lung of a human subject

Challenge 2:



F-18 FDG PET images of the pelvis region of a human subject

Significance & Novelty

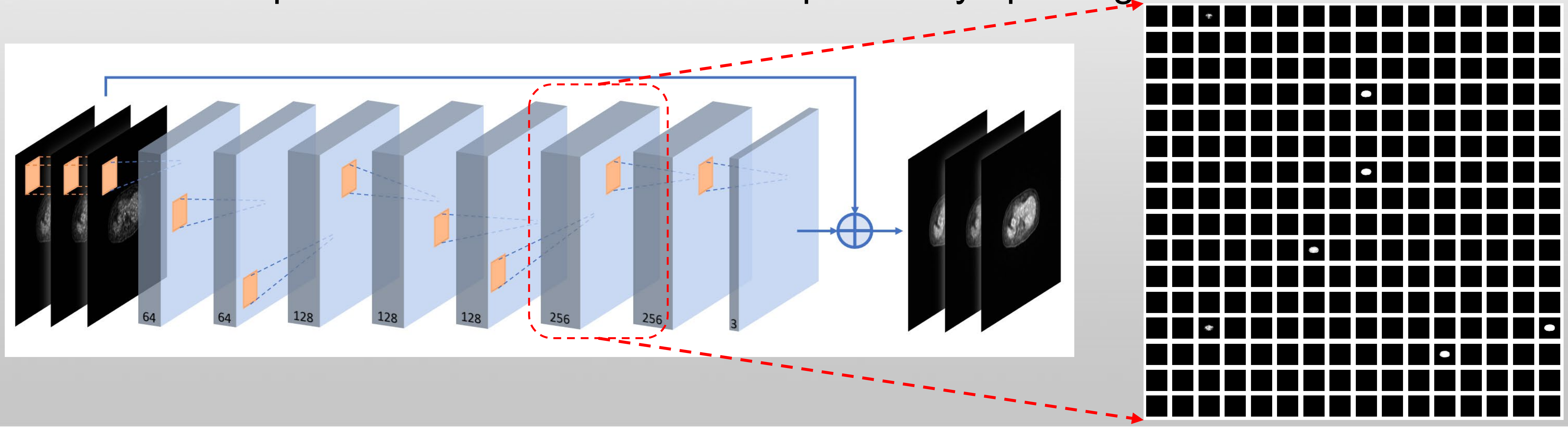
- Existing methods that can be used to avoid retraining a pretrained DNN from scratch:
  - Fine-tuning [1, 2]: Suffer from catastrophic forgetting.
  - Joint training [3, 4]: Require revisiting dataset from the previous tasks.
  - Continual learning [5, 6]: Require previous dataset, or hard to tune the hyper-parameters.
- In this work, we proposed a network fine-tuning and online-learning scheme that:
  - Adapts a pretrained DCNN to new imaging protocols with the minimum need for additional training data.
  - Adapts a pretrained DCNN to individual testing image to avoid producing artifacts on unseen features.
- We applied the proposed method on the task of F-18 FDG PET image denoising.

[1]: Amiri et al. 2019
[2]: Gong et al. 2018
[3]: Caruana 1997
[4]: Wu et al. 2018
[5]: Baweja et al. 2018
[6]: Kirkpatrick et al. 2017

Rationale: Retraining "useless" kernels

- "Useless/redundant" feature maps exists in a pretrained ConvNet:
  - ConvNet did not effectively use all of its kernels
  - Some of the kernels generate useless/redundant feature maps

- Specifically retrain these "useless" kernels:
  - Network retains the knowledge learned from the previous tasks because the kernels that produce "meaningful" feature maps were kept.
  - Network's performance on new task is improved by updating the "useless" kernels.



Method: Identifying "useless" feature maps

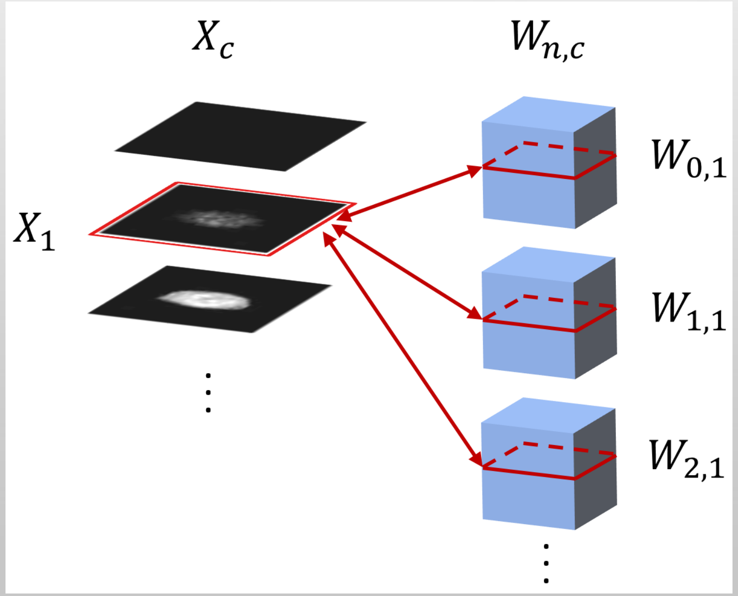
- Kernel sparsity and entropy (KSE) [7]:
  - Operates on the convolutional kernels, but it quantifies the information richness of the input feature map.
  - Kernel sparsity: l1-norm of the kernels.
    - $s_c = \sum_{n=1}^{N} |W_{n,c}|$
  - Kernel entropy: a measure of the diversity of the kernels.
    - $e_c = -\sum_{i=0}^{N-1} \frac{dm(W_{i,c})}{\sum_{i=0}^{N-1} dm(W_{i,c})} \log_2 \frac{dm(W_{i,c})}{\sum_{i=0}^{N-1} dm(W_{i,c})}$
  - KSE score:
    - $KSE = \sqrt{\frac{s_c}{1+\alpha \cdot e_c}}$
    - α a weighting parameter.
    - KSE is normalized to $[0, 1]$ in each layer.

- "Useless" feature maps are identified as those with KSE $<\varphi$.
  - φ is a predefined KSE threshold.

A convolutional operation is formulated as:

$$Y_n = \sum_{c=0}^{C-1} W_{n,c} * X_c$$

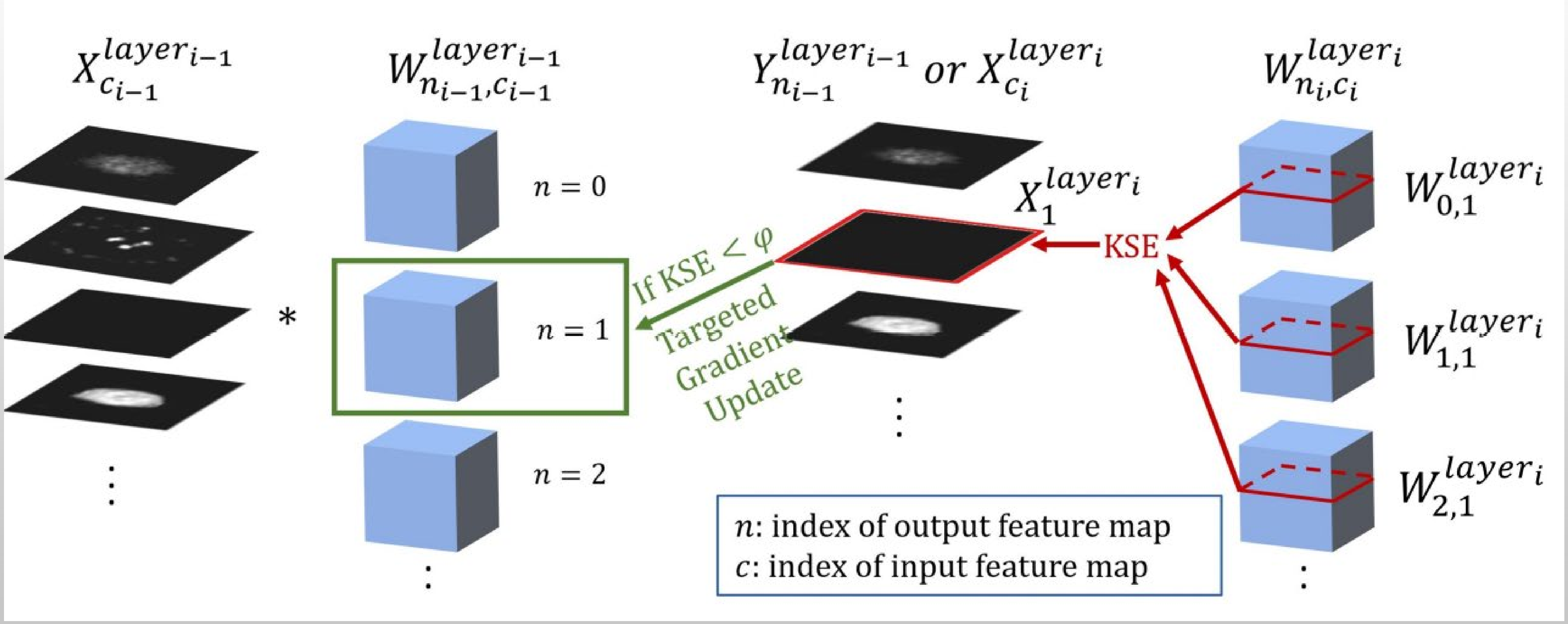
where Y is the output feature maps, X is the input feature maps, and $*$ represents convolution.



[7]: Li et al. 2019

Method: Targeted Gradient Descent

- The goal is to specifically retrain the kernels that generates these feature maps.



Method: Targeted Gradient Descent

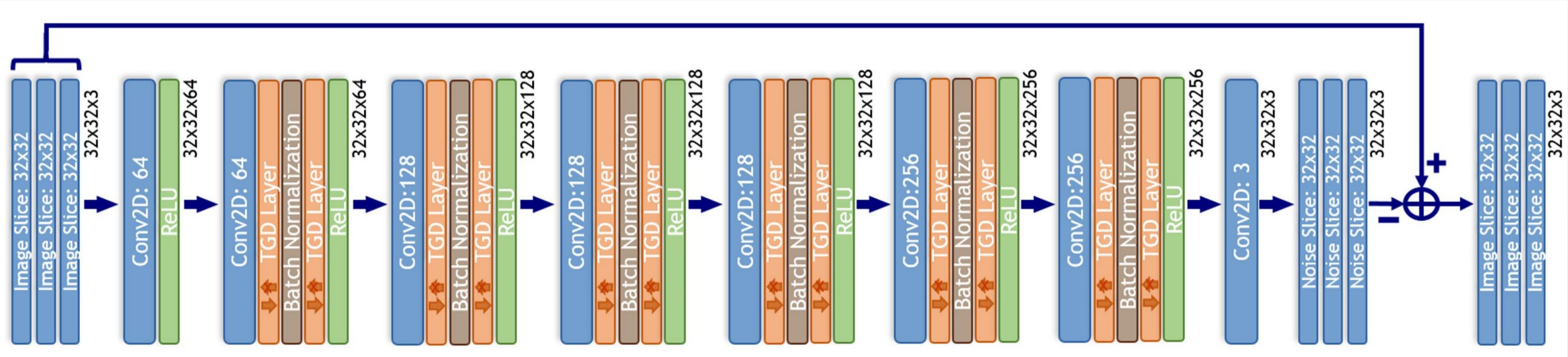
- Identify the indices of the convolution kernels that generate the “useless” feature maps. The indices were used for generating a binary mask in the gradient space:
  - $M_n = \begin{cases} 1, & if\ KSE(Y_n) < \varphi \\ 0, & if\ KSE(Y_n) \geq \varphi \end{cases}$
  - M_n zeros out the gradients for the “useful” kernels (i.e., ones with $KSE(Y_n) \geq \varphi$).

- Mathematically, the back-propagation formula with TGD is defined as:
  - $W_{n,c}^{(t+1)} = W_{n,c}^{(t)} - \eta \frac{\partial \mathcal{L}}{\partial Y_n^{(t)}} M_n X_c^{(t)} - \frac{\partial \mathcal{R}(W_{n,c}^{(t)})}{\partial Y_n^{(t)}} M_n X_c^{(t)}$

- This masking process is implemented as a novel TGD layer that only activates during backpropagation and not forward pass.

Method: Fine-tuning ConvNet with TGD

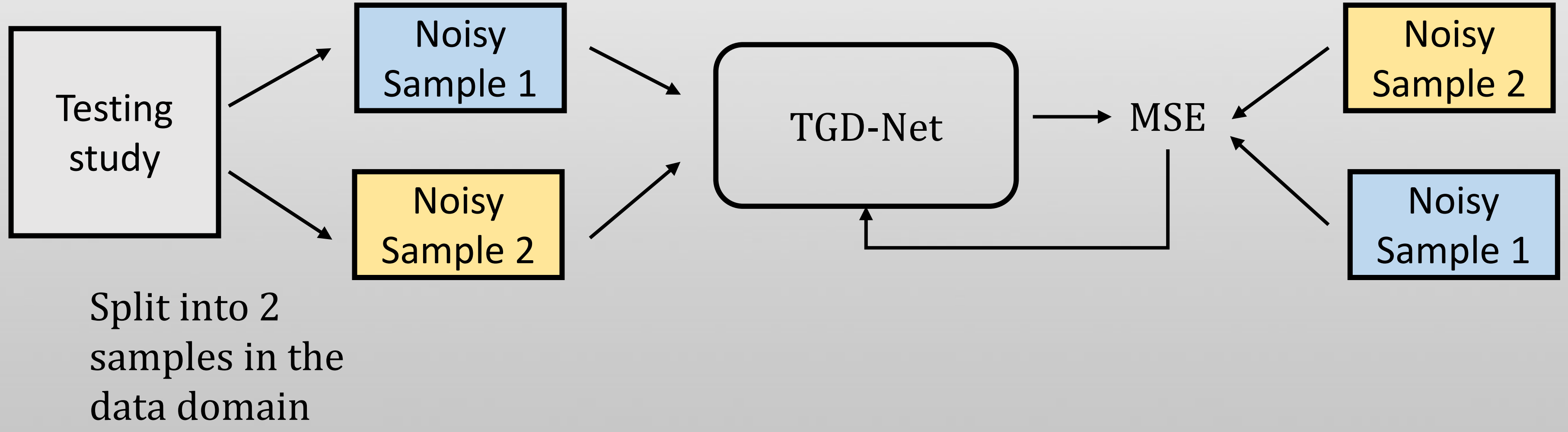
- A 2.5-dimensional DnCNN [8] that predicts noise in a given image.
  - This is based on the previous architecture for PET denoising used in [9].
- TGD layer is inserted after each convolution layer and batch normalization layer.
  - Convolution and batch norm layers contain trainable weights.
- Training with TGD layers fine-tunes the network for new tasks while maintaining the knowledge learned from prior tasks.



[8]: Zhang et al. 2017
[9]: Chan et al. IEEE MIC, 2018

Method: Online learning using TGD with N2N

- We then proposed to use TGD-network with Noise2Noise training (N2N) [9] for online learning, which helps the neural network adapt to unseen features during testing without the need to collect new training datasets to re-train the network:
    - Split the testing study into 2 i.i.d. noisy samples with nearly equivalent number of counts.
    - Using noise sample 1 as the input, noise sample 2 as the label, and vice versa.



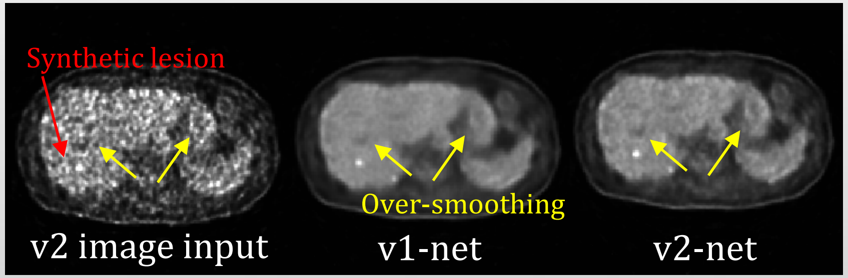
[9]: Lehtinen et al. 2018

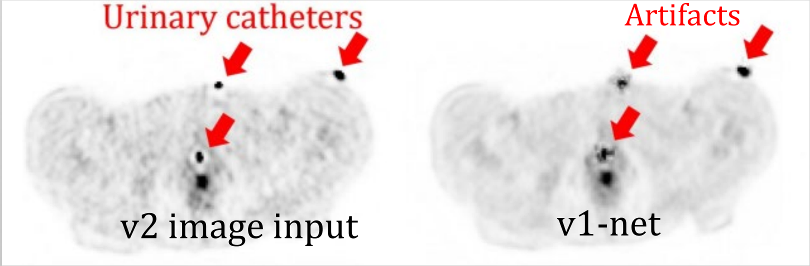
Experiment: Networks and training parameters

- Network architecture: A 8-layer DnCNN for FDG-PET image denoising
- Baseline networks:

Network Models	Number of training datasets	Noise levels generated for each training dataset	Reconstruction method
v1-net	20	5	V1 (old)
v2-net	20	5	V2 (new)

- **Fine-tuning task:**
  - FT-net: Fine-tuning the last three convolutional blocks of v1-net using 7×5 v2 images
  - TGD-net: v1-net fine-tuned using the TGD layers with 7×5 (noise levels) v2 images
- **Online-learning task:**
  - $\mathrm{TGD_{N2N}}$-net: TGD N2N applied on the v2-net
  - $\mathrm{TGD_{N2N}^2}$ -net : TGD N2N applied on the TGD-net
- All methods are trained using TensorFlow on a single NVIDIA Titan V GPU.
  - Optimizer: Adam
  - Learning rate = 0.001
  - Number of epochs = 500 for TGD, 150 for TGD N2N





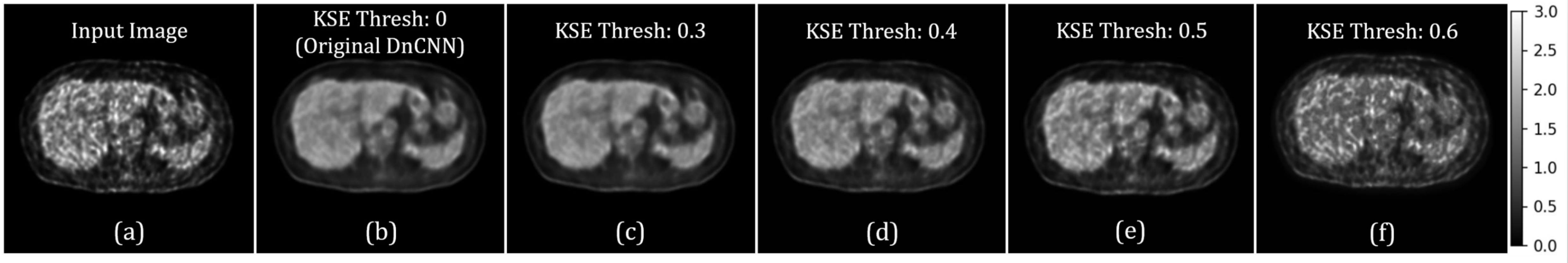
Experiment: Quantitative analysis

- **Fine-tuning task:**
  - We rebinned a 600-sec/bed F-18 FDG PET study into 10 × 60-sec/bed image i.i.d noise realizations to assess the ensemble bias on the tumor, and liver coefficient of variation (CoV) by using the 600-sec/bed image as the ground truth.
  - Ensemble bias (quantifies activity recovery in a lesion):
    - $BIAS(\%) = \frac{\frac{1}{R}\sum_r \mu_r - T}{T}$
  - Ensemble CoV (quantifies noise in a background VOI (e.g., liver)):
    - $CoV(\%) = \frac{\frac{1}{N}\sum_{i \in B} \sigma_i^R}{\bar{\mu}_B}$
- **Online-learning:**
  - Reduction of hallucination by visual assessment
  - Liver CoV (%) of the same patient

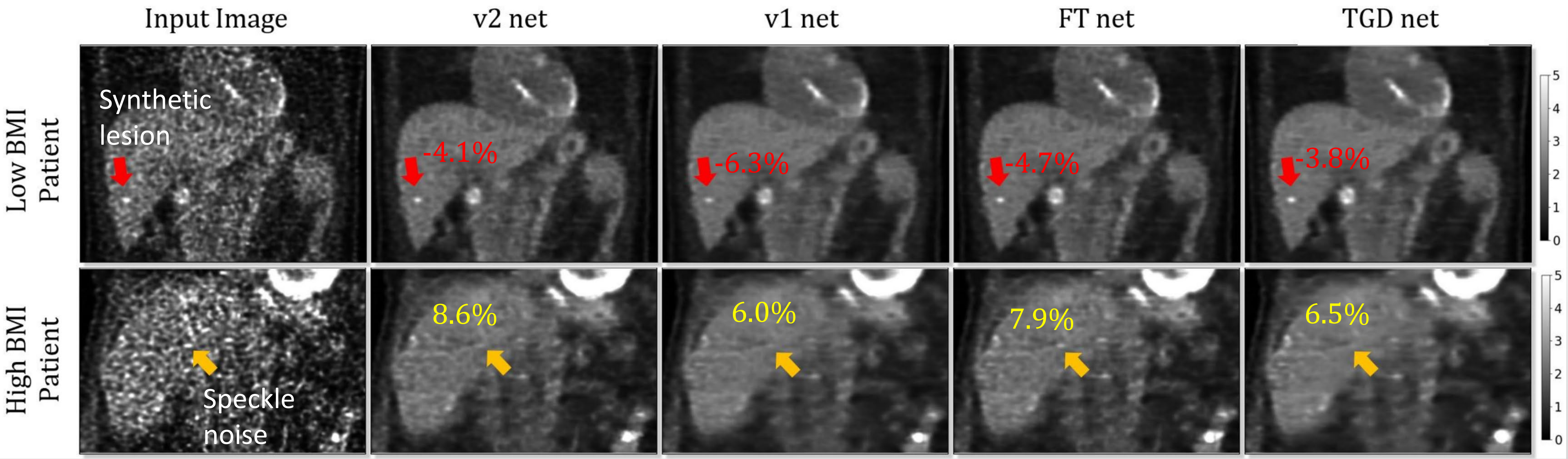
Experiment: Determine KSE threshold φ

- During the prediction, kernels in the pretrained v1 network that are recognized as "meaningless" by the KSE threshold are discarded (i.e., the weights are set to 0).

- KSE threshold values of 0.3 and 0.4 resembles the original denoising performance the best.

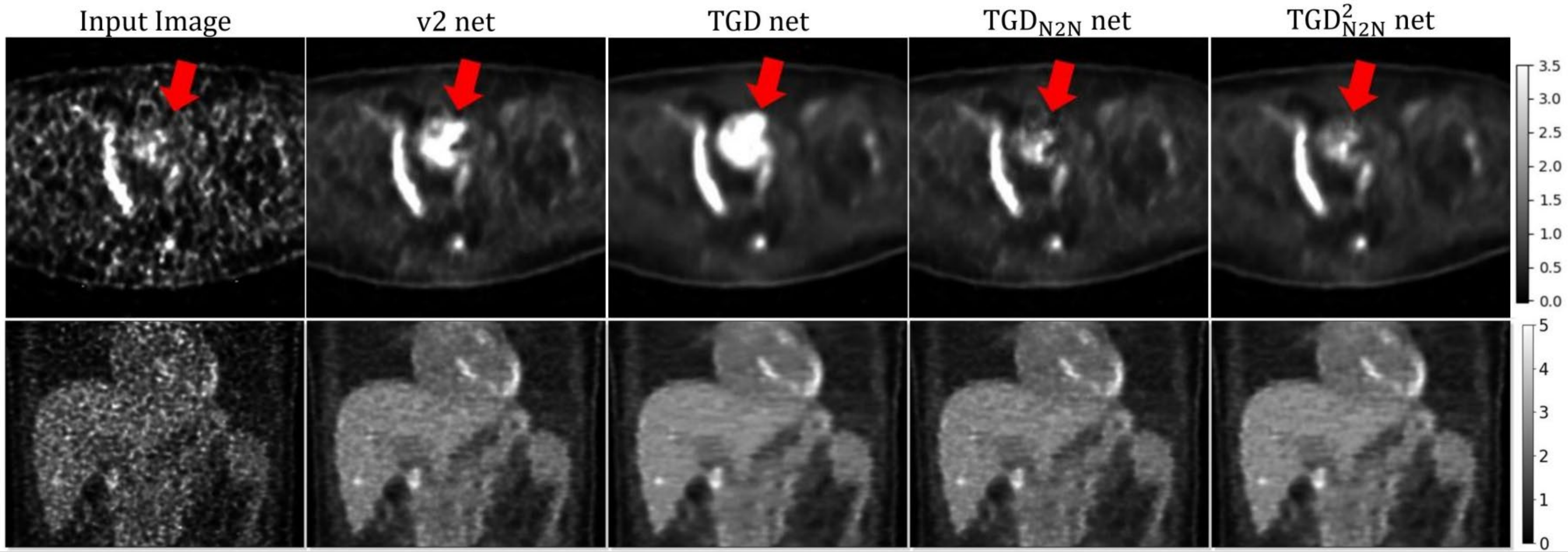


Results: TGD fine tuning results & quantifications



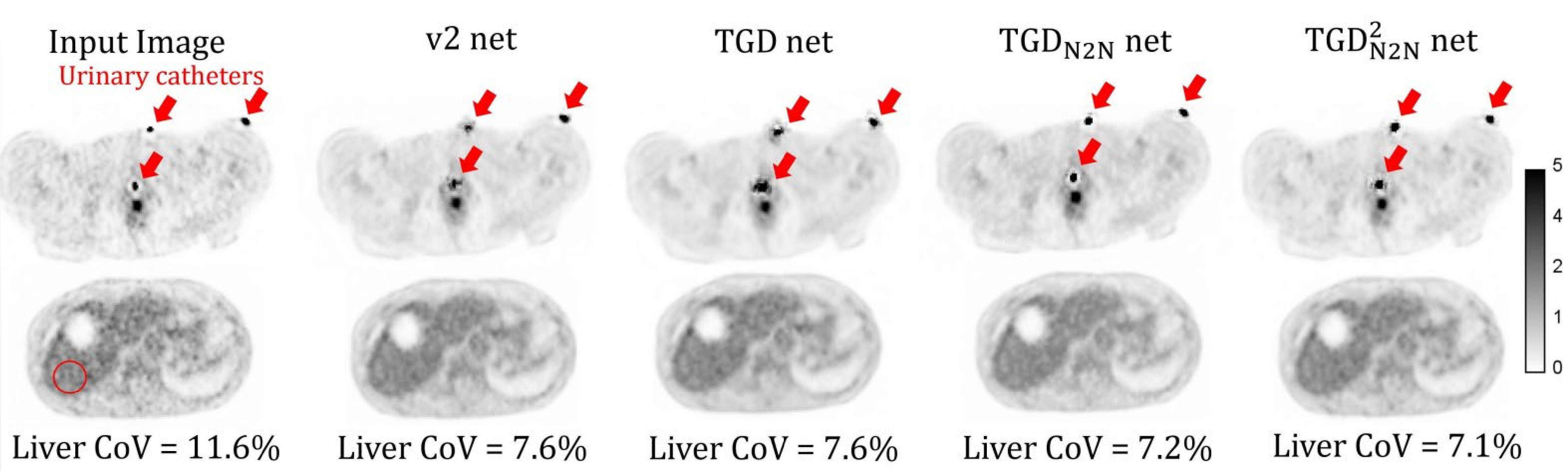
- Qualitative comparisons between the proposed TGD method and other methods on denoising two FDG patient studies
- The red numbers indicate the ensemble bias (%) comparing to the ground truth
- The yellow numbers denote the liver CoV (%)

Results: TGD N2N online learning - case 1



The red arrows indicate the artifactual feature generated by the v2-net and TGD-net around the bladder, which was not included in any training datasets. Both $\mathrm{TGD_{N2N}}$ and $\mathrm{TGD^2_{N2N}}$ yielded images which are in high fidelity to the input image on the bladder, while retaining similar denoising performance as v2-net and TGD-net.

Results: TGD N2N online learning - case 2



The red arrows indicate the urinary catheters, which was not included in any training datasets. The online learning approaches using $\mathrm{TGD_{N2N}}$ and $\mathrm{TGD^2_{N2N}}$ alleviated the artifacts while retaining similar denoising performance in terms of liver CoV in the ROI denoted by the red circle.

Conclusion:

- This work introduces Target Gradient Descent, a novel fine-tuning scheme that can effectively retrain the redundant kernels in a pre-trained network

- The proposed TGD framework can be easily incorporated into an existing network and does not require revisiting the data from previous task

- We demonstrated the effectiveness of TGD for PET image denoising

- The preliminary results show:
  - TGD enables adapting a pre-trained network to new tasks
  - TGD may allow online learning on the testing study in order to improve the network's generalization capacity in real-world applications

Reference

1. Amiri, M., Brooks, R., Rivaz, H.: Fine tuning u-net for ultrasound image segmentation: which layers? In: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data, pp. 235–242. Springer (2019)
2. Gong, K., Guan, J., Liu, C.C., Qi, J.: Pet image denoising using a deep neural network through fine tuning. IEEE Transactions on Radiation and Plasma Medical Sciences 3(2), 153–161 (2018)
3. Caruana, R.: Multitask learning. Machine learning 28(1), 41–75 (1997)
4. Wu, C., Herranz, L., Liu, X., van de Weijer, J., Raducanu, B., et al.: Memory replay gans: Learning to generate new categories without forgetting. In: Advances in Neural Information Processing Systems. pp. 5962–5972 (2018)
5. Baweja, C., Glocker, B., Kamnitsas, K.: Towards continual learning in medical imaging. arXiv preprint arXiv:1811.02496 (2018)
6. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences 114(13), 3521–3526 (2017)
7. Li, Y., Lin, S., Zhang, B., Liu, J., Doermann, D., Wu, Y., Huang, F., Ji, R.: Exploiting kernel sparsity and entropy for interpretable cnn compression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2800–28
8. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Transactions on Image Processing 26(7), 3142–3155 (2017)
9. Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., Aila, T.: Noise2noise: Learning image restoration without clean data. In: ICML (2018)
10. Chan, Chung, et al. "Noise Adaptive Deep Convolutional Neural Network for Whole-Body PET Denoising." 2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC). IEEE, 2018.