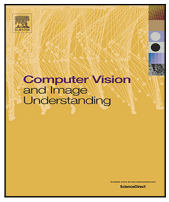
Computer Vision and Image Understanding 223 (2022) 103519

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



GAFL: Global adaptive filtering layer for computer vision

Viktor Shipitsin[1], Iaroslav Bespalov[1], Dmitry V. Dylov [*]

Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30/1, Moscow, 121205, Russian Federation

ARTICLE INFO

Communicated by Nikos Paragios

MSC:
68T01
68T05
68T10
68U10

Keywords:
Adaptive neural layer
Efficient training
Fourier filtering

ABSTRACT

We devise a universal global adaptive filtering layer, GAFL, capable of "learning" optimal frequency filter for each image in a dataset together with the weights of the base neural network that performs some computer vision task. The proposed approach takes the source image in the spatial domain, selects the best frequencies in the Fourier domain for the benefit of the global task, and prepends the inverse-transform image to the main neural network for a joint training. Remarkably, such a simple add-on layer, capable of optimizing the frequency content of an input for a specific task, dramatically improves the performance of the main network regardless of its design. We observe that the light networks gain a noticeable boost in the performance metrics; whereas, the training of the heavy ones converges faster when GAFL is prepended to the main architecture. We showcase the performance of the layer in four classical computer vision tasks: classification, segmentation, denoising, and erasing, considering popular natural and medical data benchmarks.

1. Introduction

In recent years, computer vision (CV) algorithms have advanced significantly thanks to the advent of artificial neural networks (ANN) and to the development of the computational resources capable of working with them (Szeliski, 2011). At the same time, the constantly growing volumes of data instigated a wave of research efforts involving large neural networks with a colossal number of parameters (Shazeer et al., 2017), triggering the development of approaches for efficient data processing, model optimization, and training. One promising trend is not to keep complicating the architectures, but to develop *efficient modules* that allow one to look at computer tasks from a different angle, extracting semantics from the images and ultimately demanding less effort (Guo et al., 2016; Rhu et al., 2016). What could be done with the images even *before* they enter a certain neural network is generally concerned with the task of image preprocessing (Bow, 2002) and will be the leitmotif in this work. Particularly, we are interested in developing a 'smart' preprocessing module to the following four classical CV problems: segmentation, classification, erasing, and denoising.

The *segmentation* task is one of the most popular tasks in the field of CV, as it allows to localize the object of interest in the image. When image segmentation is concerned, one naturally starts with the U-Net encoder–decoder like models (Ronneberger et al., 2015). At the moment, there are various modifications that prove more accurate than the baseline U-Net in various scenarios: Attention U-Net (Oktay et al., 2018), U-Net++ (Zhou et al., 2018; Zhou et al., 2020), U-Net 3+ (Huang et al., 2020), *etc*. Although much heavier and slower in training, ResNet and DenseNet models are also frequently employed for the purpose of segmentation (Huang et al., 2016; He et al., 2015). One naturally looks for lighter models that would reach the performance level of the heavier models with dozens of millions of parameters (He et al., 2015; Szeliski, 2011).

Image *classification* is frequently defined as the task of categorizing images into one of several predefined classes, and it is another popular problem in CV (Rawat and Wang, 2017; Szeliski, 2011). Binary classification is a precursor problem to many other CV challenges, and is an analogy to the segmentation, with the output being a single pixel. The same segmentation encoders can be employed for the classification problem to obtain an *embedding*, and then, linear layers would predict the class (Rawat and Wang, 2017).

Denoising is another important task in imaging which covers extensive range of domains and applications (McCann et al., 2017). Popular approaches, such as DnCNN (Zhang et al., 2017; Zuo et al., 2018), already became classic and can restore blurred, damaged, and noisy images exceptionally well. Naturally, denoising is also the problem where frequency decomposition of an image becomes a particularly important entity for a computer scientist (Zhang et al., 2011; McCann et al., 2017). It is interesting how the frequency spectra change in the denoising tasks. For example, one can cut out an area from an image and see how a denoising model would paint over the area in the presence of *frequency filtering*. Such erasing (Szeliski, 2011) task will be also briefly considered in this work and is adjacent to another popular problem of *super-resolution* in CV, where the missing pixels are

* Corresponding author.
E-mail address: d.dylov@skoltech.ru (D.V. Dylov).
[1] V.S. and I.B. contributed equally.

https://doi.org/10.1016/j.cviu.2022.103519
Received 4 August 2021; Received in revised form 20 April 2022; Accepted 27 July 2022
Available online 1 August 2022

processed to minimize the damage to the image or to maximize a value function such as the resolution.

The problem of frequency filtering for denoising has been studied very thoroughly in the signal processing and in the imaging physics communities (Chowdhury et al., 2017). In fact, the filtering is at the core of one of the most frequently used clinical imaging modalities — the ultrasound (Ihnatsenka and Boezaart, 2010). Its typical high/mid-range (5–15 MHz) and low (2–5 MHz) frequency probes provide either good resolution or good penetration, but not both at once. The resulting images, therefore, are extremely sensitive to the frequency tuning, with various phenomena such as reverberation, shadowing, excessive absorbance, reflection, and echo, giving the images the distinct grainy look (Song et al., 2019; Wang et al., 2019).

What operators of the medical ultrasound do with their knobs on the machine's panel to enhance the appearance of the images in real time has motivated us to mimic the similar 'live' filtering for the CV problems. Specifically, we asked ourselves, what if a pre-training block of a neural network would be capable of learning the optimal frequencies for each image in the dataset *live* during the training, effectively maximizing a value function of interest for the entire model? Can we design a universal adaptive layer that would provide the necessary frequency filter for any input image regardless of the network architecture or the CV task at hand? Herein, we present such a solution.

Using the direct and the inverse Fourier transforms, we can switch from the representation of the image in the spatial domain to that in the frequency domain and vice versa. In the frequency spectrum, particular frequencies are responsible for different properties of the image (Szeliski, 2011), which can be either enhanced or suppressed with filtering, depending on the value function of interest in one of the four CV problems described above. For example, the high-pass filter, used for the edge detection, can enhance edges and details, effectively holding promise for improving the segmentation performance if it partook in the training routine along with the main segmentation network. In this article, we devise a simple adaptive add-on layer that improves the quality and efficiency of popular neural networks in CV. The layer learns to automatically find a global filter that will leave only those frequencies that could boost the target metric in the entire dataset (for example, Dice score in the segmentation, or F_1-score in the classification problem).

The rest of the paper is structured as follows. After covering the work related to learning in the Fourier domain in Section 2, we describe the algorithm behind the global adaptive layer in Section 3. We then describe the datasets in Section 4.1: two medical (ultrasound, which has motivated our work) and four natural image data benchmarks, hypothesizing that the wave nature of the ultrasonic data would correspond to a more efficient filtering than that of the natural images. However, the rest of Section 4 reports a likewise enhancement of the baseline performance for the natural images as well. Section 4 also reports faster training convergence for the majority of models and tasks and summarizes the results of a controlled and a large-scale studies. Sections 5 and 6 conclude the paper.

Contributions of this paper are the following:

- The first *adaptive layer* to be trained alongside the main neural network to boost its performance by finding *globally optimal* filtering frequencies.
- A simple, universal, flexible, and intuitive solution for improving and accelerating neural networks.
- We show at least 6 % increase in Dice score for light U-Net-like architectures, and accelerate convergence of heavier models (such as ResNet and DenseNet).
- We report 88 experimental scenarios, 5 variations of the adaptive layer, adding them on to 5 popular architectures, and testing the outcomes on 6 (4 natural and 2 medical) dataset benchmarks in 4 CV tasks.
- Careful control and large-scale studies are reported.

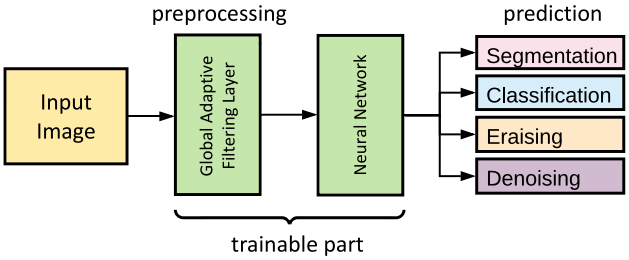


Fig. 1. Diagram of the proposed method. Global Adaptive Filtering Layer is trained together with the weights of the main neural network until the prediction is maximized in a given CV task.

2. Related work

Adoption of the learning algorithms from the *non-image* domains to improve either the target metric or the efficiency of neural networks is a rather recent trait. The Fourier space is one of such domains, where there are several works reporting the spectral transforms with consequent feature extraction to train their models (Pratt et al., 2017; Liu et al., 2018; Lin et al., 2019). Fourier analysis has also been successfully used for dynamic structure segmentation problems, where dynamic structures were distinguished using only the phase spectra (Li et al., 2009). We obviously omit a long list of works here, where the frequency data was used for feature engineering or for some domain-specific machine learning applications.

In 2020, however, there appeared a relevant work reporting semantic segmentation with domain adaptation (Yang and Soatto, 2020), where the spectral amplitudes of the source and the target images were combined to boost the performance of the model. High-frequency low-dimensional regression problems (where Fourier features improved the results of the coordinate-based multi-layer perceptions for image regression), 3D shape regression, MRI reconstruction, and inverse rendering tasks are also some very recent results (Tancik et al., 2020).

These aforementioned works have shown their effectiveness by proving that some information could be lost if one relies merely on the spatial image domain. The difficulty, however, is that the manual selection of the correct frequencies for optimizing an ANN is not a simple task. Remarkably, none of the algorithms makes effort to optimize the frequency spectra during the training routine of the main architecture. *Therefore, our solution is to automate the search for the optimal weights in the frequency spectrum until the desired metric of a given network is maximized for each CV task.* Despite being rather intuitive, such a solution has not been reported in the literature, motivating our study herein.

3. Proposed method

Data preprocessing is an essential part of any CV algorithm (Szeliski, 2011), primarily done in the image space (Buslaev et al., 2020). Proposing the same in the Fourier space, we want to dismiss the meaningless features associated with spectral frequencies brought to the scene by the image acquisition systems (an ultrasound machine or a photo camera, in our case). As such, the method we look for belongs to the class of *minimum features inductive bias* algorithms (Gordon and Desjardins, 1995). The desired 'smart' spectral preprocessing method should automatically distill the meaningful frequencies, being aware of the entire model and 'adapting' to the entire dataset.

We propose the concept of such a globally adaptive neural layer (see Fig. 1), which could be trained together with a model of interest to solve a given CV problem. By placing this layer in front of the baseline model, the algorithm should automatically select the weights for the frequency components of all images sent in as the input to carry out filtering with **one purpose only**: *improve whatever the main architecture attempts to do*.

Algorithm 1: Global Adaptive Filtering Layer

Input: I – Initial image.
$\mathcal{F}$ – Fast Fourier Transform operator.

1: $W_1, W_2, B_1, B_2 = ReLU(W_1, W_2, B_1, B_2)$;
2: $F = \mathcal{F}I$;
3: $S = W_2 * \sigma\left(W_1 * |F| + B_1\right) + B_2$;
4: $S = S * F/|F|$;
5: $I' = \mathcal{F}^{-1}S$;

Output: I' – Image after global frequency filtering.

Theoretically Blackledge (2005), Klette (2014) and Brunton and Kutz (2019), while performing the Fourier transform, it is possible to move from the frequency domain to the spatial one (and vice versa) without a loss of information. The Fourier transform has the property of linearity, preserving the accuracy of signal approximation (Parseval's theorem Blackledge, 2005), which is valid when the signal is represented using discrete vectors. Define the Fourier transform as

$$\mathcal{F}I(\mathrm{u},\mathrm{v}) = \sum_{\mathrm{x}=0}^{n-1}\sum_{\mathrm{y}=0}^{m-1}\frac{I(\mathrm{x},\mathrm{y})}{nm}\exp\left\{-\frac{2\pi i\mathrm{x}\mathrm{u}}{n} - \frac{2\pi i\mathrm{y}\mathrm{v}}{m}\right\}, \tag{3.1}$$

where $I(\mathrm{x},\mathrm{y})$ is the original image (spatial description) of size $n \times m$ and pixel coordinates (x,y), and $\mathcal{F}I(\mathrm{u},\mathrm{v})$ is the frequency domain image, with (u,v) being the coordinates of the image in the frequency domain. If the image has multiple channels, we transform each channel separately by the same formula.

For the visual analysis of Fourier transform, one usually works with the spectrum,[2] *i.e.*, the coordinate-wise absolute value $|\mathcal{F}I|$, or the energy spectrum $|\mathcal{F}I|^2$. To filter image in the frequency domain, we choose to take a function that modifies spectrum $|\mathcal{F}I|$ in a specific way. There is flexibility in designing filtering functions (Klette, 2014). For example, one can independently select the frequencies to suppress or enhance; however, due to the wide variety of options and the specifics of each task, it is very difficult to select them manually. In contrast, the proposed design of the filtering layer shown in Fig. 1 is capable of automatically forming a more sophisticated and a task-specific filter.[3]

To approximate an arbitrary nonlinear function that performs the desired filtering, a neural network with just one hidden layer is enough, yielding the *General* configuration of GAFL:

$$|\mathcal{F}I| \leftarrow W_2 * \sigma\left(W_1 * |\mathcal{F}I| + B_1\right) + B_2, \tag{3.2}$$

where W_1 and W_2 are the weight matrices with non-negative elements, B_1 and B_2 are the bias matrices with non-negative elements, $\sigma(\cdot)$ is some nonlinear activation function, and "$*$" denotes element-wise multiplication. Algorithm 1 describes the complete function of the proposed global adaptive filtering layer.

Handling small frequency values. When a base neural network is chosen, the proposed layer is pre-pended to it and, then, evaluated with several variations to experiment with the small values of non-central frequencies. These variations were studied to 'boost' the appearance of the smallest frequency pixel values in the spectra, as the intensity of a typical central frequency often 'overwhelms' the smaller values on the periphery (see insets in Fig. 3 below to see the typical spike-shaped learnt spectra). For example, the *General log* configuration

$$|\mathcal{F}I| \leftarrow W_2 * \sigma\left(W_1 * \log\left(1 + |\mathcal{F}I|\right) + B_1\right) + B_2 \tag{3.3}$$

learns the GAFL weights after the logarithmic function intensifies the high-frequency tails of the spectrum.

Similarly to General Eqs. (3.2) and (3.3), one can also experiment with a basic *Linear* configuration (a simple single-layer neural network):

$$|\mathcal{F}I| \leftarrow W * |\mathcal{F}I|, \tag{3.4}$$

and its corresponding logarithmic version, *Linear log*:

$$|\mathcal{F}I| \leftarrow \exp\left[W * \log\left(1 + |\mathcal{F}I|\right)\right] - 1, \tag{3.5}$$

where the exponential function 'undoes' the effect of the logarithm to preserve the linearity in the layer.

Number of parameters. The complex values of Fourier transform are tackled by the operation $|\mathcal{F}I|$. Yet, the important symmetry property $\mathcal{F}I(n-u, m-v) = \mathcal{F}I(-u,-v) = \mathcal{F}I(u,v)^*$ helps to compute the number of parameters added by the adaptive layers. Namely, each matrix of weights is a tensor of size $\left(C, n, \lfloor m/2 \rfloor + 1\right)$, where C is the number of channels, (n, m) is the image size. Thus, the number of learnable parameters of the proposed adaptive layer is equal to the number of weight matrices multiplied by the product of the matrix dimensions (as the operations in the frequency space are element-wise).

Computational complexity. The computational complexity for the base models is calculated using the *ptflops* software package (Sovrasov, 2019). To compute the number of operations in the direct and the inverse Fourier transforms, the Split-radix (Duhamel and Hollmann, 1984; Duhamel and Vetterli, 1990) algorithm is used. In total, the 2D discrete Fourier Transform of size (n, m) and the element-wise product of the GAFL weight matrix (size $(n, \lfloor m/2 \rfloor + 1)$) with the frequency image add up, yielding the following *MAC operations*:

$$\mathrm{MACs} \simeq 4nm\log_2(nm) - 12\ nm + 8\left[n + m\right] + n \cdot \left[\lfloor m/2 \rfloor + 1\right] \tag{3.6}$$

Eq. (3.6) can be easily generalized to multidimensional images, taking into account the linearity of the computational complexity *w.r.t.* dimensions. For those configurations of Eqs. (3.2)–(3.5) where the exponential or the logarithmic operations are employed, an upper estimation of the computational complexity is reported.

4. Experiments and results

In this section, we compare the performance of renowned neural networks *with and without* the proposed trainable layers. In all four tasks, we always choose the most popular models, common initialization strategies, and only the well-known activation and loss functions. We aim to make the existing network architectures more efficient.

Hyperparameters common to all tasks. All models are trained setting the input size to (256, 256), batch size 4, and using *Adam optimizer* (Kingma and Ba, 2014) with learning rate 0.001.

4.1. Datasets

We validate efficiency of our adaptive layer on two medical and on four natural image benchmarks.

The medical benchmarks comprise ultrasonic datasets, popular in medical vision community: *Breast Ultrasound Images* (**BUSI** Al-Dhabyani et al., 2019, 1578 images of three classes: normal (266), benign (891) and malignant (421) tumours, as well as ground-truth segmentation masks) and *Brachial Plexus Ultrasound Images* (**BPUI** BPUI, 5635 images and masks).

The natural benchmarks were selected to represent well-known datasets of various scales: from the small **Caltech Birds** (Caltech-UCSD Birds-200-2011 Wah et al., 2011, 11,788 images of 200 classes with ground-truth segmentation masks), to medium **Dogs *vs.* Cats** (kaggle dataset, 25k images for binary classification), to large-scale **CIFAR-10** (Krizhevsky and Hinton, 2009, 50k images of 10 classes) and a part of

[2] We do not centre our discussion around *phase*, which could also prove useful for some applications. See Supplementary material.

[3] Basic Fast Fourier Transform (FFT) and the element-wise multiplication functions in modern software packages are suitable.

Table 1
Segmentation results.

Model	BUSI	BPUI	Birds
U-Net	0.70	0.59	0.84
+ GAFL Linear	0.72	0.64	0.85
+ GAFL Linear log	0.71	0.64	**0.86**
+ GAFL General	**0.75**	**0.74**	**0.86**
+ GAFL General log	**0.75**	**0.74**	**0.86**
DenseNet	0.77	0.74	0.94
+ GAFL Linear	0.79	**0.77**	0.94
+ GAFL Linear log	0.80	**0.77**	0.94
+ GAFL General	**0.81**	0.74	**0.95**
+ GAFL General log	0.77	0.75	0.94
ResNet	0.80	0.71	0.93
+ GAFL Linear	**0.81**	0.72	**0.94**
+ GAFL Linear log	**0.81**	0.72	**0.94**
+ GAFL General	**0.81**	0.74	**0.94**
+ GAFL General Log	**0.81**	**0.75**	**0.94**

Validation Dice score for different models on medical (BUSI and BPUI) and natural (Caltech Birds) datasets. The best performance is highlighted in bold.

ImageNet ('Tiny' ImageNet Le and Yang, 2015, 110k images of 200 classes).

We considered medical imaging datasets separately because the ultrasound signal is known to have particular frequency bands needed for the optimal image contrast in live imaging (Ihnatsenka and Boezaart, 2010), making us hypothesize that the filtering effect would be more pronounced in these data. Dataset details and descriptions are given in the Supplementary material.

4.2. Segmentation

To test the proposed method, three different networks were studied as the base models: U-Net (Ronneberger et al., 2015), DenseNet (Huang et al., 2016), and ResNet (He et al., 2015).

For the learning process, we used the Combined Loss function of *Dice* and *Cross Entropy*, weighted as 0.6 and 0.4 respectively. The quality of segmentation is evaluated with the *Dice coefficient* (Milletari et al., 2016), which, in essence, measures the overlap between the predicted and the ground-truth masks.

The following hyperparameters were used. For U-Net, *init_features* = 32 (number of parameters in initial convolution), *depth* = 3 (number of downsteps). For DenseNet (as for densenet-121), *init_features* = 32, *growth_rate* = 32 (number of filters to add to each layer), *block_config* = 6, 12, 24, 16 (number of layers in each pooling block). For ResNet (as for resnet-18), *blocks*: 2, 2, 2, 2 (number of layers in each pooling block).

We observe improvement of the segmentation performance in all three base models. One can notice a significant increase of the metric values for the light models and an accelerated convergence for all architectures in Fig. 2. The metrics, summarized in Table 1, reveal a notable gap between the base models with and without the adaptive layer. Remarkably, these improved values originate from the images that have actually lost their clean appearance after the pre-processing step (see Fig. 3). Notice how the non-essential features and the textures disappear and how the look of the images is altered by the filters learnt by the GAFL. Ultimately, this 'ruined' look does not matter for the target task, because the quality of the segmentation task is still maximized. Fig. 3 also shows the difference in the small-value boosting configurations (Ref. Eqs. (3.2)–(3.5)), where the insets show the learnt spectral filters after the training. Comprehensive results for each dataset and each model are given in the Supplementary material.

Table 2
Classification results: Binary.

Model	BUSI norm *vs.* ben	BUSI ben *vs.* mal	D *vs.* C
CNN	0.88	0.76	0.82
+ GAFL Linear	0.89	**0.78**	**0.83**
+ GAFL Linear log	0.89	0.77	0.82
+ GAFL General	0.90	0.76	0.82
+ GAFL General log	**0.92**	0.77	**0.83**

F_1-scores for different models on BUSI validation set (normal *vs.* benign and benign *vs.* malignant classes), and Dogs *vs.* Cats datasets. The best performance is highlighted in bold.

4.3. Classification

To verify the suggested algorithm for the classification problem, a typical Convolutional Neural Network (CNN) with several convolutional blocks and fully-connected layers is used. Namely, the encoder blocks include *Conv*, *Batch Normalization*, *ReLU*, *Average Pooling*, and two *fully-connected* layers (using *init_features* = 8 and *depth* = 4). The training process is similar to the one above, with using the *weighted Cross Entropy Loss* (Ho and Wookey, 2019) combined with the F_1*-score* evaluation.

The results, presented in Table 2 and Fig. 4, demonstrate critical improvement not only in the classification task between the normal and the tumour tissues, but also in the task of distinguishing the normal tissue from the benign, and the malignant tumours. The addition of GAFL makes the prediction more sensitive to detecting these sub-classes, which can be of interest in the clinical practice (BPUI; Tuluptceva et al., 2020). Likewise, the overall quality enhancement is also evident on the natural data, where the classification experiments were performed on the large-scale datasets CIFAR-10 and ImageNet (see Table 3 and Fig. 6).

4.4. Denoising and erasing

For the problem of denoising and erasing, we considered popular model DnCNN (Zhang et al., 2016) as the baseline, the main task of which is to restore the noise, in contrast to the standard feedforward models, which restore the image. To assess the success of denoising and erasing, we used Combined Loss function of *MS-SSIM* and L_1 *Loss* with weights 0.8 and 0.2 respectively (Zhao et al., 2016) along with the *FSIM* and *PSNR* metrics (Zhang et al., 2011). We used the following hyperparameters for the architecture: *init_features* = 64, *num_layers* = 17 (number of layers).

For this task, we introduce the regular Gaussian noise and the rectangular erasing corruptions (Zhong et al., 2017) to the images, with the consequent image recovery yielding the outcomes summarized in Fig. 5. Note that despite the popularity of *PSNR* and *FSIM* metrics, they are oftentimes not representative of the true image quality. Herein, we resort to using these metrics merely to compare with the baseline models in a standard way.[4]

4.5. Control and large-scale experiments

To gauge the impact of the added layer on the base model *precisely*, a fine control of the total number of trainable parameters is desirable. Therefore, additional studies were performed for each CV problem, assuring that the number of parameters in base models was *either greater or nearly equal* of that with the pre-pended adaptive filtering layer (Table 3 and Fig. 6). In this sub-study, all experiments were run using the *Linear* configurations of the adaptive layer.

For the segmentation task, U-Net with *init_features* = 32 (467,842 parameters) was compared against U-Net with *init_features* = 16 prepended

[4] The search for the proper metrics is beyond the scope of this work.

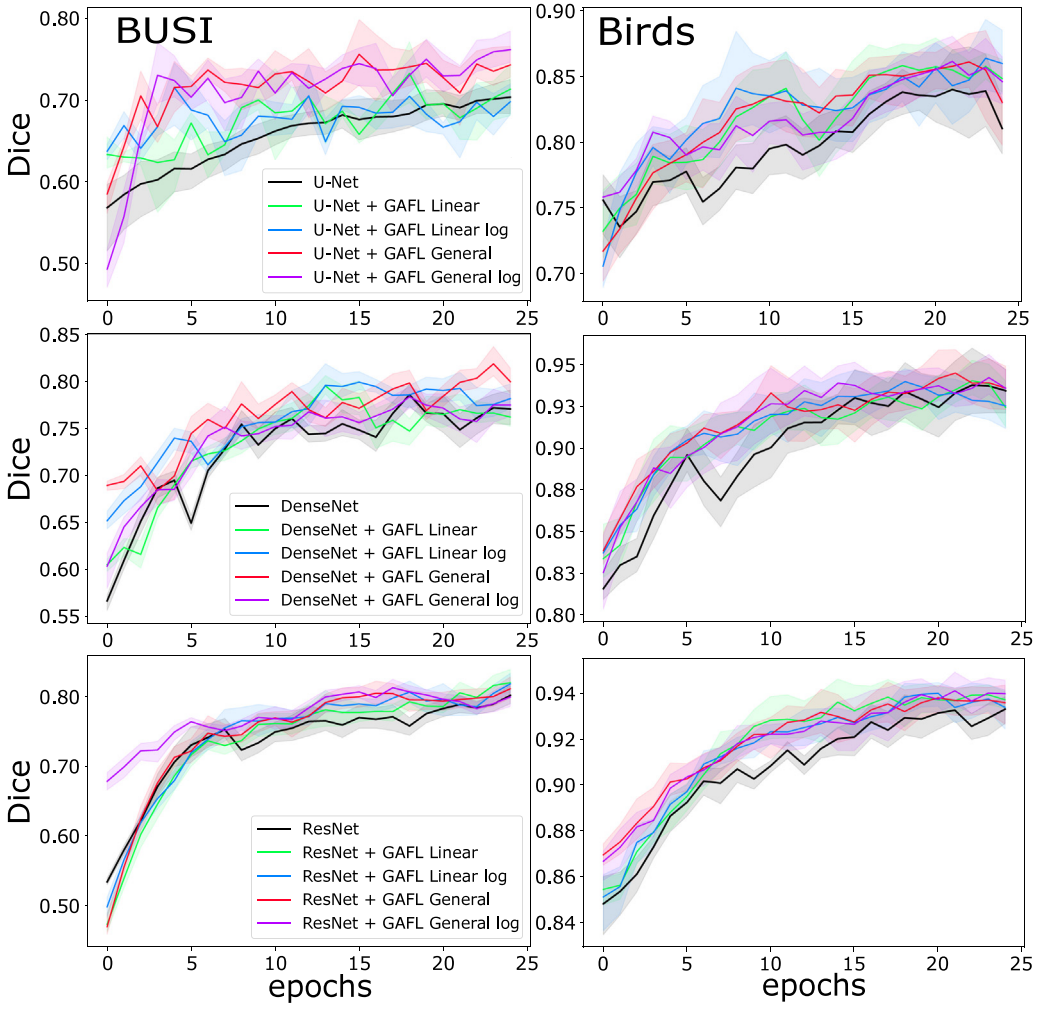


Fig. 2. Segmentation results. Average Dice coefficients on validation sets of different datasets: medical (BUSI) and natural (Birds). Top row: U-Net, middle: DenseNet, bottom: ResNet.

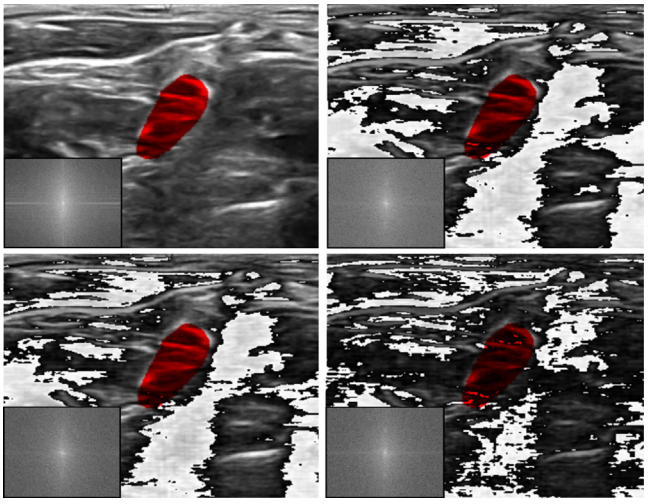


Fig. 3. Segmentation results. Examples of 'learnt' filters and their effect on the segmentation. Top left: original image and ground-truth mask; top right: *Linear* filter; bottom left: *General* filter; bottom right: *General log* filter. Corresponding 'learnt' spectra are shown in the insets in the corner. It does not matter if the learnt filter ruins the appearance of the original image; what matters is that it enhances the segmentation performance by preserving only the important frequencies.

by GAFL (248,994 parameters for BUSI with *image_size* = (512, 512) and 216,770 parameters for Caltech Birds with *image_size* = (256, 256)), which corresponds to **a reduction** of the number of model parameters by *(467,842 – 248,994)/467,842 = 46.7%* and *(467,842 – 216,770)/467,842 = 53.7* % respectively.

For the classification task, ResNet-20 was compared with the same model pre-pended by the adaptive layer for CIFAR-10 with *image_size* = (32, 32) and Tiny ImageNet with *image_size* = (64, 64), corresponding **to a reduction** of the number of model parameters by *(276,026 – 274,394) / 276,026 = 0.6%* and *(293,080 – 286,744) / 293,080 = 2.2%*, accordingly. For these experiments, *SGD optimizer* with the weight decay of 0.0001, the momentum of 0.9, and the learning rate of 0.1 was used.

For the denoising and the erasing tasks, DnCNN with *init_features* = 32 and *num_layers* = 20 was compared to itself with *init_features* = 16 and *num_layers* = 17 and with the pre-pended GAFL for BUSI with *image_size* = (512, 512), also roughly preserving the number of model parameters with *(168,225 – 167,169) / 168,225 = 0.6%*.

5. Discussion

Table 3 demonstrates the advantage of training the base model with the proposed adaptive layer. GAFL allows for both a reasonable improvement/preservation of performance in all tasks and datasets and requires fewer parameters, significantly reducing the computational complexity. The latter allows us to train the models (up to several

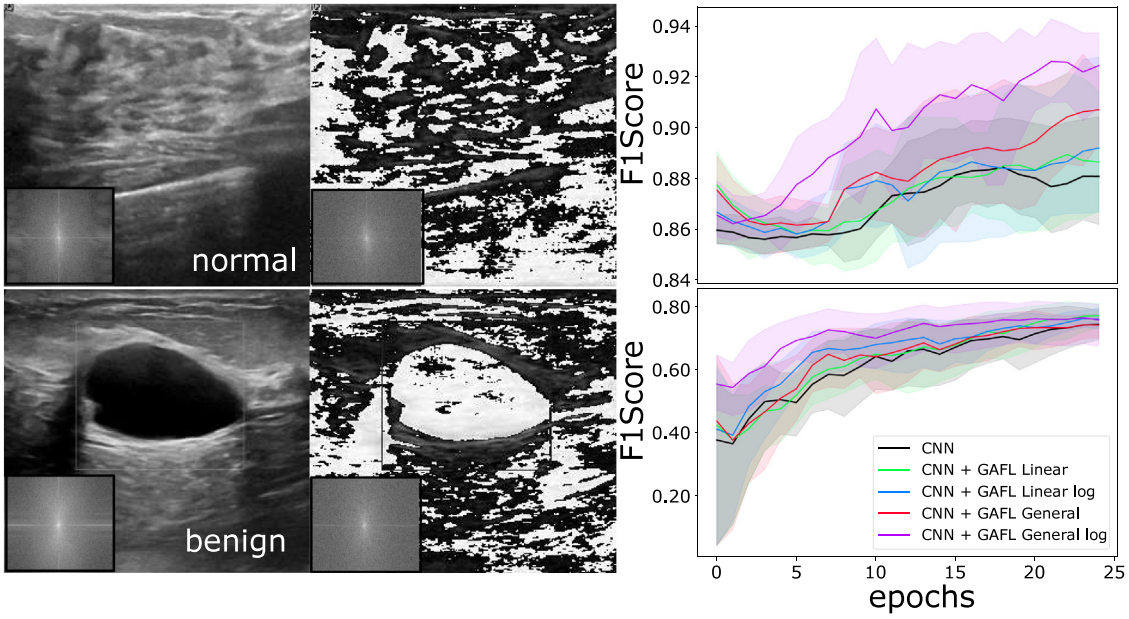


Fig. 4. Classification results: Binary. Left: Initial and filtered images using our layer in *General log* configuration. Insets show the 'learnt' optimal spectra. Right: F_1-scores on BUSI validation sets (normal *vs.* benign and benign *vs.* malignant). Despite the ruined appearance of the input image, GAFL allows to boost classification scores and accelerates convergence.

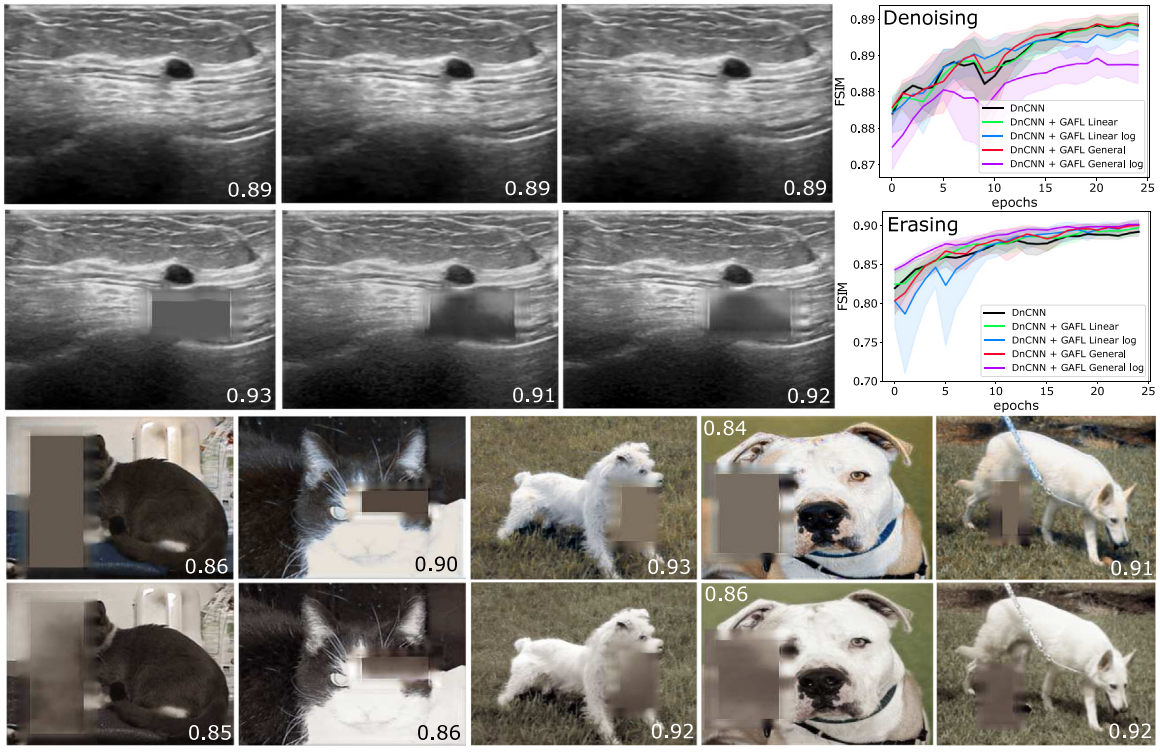


Fig. 5. Denoising/Erasing results. Plots on the right show FSIM metrics for denoising and erasing corruptions of BUSI dataset for different models. First and second row images, left to right: base DnCNN model results, the model with *Linear*, and with *General* adaptive filters on BUSI dataset. Third and fourth rows, left to right: base model and model with *Linear* adaptive filtering layer on the Dogs *vs.* Cats dataset. Note how addition of Fourier-based layer corrects for the corruptions better (for example, dog images in the third *vs.* the fourth row).

times) faster and demands less time for the inference, which favourably positions GAFL for the future deployment in various applications.

Remarkably, in the controlled experiments, we observed that the number of parameters in the base model could be reduced by half; yet, the global adaptive filtering layer allows to 'catch up' with the lost parameters and to reach the level of the base models that have twice as many parameters. The result generalize well across different datasets of various scales and across CV tasks. We did not observe a stunning improvement in the denoising problem built around DnCNN model — a result we attribute to the way the noise is handled in the Fourier domain, making our metric choice somewhat sub-optimal. Additional studies are required with the DnCNN to understand the enhancement dynamics; however, the same very model is well improved by our layer when there is a notable corruption (the erasing problem). We can

Table 3
Control experiments results.

Model	Segmentation			Classification			Denoising		Erasing
	MACs, 10^9	BUSI	Birds	MACs, 10^6	CIFAR-10	Tiny ImageNet	BUSI	MACs, 10^9	BUSI
Base model	28.79	0.64	**0.86**	44.4	0.782	0.345	**30.69**	44.26	23.28
+ GAFL Linear	7.30	0.69	**0.86**	32.0	0.792	**0.348**	30.42	9.41	22.91
+ GAFL Linear log	7.31	**0.70**	**0.86**	32.5	**0.796**	0.336	30.66	9.42	**23.92**

Comparison between different models for segmentation (Dice score), classification (Accuracy, in %), and Gaussian denoising and erasing corruptions (PSNR, in dB). The number of parameters in all experiments is controlled not to exceed the number of parameters in the corresponding base models. MACs (the same for denoising and erasing) are provided to compare the computational complexity across all models and tasks. The best performance is highlighted in bold.

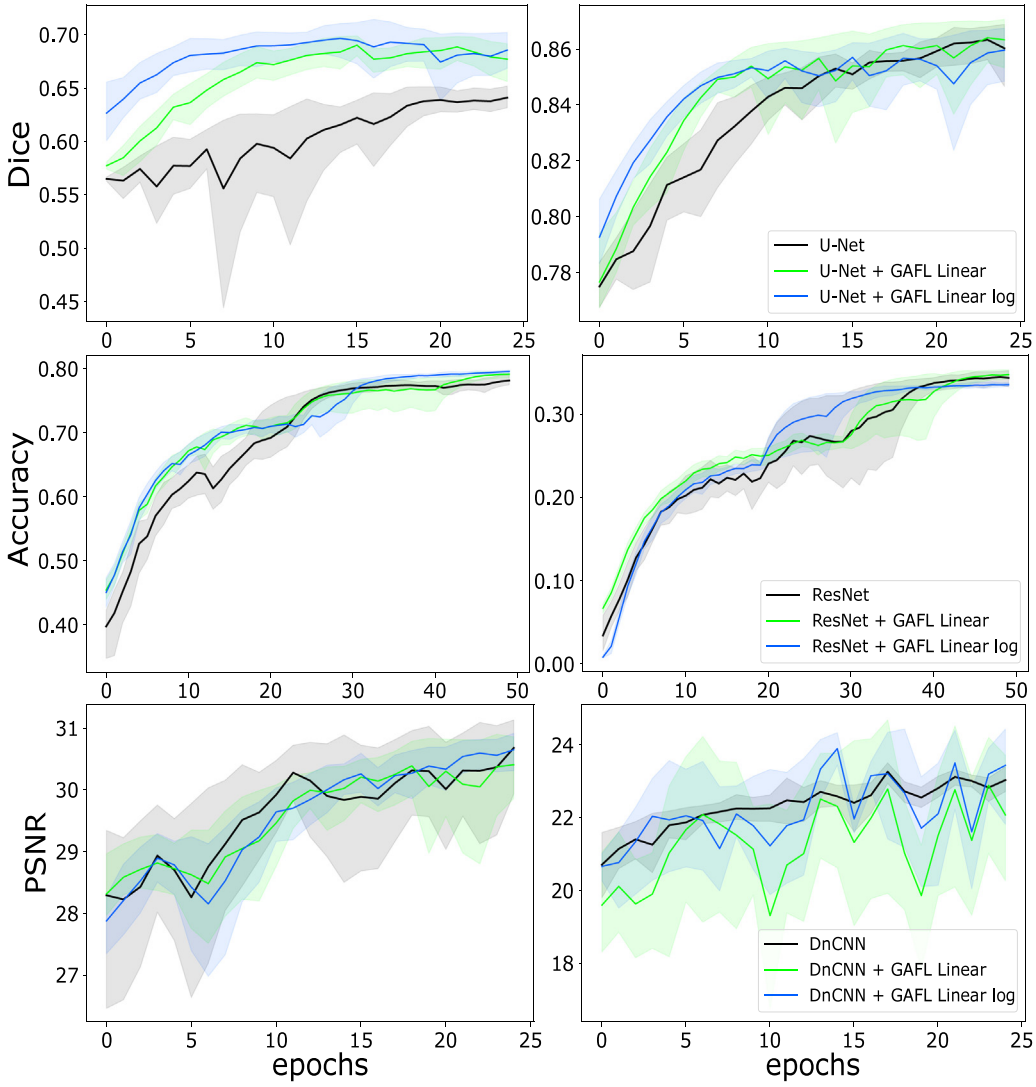


Fig. 6. Control and large-scale experiments. Metrics on validation sets for different models. Top row: Dice score for segmentation problem on medical (BUSI, left) and natural (Caltech Birds, right) datasets. Middle row: Prediction Accuracy (in %) for classification problem on CIFAR-10 (left) and Tiny ImageNet (right) datasets. Bottom row: PSNR (in dB) metric for Gaussian denoising (left) and erasing corruption (right) problems on BUSI dataset.

visually confirm that our adaptive configuration denoises and 'heals' the corruptions better than the base DnCNN model alone.

Activation functions. The activation function used in the general configuration of proposed global filtering layer is a hyperparameter that needs to be selected depending on the problem being solved and the dataset. In the provided algorithm, the function receives a non-negative matrix as input; it is since the frequency with negative weight has no physical interpretation. Therefore, several popular activation functions have been selected and investigated. As can be seen in the results in the Supplementary material, the activation function plays an important role. The average difference between the best and worst activation functions in some cases can lead to about 10 % gain by metric. It should be noted that the activation functions *Mish* and *ReLU* have shown themselves well on all datasets. So, *Mish* and *ReLU* are good for using our algorithm out of the box. Note that all the experiments reported in the main text were carried out using *ReLU* activation function.

Medical vs. Natural datasets. In datasets collected using ultrasound imaging, there are a lot of negative examples (with a zero or an empty mask). Unlike the Caltech Birds dataset, such data require more focus, dedication, and expertise to annotate them with labels. But, as with all human-tagged data, one should expect artefacts and potential errors in the markings: for example, the BPUI data have been annotated

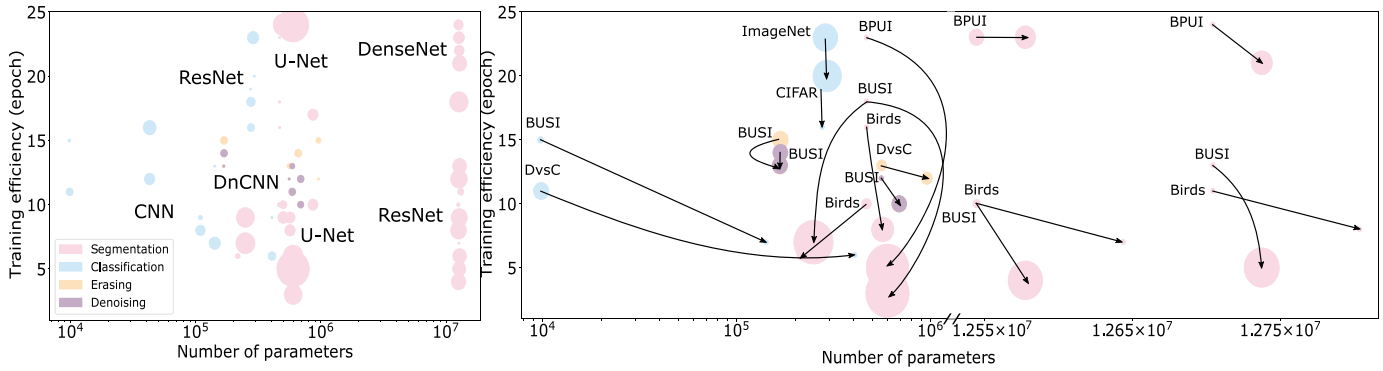


Fig. 7. Training efficiency as a function of model complexity for all 88 experiments. The bubble size encodes the gain in the corresponding metrics. Left: all experiments; right: zoomed-in areas. Colours correspond to different types of 4 CV tasks considered. Vertical and left-leaning arrows correspond to 18 control experiments run with precise control of the number of parameters in the models. The arrows indicate the correspondence of pairs (basic model → the most training-efficient model with the proposed adaptive Fourier layer). Birds: Caltech Birds (2011) dataset; DvsC: Dogs *vs.* Cats dataset; CIFAR: CIFAR-10 dataset; ImageNet: Tiny ImageNet dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

by pseudo-experts (people who have been trained and instructed by experts).

Several important observations and conclusions can be made from the results in the figures herein and from those in the Supplementary material:

- The observed improvement of Dice score on the ultrasound datasets depends on the number of parameters of the base model: the larger the architecture size, the less noticeable the increase in the metric when using the adaptive layer.
- In the natural images dataset, there is a faster convergence of all models when the proposed filtering layer is added, than in any of the base models.
- The proposed log operation applied to the spectra allows for the small pixel values and eliminates fluctuation artefacts otherwise appearing from the truncation.
- Addition of the adaptive filter to the natural images provides "smoother" convergence of the training curves. We believe this trait could be instrumental for accelerating many state-of-the-art models where predicting the behaviour of the model is important; for example, in reinforcement and/or active learning (Fig. 7).

Erasing metric choice. At the moment, there are a lot of metrics for assessing the quality of denoising of various types. However, there is no universal one that would accurately correspond to the assessment with the naked eye and is well interpreted in all cases (Ding et al., 2020). The most suitable for our task was the FSIM metric, which uses the structure of the Fourier components of the image. However, on a large number of examples (including those provided to you in Fig. 5), it was noticed that the difference in metric between the base model and the model with the proposed trainable layer should have a larger gap, since filling the inside of the cropped rectangle has greater importance than just averaging. Our experiments with other metrics, such as PSNR, present a sound but still sub-optimal alternative.

6. Conclusions

The method proposed in our work proved to be efficient on all datasets and all classical model architectures that we have considered. In all cases, the use of simple adaptive frequency filtering layer has led to faster convergence of the training process than in the case of the stand-alone model, having shown higher segmentation quality both on train and on test samples. A rather important finding is the increase of Dice score for the case of simple U-Net by around 6 % when the adaptive global filter is added. This promises an opportunity for the areas, such as medicine, where getting marked data is an acknowledged challenge, causing one to attempt learning on small datasets. In these cases, the use of heavy models with a large number of parameters is one possible solution which frequently leads to a fast overfit; whereas, addition of simple adaptive filtering layers "trims" unnecessary frequencies in the Fourier domain and makes the model learn only the vital frequencies along with the weights of the main neural network. All of this is accomplished while optimizing the targeted advantage function of interest to a particular application (for example, Dice score or F_1-score).

We believe the proposed layer can be a good add-on for a number of powerful modern preprocessing tools, including those that exist in various AutoML pipelines (Cubuk et al., 2019). In fact, there are many recently devised methods of augmentation and data preprocessing (for example, see Buslaev et al., 2020) that still await for the 'smart' frequency filtering capability. Currently, some configurations of the proposed adaptive layer could be preferred over the others, according to the data distribution and the problem being solved. Straightforward merging of these variants into a single universal configuration could be of particular practical value and will be published elsewhere. In the meanwhile, we propose to use the *General log* configuration of GAFL, as it shows consistently reliable performance in the majority of datasets and tasks and preserves the physical meaning of the Fourier filtering. We propose it for an approbation in the computer vision community as a simple add-on to a variety of modern deep learning models.[5]

The convergence speed of models plays an important role too. Areas, such as ultrasound imaging, entail big amounts of data, the labelling of which takes a lot of time and requires the involvement of highly qualified experts. Hence, the relevant methodology of active learning (Shelmanov et al., 2019) is frequently employed, requiring efficient retraining of the models and, thus, creating a welcoming setting for the adaptive pre-processing with GAFL. Same applies to the tasks that entail unsupervised segmentation and the pertinent optimization (Bespalov et al., 2020).

Initially, we anticipated that our adaptive layer would improve the convergence and the quality primarily in the ultrasound data (the echogenic nature of which is known to be prone to high sensitivity to the frequency knobs). But we were surprised to find out the improvement in the natural images as well. This expands the possible areas of application of the proposed approach and opens up a new direction of research of adaptive layers (for example, in more complex multi-layer architectures (Chowdhury et al., 2017), in generative and image translation models (Prokopenko et al., 2019), in learnable frequency kernels (Lazareva et al., 2020), in iterative anomaly detection models (Tuluptceva et al., 2020), *etc*. Lightening of these models via a simple adaptive filtering layer will be rather valuable, promising a fast and accurate solution to a variety of computer vision applications.

[5] GAFL code is available at https://github.com/cviaai/GAFL/

CRediT authorship contribution statement

Viktor Shipitsin: Methodology, Software, Visualization, Formal analysis, Validation, Writing – original draft. **Iaroslav Bespalov:** Methodology, Software, Visualization, Formal analysis, Validation, Writing – original draft. **Dmitry V. Dylov:** Conceptualization, Methodology, Visualization, Formal analysis, Supervision, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Skolkovo Institute of Science and Technology; Philips Innovation Labs RUS

Acknowledgement

We thank Nazar Buzun for helpful discussions. This research effort was supported by RFBR (grant # 21-51-12012).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2022.103519.

References

Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A., 2019. Dataset of breast ultrasound images. Data in Brief 28, 104863. http://dx.doi.org/10.1016/j.dib.2019.104863, URL https://scholar.cu.edu.eg/?q=afahmy/pages/dataset.

Bespalov, I., Buzun, N., Dylov, D.V., 2020. BRULÉ: Barycenter-regularized unsupervised landmark extraction. arXiv preprint arXiv:2006.11643.

Blackledge, J., 2005. Digital Image Processing: Mathematical and Computational Methods. pp. 1–797.

Bow, S.-T., 2002. Pattern Recognition and Image Preprocessing, 2nd Marcel Dekker, Inc., USA.

BPUI, D., 0000. Brachial Plexus Ultrasound dataset, URL http://www.kaggle.com/c/ultrasound-nerve-segmentation/data.

Brunton, S., Kutz, J., 2019. Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. http://dx.doi.org/10.1017/9781108380690.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: Fast and flexible image augmentations. Inf. 11 (2), 125. http://dx.doi.org/10.3390/info11020125, URL http://dx.doi.org/10.3390/info11020125.

Chowdhury, A., Dylov, D.V., Li, Q., MacDonald, M., Meyer, D.E., Marino, M., Santamaria-Pang, A., 2017. Blood vessel characterization using virtual 3D models and convolutional neural networks in fluorescence microscopy. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, pp. 629–632

Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V., 2019. Autoaugment: Learning augmentation policies from data. arXiv:1805.09501.

kaggle dataset, 0000. dogs and cats dataset, URL https://www.kaggle.com/c/dogs-vs-cats/data.

Ding, K., Ma, K., Wang, S., Simoncelli, E.P., 2020. Comparison of image quality models for optimization of image processing systems. arXiv:2005.01338.

Duhamel, P., Hollmann, H., 1984. Split radix'FFT algorithm. Electron. Lett. 20 (1), 14–16.

Duhamel, P., Vetterli, M., 1990. Fast Fourier transforms: a tutorial review and a state of the art. Signal Process. 19 (4), 259–299.

Gordon, D.F., Desjardins, M., 1995. Evaluation and selection of biases in machine learning. Mach. Learn. 20 (1), 5–22. http://dx.doi.org/10.1023/A:1022630017346.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S., 2016. Deep learning for visual understanding: A review. Neurocomputing 187, 27–48. http://dx.doi.org/10.1016/j.neucom.2015.09.116, URL http://www.sciencedirect.com/science/article/pii/S0925231215017634.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. CoRR, arXiv:1512.03385.

Ho, Y., Wookey, S., 2019. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE Access PP, 1. http://dx.doi.org/10.1109/ACCESS.2019.2962617.

Huang, H., Lin, L., Tong, R., Hu, H., Qiaowei, Z., Iwamoto, Y., Han, X.-H., Chen, Y.-W., Wu, J., 2020. Unet 3+: A full-scale connected unet for medical image segmentation.

Huang, G., Liu, Z., Weinberger, K.Q., 2016. Densely connected convolutional networks. CoRR arXiv:1608.06993.

Ihnatsenka, B., Boezaart, A., 2010. Ultrasound: Basic understanding and learning the language. Int. J. Shoulder Sur. 4, 55–62. http://dx.doi.org/10.4103/0973-6042.76960.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. Int. Conf. Learn. Represent.

Klette, R., 2014. Concise Computer Vision: An Introduction Into Theory and Algorithms. Springer Publishing Company, Incorporated.

Krizhevsky, A., Hinton, G., 2009. Learning multiple layers of features from tiny images. Computer Science Department, University of Toronto, Tech. Rep, 1.

Lazareva, E., Rogov, O., Shegai, O., Larionov, D., Dylov, D.V., 2020. Learnable hollow kernels for anatomical segmentation. arXiv preprint arXiv:2007.05103.

Le, Y., Yang, X., 2015. Tiny ImageNet visual recognition challenge.

Li, J., Chen, L., Cai, Y., 2009. Dynamic texture segmentation using Fourier transform. Mod. Appl. Sci. 3, http://dx.doi.org/10.5539/mas.v3n9p29.

Lin, J., Ma, L., Yao, Y., 2019. A Fourier domain training framework for convolutional neural networks based on the Fourier domain pyramid pooling method and Fourier domain exponential linear unit. IEEE Access PP, 1. http://dx.doi.org/10.1109/ACCESS.2019.2936591.

Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W., 2018. Multi-level wavelet-CNN for image restoration. CoRR arXiv:1805.07071.

McCann, M.T., Jin, K.H., Unser, M., 2017. Convolutional neural networks for inverse problems in imaging: A review. IEEE Signal Process. Mag. 34 (6), 85–95.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.C.H., Heinrich, M.P., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-net: Learning where to look for the pancreas. CoRR arXiv:1804.03999.

Pratt, H., Williams, B.M., Coenen, F., Zheng, Y., 2017. Fcnn: Fourier convolutional neural networks. ECML/PKDD.

Prokopenko, D., Stadelmann, J.V., Schulz, H., Renisch, S., Dylov, D.V., 2019. Unpaired synthetic image generation in radiology using gans. In: Workshop on Artificial Intelligence in Radiation Therapy. Springer, Cham, pp. 94–101.

Rawat, W., Wang, Z., 2017. Deep convolutional neural networks for image classification: A comprehensive review. Neural Comput. 29 (9), 2352–2449. http://dx.doi.org/10.1162/neco_a_00990, PMID: 28599112.

Rhu, M., Gimelshein, N., Clemons, J., Zulfiqar, A., Keckler, S.W., 2016. vDNN: Virtualized deep neural networks for scalable, memory-efficient neural network design. In: 2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO). pp. 1–13. http://dx.doi.org/10.1109/MICRO.2016.7783721.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J., 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv:1701.06538.

Shelmanov, A., Liventsev, V., Kireev, D., Khromov, N., Panchenko, A., Fedulova, I., Dylov, D.V., 2019. Active learning with deep pre-trained models for sequence tagging of clinical and biomedical texts. IEEE, pp. 482–489,

Song, J., Chai, Y.J., Masuoka, H., Park, S.-W., Kim, S.-j., Choi, J., Kong, H.-J., Lee, K.E., Lee, J., Kwak, N., Yi, K., Miyauchi, A., 2019. Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules. Medicine 98, e15133. http://dx.doi.org/10.1097/MD.0000000000015133.

Sovrasov, V., 2019. Flops counter for convolutional networks in pytorch framework. URL https://github.com/sovrasov/flops-counter.pytorch/.

Szeliski, R., 2011. Computer Vision Algorithms and Applications. Springer, London; New York, URL http://link.springer.com/book/10.1007%2F978-1-84882-935-0.

Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R., 2020. Fourier features let networks learn high frequency functions in low dimensional domains.

Tuluptceva, N., Bakker, B., Fedulova, I., Schulz, H., Dylov, D.V., 2020. Anomaly detection with deep perceptual autoencoders. arXiv preprint arXiv:2006.13265.

Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., 2011. The caltech-UCSD birds-200–2011 dataset. (CNS-TR-2011-001), California Institute of Technology, URL http://www.vision.caltech.edu/visipedia/CUB-200-2011.html.

Wang, L., Yang, S., Yang, S., Zhao, C., Tian, G., Gao, Y., Chen, Y., Lu, Y., 2019. Automatic thyroid nodule recognition and diagnosis in ultrasound imaging with the YOLOv2 neural network. World J. Surg. Oncol. 17, http://dx.doi.org/10.1186/s12957-019-1558-z.

Yang, Y., Soatto, S., 2020. Fda: Fourier domain adaptation for semantic segmentation.

Zhang, L., Zhang, L., Mou, X., Zhang, D., 2011. Fsim: A feature similarity index for image quality assessment. IEEE Trans. Image Process. 20, 2378–2386. http://dx.doi.org/10.1109/TIP.2011.2109730.

Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L., 2016. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Trans. Image Process. PP, http://dx.doi.org/10.1109/TIP.2017.2662206.

Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L., 2017. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Trans. Image Process. 26 (7), 3142–3155.

Zhao, H., Gallo, O., Frosio, I., Kautz, J., 2016. Loss functions for image restoration with neural networks. IEEE Trans. Comput. Imaging PP, 1. http://dx.doi.org/10.1109/TCI.2016.2644865.

Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y., 2017. Random Erasing Data Augmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, http://dx.doi.org/10.1609/aaai.v34i07.7000.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested U-net architecture for medical image segmentation. CoRR arXiv:1807.10165.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2020. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging 39 (6), 1856–1867.

Zuo, W., Zhang, K., Zhang, L., 2018. Convolutional neural networks for image denoising and restoration. In: Bertalmío, M. (Ed.), Denoising of Photographic Images and Video: Fundamentals, Open Challenges and New Trends 93–123. http://dx.doi.org/10.1007/978-3-319-96029-6_4.