Toward Efficient Image Denoising: A Lightweight Network with Retargeting Supervision Driven Knowledge Distillation

Beiji Zou[1,2], Yue Zhang[1,2], Min Wang[1,2], and Shu Liu[1,2(✉)]

[1] School of Computer Science and Engineering, Central South University, Changsha 410083, China
sliu35@csu.edu.cn

[2] Hunan Engineering Research Center of Machine Vision and Intelligent Medicine, Changsha 410083, China

Abstract. Image denoising is a fundamental but critical task. Previous works based on deep networks have made great progress, but suffer from the problem of computational overload. This paper addresses the demands by (1) a lightweight denoising network and (2) a novel knowledge distillation algorithm. The experimental results show the usefulness of the RS-KD on the proposed lightweight network and consistent gains that can be obtained on both synthetic and real-world datasets. Especially, benefiting from the retargeting supervision, our proposed distillation framework allows for arbitrary high-performance teacher networks.

Keywords: Efficient image denoising · Retargeting supervision · Arbitrary teacher · Knowledge distillation

1 Introduction

Image denoising, restoring the latent clean images from the observed noisy images, is a classic and long-lasting task in image processing. With the wide application of cameras, manufacturers such as smartphones and industrial cameras are desperately trying to upgrade their products with efficient denoising models. As a result, it is an urgent and challenging matter to implement efficient image denoising on resource-constrained devices.

In recent years, deep-learning based image denoising approaches have shown considerable success [1–4]. In contrast, to design sophisticated handcrafts, DnCNN [1] achieves impressive performance by stacking multiple convolutional layers. To take sufficient advantage of the image priors, NLRN [2] and NBNet [4] incorporates the non-local modules. Besides, the attention mechanisms have also been incorporated into the current network architecture design [3]. However, these methods are still computationally intensive and even more impractical to integrate into practice than some traditional methods. Fortunately, some techniques are proposed to overcome the computing problem, typically knowledge distillation (KD) [5]. They accomplish the model compression and computational cost reduction through the

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
N. Magnenat-Thalmann et al. (Eds.): CGI 2022, LNCS 13443, pp. 15–27, 2022.
https://doi.org/10.1007/978-3-031-23473-6_2

teaching paradigm between high-performance teacher networks usually with massive computational costs and lightweight student networks. However, these distillation methods are specifically designed for high-level tasks, and they bring no performance gains when applied to image denoising.

In this paper, we address the foregoing concerns by a lightweight network and a novel distillation algorithm with the retargeting supervision for efficient image denoising. Considering the absence of a lightweight network for RGB image denoising, we establish a lightweight deep denoising network, LUNet, by carefully considering the challenging trade-off between denoising performance and efficiency, resulting in a 14× reduction in computation cost and 10× fewer parameters. We further propose a novel distillation algorithm to improve LUNet. We first present a theoretical analysis of the image-level distillation algorithm in image denoising by modeling the distillation process as a probabilistic model. Then, we propose the retargeting supervision-driven knowledge distillation (RS-KD) algorithm to pick up the missing randomness. Specifically, we find that since the naive distillation algorithm assumes that the restored images by the teacher network are completely trustworthy, they discard randomness in the real distillation process inducing distillation failure. To overcome the deficiency, we propose the RS-KD algorithm for student networks. In contrast, to directly utilize the output of teacher networks as the supervision, we construct a multivariate Gaussian distribution with a data-adaptive variance for the prediction of teacher networks. It is tough to enable the network to learn complex distributions directly. To address this issue, we simplify the complex distribution with sampling operation. In this way, the samples shall keep moving closer to the real complex distribution as the iterations increase, hence maintaining the validity of distillation. We conduct extensive experiments to demonstrate the effectiveness of our proposed distillation methods on multiple synthetic and realistic datasets. Especially, benefiting from the retargeting supervision, our proposed distillation framework allows for arbitrary high-performance teacher networks.

In summary, our main contributions are as follows:

(1) We design a lightweight image denoising network (LUNet) for RGB image denoising, providing a baseline for the next distillation algorithm.
(2) We analyze the distillation process via a probabilistic model, theoretically uncovering the essence of the image distillation.
(3) We present a novel and flexible distillation algorithm with retargeting supervision for efficient image denoising.
(4) Extensive experiments on multiple synthetic and real-world datasets demonstrate the effectiveness of our distillation algorithm for the single image denoising task.

2 Proposed Method

2.1 Lightweight U-shaped Denoising Network: LUNet

Since a lightweight model is still absent for RGB image denoising, we present a lightweight U-shaped network, LUNet. It will also serve as a baseline model for

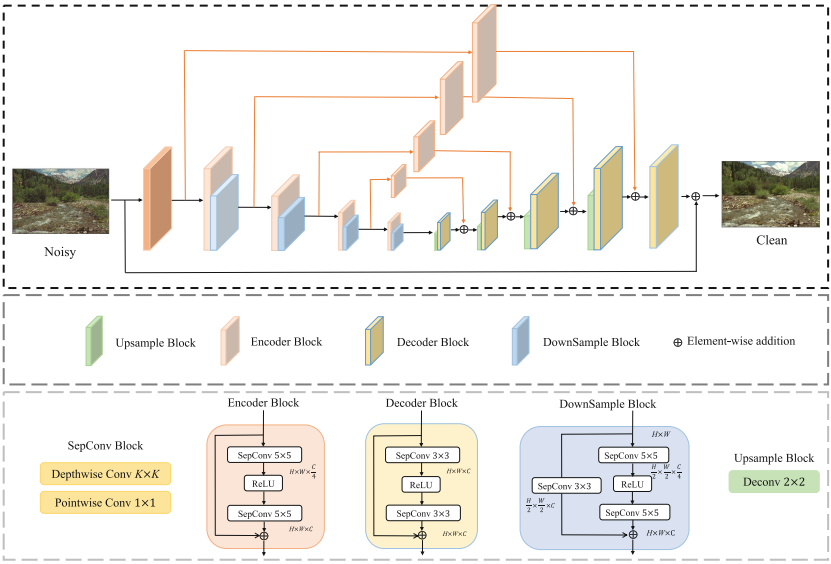

Fig. 1. Overall architecture of LUNet and structure of key building blocks. LUNet is based on UNet architecture with a depth of 4 and depth-wise separable convolutions. LUNet takes only 1.08 GMAC to process 256 × 256 inputs.

the subsequent distillation algorithm. LUNet has an encoder-decoder structure, where the input image is a noisy image and the output image is a clean image. As shown in Fig. 1, LUNet has four encoding stages and corresponding decoding stages. Given an input noisy image $\mathbf{I} \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}\times 3}$, the network first applies a 3×3 convolutional layer with step size 1 to project the input image into the feature space. Then, the subsequent encoders encode the projected features. In the latter step of each encoding stage, the downsample block subsamples the feature maps to reduce the memory consumption. Specifically, the input feature $\mathbf{X} \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}\times\mathrm{C}}$ of each encoder is downsampled to $\mathbf{X} \in \mathbb{R}^{\frac{1}{2}\mathrm{W}\times\frac{1}{2}\mathrm{H}\times 2\mathrm{C}}$. After encoding, the feature map with the smallest spatial size feature is gradually decoded to the original size. The input feature map is up-sampled by a 2×2 deconvolutional layer, which up-samples the spatial resolution by a factor of 2, and then compresses the number of channels of the feature map. The input and output of the downsampling operation are exactly opposite to the input and output of the upsampling operation so that the upsampled feature map at the decoder shall be decoded together with the input feature map from the encoder. After decoding, we obtain a feature map of the same size as the input map. Finally, the feature map is projected into the pixel space using a 3×3 convolutional kernel layer with step size 1 and output by adding the input noise image. Practically, for 256×256 inputs, LUNet only takes 1.08 GMAC, which is $14\times$ lesser than the original UNet [6].

2.2 Retargeting Supervision Driven Knowledge Distillation

Analysis. Given an image pair, $(\boldsymbol{x},\ \boldsymbol{y})$, where $\boldsymbol{x}$ is the corrupted image with noise and $\boldsymbol{y}$ is the ground truth, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{\mathrm{W}\times\mathrm{H}\times 3}$. Here, we denote the restored images of the student and teacher networks as $\hat{\boldsymbol{y}}$, $\bar{\boldsymbol{y}}$, respectively.

The naive distillation algorithm usually take the L1 loss for the teaching process, *i.e.*, $||\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}||_1$, which is based on the assumption that the teacher can perfectly reconstruct the degraded image. However, image denoising is a typical ill-posed task, which means that the optimal solution is not singular. Therefore, it is too ideal and intuitive to use only a single image as the optimal recovery. In contrast, we leave this assumption behind and turn to the probabilistic approach to explore the real principle.

From a statistical viewpoint, the images recovered by both the teacher network and the student network are typically random variables. The primary goal is to maximize the joint probability distribution:

$$P(\hat{\boldsymbol{y}}, \bar{\boldsymbol{y}}|\boldsymbol{x}) = P(\hat{\boldsymbol{y}}|\bar{\boldsymbol{y}})P(\hat{\boldsymbol{y}}|\boldsymbol{x}) \tag{1}$$

where $\bar{\boldsymbol{y}}$ serves as the given knowledge and plays a crucial role in (1). It was noted that the naive knowledge distillation algorithm supposes that we have got a perfect image restored by the teacher network. In other words, the images recovered by the teacher network subject to a probability $P(\bar{\boldsymbol{y}}|\boldsymbol{x})) = \mathbf{1}$. The desperation they strive for is merely a matter of simplification *i.e.*, $\max P(\hat{\boldsymbol{y}}|\bar{\boldsymbol{y}})$. As we have discussed above, it is unreasonable and also is the essential explanation for the failure of naive knowledge distillation.

In contrast, we consider the recovered image of the teacher network as a true random variable with probability density $P(\hat{\boldsymbol{y}}|\bar{\boldsymbol{x}})$. According to the central limit theorem, we shall model the image restored by the teach-er network as a multivariate Gaussian distribution:

$$P(\bar{\boldsymbol{y}}|\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2}$$

where $\boldsymbol{\mu}$ is the mean value, which is equivalent to the output of the teacher output. $\boldsymbol{\Sigma}$ is the variance term and it needs to be set delicately. It is worth noting that the distribution of $P(\bar{\boldsymbol{y}}|\bar{\boldsymbol{x}})$ is not required to be Gaussian. Actually, it is more close to Laplace. Fortunately, the Laplace distribution might be re-parameterized as $\bar{\boldsymbol{y}} - \boldsymbol{\Sigma} * \mathrm{sgn}(\boldsymbol{z}) * \ln(1 - 2|\boldsymbol{z}|)$, where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. For ease of illustration, we use the Gasussian distribution as a typical example.

Our goal is to allow the student network to learn from this distribution of teachers. According to [7,8], $P(\hat{\boldsymbol{y}}|\bar{\boldsymbol{y}})$ can be modeled as Boltzmann distribution:

$$p(\hat{\boldsymbol{y}}|\bar{\boldsymbol{y}}) \propto exp(-\frac{|\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}|}{kT}) \tag{3}$$

where the kT is a constant, and it is the product of Boltzmann's constant k and thermodynamic temperature T.

Therefore, we have obtained the explicit probability densities of all the demanded distributions, $P(\bar{\boldsymbol{y}}|\boldsymbol{x})$ and $P(\hat{\boldsymbol{y}}|\bar{\boldsymbol{y}})$, in Eq. 1. It would be quite preferable to optimize the above joint probability density distribution directly, however

it is still intractable to enable the network to learn the probability density function. To solve this problem, we propose a simplified method to ease the learning process. In each iteration, we approximate the conditional probability distribution $P(\bar{\boldsymbol{y}}|\boldsymbol{x})$ by some instances in it. Specifically, we first sample some variables in the distribution $P(\bar{\boldsymbol{y}}|\boldsymbol{x})$, replacing the whole distribution in $P(\hat{\boldsymbol{y}}, \bar{(\boldsymbol{y})})$. Fortunately, with the increasing number of training iterations, this approximation would be safe because this approximation keeps getting closer to the real distribution. Thus, we shall use the sampled joint probability density function, $P(\hat{\boldsymbol{y}}|\bar{\boldsymbol{y}})$, as supervision for the student network. It should be emphasized that the $\bar{\boldsymbol{y}}$ here is randomly sampled, which is intrinsically dissimilar to the one in naive KD. Finally, we apply the negative log-likelihood as our optimization goal:

$$\min \mathbb{E}_{\bar{\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[\|\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\|_1] \tag{4}$$

We shall denote Eq. 4 as $\mathcal{L}_{distil}$ for notation convenience. In addition, according to the Jensen's inequality, we have $\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$ for any convex function $f(\cdot)$. Because the p-norm is convex and thus we shall get

$$\mathbb{E}_{\bar{y}}[\|\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\|_1] \geq \|\mathbb{E}_{\bar{y}}[\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}]\|_1 = \|\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}\|_1 \tag{5}$$

which suggests that our optimization objective is an upper bound for naive knowledge distillation. In other words, naive knowledge distillation suffers from a degraded problem. It loses the randomness during training and the supervision of the student network is unfortunately restricted to the output of the teacher network. Therefore, to solve this problem, we take the sampled joint probability distribution described in the last section as a new learning objective for the student network.

The Retargeting Supervision. The analysis demonstrates an probabilistic model for Knowledge distillation. The prior distribution $P(\bar{\boldsymbol{y}})$ determines the supervision quality of the students' network, especially the variance term. An intuitive idea is to set the $\boldsymbol{\Sigma}$ as a small constant value, kI, only for introducing randomness. I is the identity matrix. However, such an operation is equivalent to adding a small random noise $\boldsymbol{z} \sim \mathcal{N}(0, kI)$ to the output of the teacher network, which leads to worse results. In contrast, we introduce retargeting supervision to help the student network learning efficiently. Especially, we propose a data-adaptive $\boldsymbol{\Sigma}$ to acquire randomness.

$$\boldsymbol{\Sigma} = |\bar{\boldsymbol{y}} - \boldsymbol{y}| \tag{6}$$

where $|\cdot|$ refers to element-wise absolute function. In particular, as shown in Fig. 2. We introduce an auxiliary network branch attached to the student network to learn the variance. This would allow the student network to capture the distribution completely. Especially, Eq. 6 may introduce a denoising bias. However, the sampling of multivariate Gaussian distributions can alleviate the problem to some extent. We use L1 loss to enable the auxiliary network to capture the data-adaptive variance

$$\mathcal{L}_{aux} = \||\bar{\boldsymbol{y}} - \boldsymbol{y}| - \boldsymbol{\sigma}\|_1 \tag{7}$$

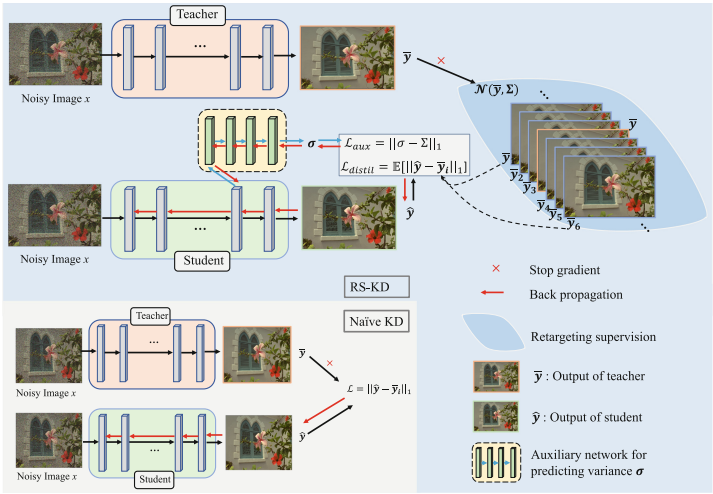

Fig. 2. The pipeline of our proposed Retargeting Supervisionis driven Knowledge Distillation (RS-KD) framework. During training, RS-KD distills the student network via retargeting supervision with the assistance of an auxiliary network. For testing, the only required network is the original student network, leaving no auxiliary network.

where $\boldsymbol{\sigma}$ is the output of the auxiliary network. Incorporating Eq. 4 and Eq. 7, the final loss function is formulated as:

$$\min \ \mathcal{L}_{total} = \mathcal{L}_{distil} + \alpha \mathcal{L}_{aux} \tag{8}$$

where α is the hyper-parameter to balance different aspects of loss. As a result, the multivariate Gaussian distribution composed by the output of the teacher network and the data-dependent variance term forms the new retargeting supervision.

Note that, in Eq. 8, we do not directly apply the given ground truth as the supervision. This brings an advantage in that we do not have to carefully weigh the ground truth and the output of the teacher while training the student network. In addition, since $|\bar{\boldsymbol{y}} - \boldsymbol{y}|$ contains the information passed by both ground truth and the teacher, the supervision from teachers dominates Eq. 8. This enables an arbitrary high-performance teacher network to be selected in our distillation framework. This paper uses MIRNet [9] as the teacher. After training, we only need to keep the original student network, LUNet, for testing. Therefore, we do not introduce additional computation in the student network.

3 Experimental Results

3.1 Experimental Settings

Implementation Details. We train all the models with the Adam optimizer with momentum terms (0.9, 0.999). During training, we crop images into 256×256 patches and training networks with a batch size of 32 for $400,000$ iterations. The initial learning rate is $2{\times}10^{-4}$, and it steps at $240,000$ and $360,000$ iterations with scale 0.1. We apply random rotation and flipping to augment the training data. For distillation experiments, we do not perform a pretraining process on LUNet. We perform all experiments on Nvidia 2080Ti GPUs. Specifically, we take PSNR and SSIM as the metrics.

Datasets. For synthetic datasets, the training dataset are consisted of 432 images from BSD [10], 400 images from the validation set of ImageNet [11] and 4,774 images from the Waterloo dataset [12]. We follow the same setting in [13] to generate non-i.i.d Gaussian noise as following,

$$\boldsymbol{n} = \boldsymbol{M} * \boldsymbol{n}^1, \boldsymbol{n}^1_{ij} \sim \mathcal{N}(0, 1) \tag{9}$$

where $\boldsymbol{M}$ is a spatially variant mask with the same size as the clean image. In this paper, we choose a Gaussian window function with a variance being 10. We utilize the above generated data as training data and test models with regular Gaussian noise. While testing, we consider three noise levels, namely $sigma = 15, 25, 50$. Then we evaluate on Set5 [14], LIVE1 [15] and BSD68 [16].

For real-world datasets, we conduct experiments on Smartphone Image Denoising Dataset (SIDD) [17]. SIDD is composed of about 30,000 noisy images from 10 scenes under different lighting conditions. It employs five representative smartphone cameras and generates their gro-und truth images through a systematic procedure. SIDD is available to measure the denoising performance of smartphone cameras. As a benchmark, SIDD splits 1280 color images for validation.

3.2 Quantitative and Qualitative Results

Comparisons of Efficient Denoising. We report the quantitative results on both synthetic and real-world datasets. The real-world denoising performance on SIDD presents in Table 1. We also compare the computational complexity of the latest methods, including the model parameters as well as the practical running time. FAN is chosen as the efficient baseline model. It is worth noting that our LUNet has a higher performance, nearly 2 dB. Our LUNet effectively weighs computational cost and performance. In comparison with the original UNet [6], LUNet has comparable performance with less computation.

Table 1. Quantitative results on real-world dataset SIDD [17]. LUNet† is the LUNet trained in the standard fashion. LUNet* means the LUNet is trained with our distillation scheme. The time tests perform on all performed on a single Nvidia 2080Ti GPU.

Method	MAC(G) ↓	Param(G) ↓	Runtime(ms) ↓	PSNR ↑	SSIM ↑
DnCNN [1]	68.15	0.56	21.69	23.66	0.583
BM3D [18]	-	-	41.56	25.65	0.685
WNNM [19]	-	-	-	25.78	0.809
NLM [20]	-	-	-	26.76	0.699
KSVD [21]	-	-	-	26.88	0.842
CBDNet [22]	40.38	4.37	80.76	30.78	0.754
RIDNet [23]	40.34	1.49	98.13	38.71	0.914
VDN [13]	41.88	7.70	99.00	39.28	0.909
DANet+ [24]	14.85	9.15	65.62	39.47	0.918
MIRNet [9]	786.43	31.79	192.61	39.72	0.959
MPRNet [3]	588.14	15.74	180.00	39.71	0.958
NBNet [4]	354.80	13.3	37.44	**39.75**	**0.973**
UNet [6]	14.85	9.15	4.1	36.71	0.913
FAN [25]	2.67	0.26	3.6	34.59	0.901
LUNet†	1.08	0.95	3.6	36.39	0.912
LUNet*	**1.08**	**0.95**	**3.6**	36.56	0.914

Table 2. Quantitative results PSNR on synthetic datasets with i.i.d Gaussian noise. LUNet† is the LUNet trained in the standard fashion. LUNet* means the LUNet is trained with our distillation scheme.

Dataset	*sigma*	CBM3D [26]	WNNM [19]	DnCNN [1]	MemNet [27]	FFDNet [28]	UDNet [29]	VDN [13]	**NBNet** [4]	UNet [6]	FAN [25]	LUNet† [25]	LUNet* ours
Set5 [14]	15	33.42	32.92	34.04	34.18	34.30	34.19	34.34	**34.64**	32.30	30.28	30.27	31.52
	25	30.92	30.61	31.88	31.98	32.10	31.82	32.24	**32.51**	27.43	26.24	26.34	28.40
	50	28.16	27.58	28.95	29.10	29.25	28.87	29.47	**29.70**	23.60	21.06	21.02	21.50
LIVE1 [15]	15	32.85	31.70	33.72	33.84	33.96	33.74	33.94	**34.25**	31.12	30.87	30.93	31.42
	25	30.05	29.15	31.23	31.26	31.37	31.09	31.50	**31.73**	27.38	27.50	27.43	27.85
	50	26.98	26.07	27.95	27.99	28.10	27.82	28.36	**28.55**	23.44	20.90	20.99	21.85
BSD68 [16]	15	32.67	31.27	33.87	33.76	33.85	33.76	33.90	**34.15**	31.74	31.16	31.07	31.97
	25	29.83	28.62	31.22	31.17	31.21	31.02	31.35	**31.54**	28.62	27.15	27.27	27.79
	50	26.81	25.86	27.91	27.91	27.95	27.76	28.19	**28.35**	22.64	21.85	21.68	22.48
Average	-	30.18	29.30	31.19	31.24	31.34	31.11	31.47	**31.71**	27.58	26.34	26.33	27.19

The Effectiveness of RS-KD. Our distillation algorithm can enhance the performance of LUNet on both synthetic and real-world datasets. As shown in Table 1, the distilled LUNet has a higher SSIM value compared to the original UNet, which means that the proposed optimization introduces a trade-off between quality and speed. We suppose that the introduced randomness allows

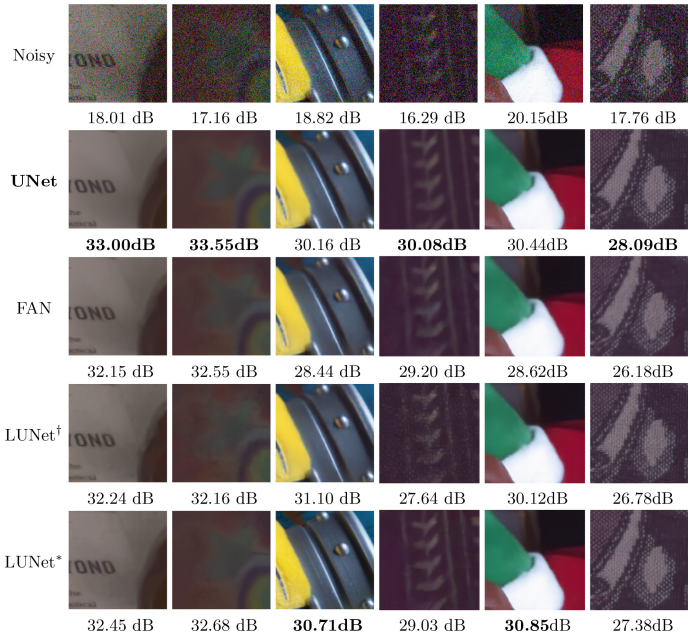

Fig. 3. Qualitative results on SIDD [17]. Our distillation algorithm enables LUNet to produce visual-pleasing denoised images. Please enlarge the screen for more detailed information.

Table 3. Ablation study of the KD methods. Naive KD is shown in the bottom left in Fig. 2. RS means our proposed retargeting supervision driven KD.

KD methods	Loss	PSNR	SSIM
Naive	L1	36.41(+0.02)	0.9120(+0.0001)
RS	Eq. 8	36.56(**+0.17**)	0.9141(**+0.0022**)

the student network to capture more probable textures in the learning process. This further demonstrates the effectiveness of our proposed distillation algorithm. As shown in Table 2, our RS-KD algorithm an average gain of 0.85 dB to LUNet for synthetic experiments on Gaussian noise. This further demonstrates the superior performance of our method for simple noise.

Qualitative Results. We show the visual comparisons in Fig. 3. Compared with the results of FAN, LUNet can produce favorable recovery results. Moreover, our distillation algorithm makes the original LUNet more effective in recovering texture information without color distortion.

3.3 Ablation Studies

We present the ablation studies to analyze the contribution of each component of our model. The evaluations are performed on the intractable SIDD dataset.

Comparison with the KD Methods. As shown in Table 3, compared to the naive KD supervision, our retargeting supervision drive KD design can provide effective enhancement to the baseline. This phenomenon is consistent with that in the classification task, meaning that naive KD do not convey the knowledge of teachers properly.

The Hyperparameter α. We explore the importance of the information in the retargeting supervision. As shown in Table 4, when $\alpha = 0.0001$, the performance is optimal. In particular, when $\alpha = 0$, *i.e.*, using only the naive KD for the distillation supervision, the results are also relatively poor. This shows the necessity of the existence of the auxiliary network.

The Variance Term. We further explore the different choices of the variance item. We generate them in two ways, *i.e.*, $|\boldsymbol{y} - \bar{\boldsymbol{y}}|$ and $|\boldsymbol{y} - \hat{\bar{\boldsymbol{y}}}|$ respectively. As shown in Table 5, $|\boldsymbol{y} - \bar{\boldsymbol{y}}|$ performs better for PSNR. We suppose that this approach enables the student network to identify the shortcomings of the teacher network and thereby learn the teacher network thoroughly. Besides, both of them have higher SSIM than that of the original UNet. This may means that our approach can effectively leverage the knowledge of the teacher network to enhance the performance of the student network.

Table 4. Ablation study of the hyperparameter α.

α	0.01	0.005	0.001	0.0001	0.00001	0
PSNR	36.39	36.5	36.51	**36.56**	36.46	36.45
SSIM	0.9097	0.9119	0.9137	**0.9141**	0.9109	0.9119

Table 5. Ablation study of the variance item.

Variance	α	PSNR	SSIM
$\boldsymbol{\Sigma} = \lvert\boldsymbol{y} - \bar{\boldsymbol{y}}\rvert$	0.0001	**36.56**	**0.9141**
$\boldsymbol{\Sigma} = \lvert\boldsymbol{y} - \hat{\bar{\boldsymbol{y}}}\rvert$	0.0001	36.50	0.9138

The Teacher Network. We explore the contribution of different teacher networks in the proposed distillation framework as shown in Table 6. In contrast to previous studies [30,31], our distillation algorithm has no restrictions on the teacher network, offering considerable flexibility.

Table 6. Ablation study of the teacher model.

Teacher model	Student model	Baseline	Distilled
VDN [13]	LUNet	36.39	36.52(+**0.13**)
DANet+ [24]	LUNet	36.39	36.51(+**0.12**)
MPRNet [3]	LUNet	36.39	36.49(+**0.10**)
MIRNet [9]	LUNet	36.39	36.56(+**0.17**)

4 Conclusions

In this paper, we make contributions for efficient image denoising from two aspects, efficient network structure and distillation algorithm respectively. We first design a lightweight U-shaped network, LUNet, which has 14× lower computation cost and 10× fewer parameters than the original UNet. Then, we propose a novel distillation algorithm to improve the performance of LUNet. Finally, supported by the RS-KD algorithm, LUNet accomplishes efficient image denoising. We expect that our work will encourage further research on the knowledge distillation algorithms for other low-level vision tasks.

Acknowledgements. This work was supported by the National Science and Technology Major Project under Grant 2018AAA0102100, and the National Natural Science Foundation of China under Grant 61902435. We are grateful for resources from the High Performance Computing Center of Central South University.

References

1. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017)
2. Liu, D., Wen, B., Fan, Y., Loy, C.C., Huang, T.S.: Non-local recurrent network for image restoration. arXiv preprint arXiv:1806.02919 (2018)
3. Zamir, S.W., et al.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14821–14831 (2021)
4. Cheng, S., Wang, Y., Huang, H., Liu, D., Fan, H., Liu, S.: NBNet: noise basis learning for image denoising with subspace projection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4896–4906 (2021)

5. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
6. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
7. Bruna, J., Sprechmann, P., LeCun, Y.: Super-resolution with deep convolutional sufficient statistics. arXiv preprint arXiv:1511.05666 (2015)
8. He, X., Cheng, J.: Revisiting L1 loss in super-resolution: a probabilistic view and beyond. arXiv preprint arXiv:2201.10084 (2022)
9. Zamir, S.W., et al.: Learning enriched features for real image restoration and enhancement. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12370, pp. 492–511. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58595-2_30
10. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 898–916 (2010)
11. Deng, J., Russakovsky, O., Krause, J., Bernstein, M.S., Berg, A., Fei-Fei, L.: Scalable multi-label annotation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3099–3102 (2014)
12. Ma, K., et al.: Waterloo exploration database: new challenges for image quality assessment models. IEEE Trans. Image Process. **26**(2), 1004–1016 (2016)
13. Yue, Z., Yong, H., Zhao, Q., Zhang, L., Meng, D.: Variational denoising network: Toward blind noise modeling and removal. arXiv preprint arXiv:1908.11314 (2019)
14. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
15. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans. Image Process. **15**(11), 3440–3451 (2006)
16. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision, ICCV 2001, vol. 2, pp. 416–423. IEEE (2001)
17. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1692–1700 (2018)
18. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising with block-matching and 3D filtering. In: Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning, vol. 6064, p. 606414 (2006)
19. Gu, S., Zhang, L., Zuo, W., Feng, X.: Weighted nuclear norm minimization with application to image denoising. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2862–2869 (2014)
20. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 2, pp. 60–65. IEEE (2005)
21. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11), 4311–4322 (2006)
22. Shi, G., Zifei, Y., Kai, Z., Wangmeng, Z., Lei, Z.: Toward convolutional blind denoising of real photographs. arXiv preprint arXiv:1807.04686 (2018)

23. Anwar, S., Barnes, N.: Real image denoising with feature attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3155–3164 (2019)
24. Yue, Z., Zhao, Q., Zhang, L., Meng, D.: Dual adversarial network: toward real-world noise removal and noise generation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12355, pp. 41–58. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58607-2_3
25. Young, L.D., et al.: Feature-align network with knowledge distillation for efficient denoising. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 709–718 (2022)
26. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1712–1722 (2019)
27. Tai, Y., Yang, J., Liu, X., Xu, C.: Memnet: a persistent memory network for image restoration. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4539–4547 (2017)
28. Zhang, K., Zuo, W., Zhang, L.: FFDNet: toward a fast and flexible solution for CNN-based image denoising. IEEE Trans. Image Process. **27**(9), 4608–4622 (2018)
29. Lefkimmiatis, S.: Universal denoising networks: a novel CNN architecture for image denoising. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3204–3213 (2018)
30. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1365–1374 (2019)
31. Guo, Q., et al.: Online knowledge distillation via collaborative learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11020–11029 (2020)