Less Is More Important: An Attention Module Guided by Probability Density Function for Convolutional Neural Networks

Jingfen Xie, Jian Zhang

Peking University Shenzhen Graduate School, Shenzhen, China
xiejf@stu.pku.edu.cn, zhangjian.sz@pku.edu.cn

Abstract

Attention modules, which adaptively weight and refine features according to the importance of the input, have become a critical technique to boost the capability of convolutional neural networks. However, most existing attention modules are heuristic without a sound interpretation, and thus, require empirical engineering to design structure and operators within the modules. To handle the above issue, based on our 'less is more important' observation, we propose an Attention Module guided by Probability Density Function (PDF), dubbed PdfAM, which enjoys a rational motivation and requires few empirical structure designs. Concretely, we observe that pixels with less occurrence are prone to be textural details or foreground objects with much importance to aid vision tasks. Thus, with PDF values adopted as a smooth and anti-noise alternative to the pixel occurrence frequency, we design our PdfAM by first estimating the PDF based on some distribution assumption, and then predicting a 3D attention map via applying a negative correlation between the attention weights and the estimated PDF values. Furthermore, we develop learnable PDF-rescale parameters so as to adaptively transform the estimated PDF and predict a customized negative correlation. Experiments show that our PdfAM consistently boosts various networks under both high- and low-level vision tasks, and also performs favorably against other attention modules in terms of accuracy and convergence.

Introduction

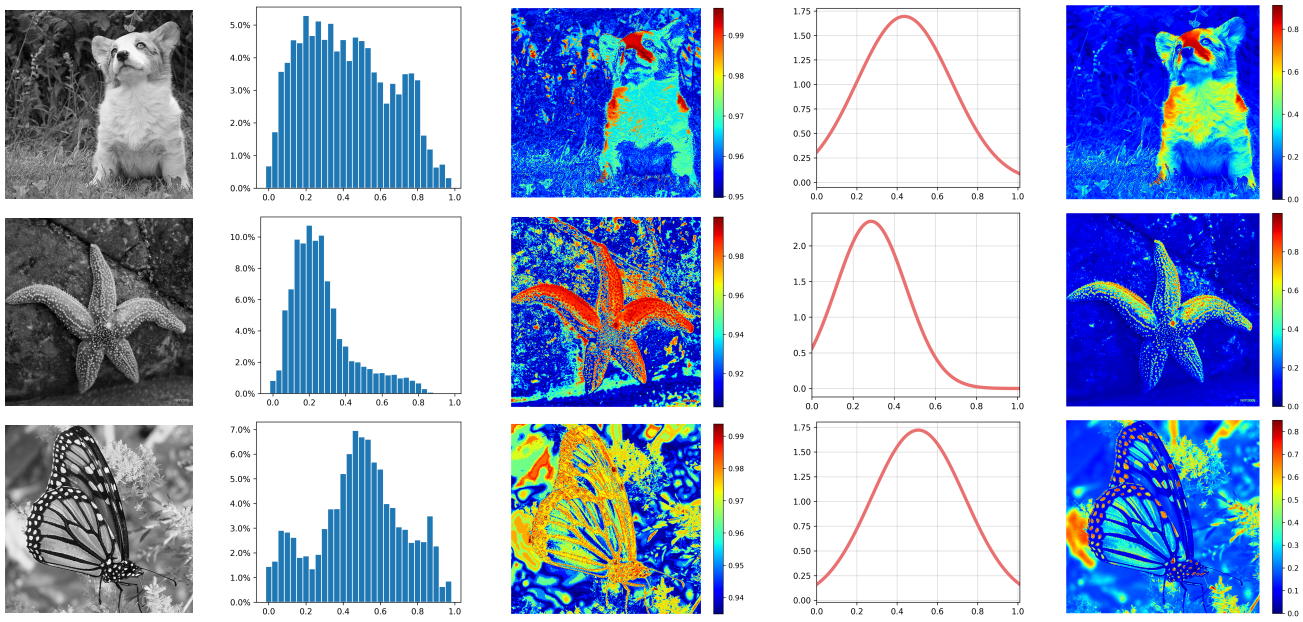
Convolutional Neural Networks (CNNs) are incredibly powerful in exploring visual imagery and have brought great success in many computer vision problems, including high-level tasks (*e.g.*, image classification (Krizhevsky, Sutskever, and Hinton 2012) and video understanding (Tran et al. 2018)) and low-level tasks (*e.g.*, image restoration (Zhang et al. 2017) and video restoration (Wang et al. 2019)). Various studies have explored better designs on CNN structures and shown high improvements, thus demonstrating the great significance of constructing a strong CNN in vision research.

A modern CNN usually consists of several stages, and each stage contains a few blocks. Such a block is built with operators, *e.g.*, convolution, activation, normalization, pooling or customized meta-structure (referred to as **module** in this paper). Rather than devising the entire architecture, many efforts have focused on developing advanced blocks to boost CNNs, *e.g.*, stacked convolutions (Szegedy et al. 2015) and residual units (He et al. 2016). However, constructing these blocks needs empirical expert knowledge, and costs enormous trials and time. Recently, researchers present automated machine learning (AutoML) techniques to allow automatic architecture construction (Zoph and Le 2016).

In addition to devising blocks, another important research area is plug-and-play attention modules (Hu, Shen, and Sun 2018; Woo et al. 2018), which can tell where to focus and how to recalibrate feature maps within a block, thus enabling the network to learn more informative features. Besides, attention modules are designed independently from architecture, and thus, can be plugged into various networks. Overall, attention modules enjoy the merits of representation effectiveness and application universality. Many studies explore how to generate attention weights for each channel or each spatial region, respectively categorized as channel attention (Hu, Shen, and Sun 2018; Wang et al. 2020) and spatial attention (Jaderberg et al. 2015; Wang et al. 2018). To integrate the above two dimensions, some works (Woo et al. 2018; Fu et al. 2019) propose to first separately predict channel and spatial weights, and then conduct a fusion operator to produce the final attention. There also exists literature (Wang et al. 2017; Yang et al. 2021) that explores a more direct way and jointly generates a 3D (height, width, channel) attention map in both spatial and channel dimensions.

However, most existing attention modules are heuristic without a sound interpretation, and thus, require significant empirical engineering to design structure and operators within the module. To deal with the above defects, we propose a Probability Density Function (PDF) guided Attention Module, dubbed PdfAM, which enjoys a sound motivation and requires few empirical structure designs. Specifically, the proposed design is based on our interesting 'less is more important' observation that pixels with less occurrence are prone to highlight foreground objects or textural details, both with great importance for computer vision tasks, as shown in Fig. 1 (c) and (e). Considering that it's heavy and noise-sensitive to directly use histograms to calculate the occurrence frequency of pixel values, we adopt PDF values to

This work was supported in part by Shenzhen General Research Project under Grant JCYJ20220531093215035. *Corresponding author: Jian Zhang*.



(a) Grey Images (b) Histograms (c) Histogram-based attention (d) Estimated PDF (e) PDF-based attention

Figure 1: Motivations. (a) Grey images $\mathbf{X}$ with N pixels $\{x_n\}_{n=1}^{N}$. (b) Histograms (30 bins). (c) Histogram-based attention, which uses histograms in (b) to get pixel occurrence $\mathcal{H}(x_n)$, and assigns attention weights via $m_n = 1/(\mathcal{H}(x_n)+1)$. We can see that pixels with less occurrence highlight foreground objects or textural details, both critical for vision tasks. (d) Estimated PDF based on Gaussian distribution assumption. (e) PDF-based attention, which uses PDF values $\mathcal{P}(x_n)$ in (d) to assign attention weights via $m_n = 1/(\mathcal{P}(x_n) + 1)$. Compared to (c), results in (e) effectively reduce noise and get more focused attention.

act as a more smooth, focused and noise-resistant alternative. Therefore, based on some rational distribution assumption (*e.g.*, a single Gaussian or a more complicated Gaussian Mixture Model), we propose to first estimate the PDF values, and then generate the attention map via applying a negative correlation mapping from the estimated PDF values to the attention weights for each pixel. Furthermore, we develop learnable PDF-rescale parameters so as to improve the adaptability of predicting a customized negative correlation. Overall, the main contributions of this paper are three-fold:

- Based on our 'less is more important' observation, we propose a Probability Density Function guided Attention Module, dubbed PdfAM, to get a 3D attention map with a proper explanation and few empirical structure designs.
- We further develop learnable PDF-rescale parameters to adaptively modulate the estimated PDF and predict customized negative correlations, thus enabling adaptability with respect to different blocks, networks and tasks.
- Experiments show that our PdfAM boosts various networks for both high- and low-level vision tasks, and also performs favorably against other popular attention modules in terms of performance and convergence.

Related Work

Network Architectures

Since the success of large-scale CNNs, many works have been proposed to strengthen CNNs. VGG (Simonyan and Zisserman 2014) and InceptionNet (Szegedy et al. 2015) use stacked convolutions to build deeper networks. To handle the optimization issue of deep CNNs, ResNet (He et al. 2016) proposes skip connections. Aside from network depth, WideResNet increases the number of filters (network width). To increase network cardinality, ResNeXt (Xie et al. 2017) adopts grouped convolutions. DenseNet (Huang et al. 2017) concatenates features from different layers. MobileNet (Howard et al. 2017) and ShuffleNet (Zhang et al. 2018) respectively emphasize the efficiency of depthwise convolution and shuffle operation between group convolutions. AutoML (Zoph and Le 2016; Tan, Pang, and Le 2020) automatically searches better architectures with fewer manual designs. Different from these works, we aim at designing a plug-and-play lightweight attention module, which can be plugged into multiple networks and boost various tasks.

Attention Modules

Human visual attention is a critical mechanism that diverts focus to the most relevant regions of an image and disregards less important parts (Corbetta and Shulman 2002), which inspires researchers to devise similar attention modules in CNNs. Typically, Squeeze-and-Excitation (SE) (Hu, Shen, and Sun 2018), uses global average-pooling and fully-connected (FC) layers to capture interactions between channels and tell what channels to attend. Similar **channel attention** includes gated channel transformation (Yang et al. 2020) and higher-order attention (Dai et al. 2019, 2021). Apart from channel attention, some works use adaptive spa-

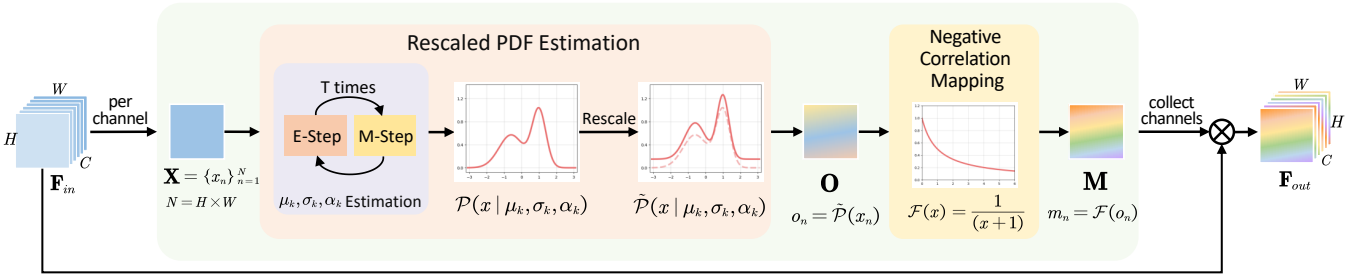


Figure 2: Illustration of PdfAM based on GMM. Taking a multi-channel feature map $\mathbf{F}_{in}$ as input, the attention is produced *independently* for each channel $\mathbf{X}$, and finally multiplied to $\mathbf{F}_{in}$ to output $\mathbf{F}_{out}$. Two steps are included: 1) Rescaled PDF estimation, which first uses unrolled EM algorithm to estimate GMM parameters $\mu_k, \sigma_k, \alpha_k$ and PDF $\mathcal{P}(x)$, and then adopts learnable rescaling to get rescaled PDF $\tilde{\mathcal{P}}(x)$, which is applied to each input pixel x_n in $\mathbf{X}$ to get occurrence frequency o_n. 2) Negative correlation mapping, which maps o_n to attention weights m_n for each pixel. Codes: https://github.com/chuan1093/PdfAM.

tial selection to get **spatial attention** telling where to attend, *e.g.*, spatial transformer (Jaderberg et al. 2015), Gather-Excite (Hu et al. 2018) and non-local networks (Wang et al. 2018). To consider both dimensions, (Woo et al. 2018; Fu et al. 2019) first predict channel and spatial weights separately, and then fuse them to get the final attention. Other works (Wang et al. 2017; Zhao et al. 2020; Mou et al. 2021) directly produce a 3D attention map to get informative features in both spatial and channel dimensions. However, most existing attention modules are hand-crafted designs based on ungrounded heuristics and need empirical engineering on structural choices. In contrast, our PdfAM has a rational explanation with few empirical structure designs.

Proposed Method

Motivation

We present a histogram-based attention in Fig. 1 (c), which first uses the histograms in Fig. 1 (b) to get pixel occurrence $\mathcal{H}(x_n)$, and then assigns attention weights $m_n = 1/(\mathcal{H}(x_n) + 1)$. Here, x_n denotes a pixel in image $\mathbf{X}$ (N pixels in total). The results basically verify that pixels with less occurrence tend to be foreground objects or textural details, both critical for vision tasks. However, one can observe that histogram-based attention is noise-sensitive and not so focused. Besides, histograms calculation is heavy and not differentiable (Avi-Aharon, Arbelle, and Raviv 2020), with the number of bins as a hand-crafted hyper-parameter.

To handle the above defects, we adopt PDF values as a differentiable, smooth and noise-resistant alternative. Based on the histograms in Fig. 1 (b), we rationally make a Gaussian distribution assumption, and use the estimated PDF values $\mathcal{P}(x_n)$ in Fig. 1 (d) to assign attention weights $m_n = 1/(\mathcal{P}(x_n)+1)$. As shown in Fig. 1 (e), the PDF-based attention effectively reduces noise and gets more focused on foreground objects or textural details, thanks to the smoothness of PDF. Concretely, one can see that, for the Dog image with a clear foreground differing a lot from the large background, the attention highlights the target object. For the Starfish image, though the difference between foreground and background is not so obvious, the foreground is still highlighted with clear texture patterns. As for the complex Butterfly image, the attention remarkably focuses on textural details.

Note that, such 'less is more important' rough observation might not be fully convincing for all images, but it might be more obvious for feature maps in the network, due to the feature extraction process. Overall, it's verified and also rational that pixels with smaller PDF values tend to present foreground objects or detailed textures, and thus, should be allocated larger attention weights.

PdfAM Based on GMM

Based on the above motivations, we present a Probability Density Function (PDF) guided Attention Module, dubbed PdfAM. In this section, focusing on more generalization, we adopt the Gaussian Mixture Model (GMM) distribution assumption, rationally based on our observations about feature value histograms. As shown in Fig. 2, the proposed PdfAM inputs and outputs multi-channel feature maps, and generates attention **independently** for each input channel, where two steps are included: 1) rescaled PDF estimation for $\tilde{\mathcal{P}}(x)$, and 2) negative correlation mapping via $\mathcal{F}(x)$. For statement convenience, we describe how to predict attention weights for **one input channel** in the following.

Rescaled PDF Estimation: Denoting the input channel as $\mathbf{X} \in \mathbb{R}^{H \times W}$ with N pixels $\{x_n\}_{n=1}^{N}$ ($N = H \times W$), we assume that x_n obeys a GMM composed of K single Gaussian distributions, and thus, follows the PDF given by:

$$\begin{aligned} p(x \mid \Theta) &= \sum_{k=1}^{K} \alpha_k \mathcal{N}(x \mid \mu_k, \sigma_k) \\ &= \sum_{k=1}^{K} \alpha_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right], \end{aligned} \tag{1}$$

where μ_k, σ_k and α_k respectively denote mean, standard deviation and mixture coefficient of the k-th Gaussian distribution, with α_k constrained via $\alpha_k \in [0, 1]$ and $\sum_{k=1}^{K} \alpha_k = 1$. All the parameters to be estimated are collectively denoted by Θ, *i.e.*, $\Theta = \{\mu_k, \sigma_k, \alpha_k\}_{k=1}^{K}$.

We adopt the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) to estimate GMM-PDF

parameters. Specifically, EM introduces a latent variable z ($z \in \{1, 2, ..., K\}$) to describe the probability of a given sample x belonging to which component in the Gaussian mixture. Starting from an initial estimate, EM iteratively updates Θ till convergence. Each iteration contains an E-step and an M-step. In the E-step, given observations x_n and parameters Θ, the posterior distribution of z is updated by:

$$z_{n,k} = \frac{\alpha_k \mathcal{N}(x_n \mid \mu_k, \sigma_k)}{\sum_{k=1}^{K} \alpha_k \mathcal{N}(x_n \mid \mu_k, \sigma_k)}. \tag{2}$$

In the M-step, given observations x_n and distribution $z_{n,k}$, the likelihood is maximized by making partial derivative over the parameters Θ, obtaining update formulas as:

$$\hat{\mu}_k = \frac{\sum_{n=1}^{N} (z_{nk} x_n)}{\sum_{n=1}^{N} z_{nk}}, \tag{3}$$

$$\hat{\sigma}_k^2 = \frac{\sum_{n=1}^{N} z_{nk} (x_n - \mu_k)^2}{\sum_{n=1}^{N} z_{nk}}, \tag{4}$$

$$\hat{\alpha}_k = \frac{\sum_{n=1}^{N} z_{nk}}{N}. \tag{5}$$

Traditionally, EM needs to iterate between E-step and M-step till convergence. In order to implement the iteration process as an end-to-end differentiable operator in our attention module, inspired by (Greff, Van Steenkiste, and Schmidhuber 2017), we fix the iteration times as T and unroll the EM iterations to form a GMM-PDF parameter estimation operator, as shown in Fig. 2. We set $K = 2$ and $T = 3$ by default.

Considering that designing a fixed relation between attention and PDF values lacks flexibility and adaptability, we devise learnable parameters to rescale PDF without changing its peak axes and general shape before converting PDF values to attention. Concretely, we adopt three scaling parameters S^M, S^A and S^E to respectively conduct multiplicative, additive and exponential scaling of the GMM-PDF in Eq. (1). The rescaled GMM-PDF is formulated as:

$$\tilde{\mathcal{P}}(x \mid \Theta) = S^A + \sum_{k=1}^{K} \alpha_k \frac{S^M}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{S^E \cdot (x - \mu_k)^2}{2\sigma_k^2}\right]. \tag{6}$$

To realize positive rescaling to PDF, the three scaling parameters are constrained to be positive via applying a *Softplus* function to unconstrained learnable parameters. Fig. 3 shows the effects of three scaling parameters on two functions: 1) the rescaled PDF $\tilde{\mathcal{P}}(x)$, and 2) the entire attention function $\mathcal{M}(x) = \mathcal{F}(\tilde{\mathcal{P}}(x))$. One can observe that, in general, S^M and S^E rescale the height and width respectively, and S^A realizes vertical movements. Note that S^E does not change the peak values for each Gaussian component, but affects the overall GMM peak values, due to the change in steepness. Later analysis on Fig. 7 verifies that our PDF-rescale design can learn adaptive scaling parameters with respect to different blocks, networks and tasks. We use the estimated PDF values as the pixel occurrence frequencies $\mathbf{O} \in \mathbb{R}^{H \times W}$, with each pixel occurrence $o_n = \tilde{\mathcal{P}}(x_n)$.

Negative Correlation Mapping: Based on our previous discussed 'less is more important' observation, we transform the pixel occurrence frequencies $\mathbf{O}$ to get attention weights

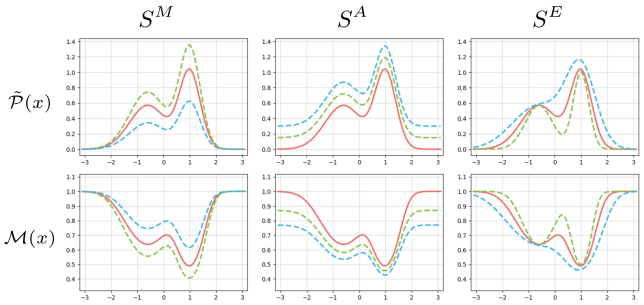


Figure 3: Effects of three scaling parameters on two functions: 1) the rescaled PDF $\tilde{\mathcal{P}}(x)$, and 2) the entire attention function $\mathcal{M}(x) = \mathcal{F}(\tilde{\mathcal{P}}(x))$. Solid and dotted lines respectively denote un-rescaled and rescaled version. S^M, S^A and S^E respectively realize height rescale, vertical movements and width rescale.

$\mathbf{M} \in \mathbb{R}^{H \times W}$ by a **negative** correlation mapping, with each pixel m_n obtained via:

$$m_n = \mathcal{F}(o_n) = \frac{1}{o_n + 1}, \tag{7}$$

which outputs $m_n \in [0, 1]$, due to $o_n = \tilde{\mathcal{P}}(x_n) > 0$. The attention is produced independently for each channel, and finally collected and multiplied to the multi-channel input, as shown in Fig. 2. Formally, denoting the input and output multi-channel features as $\mathbf{F}_{in}$ and $\mathbf{F}_{out}$, and the collected multi-channel attention map as $\mathbf{A}$, we have:

$$\mathbf{F}_{out} = \mathbf{F}_{in} \odot \mathbf{A}. \tag{8}$$

When integrating PdfAM into various residual blocks, following (Howard et al. 2019), we apply PdfAM after the second convolution layer.

PdfAM Based on Gaussian Distribution

Considering that PdfAM based on GMM requires high computation and memory consumption due to the unrolled EM algorithm, we provide a fast special case when $K = 1$, *i.e.*, based on the single Gaussian distribution assumption. These two versions of PdfAM are respectively denoted by **PdfAM-Gmm** and **PdfAM-Gau**. In PdfAM-Gau, the input pixels x_n have a simpler PDF as:

$$\mathcal{P}(x \mid \Theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right], \tag{9}$$

where $\Theta = \{\mu, \sigma\}$ can be easily estimated by $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$ and $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_k)^2$. The rescaled PDF is formulated as:

$$\tilde{\mathcal{P}}(x \mid \Theta) = S^A + \frac{S^M}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{S^E \cdot (x - \mu)^2}{2\sigma^2}\right]. \tag{10}$$

Other settings in PdfAM-Gau are the same as PdfAM-Gmm. We can easily implement PdfAM-Gau in a few lines of codes. Apart from GMM and Gaussian, our PdfAM enjoys great generality of being applied to other assumptions, *e.g.*, Laplace distribution.

T	K	C10	C100
3	1	92.59	68.74
3	2	92.68	69.22
3	3	92.64	69.23
3	4	92.84	69.28
3	5	92.78	69.37

T	K	C10	C100
1	2	92.58	69.04
2	2	92.60	69.08
3	2	92.68	69.22
4	2	92.68	69.21
5	2	92.75	69.28

Table 1: Ablation on T and K in PdfAM-Gmm.

Rethinking Existing Attention Modules

Interestingly, we find that a recently proposed interpretable attention module, called SimAM (Yang et al. 2021), is a special case of our PdfAM-Gau. Concretely, based on well-known neuroscience theories, SimAM proposes to optimize an energy function with a regularizer to find the importance of each neuron and derives a fast closed-form solution, with the final attention weights m_n calculated as:

$$m_n = 1/\left\{1 + \frac{1}{\sqrt{e}} \exp\left[-\frac{(x_n - \hat{\mu})^2}{4\,(\hat{\sigma}^2 + \lambda)}\right]\right\}. \tag{11}$$

Here, m_n is the attention weight allocated to the neuron with data x_n. $\hat{\mu}$ and $\hat{\sigma}$ are the same as our PdfAM-Gau, calculated within input feature channel. λ is a small constant produced by the regularizer, which is the same as our implementation of adding a small constant to $\hat{\sigma}^2$ to avoid division by zero. When $S^M = \sqrt{2\pi\sigma^2/e}$, $S^A = 0$ and $S^E = 0.5$, SimAM is a strict special case of PdfAM-Gau. Such finding provides another angle of explanation for both SimAM and PdfAM.

In addition, from the perspective of using global statistics for attention, our PdfAM provides a more efficient and rational manner. One category of existing channel attention uses global statistics to capture global information within each channel. For instance, SE (Hu, Shen, and Sun 2018) exploits first-order statistics (mean), while SRM (Lee, Kim, and Nam 2019) and MAN (Dai et al. 2021) further adopt second-order (variance) and higher-order statistics (Skewness and Kurtosis) respectively. Compared to the above works, our proposed PdfAM not only efficiently generates a 3D attention map, but also explicitly utilizes first- and second-order global statistics in a PDF-based interpretable way.

Experiments

Ablation Study

Ablation experiments are based on CIFAR classification, including 10 and 100 categories (C10 and C100). We adopt the same experimental settings as (Yang et al. 2021) , and report average Top-1 accuracy(%) over 5 trials. (Standard derivation is not reported due to space limitation.) Ablation studies are all conducted under ResNet-20 (He et al. 2016).

First, for PdfAM-Gmm, to choose proper settings for the number of Gaussian components K and the EM iteration times T, we apply PdfAM-Gmm with various K and T to ResNet-20. As shown in Table 1, one can observe that, in general, the classification accuracy is improved as K or T increases. Considering the trade-off between complexity and performance, we set $K = 2$ and $T = 3$ by default.

Second, we conduct an ablation study on the three scaling parameters of our learnable PDF-rescale design, so as to weigh their contributions. The results are provided in Table 2. One can see that, for PdfAM-Gmm, the multiplicative and additive scaling parameters (S^M and S^A) are relatively more important, while for PdfAM-Gau, the exponential parameter S^E contributes more. Besides, the entire learnable PDF-rescale design brings about 0.2% and 0.3% accuracy improvements on average for C10 and C100 respectively. Overall, each scaling parameters plays an important role in PdfAM, and the entire adoption gets the best performance.

S^M	S^A	S^E	PdfAM-Gmm		PdfAM-Gau	
			C10	C100	C10	C100
✗	✗	✗	92.49	68.89	92.45	69.09
✓	✗	✗	92.63	69.02	92.35	69.30
✗	✓	✗	92.58	69.13	92.52	69.30
✗	✗	✓	92.49	69.00	92.63	69.45
✓	✓	✗	92.62	69.23	92.51	69.30
✓	✗	✓	92.61	69.03	92.56	69.36
✗	✓	✓	92.52	69.24	92.69	69.46
✓	✓	✓	92.68	69.22	92.79	69.48

Table 2: Ablation on the three scaling parameters.

	PdfAM-Gmm		PdfAM-Gau	
	C10	C100	C10	C100
Tanh	92.38	68.82	92.39	68.91
Tanh Variant	92.47	69.06	92.50	68.98
$\mathcal{F}(x)$ (ours)	92.68	69.22	92.79	69.48

Table 3: Ablation on the mapping function.

Finally, to validate the design of our mapping function $\mathcal{F}(x)$, we conduct ablation comparisons with two alternates: 1) Tanh function $m_n = \tanh(o_n)$ adopting a positive correlation, and 2) Tanh Variant $m_n = -\tanh(o_n) + 1$ with a negative correlation. Both functions constrain $m_n \in [0, 1]$. Comparison results under PdfAM-Gmm and PdfAM-Gau are shown in Table 3. We can see that, Tanh function gets the lowest accuracy, validating the importance of negative correlation between $\mathbf{M}$ and $\mathbf{O}$. Additionally, our $\mathcal{F}(x)$ achieves higher accuracy than Tanh Variant, thus verifying the effectiveness of our simple design on the mapping function.

ImageNet Classification

Here, PdfAM is evaluated on ImageNet classification with 1000 categories, which consists of around 1.2M training and 50K validation images. Our PdfAM-Gau is compared to five attention modules, *i.e*, SE (Hu, Shen, and Sun 2018), CBAM (Woo et al. 2018), ECA (Wang et al. 2020), SRM (Lee, Kim, and Nam 2019) and SimAM (Yang et al. 2021), under multiple architectures including ResNet (He et al. 2016), ResNeXt (Xie et al. 2017), and MobileNetV2 (Sandler et al. 2018). Since we observe that the feature histograms for ImageNet classification tend to obey Gaussian distribution, we do not conduct experiments for PdfAM-Gmm here. We adopt the same settings as (Yang et al. 2021) and use its results for the compared five attention modules.

From Table 4, we can see that PdfAM-Gau achieves noticeable accuracy enhancements across multiple networks. Concretely, PdfAM-Gau obtains remarkably leading perfor-

Attention Module	ResNet-18			ResNet-34			ResNet-50		
	Top-1 Acc.	Top-5 Acc.	# Para ↑	Top-1 Acc.	Top-5 Acc.	# Para ↑	Top-1 Acc.	Top-5 Acc.	# Para ↑
Baseline	70.33 %	89.58 %	0	73.75 %	91.60 %	0	76.34 %	93.12 %	0
+ SE	71.19 %	90.21 %	0.087 M	74.32 %	91.99 %	0.157 M	77.51 %	93.74 %	2.515 M
+ CBAM	71.24 %	90.04 %	0.090 M	74.41 %	91.85 %	0.163 M	**77.63 %**	**93.88 %**	2.533 M
+ ECA	70.71 %	89.85 %	36	74.03 %	91.73 %	74	77.17 %	93.52 %	88
+ SRM	71.09 %	89.98 %	0.004 M	**74.49 %**	92.01 %	0.008 M	77.51 %	93.06 %	0.030 M
+ SimAM	71.31 %	89.88 %	0	74.46 %	92.02 %	0	77.45 %	93.66 %	0
+ PdfAM-Gau	**71.32 %**	**90.22 %**	0.006 M	74.47 %	**92.05 %**	0.011 M	77.41 %	93.63 %	0.011 M
Attention Module	ResNet-101			ResNeXt-50			MobileNetV2		
	Top-1 Acc.	Top-5 Acc.	# Para ↑	Top-1 Acc.	Top-5 Acc.	# Para ↑	Top-1 Acc.	Top-5 Acc.	# Para ↑
Baseline	77.82 %	93.85 %	0	77.47 %	93.52 %	0	71.90 %	90.51 %	0
+ SE	78.39 %	94.13 %	4.743 M	77.96 %	93.93 %	2.51 M	72.46 %	**90.85 %**	0.028 M
+ CBAM	78.57 %	**94.18 %**	4.781 M	78.06 %	**94.07 %**	2.53 M	**72.49 %**	90.78 %	0.032 M
+ ECA	78.46 %	94.12 %	171	77.74 %	93.87 %	86	72.01 %	90.46 %	59
+ SRM	78.58 %	94.15 %	0.065 M	78.04 %	93.91 %	0.030 M	72.32 %	90.70 %	0.003 M
+ SimAM	78.65 %	94.11 %	0	78.00 %	93.93 %	0	72.36 %	90.74 %	0
+ PdfAM-Gau	**78.68 %**	94.17 %	0.024 M	**78.13 %**	93.98 %	0.023M	72.39 %	90.65 %	0.021 M

Table 4: Top-1 and top-5 accuracies (%) for various networks with 5 attention modules and our PdfAM on ImageNet classification. The number of parameters added to baseline is also shown. The best and second best are bold and underlined, respectively.

mance in ResNet-18, ResNet-34, ResNet-101 and ResNeXt-50. As for ResNet-50, and MobileNetV2, our PdfAM-Gau still performs favorably against other attention modules. Besides, the number of parameters introduced by PdfAM-Gau is quite small, especially compared to SE and CBAM. Overall, PdfAM can be integrated into multiple CNN structures and boosts capability with negligible additional parameters.

Fig. 4 visualizes the attention maps in ResNet-18 trained with PdfAM-Gau. Taking three validation images as examples, Fig. 4 provides the last three attention heatmaps generated in the network, *i.e.*, the attention in the last three ResBlocks (RBs): Layer#3-RB#2, Layer#4-RB#1 and Layer#4-RB#2. The heatmaps are generated by averaging the 3D attention weights along the channel dimension. As can be seen, via gradually correcting the attention, PdfAM refines the features to focus on main objects. Both the above quantitative and qualitative results verify the effectiveness of PdfAM in enhancing the capability of various networks.

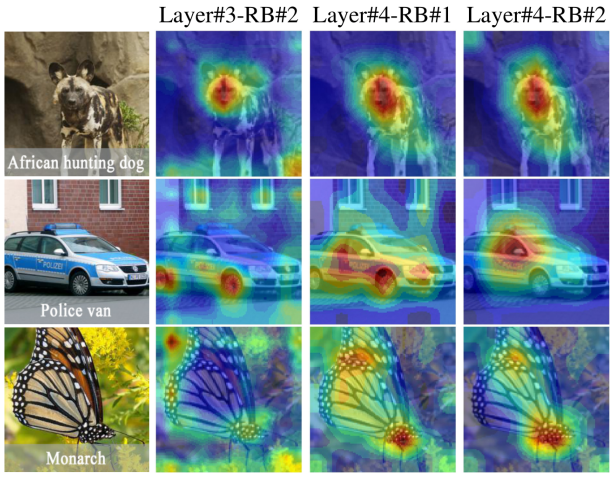


Figure 4: Visualization of attention maps in PdfAM-Gau of ResNet-18 on ImageNet classification.

MRI Reconstruction

We evaluate in a low-level vision task: MRI reconstruction (Lustig et al. 2008; Xie et al. 2022; Zhang et al. 2022), which adopts sub-sampling to accelerate imaging, and aims to reconstruct de-aliased images from sub-sampled data. We use Brain dataset (Yang et al. 2016) with 100 training and 50 testing images. Results are evaluated with PSNR and SSIM. PdfAM is compared with four attention modules, *i.e.*, SE, ECA, CBAM and PA (Zhao et al. 2020), under three reconstruction networks with RBs, including: 1) ISTA-RB, which replaces the denoising network in ISTA-Net (Zhang and Ghanem 2018) as two RBs, 2) HC-PGD (Hosseini et al. 2020), and 3) MADUN (Song, Chen, and Zhang 2021).

Results are shown in Table 5, where PdfAM and PdfAM* stand for PdfAM-Gau and PdfAM-Gmm, respectively. Note that we do not integrate PdfAM-Gmm into the heavy HC-PGD (1.11 M) and MADUN (3.02 M) to avoid high computation and memory consumption. We can see that, PdfAM consistently gets dramatic improvements across various networks and sampling ratios. In ISTA-RB, PdfAM-Gau achieves remarkably leading accuracy across three sub-sampling ratios, and PdfAM-Gmm bahaves even slightly better. In HC-PGD and MADUN, PdfAM-Gau performs favorably against other attention modules.

Fig. 5 further visualizes the attention maps for PdfAM-Gau and PdfAM-Gmm on ISTA-RB with 10% sub-sampling ratio for two example testing images. As can be seen, both two PdfAMs highlights important textural patterns to aid image reconstruction. Note that PdfAM-Gmm seems to highlight more details, which might be attributed to adopting GMM as a more precise distribution assumption based on observations about feature histograms.

Furthermore, Fig. 6 shows the PSNR convergence curves with respect to epoch numbers for various attention modules on ISTA-RB. One can observe that, PdfAM-Gmm and

Model	Ratio 5%	Ratio 10%	Ratio 15%
ISTA-RB	30.74/0.8227	34.91/0.9074	37.12/0.9348
+ SE	31.10/0.8324	35.08/0.9096	37.27/0.9361
+ ECA	31.06/0.8291	35.10/0.9096	37.25/0.9357
+ CBAM	31.19/0.8348	35.17/0.9110	37.37/0.9370
+ PA	30.77/0.8233	35.02/0.9090	37.28/0.9361
+ PdfAM	31.24/0.8378	35.15/0.9105	37.35/0.9369
+ PdfAM*	**31.26/0.8381**	**35.20/0.9113**	**37.39/0.9374**
HC-PGD	30.52/0.8122	34.75/0.9015	36.77/0.9297
+ SE	31.64/0.8380	35.16/0.9098	**37.25/0.9353**
+ ECA	31.73/0.8390	35.13/0.9090	37.24/0.9342
+ CBAM	31.65/0.8381	35.07/0.9078	37.20/0.9342
+ PA	31.57/0.8329	34.83/0.9064	36.80/0.9295
+ PdfAM	**31.91/0.8482**	**35.17/0.9122**	37.15/0.9349
MADUN	32.82/0.8757	36.17/0.9246	38.00/0.9430
+ SE	32.94/0.8775	**36.35/0.9266**	**38.12**/0.9440
+ ECA	32.91/0.8778	36.28/0.9255	38.06/0.9434
+ CBAM	32.82/0.8750	36.25/0.9255	**38.12/0.9441**
+ PA	**33.07**/0.8792	36.29/0.9256	38.05/0.9434
+ PdfAM	32.99/**0.8809**	36.28/0.9256	38.09/0.9430

Table 5: PSNR/SSIM for various networks with 4 attention modules and our PdfAM on Brain MRI dataset. The best and second best results are bold and underlined, respectively.

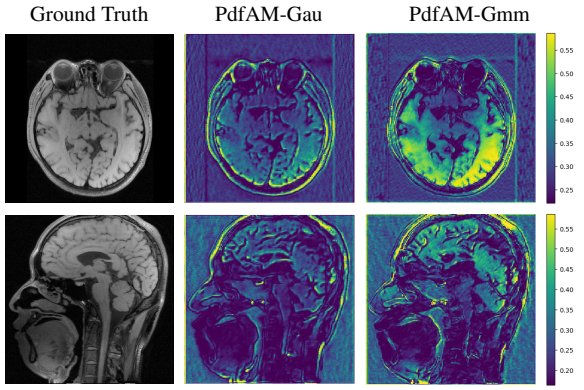


Figure 5: Visualization of attention maps in PdfAM-Gau of ISTA-RB on MRI reconstruction under ratio 10%.

PdfAM-Gau achieves the fastest and second fastest convergence speed, respectively. Overall, apart from high-level vision tasks, our PdfAM can also enhance various networks in low-level vision tasks with negligible additional parameters, and performs favorably against other popular attention modules in terms of both performance and convergence.

Analysis on Learnable PDF-rescale Design

Here, we collect multiple models from previous experiments, and analyse the learned scaling parameters under various settings, so as to validate the adaptability of our learnable PDF-rescale design. Fig. 7 presents box plots for the three learned scaling parameters S^M, S^A and S^E under eight various settings. Specifically, the first row of Fig. 7 provides box plots for different RBs in ResNet-20 with PdfAM-Gau for C10 classification task, where 'L3RB1' stands for Layer#3-RB#1, *etc*. We can see that, the multiplicative and additive scaling parameters (S^M and S^A) slightly differ

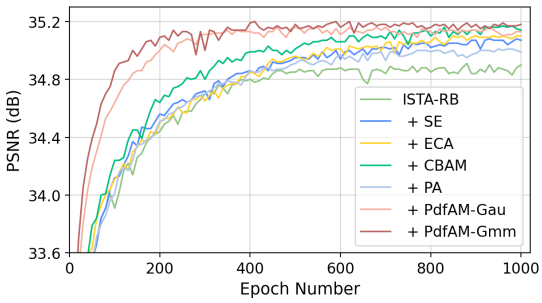


Figure 6: Convergence curves for various attention modules.

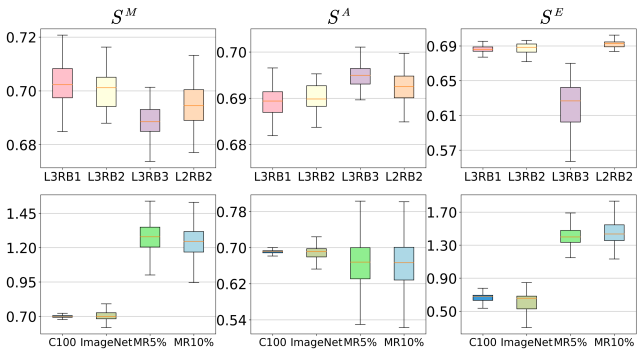


Figure 7: Boxplots of scaling parameters on various settings.

across different RBs, while the exponential scaling S^E behaves more diversity. The second row of Fig. 7 shows box plots for four networks and tasks: 1) ResNet-20 for C100 classification, 2) ResNet-18 for ImageNet classification, 3) ISTA-RB for MRI reconstruction under ratio 5%, and 4) ratio 10%, all with our PdfAM-Gau. One can see that, within tasks of the same kind (*e.g.*, two classification tasks, or two reconstruction tasks), the scaling parameters behave similarity. In contrast, dissimilarity is expressed for tasks of different kinds. Overall, the above results verify that our learnable PDF-rescale design learns adaptive scaling parameters and customized negative correlations between attention and PDF values, with respect to different blocks, networks and tasks.

Conclusion

In this paper, based on our 'less is more important' observation, we propose a novel Probability Density Function (PDF) guided Attention Module, dubbed PdfAM, to produce a 3D attention map with a rational explanation and few empirical structure designs. PdfAM first estimates the PDF values, which are then transformed to the attention weights via applying a negative correlation mapping. We further develop a learnable PDF-rescale design to transform the PDF estimation with high adaptability concerning different blocks, networks and tasks. Experiments show that PdfAM boosts the capability of various networks under both high- and low-level vision tasks, and also performs favorably against other popular attention modules in terms of performance and convergence. Theoretical analysis and more extensive applications are both worthy of further exploration in the future.

References

Avi-Aharon, M.; Arbelle, A.; and Raviv, T. R. 2020. DeepHist: Differentiable joint and color histogram layers for image-to-image translation. *arXiv preprint arXiv:2005.03995*.

Corbetta, M.; and Shulman, G. L. 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3): 201–215.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dai, T.; Lv, Y.; Chen, B.; Wang, Z.; Zhu, Z.; and Xia, S.-T. 2021. Mix-order attention networks for image restoration. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Greff, K.; Van Steenkiste, S.; and Schmidhuber, J. 2017. Neural expectation maximization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hosseini, S. A. H.; Yaman, B.; Moeller, S.; Hong, M.; and Akçakaya, M. 2020. Dense recurrent neural networks for accelerated MRI: History-cognizant unrolling of optimization algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 14(6): 1280–1291.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. 2019. Searching for MobileNetV3. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, J.; Shen, L.; Albanie, S.; Sun, G.; and Vedaldi, A. 2018. Gather-Excite: Exploiting feature context in convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Lee, H.; Kim, H.-E.; and Nam, H. 2019. SRM: A style-based recalibration module for convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Lustig, M.; Donoho, D. L.; Santos, J. M.; and Pauly, J. M. 2008. Compressed sensing MRI. *IEEE Signal Processing Magazine*, 25(2): 72–82.

Mou, C.; Zhang, J.; Fan, X.; Liu, H.; and Wang, R. 2021. COLA-Net: Collaborative attention network for image restoration. *IEEE Transactions on Multimedia*, 24: 1366–1377.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, J.; Chen, B.; and Zhang, J. 2021. Memory-augmented deep unfolding network for compressive sensing. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tan, M.; Pang, R.; and Le, Q. V. 2020. EfficientDET: Scalable and efficient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; and Hu, Q. 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. EDVR: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional block attention module. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Xie, J.; Zhang, J.; Zhang, Y.; and Ji, X. 2022. PUERT: Probabilistic Under-sampling and Explicable Reconstruction Network for CS-MRI. *IEEE Journal of Selected Topics in Signal Processing*, 16(4): 737–749.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, L.; Zhang, R.-Y.; Li, L.; and Xie, X. 2021. SimAM: A simple, parameter-free attention module for convolutional neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*.

Yang, Y.; Sun, J.; Li, H.; and Xu, Z. 2016. Deep ADMM-Net for compressive sensing MRI. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Yang, Z.; Zhu, L.; Wu, Y.; and Yang, Y. 2020. Gated channel transformation for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, J.; and Ghanem, B. 2018. ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, J.; Zhang, Z.; Xie, J.; and Zhang, Y. 2022. High-Throughput Deep Unfolding Network for Compressive Sensing MRI. *IEEE Journal of Selected Topics in Signal Processing*, 16(4): 750–761.

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a Gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7): 3142–3155.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, H.; Kong, X.; He, J.; Qiao, Y.; and Dong, C. 2020. Efficient image super-resolution using pixel attention. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Zoph, B.; and Le, Q. V. 2016. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.