This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SwinIA: Self-Supervised Blind-Spot Image Denoising without Convolutions

Mikhail Papkov*[1] mikhail.papkov@ut.ee

Pavel Chizhov*[1,2] pavel.chizhov@thws.de

Leopold Parts[1] leopold.parts@ut.ee

[1]Institute of Computer Science, University of Tartu, Estonia
[2]CAIRO, Technical University of Applied Sciences Würzburg-Schweinfurt, Germany

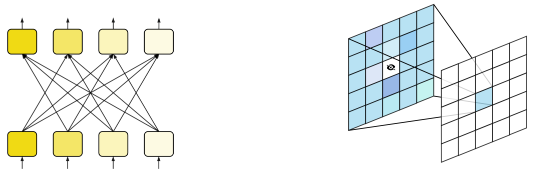
Abstract

Self-supervised image denoising implies restoring the signal from a noisy image without access to the ground truth. State-of-the-art solutions for this task rely on predicting masked pixels with a fully-convolutional neural network. This most often requires multiple forward passes, information about the noise model, or intricate regularization functions. In this paper, we propose a Swin Transformer-based Image Autoencoder (SwinIA), the first fully-transformer architecture for self-supervised denoising. The flexibility of the attention mechanism helps to fulfill the blind-spot property that convolutional counterparts normally approximate. SwinIA can be trained end-to-end with a simple mean squared error loss without masking and does not require any prior knowledge about clean data or noise distribution. Simple to use, SwinIA establishes the state of the art on several common benchmarks.

1. Introduction

Image denoising methods aim to reconstruct true signal given corrupted input. The corruption depends on the camera sensor, signal processor, and other aspects of the image acquisition procedure, and can take various forms such as Gaussian noise, Poisson noise, salt-and-pepper noise, *etc.* Noise levels also vary with illumination and exposure, and some amount is always present in any image. This makes denoising an integral part of image processing pipelines.

As in other fields in computer vision, deep learning solutions have superseded classical methods [3,5,22] for denoising. However, when approached naively, neural networks require huge amounts of paired noisy and clean images for supervised learning. Collecting such a dataset is usually impracticable. Lehtinen *et al.* [18] proposed Noise2Noise showing that supervision with independently corrupted data is equivalent to supervision with clean data. However, this approach still requires collecting multiple image copies, which may not be present in existing datasets.



(a) Each token representation is based on the rest of the sequence, excluding the token itself.

(b) Each pixel representation is based on all other pixels, making its own value a blind spot.

Figure 1. Self-unaware autoencoding in text and images.

Self-supervised denoising avoids the demand for paired data since it learns only from noisy images. Developing Noise2Noise ideas, Noisier2Noise [24] and Recorrupted2Recorrupted [26] applied additional noise on training images to emulate the strongly supervised Noise2Noise scenario. This can be accomplished by assuming a prior knowledge of the noise model or test-time aggregation. Self2Self [28] proposed training a network with Bernoulli input dropout and inference by ensembling multiple outputs. Neighbor2Neighbor [14] sub-sampled the input image and treated the result as independently corrupted copies.

Noise2Void [15] and Noise2Self [2] introduced a different approach to self-supervised denoising — a blind-spot network (BSN). This type of network reconstructs a pixel from its neighborhood, assuming spatially independent zero-mean noise. It is practically difficult to dissect the continuous receptive field of convolutional neural networks (CNN), so BSN is usually emulated by a masking procedure to hide a small portion of pixels by substitution or random noise and learn solely from them. However, learning from a few data points per image slows down convergence, and different masking approaches may produce drastically different results [37]. Laine *et al.* [16] proposed to restrict blindness by constructing four denoising branches with unidirectional receptive fields. In practice, this was achieved by passing four differently rotated input copies through the network. Later, Honzátko *et al.* [13]

*equal contribution

and Wu *et al.* [36] adopted dilated convolutions to create a true BSN which does not require masking. We further discuss these methods below. Subsequent works abandoned the idea of strict pixel blindness and adopted multiple forward passes through the network. Noise2Same [37] and its modification Noise2Info [33] make two forward passes (one with a random mask, one without) and regularize the training with invariance loss in masked pixel locations. DCD-Net [43] combines Noise2Noise and Noise2Void in an iterative denoise-corrupt-denoise pipeline. Blind2Unblind [35] utilizes global masking and combines the denoising results from passing 17 image copies[1] through the network. While superior in performance, this approach is time-consuming, requires tuning multiple hyperparameters for each dataset, and exhibits unstable training (Tab. 2).

Most recently, vision transformers started to outperform CNNs across a variety of benchmarks, including supervised denoising. SwinIR [19], based on Swin Transformer [21], achieved state-of-the-art results in JPEG compression artifact reduction. Uformer [34] and Restormer [38] concurrently excelled in camera noise removal. Evolution of vision transformers closely followed their path in natural language processing [8]. Bao *et al.* [1] introduced BERT-style pre-training for image datasets, and He *et al.* [11] showed that transformers can confidently reconstruct up to 95% of hidden data in a masked autoencoder fashion. These results hint that we could use transformers to design blind-spot self-supervised denoising models.

In this paper, we propose **SwinIA** — Swin Transformer-based Image Autoencoder, the first fully transformer-based architecture for self-supervised image denoising. SwinIA does not require any prior knowledge of noise distribution. It also does not have access to clean images, either through pre-training or knowledge distillation. Neither does it use input masking, auxiliary regularization losses, or multiple forward passes. Instead, SwinIA is trained as a plain autoencoder by minimizing the mean squared error (MSE) computed over the full image. To our knowledge, it is the first precise implementation of the original BSN idea. We rigorously test our SwinIA method on a variety of synthetic and real-world datasets and demonstrate its competitiveness against state-of-the-art self-supervised denoising solutions.

2. Related work

We further describe the foundational ideas of blind-spot networks and denoising vision transformers that our model is related to. We introduce their properties and usual training schemes and give the context for our advances.

The blind-spot property is usually achieved by masking [15] or multiple forward passes through the network [16]. These techniques overcome the continuous receptive field issue of CNNs. It is also possible to maintain blindness with increasingly dilating convolutions. Honzátko *et al.* [13] proposed a blind-spot convolutional layer with a virtual "hole" in the kernel center. Their work followed the training setup of Laine *et al.* [16] and demonstrated similar performance on sRGB datasets. The main limitation of both approaches is the assumption that the noise distribution is known and the predictions are refined with probabilistic post-processing (posterior mean estimation, or PME). Finally, Wu *et al.* [36] also utilized a dilated blind-spot network (DBSN) in a multi-stage pipeline with clean images provided via knowledge distillation.

Transformers are widely used for image restoration in the supervised setting [19, 34, 38], but rarely for self-supervised denoising. Zhang *et al.* [40] proposed a Context-aware Denoise Transformer (CADT) based on SwinIR [19] and masking scheme of Blind2Unblind [35]. They used Swin Transformer [21] blocks in the global branch of the network and trained them with patch embeddings, not pixel embeddings. However, they argued that a transformer alone is not suitable for the task and thus complemented it with convolutional local feature extraction. CADT used sixteen masked forward passes from Blind2Unblind in both training and inference. Liu *et al.* [20] built a single-image denoising transformer (DnT) from self-attention blocks interleaved with convolutional layers. This architecture was not tested for self-supervised denoising on larger datasets.

3. Design

BSN was proposed many years ago [2, 15], but to date, there is no implementation strictly adhering to the original idea. Existing solutions use masking [2, 15, 17, 35], assume known noise distributions [13, 16], or learn from clean data through knowledge distillation [36]. Thus, creating an assumption-free BSN that is trained end-to-end as an autoencoder with hyperparameter-free MSE loss between input and output remains an open challenge. This learning objective can be formulated as follows:

$$\mathcal{L}(f|\theta) = \mathbb{E}_x \| f(x|\theta) - x \|^2. \quad (1)$$

Here x is a noisy input image and f is a model with a set of parameters θ. We hypothesize that transformers could be suitable for this task because it is possible to control pixel interaction through the attention mechanism.

Shin *et al.* [31] introduced the idea of self-unaware text autoencoding using transformers (T-TA). They modified the transformer model so that each token representation is built based on all the other tokens, except itself, as in Fig. 1a. T-TA builds text representations in one iteration without access to a token's own value, as opposed to the masked language modeling objective of BERT [7], where the tokens are processed one at a time. We transfer this idea to the im-

[1] In the official repository, authors set mask window width to 4, creating 16 masked copies in addition to the unmasked image.

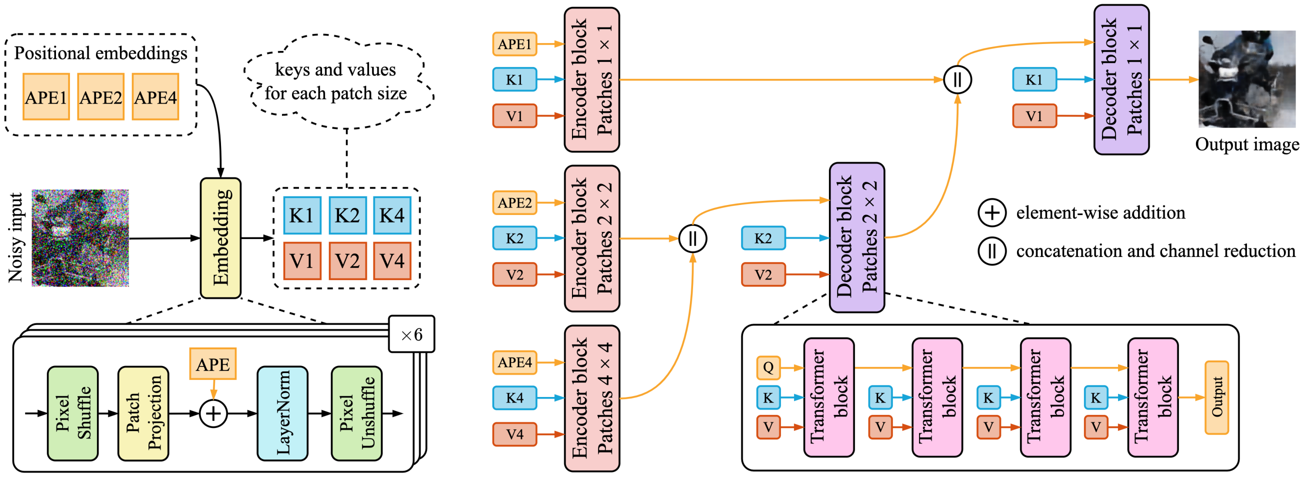


Figure 2. SwinIA model. Multiscale positional embeddings act as queries for the encoder (three parallel blocks) and are added to patch embeddings to create constant keys and values. The decoder (two remaining blocks) fuses the extracted features into the final output image. Encoder and decoder blocks have an identical structure and consist of four transformer blocks with cyclically shifted window attention.

age domain to create a vision transformer autoencoder with self-unaware pixel-level tokens (Fig. 1b).

In order to efficiently exploit a transformer-based model at the pixel level for high-dimensional image data, we need to use local attention. Swin Transformer [21] is a powerful multi-purpose vision model, which uses the window multi-head self-attention (MSA) mechanism that restricts self-attention to windows of fixed size. The windows are shifted from block to block to avoid bordering artifacts and spread the field of view. Swin Transformer was already efficiently utilized at the pixel level in SwinIR [19], where individual pixels were embedded into tokens.

Combining the ideas above, we formulate a list of requirements to design a blind-spot transformer.

Pixel level. The network should process images at single-pixel level. The absence of pixel processing would impede the understanding of random pixel-level noise and the reconstruction of high-frequency details.

Self-unawareness. At any stage of the blind-spot network, individual pixels should not have access to their own state on the previous levels. This will prevent it from learning an identity function by minimizing MSE loss.

Unblinding. Blind-spot training inevitably leads to information loss from the most significant source — the actual value of the pixel. Therefore, it is important to unhide these values during inference without disturbing the learned modality of the model.

Long-range interactions. In our setting, downsampling pooling operations in the encoder would disrupt input isolation by mixing together feature vectors of individual pixels. Therefore, we need a downsampling operation that enables attention between groups of pixels and at the same time, maintains the independence of each pixel.

4. Methods

4.1. Input embedding

In contrast with the explicit pixel-level processing in SwinIR [19], we propose to operate on shuffled square patches of size $p \times p$ pixels (see the left part of Fig. 2). An example of pixel shuffle is illustrated in Fig. 3. The queries are set to learnable absolute positional embeddings (APE), separate for each patch size. The input image is projected into keys and values only once for each patch size $p \in \{1, 2, 4\}$ as follows:

$$K_p = h_p^{-1} \left(\mathrm{LayerNorm} \left(h_p \left(X \right) W_{kp} + \mathrm{APE}_p \right) \right), \quad (2)$$

$$V_p = h_p^{-1} \left(\mathrm{LayerNorm} \left(h_p \left(X \right) W_{vp} + \mathrm{APE}_p \right) \right). \quad (3)$$

Here W_{kp}, W_{vp} are linear projection parameter matrices, and h_p is an operation of shuffling into patches of size $p \times p$. Maintaining keys and values intact throughout the architecture is essential for self-unawareness [31].

4.2. SwinIA model

SwinIA is an encoder-decoder model consisting of three encoder and two decoder blocks, the architecture is presented in the left part of Fig. 2. Three encoder blocks encode the inputs for each of the three patch sizes $p \in \{1, 2, 4\}$ and are computed separately, each using a corresponding set of positional embeddings (APE) as queries. The encoded representations are fused up through the decoders of corresponding patch sizes with skip connections by concatenation and linear projection:

$$\mathrm{shortcut}(\mathrm{X}_1, \mathrm{X}_2) = \left(X_1 \parallel X_2 \right) W + b. \quad (4)$$

Each encoder/decoder block consists of four transformer

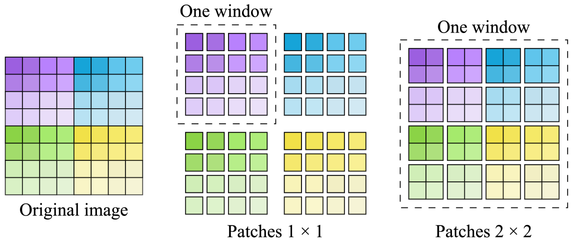


Figure 3. Pixel shuffle example of an image of size 8×8 with window size 4 into patches of sizes 1×1 and 2×2.

blocks. As in Swin Transformer [21], the attention is computed in square windows of fixed size. To avoid bordering artifacts, the images have to be diagonally shifted in every second block. The shift size along one dimension is equal to half of a window size.

4.3. Transformer block

Transformer blocks in SwinIA are comprised of multi-head self-attention (MSA) and multi-layer perceptron (MLP) submodules with pre-normalization and additive shortcuts, as in Fig. 4. Since we compute patch-level attention, the inputs are first shuffled into patches as 2D matrices $\mathbb{R}^{p^2 \times d}$, where p is patch size along one dimension and d is the embedding dimensionality in the model.

SwinIA transformer block utilizes a window multi-head self-attention (MSA) mechanism with a masked main diagonal of the attention matrix. The masking is performed by subtracting a large constant from the main diagonal of the dot-product, therefore the SoftMax values there become infinitesimal:

$$\mathrm{MSA}\left(Q, K, V\right) = \mathrm{SoftMax}\left(\frac{QK^T - I_n \cdot 10^9}{\sqrt{d_h} \cdot p}\right) V. \quad (5)$$

Here d_h is embedding dimensionality per attention head, p is the current patch size, and I_n is an identity matrix where n is the attention sequence length.

The MSA is followed by a layer normalization and pixel unshuffle operation. The unshuffled outputs are fed into a two-layer MLP with 4 times increased hidden dimensionality and GELU activations [12].

4.4. Architecture justification

In this part, we will analyze how the proposed architecture addresses the design requirements formulated in Sec. 3.

SwinIA operates on a **pixel level**, because its smallest patch size is 1×1. Therefore, our model can capture valuable pixel-to-pixel interactions.

Self-unawareness in SwinIA is ensured by a combination of the diagonal attention mask and input isolation. The diagonal mask restricts the attention so that none of the pixels have access to their value from the previous layer. However, this restriction could be bypassed by a simple permutation learning within two consequent transformer blocks. Input isolation makes it impossible: in every dot-product of the attention, only one of the components is aware of its surroundings, as keys and values are projected with a single-patch field of view and frozen from the beginning.

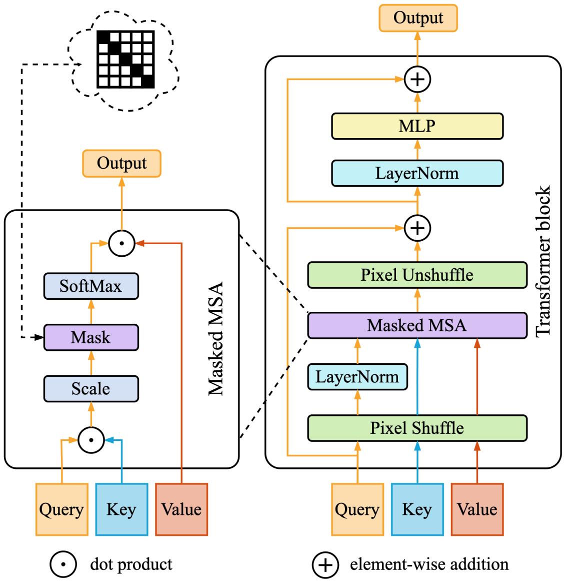


Figure 4. SwinIA transformer block architecture. MSA and MLP are preceded by layer normalization and complemented with a shortcut by addition. Only queries are normalized before the MSA because keys and values are normalized upon creation. The attention is performed between shuffled patches. The attention matrix is diagonally masked to maintain pixel self-unawareness.

Additionally, unlike in standard encoder-decoder architectures, encoder blocks in SwinIA run in parallel. Since the patch size increases in the encoder flow, bigger patches would consist of context-aware smaller patches from the previous level. As a result, the noise would leak, and the model would learn a simple identity function (see Tab. 5).

Unblinding is achieved during inference by removing the diagonal mask and thus applying the complete set of attention weights to values:

$$\mathrm{MSA}_{\mathrm{eval}}\left(Q, K, V\right) = \mathrm{SoftMax}\left(\frac{QK^T}{\sqrt{d_h} \cdot p}\right) V. \quad (6)$$

Intuitively, this allows to propagate pixel's own signal by iteratively re-weighting it with the most similar neighbors throughout the network. Since attention matrix does not contain learnable components and simply reflects (self-)similarity of pixel embeddings, unmasking the main diagonal maintains the learned modality and does not disrupt the forward pass.

	Method	train	test	Gaussian $\sigma = 15$		Gaussian $\sigma = 25$		Gaussian $\sigma = 50$	
				BSD68	Set12	BSD68	Set12	BSD68	Set12
Supervised	Noise2Clean	1	1	31.58/0.889	32.60/0.899	29.02/0.822	30.07/0.852	26.08/0.715	26.88/0.777
	SwinIR	1	1	31.97/ —	33.36/ —	29.50/ —	31.01/ —	26.58/ —	27.91/ —
Self-sup.	R2R	1	50	31.54/0.885	**32.54/0.897**	28.99/0.818	30.06/0.851	26.02/0.705	26.86/0.771
	Noise2Self[†]	1	1	30.63/0.843	29.88/0.840	28.88/0.789	28.37/0.799	26.19/0.664	25.56/0.692
	Noise2Same[†]	2	1	30.85/0.850	30.02/0.849	29.13/0.800	28.54/0.814	26.75/0.714	26.13/0.744
	Blind2Unblind	17	1	31.44/0.884	32.46/**0.897**	28.99/0.820	**30.09/0.854**	26.09/0.715	**26.91/0.776**
Self-sup. (true BSN)	Laine19	4	4	—	—	28.84/0.814	—	25.78/0.698	—
	SwinIA (ours)[†]	1	1	**31.84/0.885**	31.04/0.882	**30.01/0.837**	29.61/0.848	**27.23/0.743**	26.88/0.772

Table 1. Grayscale image denoising results for BSD68 and Set12 with synthetic noise along with the method description (supervision type and number of train and test passes). The highest PSNR(dB)/SSIM among self-supervised denoising methods is highlighted in **bold**, the second-best is underlined. [†] denotes the models that we implemented and trained ourselves.

5. Experimental results

For our experiments, we chose a model configuration with embeddings of dimensionality 144 throughout the network, 16 attention heads in each block, and windows of size 8×8. We extensively test SwinIA against state-of-the-art self-supervised denoising methods on synthetic and real-world data. Since SwinIA is a BSN and inevitably loses information because of hiding pixels from themselves, we separately focus on comparison with methods with similar properties [16]. We use peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) for evaluation. Training process is described in detail in Appendix C.

5.1. Synthetic noise (grayscale)

Following Wang *et al.* [35], we use BSD400 [41] for training and test on Set12 and BSD68 [30].We apply Gaussian noise with $\sigma = \{15, 25, 50\}$ to the images. For evaluation, we repeat BSD68 4 times, and Set12 20 times. This results in 512 (272 and 240) testing images in total.

The results are summarized in Tab. 1. On BSD68, SwinIA ranked first for all noise levels. Apart from self-supervised methods, SwinIA beats SwinIR — a supervised denoising transformer. Interestingly, we obtained lower scores on Set12 but still consistently outperformed Noise2Self and Noise2Same and tailed the scores of R2R and Blind2Unblind. We also improved over the other BSN [16] by +1.31dB PSNR on BSD68 on average.

5.2. Mixture synthetic noise

We experiment with the sRGB natural images dataset (ImageNet) and grayscale Chinese characters dataset (HànZì) with a mixture of multiple noise modalities, following Xie *et al.* [37]. ImageNet dataset was generated by randomly cropping 60 000 patches of size 128×128 from the first 20 000 images in ILSVRC2012 [6] validation set that consists of 50 000 instances. We use 978 images for testing. Poisson noise ($\lambda = 30$), additive Gaussian noise ($\sigma = 60$), and Bernoulli noise ($p = 0.2$) were applied to the clean images before the training.

Method	train	test	ImageNet	HànZì
NLM	-	-	18.04/ —	8.41/ —
BM3D	-	-	18.74/ —	10.90/ —
Noise2Clean	1	1	23.39/ —	15.66/ —
Noise2Noise	1	1	23.27/ —	14.30/ —
Noise2Void	1	1	21.36/ —	13.72/ —
Noise2Self[†]	1	1	21.33/0.574	14.16/0.512
Noise2Same[†]	2	1	22.85/0.625	14.85/0.542
Blind2Unblind*	17	1	23.74/0.649	13.87/0.509
Noise2Info	2	1	22.60/ —	14.43/ —
Laine19	4	4	20.89/ —	10.70/ —
SwinIA (ours)[†]	1	1	**23.91/0.668**	**14.92/0.574**

Table 2. Denoising results on datasets with mixed synthetic noise along with the method description (number of train and test passes). The highest PSNR(dB)/SSIM among self-supervised denoising methods is in **bold**, while the second-best is underlined. [†] denotes the models that we implemented and trained ourselves. *Blind2Unblind diverged on HànZì with different learning rates, so we provide average metrics of three runs after the 20th epoch.

HànZì dataset consists of 78 174 noisy images with 13 029 different Chinese characters of size 64×64. Each noisy image is generated by applying Gaussian noise ($\sigma = 0.7$) and Bernoulli noise ($p = 0.5$) to a clean image. We select 10% of images for testing and use the rest for training.

We present the results in Tab. 2. SwinIA showed state-of-the-art performance on both datasets, outperforming not only Noise2Same and its recent modification Noise2Info but also Blind2Unblind. It also outperformed another BSN by Laine *et al.* [16] by +3.62dB PSNR on average.

		train	test	Gaussian $\sigma = 25$			Gaussian $\sigma \in [5, 50]$		
	Method			KODAK	BSD300	SET14	KODAK	BSD300	SET14
Traditional	CBM3D	-	-	31.87/0.868	30.48/0.861	30.88/0.854	32.02/0.860	30.56/0.847	30.94/0.849
Supervised	Noise2Clean	1	1	32.43/0.884	31.05/0.879	31.40/0.869	32.51/0.875	31.07/0.866	31.41/0.863
	Noise2Noise	1	1	32.41/0.884	31.04/0.878	31.37/0.868	32.50/0.875	31.07/0.866	31.39/0.863
*Self-sup. (Gaussian)**	Laine19-pme	4	4	32.40/0.883	30.99/0.877	31.36/0.866	32.40/0.870	30.95/0.861	31.21/0.855
	Honzatko-pme	1	1	32.45/ —	31.02/ —	31.25/ —	32.46/ —	31.18/ —	31.25/ —
	Noisier2Noise	1	1	30.70/0.845	29.32/0.833	29.64/0.832	—	—	—
Self-sup.	Self2Self	1	50	31.28/0.864	29.86/0.849	30.08/0.839	31.37/0.860	29.87/0.841	29.97/0.849
	Noise2Void	1	1	30.32/0.821	29.34/0.824	28.84/0.802	30.44/0.806	29.31/0.801	29.01/0.792
	Noise2Same†	2	1	30.77/0.841	29.50/0.834	29.53/0.827	30.78/0.835	29.49/0.823	29.34/0.817
	DBSN	2	1	31.64/0.856	29.80/0.839	30.63/0.846	30.38/0.826	28.34/0.788	29.49/0.814
	R2R	1	50	32.25/0.880	30.91/0.872	**31.32/0.865**	31.50/0.850	30.56/0.855	30.84/0.850
	NBR2NBR	2	1	32.08/0.879	30.79/0.873	31.09/0.864	32.10/0.870	30.73/0.861	31.05/**0.858**
	B2UB	17	1	**32.27**/0.880	30.87/0.872	31.27/0.864	32.34/**0.872**	30.86/0.861	**31.14**/0.857
	DCD-Net	3	1	**32.27/0.881**	**31.01/0.876**	31.29/0.862	**32.35/0.872**	**31.09/0.866**	31.09/0.855
Self-sup. (true BSN)	Laine19	4	4	30.62/0.840	28.62/0.803	29.93/0.830	30.52/0.833	28.43/0.794	29.71/0.822
	SwinIA (ours)†	1	1	31.43/0.863	29.94/0.853	30.56/0.856	31.54/0.859	30.00/0.847	30.55/0.849

		train	test	Poisson $\lambda = 30$			Poisson $\lambda \in [5, 50]$		
	Method			KODAK	BSD300	SET14	KODAK	BSD300	SET14
Traditional	Anscombe	-	-	30.53/0.856	29.18/0.842	29.44/0.837	29.40/0.836	28.22/0.815	28.51/0.817
Supervised	Noise2Clean	1	1	31.78/0.876	30.36/0.868	30.57/0.858	31.19/0.861	29.79/0.848	30.02/0.842
	Noise2Noise	1	1	31.77/0.876	30.35/0.868	30.56/0.857	31.18/0.861	29.78/0.848	30.02/0.842
*Self-sup. (Poisson)**	Laine19-pme	4	4	31.67/0.874	30.25/0.866	30.47/0.855	30.88/0.850	29.57/0.841	28.65/0.785
	Honzatko-pme	1	1	31.67/ —	30.25/ —	30.14/ —	—	—	—
Self-sup.	Self2Self	1	50	30.31/0.857	28.93/0.840	28.84/0.839	29.06/0.834	28.15/0.817	28.83/0.841
	Noise2Void	1	1	28.90/0.788	28.46/0.798	27.73/0.774	28.78/0.758	27.92/0.766	27.43/0.745
	Noise2Same†	2	1	27.73/0.747	26.69/0.714	26.78/0.735	27.44/0.738	26.36/0.700	26.37/0.721
	DBSN	2	1	30.07/0.827	28.19/0.790	29.16/0.814	29.60/0.811	27.81/0.771	28.72/0.800
	R2R	1	50	30.50/0.801	29.47/0.811	29.53/0.801	29.14/0.732	28.68/0.771	28.77/0.765
	NBR2NBR	2	1	31.44/0.870	30.10/0.863	30.29/0.853	30.86/0.855	29.54/0.843	29.79/0.838
	B2UB	17	1	31.64/0.871	30.25/0.862	30.46/0.852	**31.07/0.857**	29.92/0.852	**30.10/0.844**
	DCD-Net	3	1	**32.35/0.872**	**31.09/0.866**	**31.09/0.855**	31.00/**0.857**	**29.99/0.855**	29.99/0.843
Self-sup. (true BSN)	Laine19	4	4	30.19/0.833	28.25/0.794	29.35/0.820	29.76/0.820	27.89/0.778	28.94/0.808
	SwinIA (ours)†	1	1	31.01/0.857	29.61/0.847	29.98/0.847	30.29/0.835	28.84/0.818	29.35/0.827

Table 3. Denoising results on synthetic sRGB datasets along with the method description (supervision type and number of train and test passes). The highest PSNR(dB)/SSIM among self-supervised denoising methods is highlighted in **bold**, the second-best is underlined. * denotes assuming known noise model. † denotes the models that we implemented and trained ourselves.

5.3. Synthetic noise (sRGB)

We follow Huang *et al.* [14] to create training and test sRGB datasets. For training, we select 44 328 images between 256 × 256 and 512 × 512 pixels from the ILSVRC2012 [6] validation set. For testing, we use Kodak [9], BSD300 [23], and Set14 [39], repeated by 10, 3, and 20 times, respectively. This adds up to 780 (240, 300, and 240) test images. We apply four types of noise in sRGB: Gaussian noise with (1) $\sigma = 25$ and (2) $\sigma \in [5, 50]$, Poisson noise with (3) $\lambda = 30$ and (4) $\lambda \in [5, 50]$.

We present the results in Tab. 3. SwinIA consistently supersedes the other BSN method by Laine *et al.* [16] and most of the mask-based methods, especially with Poisson noise, where we beat R2R by 0.5dB PSNR on average. However, it did not compete with the state-of-the-art methods employing multiple passes in training and inference.

5.4. Natural noise in fluorescent microscopy

We use Confocal Fish, Confocal Mice, and Two-Photon Mice datasets from the Fluorescent Microscopy Denoising

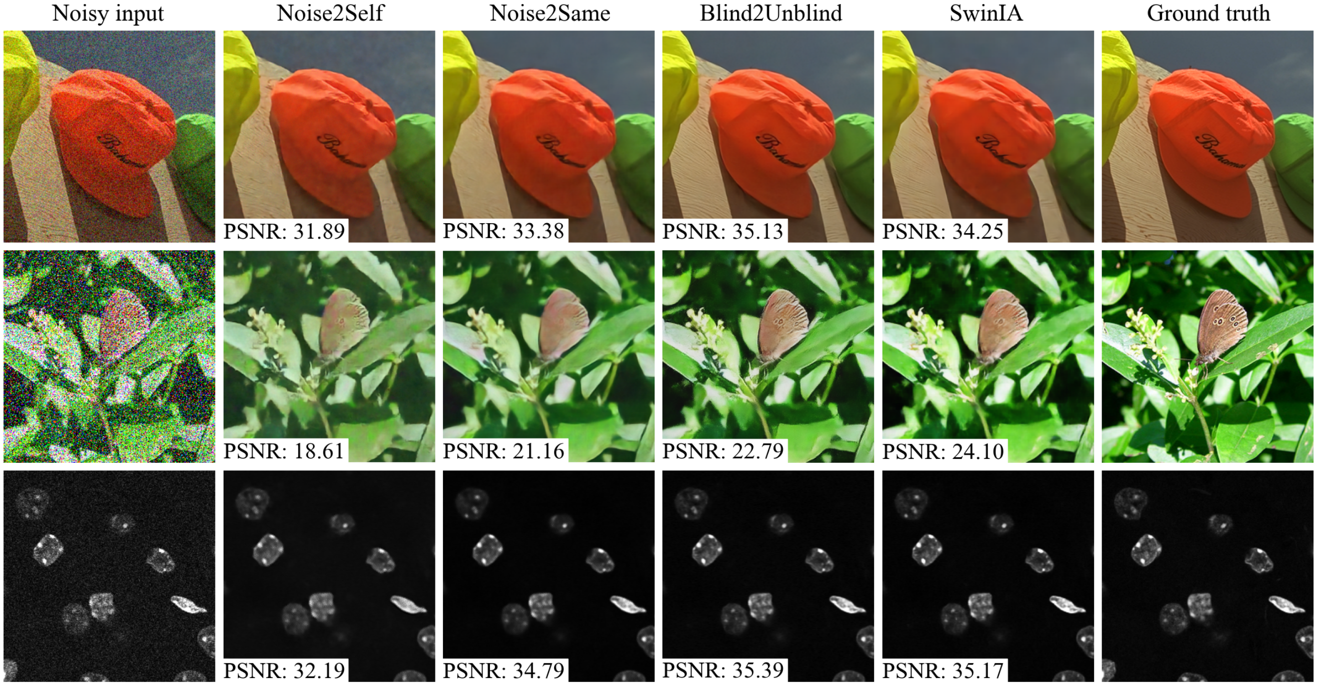


Figure 5. Kodak [9] (top), ImageNet [37] (middle), and FMD Two-Photon Mice [42] (bottom) denoising examples. Every predicted image is cropped to a square for visualization presented with the corresponding PSNR score (in dB).

Dataset [42]. Each dataset consists of 20 views, each comprising 50 grayscale images of size 512×512. Each image contains a **natural** mixture of Poisson and Gaussian noise. Ground truth is obtained by averaging all images in the view. We follow Wang *et al.* [35] and select the 19th view for testing and the rest for training.

The results are presented in Tab. 4. SwinIA performed competitively across all datasets, yielding either best or second-best scores. Another BSN by Laine *et al.* [16] required knowledge about the noise model and performed worse for the real-world data with noise mixture: SwinIA was better by +0.91dB PSNR and +0.025 SSIM on average for both Gaussian and Poisson assumed distributions.

5.5. Ablation study

We ran ablation experiments on synthetic noise grayscale datasets to validate our key architecture design elements. The results are summarized in Tab. 5. In particular, we experimented with alternative architectures without pixel shuffle: dilated attention [10] and flat architecture without encoder-decoder separation [19]. Both not only harmed the performance but also considerably increased training time. Removing the mask on inference led to a valuable performance increase, which proves the necessity of unblinding.

Methods	Confocal Fish	Confocal Mice	Two-Photon Mice
BM3D	32.16/0.886	37.93/0.963	33.83/0.924
Noise2Clean	32.79/0.905	38.40/0.966	34.02/0.925
Noise2Noise	32.75/0.903	38.37/0.965	33.80/0.923
Laine19-pme (G)	23.30/0.527	31.64/0.881	25.87/0.418
Laine19-pme (P)	25.16/0.597	37.82/0.959	31.80/0.820
Noise2Void	32.08/0.886	37.49/0.960	33.38/0.916
NBR2NBR	32.11/0.890	37.07/0.960	33.40/**0.921**
Noise2Self[†]	31.96/0.877	36.45/0.960	31.61/0.910
Noise2Same[†]	32.36/0.893	37.64/0.960	33.55/0.917
Blind2Unblind	**32.74**/0.897	**38.44**/0.964	**34.03**/0.916
CADT	32.52/0.895	38.21/0.962	33.64/0.914
Laine19 (G)	31.62/0.849	37.54/0.959	32.91/0.903
Laine19 (P)	31.59/0.854	37.30/0.956	33.09/0.907
SwinIA (ours)[†]	32.65/**0.904**	38.21/**0.966**	33.90/0.920

Table 4. Denoising results on Fluorescent Microscopy datasets. The highest PSNR(dB)/SSIM among self-supervised denoising methods is highlighted in **bold**, while the second-best is underlined. For Laine *et al.* [16], G — Gaussian, P — Poisson. [†] denotes the models that we implemented and trained ourselves.

We separately tested attention masking and input isolation as unavoidable restrictions for self-unawareness, and the full encoder flow — a configuration where the input is sequentially propagated down the encoder blocks allowing context awareness inside of patches. All three experiments resulted in learning the identity function and poor scores.

Experiment	1 epoch (min)	BSD68	Set12
Our best	11.5	**30.01/0.837**	29.61/**0.848**
Dilated attention	14.5	29.90/0.830	**29.63**/0.842
Flat architecture	14.5	29.87/0.826	29.62/0.840
Masked inference	11.5	29.35/0.816	28.92/0.832
No attention mask*	11.5	20.45/0.380	20.33/0.394
No input isolation*	12	20.51/0.379	20.39/0.383
Full encoder*	11.5	21.22/0.396	21.01/0.402
Larger window (12)	47	30.09/0.840	29.69/0.850
Smaller window (6)	7	29.79/0.830	29.37/0.842

Table 5. Ablation results on grayscale data with synthetic Gaussian noise ($\sigma = 25$). The experiments marked with * ended up learning the identity function.

We also experimented with the attention window size. Larger window size w requires larger training crops p because of downsampling and is computationally expensive since the attention computation complexity is $\Theta(w^4)$. For $w = 12, p = 96$, the training was four times longer. The increased context provided a marginal gain of +0.08/+0.003 PSNR/SSIM on average, while decreasing it to $w = 6, p = 48$ reduced the scores by -0.23/-0.006 (see Tab. 5).

6. Discussion

The flexibility of transformer architecture allowed us to build an assumption-free SwinIA model that has many strengths. First, it is robust across various noise types and image modalities. Most notably, it achieves state-of-the-art performance for the most complex synthetic mixed noise datasets and several others. Second, our model is optimized by minimizing a simple loss function without tunable hyperparameters. The main competitors, Blind2Unblind [35] and DCD-Net [43], have multiple empirically set loss constants changing according to the selected training schedule. Finally, SwinIA uses a single forward pass in both training and inference allowing to decrease time and compute, which is especially important for a transformer-based model (most competitor models use multiple passes, as we report in our tables with results). In our experiments, SwinIA trains twice faster than Blind2Unblind on the same hardware (see Appendix C for the details).

The versatility of a true blind-spot model comes with limitations. The pixel itself contains the most useful information about its true signal, which is inevitably lost in the training process. However, we are able to remove the attention mask during inference and allow pixels to attend to their initial values. This is not possible in a convolutional BSN where hiding is done through zeroing trainable weights [13,36] or excluding the central pixel from the field of view [16]. We further discuss and visualize our unmasking in Appendix A. Also, a BSN assumes spatially uncorrelated noise, which is not the case for many digital photography datasets because of hardware pixel interpolation. This problem can be mitigated with increased patch size or dilated attention resembling the approach by Lee *et al.* [17].

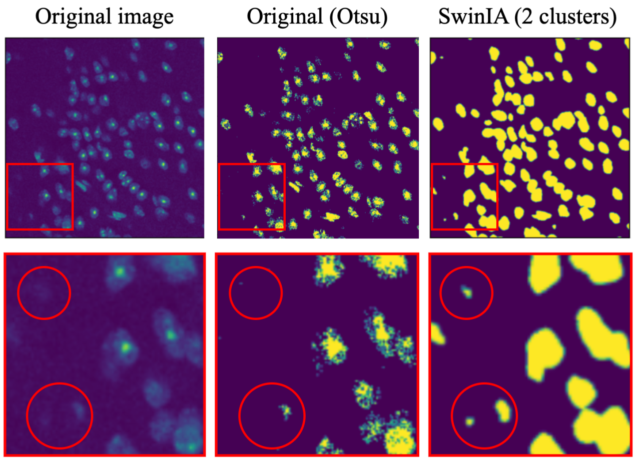


Figure 6. Binary thresholding on FMD Confocal Mice. We apply Otsu thresholding [25] to the original image (middle column) and k-means with $k = 2$ to the final feature map of SwinIA (right column). A part of each image is zoomed (bottom row), and the blurry and half-light cells are highlighted with red circles.

Transformers are known for their ability to extract rich representations from large datasets, and we expect our method to improve with increasing training set size. Besides, being conceptually similar to the language modeling objective [31], our solution could be used in self-supervised pre-training to produce **pixel** embeddings for downstream image tasks. Fig. 6 shows an example of SwinIA embeddings clustering into segmentation masks. Fig. 6 also features Otsu thresholding [25] to ensure that quality masks are not simply obtainable straight from the noisy input. In Appendix B, we show more examples of feature clustering, also comparing to other models. We leave further investigation of the feature extraction abilities for future work.

7. Conclusion

We propose SwinIA, the first convolution-free transformer architecture for blind-spot self-supervised denoising. Unlike its counterparts, it does not require access to clean data or assume any noise distribution. SwinIA also does not use input masking and can be trained in an autoencoder fashion with a single forward pass and an MSE loss. Finally, it does not require multivariate hyperparameter tuning and achieves competitive results, outperforming state-of-the-art methods on several common benchmarks and showing robustness to different kinds of synthetic and natural noise in images of various modalities.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[2] Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 524–533. PMLR, 09–15 Jun 2019. 1, 2, 13, 14, 15, 16, 17

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 60–65. Ieee, 2005. 1

[4] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023. 12

[5] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Color image denoising via sparse 3d collaborative filtering with grouping constraint in luminance-chrominance space. volume 1, pages I – 313, 09 2007. 1

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5, 6, 16

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Rich Franzen. Kodak lossless true color image suite. *source: http://r0k. us/graphics/kodak*, 4(2), 1999. 6, 7, 17

[10] Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. *arXiv preprint arXiv:2209.15001*, 2022. 7

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[12] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). 2016. 4

[13] David Honzátko, Siavash A Bigdeli, Engin Türetken, and L Andrea Dunbar. Efficient blind-spot neural network architecture for image denoising. In *2020 7th Swiss Conference on Data Science (SDS)*, pages 59–60. IEEE, 2020. 1, 2, 8

[14] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14781–14790, 2021. 1, 6

[15] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2129–2137, 2019. 1, 2

[16] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 5, 6, 7, 8

[17] Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. Apbsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17725–17734, 2022. 2, 8

[18] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018. 1

[19] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021. 2, 3, 7, 12

[20] Xiaolong Liu, Yusheng Hong, Qifang Yin, and Shuo Zhang. Dnt: Learning unsupervised denoising transformer from single noisy image. In *Proceedings of the 4th International Conference on Image Processing and Machine Vision*, IPMV '22, page 50–56, New York, NY, USA, 2022. Association for Computing Machinery. 2

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 4

[22] Markku Makitalo and Alessandro Foi. Optimal inversion of the anscombe transformation in low-count poisson image denoising. *IEEE Transactions on Image Processing*, 20(1):99–109, 2011. 1

[23] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, volume 2, pages 416–423, 2001. 6, 17

[24] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12064–12072, 2020. 1

[25] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. 8, 11

[26] Tongyao Pang, Huan Zheng, Yuhui Quan, and Hui Ji. Recorrupted-to-recorrupted: unsupervised deep learning for

image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2043–2052, 2021. 1

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 12, 16

[28] Yuhui Quan, Mingqin Chen, Tongyao Pang, and Hui Ji. Self2self with dropout: Learning self-supervised denoising from single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1890–1898, 2020. 1

[29] Alex Rogozhnikov. Einops: Clear and reliable tensor manipulations with einstein-like notation. In *International Conference on Learning Representations*, 2022. 12, 16

[30] Stefan Roth and Michael J Black. Fields of experts: A framework for learning image priors. In *CVPR*, pages 860–867. IEEE, 2005. 5, 17

[31] Joongbo Shin, Yoonhyung Lee, Seunghyun Yoon, and Kyomin Jung. Fast and accurate deep bidirectional language representations for unsupervised learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 823–835, Online, July 2020. Association for Computational Linguistics. 2, 3, 8

[32] Leslie N. Smith and Nicholay Topin. Super-convergence: very fast training of neural networks using large learning rates. In *Defense + Commercial Sensing*, 2018. 12

[33] Jiachuan Wang, Shimin Di, Lei Chen, and Charles Wang Wai Ng. Noise2info: Noisy image to information of noise for self-supervised image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16034–16043, 2023. 2

[34] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 2

[35] Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2unblind: Self-supervised image denoising with visible blind spots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2027–2036, 2022. 2, 5, 7, 8, 11, 12, 13, 14, 15

[36] Xiaohe Wu, Ming Liu, Yue Cao, Dongwei Ren, and Wangmeng Zuo. Unpaired learning of deep image denoising. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, pages 352–368. Springer, 2020. 2, 8

[37] Yaochen Xie, Zhengyang Wang, and Shuiwang Ji. Noise2Same: Optimizing a self-supervised bound for image denoising. In *Advances in Neural Information Processing Systems*, volume 33, pages 20320–20330, 2020. 1, 2, 5, 7, 11, 12, 13, 14, 15, 16, 17

[38] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 2

[39] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730, 2010. 6, 17

[40] Dan Zhang and Fangfang Zhou. Self-supervised image denoising for real-world images with context-aware transformer. *IEEE Access*, 2023. 2

[41] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP*, 26(7):3142–3155, 2017. 5, 16

[42] Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard. A poisson-gaussian denoising dataset with real fluorescence microscopy images. In *CVPR*, 2019. 7, 11, 15, 16, 17

[43] Yunhao Zou, Chenggang Yan, and Ying Fu. Iterative denoiser and noise estimator for self-supervised image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13265–13274, October 2023. 2, 8