SEGMENTATION-AWARE IMAGE DENOISING WITHOUT KNOWING TRUE SEGMENTATION

A Thesis

by

SICHENG WANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair Name,	Zhangyang Wang
Committee Members,	Theodora Chaspari
	Nima Kalantari
	Chao Tian
Head of Department,	Scott Schaefer

December 2020

Major Subject: Computer Science

ABSTRACT

Recent works have discussed application-driven image restoration neural networks capable of not only removing noise in images but also preserving their semantic-aware details, making them suitable for various high-level computer vision tasks as the pre-processing step. However, such approaches require extra annotations for their high-level vision tasks in order to train the joint pipeline using hybrid losses, yet the availability of those annotations is often limited to a few image sets, thereby restricting the general applicability of these methods to simply denoise more unseen and unannotated images. Motivated by this, we propose a segmentation-aware image denoising model dubbed **U-SAID**, based on a novel **unsupervised** approach with a pixel-wise uncertainty loss. U-SAID does not require any ground-truth segmentation map, and thus can be applied to any image dataset. It is capable of generating denoised images with comparable or even better quality than that of its supervised counterpart and even more general "application-agnostic" denoisers, and its denoised results show stronger robustness for subsequent semantic segmentation tasks. Moreover, plugging its "universal" denoiser without fine-tuning, we demonstrate the superior **generalizability** of U-SAID in three-folds: (1) denoising unseen types of images; (2) denoising as pre-processing for segmenting unseen noisy images; and (3) denoising for unseen high-level tasks. Extensive experiments were conducted to assess the effectiveness and robustness of the proposed U-SAID model against various popular image sets.

DEDICATION

To my parents, my grandmother, my uncle & aunt and Haotian.

ACKNOWLEDGMENTS

Foremost, I would like to express my deepest appreciation to my advisor Prof. Zhangyang Wang for the continuous support of my master study and research, for his patience, motivation, enthusiasm, and immense knowledge.

I would also like to extend my deepest gratitude to Prof. Bihan Wen. The completion of this study could not have been possible without the expertise of Prof. Wen.

Thanks also to the rest of my thesis committee members who provided help and support for this project.

Moreover, I had great pleasure of working and learning with all my friends in VITA lab and Extreme CV team at Walmart. Special thanks to Andrew Stevens, Emilio Torres Jr. and Haotian Zhong for giving me valuable advice in my thesis writing and defense presentation.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professor Zhangyang Wang (chair), Professor Theodora Chaspari and Professor Nima Khademi Kalantari of the Department of Computer Sicence and Engineering and Professor Chao Tian of the Department of Electrical Computer Engineering.

The analyses depicted in Chapter 4 were conducted in part by Professor Bihan Wen of School of Electrical and Electronic Engineering of Nanyang Technological University.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was supported by Graduate Merit Scholarship from Texas A&M University.

NOMENCLATURE

NIQE	Naturalness Image Quality Evaluator
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure
CNN	Convolutional Neural Network
DNN	Deep Neural Network
MSE	Mean Squared Error
U-SAID	Unsupervised Segmentation-aware Image Denoising Model
S-SAID	Supervised Segmentation-aware Image Denoising Model
GT	Groundtruth
MRI	Magnetic Resonance Imaging
FPN	Feature Pyramid Network
ResNet	Residual Neural Network

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Image denoising aims to recover the underlying clean image signal from its noisy measurement. Traditionally, it has been treated as an independent signal recovery problem, focusing on either single-level fidelity (e.g., PSNR) or human perception quality of the recovery results. However, once high-level vision tasks are conducted on noisy images, and such a separate image denoising step is typically applied as preprocessing, it becomes suboptimal due to its unawareness of semantic information. A series of recent works [1, 2, 3, 4, 5, 6] discuss *application-driven image restoration models* that are capable of simultaneously removing noise and preserving semantic-aware details for certain high-level vision tasks. Those models achieve visually-promising denoising results with richer details, in addition to better utility when supplied for high-level task pre-processing.

However, a common drawback of these models is their demand for *extra annotations* for the high-level vision tasks, which they require in order to train the joint pipeline with hybrid low-level and high-level supervisions. On the one hand, such annotations (e.g., object bounding boxes, semantic segmentation maps) are often highly non-trivial to obtain for real images, thus limiting current works to synthesizing noise on existing annotated, clean datasets, to demonstrate the effectiveness of their methods. On the other hand, training with only one annotated dataset runs the risk of overly tying the resulting denoiser with the semantic information of this specific dataset, causing a lack of universality and the potential to exhibit various artifacts due to overfitting when attempting to denoise other substantially different images.

This paper attempts to break the aforementioned hurdles of existing application-driven image restoration models. We propose a novel *unsupervised segmentation-aware image denoising* (**U-SAID**) model that enforces segmentation awareness and the discriminative ability of denoisers **without actually needing any segmentation groudtruth during training**. It is implemented by creating a novel loss term that penalizes the *pixel-wise uncertainty* of the denoised outputs for segmentation. Our contributions are twofold:

- On the *low-level* vision side, to the best of our knowledge, U-SAID is the first unsupervised (or "self-supervised") application-driven image restoration model. In contrast to the existing peer work [1], U-SAID can be trained on any image datatset, without needing ground-truth (GT) segmentation maps. That greatly extends the applicability of U-SAID as a more "universal" denoiser, that can be applied to denoise images with few semantic annotations while being substantially different from natural images in existing segmentation datasets. Compared to standard "application-agnostic" denoisers such as [7], U-SAID is observed to provide better visual details, that are also more favored under perception-driven metrics [8].

- On the *high-level* vision side, the U-SAID denoising network is shown to be robust and "universal" enough, when applied to denoising different noisy datasets, as well as when used towards boosting the segmentation task performance on unseen noisy datasets, thanks to its less semantic association with any dataset annotation. Furthermore, U-SAID trained with segmentation awareness generalizes well to unseen high-level vision tasks, and can be plugged into without fine-tuning, which reduces the training effort when applied to various high-level tasks.

This paper is constructed as below: section 2 reviews related literature from deep neural network based denoiser to more recent application-driven denoising works, which serve as the base to our method. In section 3, we introduce our proposed unsupervised segmentation-aware image denoising network, including the network architecture and the loss function definitions. In addition, we provide proof of concept experiments and analysis to show why our method works without strong label supervision. In section 4, we validate the proposed method with extensive experiments on various popular image sets to demonstrate its outstanding effectiveness, robustness, and universality. Finally, section 5 concludes the thesis by advocating that our methodology is (almost) a *free lunch* for image denoising, and has a plug-and-play nature to be incorporated with existing deep denoising models.

2. LITERATURE REVIEW

Image denoising has been studied with intensive efforts for decades. Earliest methods refer to various image filters [9]. Later on, many model-based method with various priors have been introduced to this topic, in either spatial or transform domain, or their hybrid, such as spatial smoothness [10], non-local patch similarity [11], sparsity [12, 13, 14] and low-rankness [15]. More recently, a number of deep learning models have demonstrated superior performance for image denoising [16, 17, 7]. Despite their encouraging process, most existing denoising algorithms reconstruct images by minimizing the mean square error (MSE), which is well-known to be mis-aligned with human perception quality and often tends to over-smooth textures [18]. Moreover, while image denoising algorithms are often needed as the pre-processing step for the acquired noisy visual data before subsequent high-level visual analytics, their impact on the semantic visual information was much less explored.

Lately, a handful of works are devoted to closing the gap between the low-level (e.g., image denoising, as a representative) and high-level computer vision tasks. Such marriage leads to, not only better utility performance for high-level target tasks, but also the denoising outputs with richer visual details after receiving the extra semantic guidance from the high-level tasks, the latter being first revealed in [19, 20]. [1] presented a systematical study on the mutual influence between the low-level and high-level vision networks. The authors cascaded a fixed pre-trained semantic segmentation network after a denoising network, and tuned the entire pipeline with a joint loss function of MSE and segmentation loss. The overview of cascaded denoising network is illustrated in 2.1. In that way, the authors showed the denoised images to have sharper edges and clearer textual details, as well as higher segmentation and classification accuracies when feeding such denoised images for those tasks. A similar effort was described in [4], where a segmentation-aware deep fusion network was proposed to utilize the segmentation labels in MRI datasets to aid MRI compressive sensing recovery. [3] considered a joint pipeline of image dehazing and object detection. [21] proposed to incorporate global semantic priors (e.g., eyes and mouths) as an

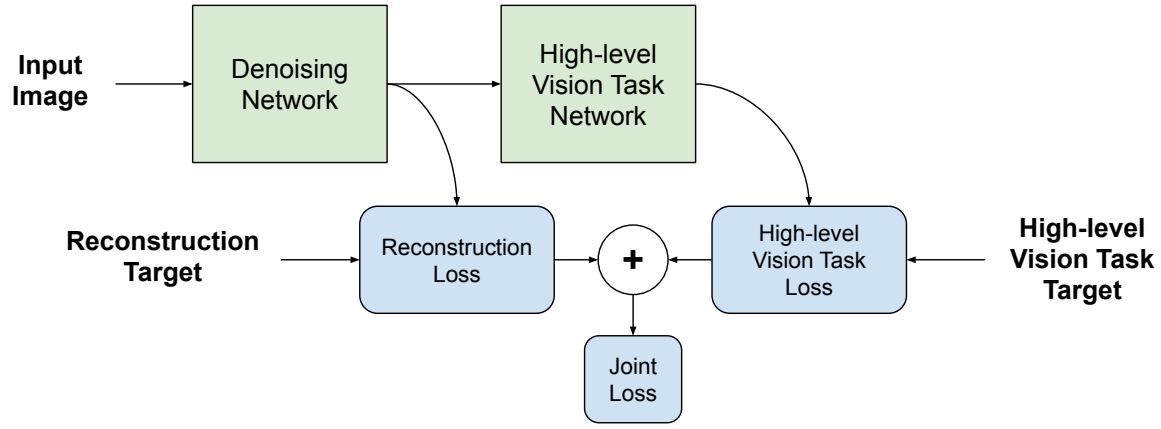


Figure 2.1: Overview of high-level supervised denoiser.

input to deblur the highly structured face images. This field is now rapidly growing, with a few benchmarks launched recently [22, 23, 24, 25].

Following [1, 4], we also adopt segmentation as our high-level task, because it can supply pixel-wise feedback and is thus considered to be more helpful for dense regression tasks. As pointed out by [26], the availability of segmentation information can compromise the over-smoothening effects of CNNs across regions and increases their spatial precision. However, we would like to emphasize (again) that while [1, 26, 4] all exploit GT segmentation maps as extra strong *supervision* information during training, we have only a weaker form of *feedback* available from the segmentation task, due to the absence of its GT as extra information. Straightforwardly, our methodology is applicable when cascaded with other high-level tasks as well.

Our work is also broadly related to training deep network with noisy or uncertain annotations [27, 28]. Especially for the segmentation task, existing supervised models require manually labeled segmentations for training. But pixel-based labeling for high-resolution images is often time-consuming and error-prone, causing incorrect pixel-wise annotations. Existing works often consider them as label noise [29]. For example, [30] proposed a noise-tolerant deep model for histopathological image segmentation, using the label-flip noise models proposed in [31]. However, those algorithms still need to be given segmentation maps (though inaccurate), and often

demand more statistical estimations of the label noise.

3. THE PROPOSED MODEL: U-SAID

Our proposed unsupervised segmentation-aware image denoising (U-SAID) network follows the same cascade idea of the segmentation-guided denoising framework proposed by [1]. We replace their self-designed U-Net denoiser with the classical deep denoiser DnCNN [7], using the 20-layer blind color image denoising model referred to as CDnCNN-B[1], since we favor more robustness to varying noise labels. Note that the choice of denoiser network should not affect much our obtained conclusions. Its loss L_{MSE} is the reconstruction MSE between the denoised output and the clean image. The network architecture is illustrated in Figure 3.1.

3.1 Design of USA Module

The USA module is composed of a feature embedding sub-network for transforming the input (denosied image) to the feature space, followed by an unsupervised segmentation sub-network that calculates the *pixel-wise uncertainty* of semantic segmentation.

3.1.1 Feature embedding sub-network

We use a Feature Pyramid Network (FPN) [32], with a ResNet-101 backbone as the feature encoder, which is the bottom-up path in yellow blocks as illustrated in Figure 3.1. M2-M5 are feature maps which undergo 1×1 convolutions and element-wisely added to the feature maps from top-down pathway. The final feature maps (P2-P5) are generated through two layers of 3×3 convolutions from M2-M5. We use ImageNet-pretrained weights[2] for the backbone, and keep all default architecture details of FPN/ResNet-101 unchanged. During training, the ResNet-101 backbone is frozen as a fixed feature extractor, and the top-down feature pyramid part of FPN started with random Gaussian initializations and also keeped fixed.

[1] https://github.com/cszn/DnCNN

[2] https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py

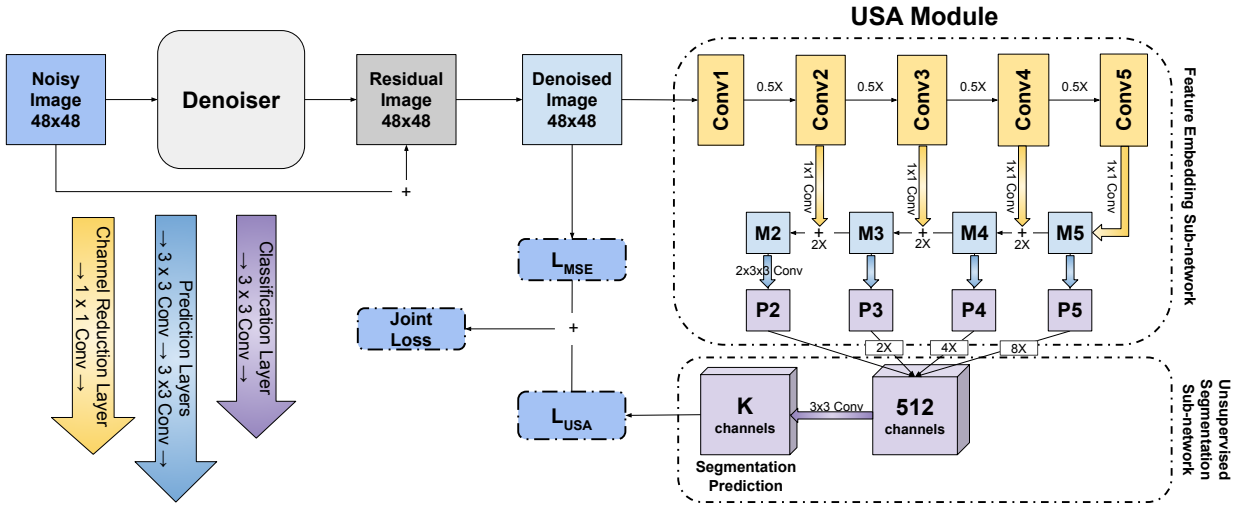


Figure 3.1: The architecture of the proposed U-SAID network

3.1.2 Unsupervised segmentation sub-network

We assume the input image resolution to be $M \times N$ and to contain at most K different semantic classes. After FPN, we obtain 512 channels of feature maps $\in \mathbb{R}^{\frac{M}{4} \times \frac{N}{4}}$. We then apply K 3 $\times$ 3 convolutions to re-organize the output feature maps into K channels, eventually leading to a (resized) K-class segmentation map.

Since the image segmentation task can be cast as pixel-wise classification, classical segmentation networks will adopt pixel-wise softmax function to generate a K-class probability vector $p_{i,j}$, for the (i, j)-th $\mathbb{R}^K$ vector (i, j range from 1 to M, N, respectively), choosing the highest probability class and producing the final segmentation map $\in \mathbb{R}^{M \times N}$. However, since we have no GT pixel labels in the unsupervised case, we instead minimize the average entropy function of all predicted class vectors $p_{i,j}$, denoted as L_{USA}, to encourage confident predictions at all pixels:

$$L_{USA} = \frac{1}{MN} \sum_{1 \le i \le M, 1 \le j \le N} -p_{i,j} \log p_{i,j}$$

All layer-wise weights in the unsupervised segmentation sub-network are random Gaussian initialized, and the ResNet-101 backbone uses the pre-trained ImageNet weights.

3.2 Analysis of L_{USA}

We train the cascade of denoising network and USA module in an end-to-end manner, while fixing the weights in the latter. If more components of the USA modules are trainable, the model would easily collapse to a bad local minimum quickly, which leads to only a single class being predicted throughout the image. In that way, no segmentation guidance can be provided by the USA module. We thereby performed a simple experiment to show the collapse of the USA module by allowing more parts of the USA module to train. Meanwhile, we plot the training loss L_{USA} against the steps to show the relationship between loss and the learnable parts. Please note the training of the USA module remains unsupervised in this experiment. The plot in Fig. 3.2 reveals the loss drops more quickly if more parts are trained. We train the USA module using only half

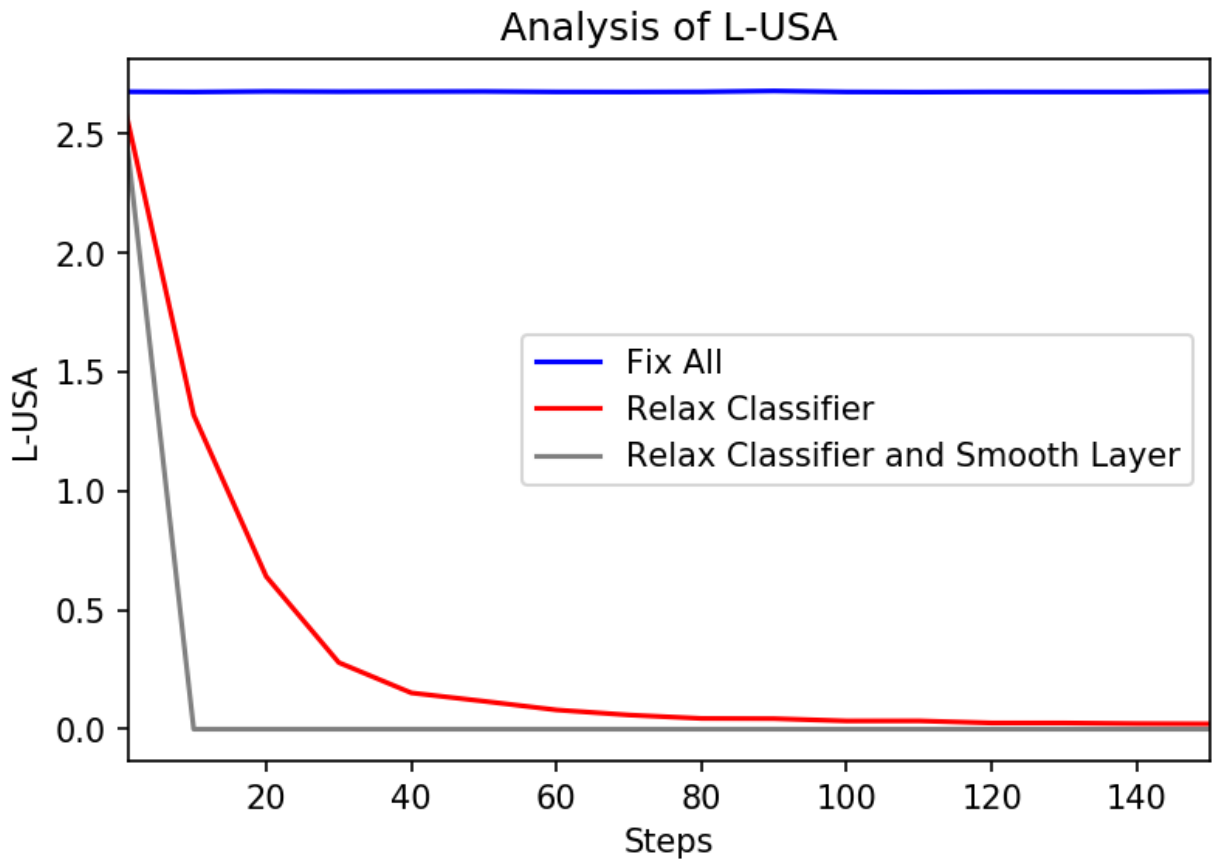


Figure 3.2: USA Module Training Loss L_{USA} Plot

epoch of data by (1) fixing all parts, (2) relaxing the last classification layer (the last 3×3 convolutional layer), or (3) relaxing the classification layer and the smooth layers (two 3×3 convolutional layers). It is clear from the plot that the training loss of (3) decrease to zero within 10 steps. We have tested to relax more layers, but the loss plot overlaps with the (3) as their loss show a sharp downturn in loss. Although loss of (2) decrease slower than (3), it still drops to zero within 100 steps. However, this phenomenon does not happen if the entire USA module is fixed. The proposed U-SAID is highly regularized by the nontrainable segmentation parameters, which effectively prevents the U-SAID segmentation model to collapse, i.e., avoiding the bad local minimum in the training.

The overall loss for U-SAID is: $L_{MSE} + \gamma L_{USA}$, with the default $\gamma = 1$ unless otherwise specified. The training dataset for U-SAID could be any image set and is unnecessary to have segmentation annotations, overcoming the limitations in [1, 4]. That said, we need an estimate of segmentation class numbers K to construct L_{USA}: an ablation study of estimated K will follow.

3.3 Why it works?

A noteworthy feature of U-SAID is freezing the high-level network while only training the denoiser. Without strong label supervision, one may wonder why it can regularize the denoiser training effectively, since it is high level features include the random initialization keep fixed, and the ResNet-101 ImageNet features can still be regressed into some unknown map, that is only required to be low-entropy pixel-wise. In fact, if the network itself holds large enough capacity, one may expect to be able to find parameters that can fit with any given pixel-wise map (low-entropy or not), that conveys little semantical information (e.g., random maps).

That might have reminded the *deep image prior* proposed in [33]: the authors first trained a convolutional network from random scratch, to regress from a random vector to a given corrupted image, and then used the trained network as a regularization. Since no aspect of the network is pre-trained from data, such deep image prior is effectively handcrafted and was shown to work well for various image restoration tasks. The authors attributed the success to the convolutional architecture itself, that appeared to possess high noise impedance. In our case, the ImageNet features are

thought as highly relevant to image semantics. Therefore, we make the similar hypothesis with the authors of [33]: although the parameterization may regress to any random unstructured label map, it does so very reluctantly.

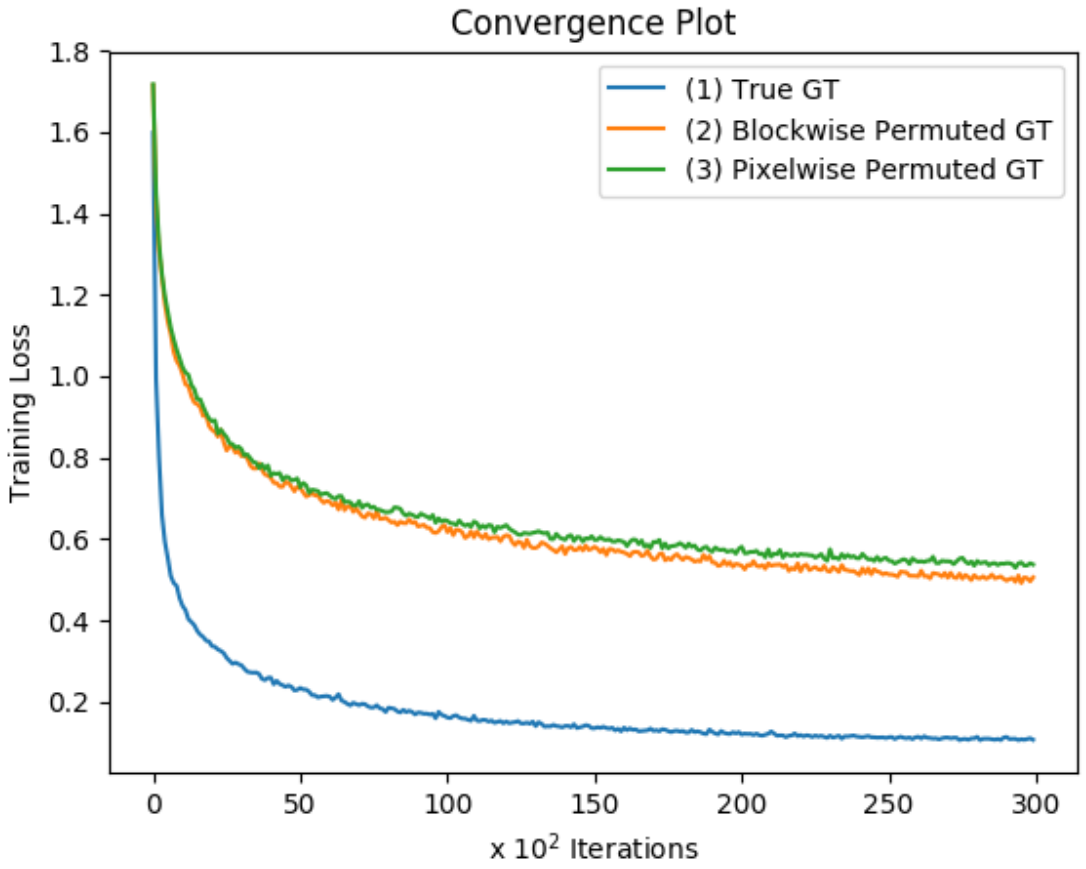


Figure 3.3: Supervised Segmentation Model Convergence Plot

To verify our hypothesis, we conduct a simple proof-of-concept experiment inspired by [34]. In the USA module, we replace L_{USA} with a standard pixel-wise cross entropy loss, having ResNet-101 fixed with ImageNet weights and other parts initialized randomly. We then use PASCAL VOC 2012 training set to train this modified USA module, in a supervised way, but with three different choices for the supervision: 1) the GT segmentation maps; 2) evenly cutting each GT map into 4

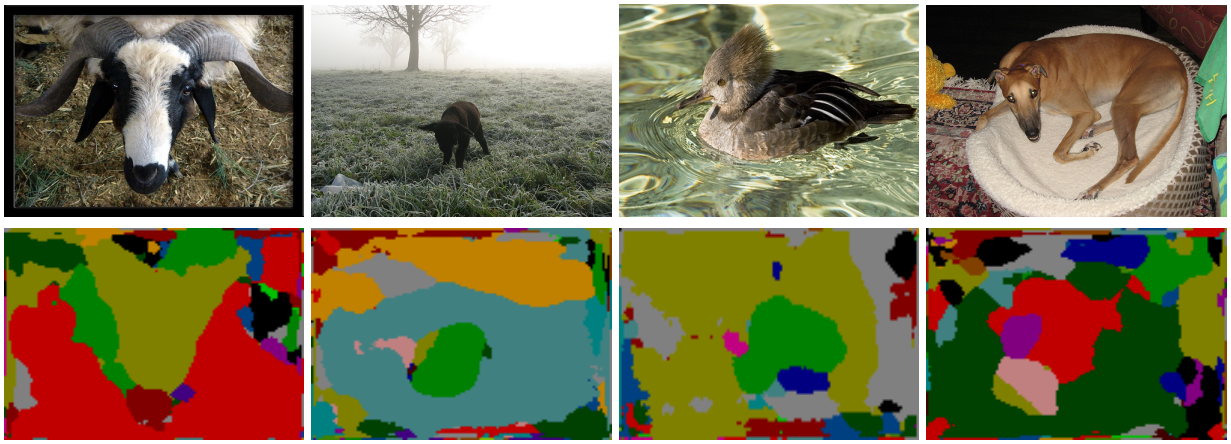


Figure 3.4: Images (top row) and their segmentation maps (second row) produced by USA module on PASCAL VOC.

sub-images, and randomly permuting their locations; 3) randomly permuting all pixel locations in each GT map. Notice that if we compute L_{USA} values for the three target maps, they should be the same.

We show in Figure 3.3 the value of training loss, as a function of the gradient descent iterations for three supervisions. Apparently, the network can converge much faster to GT maps; the more GT maps were permuted, the more convergence "inertia" we observe. In other words, the network descends much more quickly towards semantically meaningful maps, and resists "bad" solutions with fewer semantics, although their entropies might have been the same.

Another question raised is how the segmentation map produced by USA module would enhance the image denoising quality. [1] has proved the effectiveness of high-level guidance in image denoising, we therefore show how the unsupervised segmentation map is helpful. Figure 3.4 visualizes these segmentation maps outputted by the USA module. Please note that we use clean images here instead of noisy images to avoid any interruption by noise. While the segmentation itself is apparently inaccurate because it never sees any supervision, we observe those images to be partitioned into multiple segments (especially the salient objects), and each segment to usually contain pixels of the same semantic characteristics. Therefore, despite using no supervised label, the USA modules still manages to learn object saliency, as well as pixel-level semantic coherency,

which are useful to guide the denoising module to preserve important details.

4. PROPOSED EXPERIMENTS, DATASETS AND RESULTS

4.1 Datesets

PASCAL-VOC 2012[1] is a standardized dataset that is widely used for object class recognition and for assessing and comparing different methods. This dataset includes 20 object classes and one background class with 11,530 images that contain 6,929 segmentations.

We propose to use the PASCAL-VOC 2012 training set as the input of the U-SAID denoiser. Gaussian i.i.d. noise with zero mean and standard deviation σ is added to the images to synthesize the noisy input image during training. The testing set is generated similarly by adding noise on the PASCAL-VOC 2012 validation set.

4.2 Training Strategy

We train the cascade of denoising network and USA module in an end-to-end manner, while fixing the weights in the latter module. The overall loss for U-SAID is: $L_{MSE} + \gamma L_{USA}$, with the default $\gamma = 1$ unless otherwise specified. We use the Adam solver to train both the denoiser part and the USA module. The batch size is 16. The input patches are set to be 48×48 pixels (patches are randomly sampled from images with a stride of 1). The initial learning rate is set as 1e-3 for all learnable parts of U-SAID, using a multi-step learning decay strategy, i.e. dividing the learning rate by 10 at epoch 10, 40 and 80, respectively. The training is terminated after 100 epochs.

4.3 Experiments

4.3.1 Denoising Study on PASCAL-VOC

We compare U-SAID with the original CDnCNN-B (re-trained on our training set) [7], which requires no segmentation information at all. We further create another denoiser following the same idea of [1]: cascading CDnCNN-B with the supervised segmentation network (i.e., replacing L_{USA} with a standard pixel-wise softmax loss), with all other training protocols and initialization the same as U-SAID. We call it *supervised segmentation-aware image denoising* (**S-SAID**), and

[1] `http://host.robots.ox.ac.uk/pascal/VOC/voc2012/`

Table 4.1: The average image denoising performance comparison on PASCAL-VOC 2012 validation set, with σ = 15, 25, 35.

		CDnCNN-B	S-SAID	U-SAID
σ=15	PSNR (dB)	**33.56**	33.40	33.50
	SSIM	**0.9159**	0.9136	0.9153
	NIQE	4.3290	4.0782	**4.0049**
σ=25	PSNR (dB)	**31.18**	31.01	31.13
	SSIM	**0.8725**	0.8698	0.8724
	NIQE	4.2247	**3.8508**	3.8975
σ=35	PSNR (dB)	**29.65**	29.47	29.59
	SSIM	0.8344	0.8312	**0.8347**
	NIQE	4.1022	**3.6679**	3.7612

train it with the hybrid MSE-segmentation loss (the two losses are weighted equally), using the ground-truth segmentation maps available on the PASCAL training set. **Note that S-SAID is the only method that exploits "true" segmentation information**, making it a natural baseline for U-SAID to show the effect of such *extra information*. We do not include other denoising methods such as [11, 16, 15] because: 1) their average performance was shown to be worse than CDnCNN; and 2) most of them are not designed for the blind denoising scenario, thus hard to make fair comparisons. We have exhaustively tuned the hyper-parameters (learning rates, etc.) for CDnCNN-B and S-SAID, to ensure the optimal performance of either baseline.

The typical metric used for image denoising is PSNR, which has been shown to correlate poorly with human assessment of visual quality [35]. On the other hand, in the metric of PSNR, a model trained by minimizing MSE on the image domain should always outperform a model trained by minimizing a hybrid weighted loss. Therefore, we emphasize that the goal of our following experiments is not to pursue the highest PSNR, but to quantitatively demonstrate the different behaviors between models with and without segmentation awareness.

Table 4.1 reports the denoising performance in terms of PNSR, SSIM and Naturalness Image Quality Evaluator (NIQE) [8]. The last one is a well-known no-reference image quality score to indicate the perceived "naturalness" of an image: a smaller score indicates better perceptual quality.

Please note that the values in red and bold in the table indicate the best performance and the blue underlined values indicate the second best performance (the same hereinafter). Our observations from Table 4.1 are summarized as below:

- Since CDnCNN-B is optimized towards the MSE loss, it is not surprising that it consistently achieves the best PSNR results among all. However, U-SAID is able to achieve *only marginally inferior* PSNR/SSIMs to CDnCNN-B, which usually surpass S-SAID.
- The two methods with segmentation awareness (U-SAID and S-SAID) are significantly more favored by NIQE, showing a large margin over CDnCNN-B (e.g., nearly 0.4 at $\sigma = 25$). That testifies the benefits of considering high-level tasks for denoising.
- While not exploiting the true segmentation maps during training as S-SAID did, the performance of U-SAID is almost as competitive as S-SAID under the NIQE metric. In other words, *we did not lose much without using the true segmentation as supervision*.

4.3.2 Ablation Study on "Unsupervised Segmentation"

In training U-SAID above, we have used the "true" class number $K = 21$. It is then to our curiosity that: is this ground-truth value really best for training denoisers? Or, if the class number information cannot be accurately inferred when tackling general images, how much the denoising performance might be affected?

K	10	15	20	21 (default)	22	25	40
NIQE	3.9878	**3.8320**	4.0783	3.8975	3.8455	4.1139	3.9746
PSNR	31.00	31.06	30.99	**31.13**	31.01	30.99	30.98

Table 4.2: Ablation study of varying K in U-SAID training.

We hereby present an ablation study, by training several U-SAID models with different K values (all else remain unchanged), and compare their denoising performance on the testing set,

as displayed in Table 4.2. It is encouraging to observe that, the U-SAID denoising performance (PSNR and SSIM) consistently increase as K grows from smaller values (10, 15) towards the true value (21), and then gradually decreases as K get further larger. The NIQE values show the similar first-go-up-then down trend, except the peak slightly shifted to 15. That acts as a side evidence that rather than learning a semantically blind discriminator, the USA module indeed picks up the semantic class information and benefits from the correct K estimate. On the other hand, the variations of denoising performance w.r.t K are mild and smooth, showing certain robustness to inaccurate Ks too.

4.3.3 More Comparison to Relevant Methods

To solidify our results, we include more off-the-shelf denoising methods for comparison. We performed these experiments on Kodak dataset with three test sigmas 15, 25 and 35. A detailed comparison for each method we use is shown in Table 4.3. However, all methods we mentioned previously, i.e. CDnCNN-B, S-SAID and U-SAID, are blind to the noise level, the competing methods are non-blind denoisers. Therefore, we created two settings to simulate blind denoising:

- Applying the median sigma as denoising input, i.e. $\sigma = 25$;
- Assuming the oracle sigma is known in denoising

The second setting is apparently unfair to our blind model. Even so, we demonstrate the results in Table 4.4, from which U-SAID constantly yields the best performance.

4.3.4 Segmentation Study on PASCAL-VOC

We next investigate the effectiveness of denoising as a pro-processing step for the semantic segmentation over noisy images, which follows the setting in [1]. We first pass the noisy images in the PASCAL-VOC testing set through each of the three learned denoisers (CDnCNN-B, S-SAID, and U-SAID). We then apply a FPN *pre-trained on the clean PASCAL-VOC 2012 training set*, on the denoised testing sets, and evaluate the segmentation performance in terms of mean intersection-over-union (mIOU).

Table 4.3: Comparison of different methods if it is i) deep learning based, ii) semantic-aware denoising methods, or iii) requires extra segmentation annotation.

	Deep Learning	Semantic -Aware	Segmentation Annotation
U-SAID	✓	✓	
S-SAID	✓	✓	✓
CDnCNN-B	✓		
MLP [16]	✓		
MC-WNNM [36]			
CBM3D [37]			

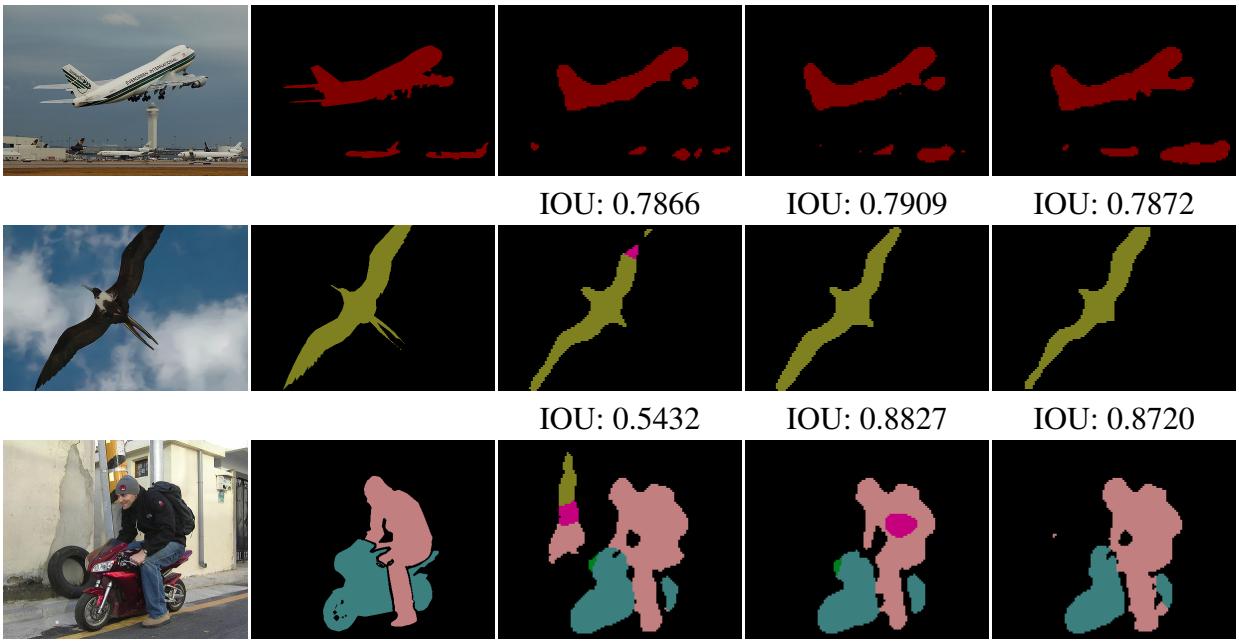
Table 4.4: The average Image denoising performance comparison in NIQE/ PSNR on the Kodak dataset, with noise σ = 15, 25, 35, respectively.

Setting I			
	σ=15	σ=25	σ=35
MLP [16]	4.3924/ 29.83	3.0205/ 30.09	6.5367/ 23.50
MC-WNNM [36]	5.6334 / 31.04	3.6731/ 31.35	8.6496/ 21.53
CBM3D [37]	3.7707/ 32.60	2.6152/ 31.81	6.7044/ 25.29
Setting II			
	σ=15	σ=25	σ=35
MLP [16]	4.675/ 29.11	3.008/ 30.09	3.070/ 28.67
MC-WNNM [36]	3.302/ 33.94	3.673/ 31.35	4.039/ 29.70
CBM3D [37]	2.6360/ 34.40	2.6620/ 31.81	2.6786/ 30.04

As compared in Table 4.1, when we apply the CDnCNN-B denoiser without considering high-level semantics, it easily fails to achieve high segmentation accuracy due to the artifacts introduced during denoising (even though those artifacts might not be reflected by PSNR or SSIM). With their segmentation awareness, both S-SAID and U-SAID have led to remarkably higher mIOUs. Most impressively, U-SAID is comparable to S-SAID, in spite of *the former never having seen true segmentation information on this dataset (training set)*, whilst the latter has. Figure 4.1 has visually confirmed the impact of denoisers on the segmentation performance.

Table 4.5: Segmentation results (mIoU) after denoising noisy image inputs, averaged over Pascal VOC 2012 validation dataset.

	noisy	CDnCNN-B	S-SAID	U-SAID
σ=15	0.4227	0.4238	**0.4349**	0.4336
σ=25	0.4007	0.4003	**0.4084**	0.4047
σ=35	0.3667	0.3724	**0.3802**	0.3785



(a) Original (b) Groundtruth (c) C-DnCNNB (d) S-SAID (e) U-SAID

Figure 4.1: Visualized semantic segmentation examples from Pascal VOC 2012 validation set.

The first row is added with noise of σ = 15, the second row σ = 25 and the third row σ = 35. Columns (a) - (b) are the ground truth images and true segmentation maps; (c) -(e) are the results by applying the pre-trained segmentation model on the denoised images using (c) C-DnCNNB; (d) S-SAID; and (e) U-SAID.

4.3.5 Generalizability Study: Data, Semantics, and Task

In this section, we define and compare three aspects of general usability, which were often overlooked in previous research of learning-based denoisers:

- **Data Generalizability**: whether a denoiser trained on one dataset can be applicable to restoring another.

- **Semantic Generalizability**: whether a denoiser trained on one dataset can be effective in preserving semantics, as the preprocessing step for applying semantic segmentation over another noisy dataset (with unseen classes).

- **Task Generalizability**: whether a denoiser trained with segmentation awareness can also be effective as preprocessing for other high-level tasks over noisy images.

Throughout the whole section below, all three denoisers used are the same models trained on PASCAL-VOC 2012 above. **There is no re-training involved**. Our hypothesis is that since U-SAID is not trained with any annotation on the original training set, it may less likely overfit the training set's semantics than S-SAID, while still preserving discriminative features, and hence could generalize better to various unseen data, semantics and tasks.

4.3.5.1 Denoising Unseen Noisy Datasets

We evaluate the denoising performance over the widely used Kodak dataset[2], consisting of 24 color images. Table 4.6 reports the quantitative results, which show strong consistency across all three noise levels: CDnCNN-B achieves the highest PSNR and SSIM values, while S-SAID performs the best in terms of NIQE. Interestingly, U-SAID seems to be the *"balanced"* solution in terms of data generalizability: it tends to obtain very close PSNR and SSIM values compared to CDnCNN-B, while producing comparable or even better NIQE values to S-SAID (especially at smaller σs). We further observe that U-SAID is usually able to preserve sharper edges and textures than CDnCNN-B, sometimes even better than S-SAID. Figure 4.2 displays a group of examples, where U-SAID finds clear advantages in preserving local fine details on the sail. Please referring more visualizations to 4.3.

PSNR = 34.31	PSNR = 34.02	PSNR = 34.27	
NIQE = 3.01	NIQE = 2.76	NIQE = 2.71	$\sigma = 25$
(a) CDnCNN-B	(b) S-SAID	(c) U-SAID	(d) Ground Truth

Figure 4.2: Visual comparison on one Kodak image at σ = 25.

We show the full images (top) and zoom-in regions (bottom) of the ground truth as well as three denoised images by CDnCNN-B, S-SAID and U-SAID (Best viewed on high-resolution color display, lower NIQE is better). Their corresponding segmentation label maps are shown below. The zoom-in region is displayed in the green box.

4.3.5.2 Denoising for Unseen Dataset Segmentation

We choose two real-world datasets, whose class categories are substantially different from PASCAL VOC: i) The ISIC 2018 dataset [38][3]. We choose the validation set of Task 1: Lesion Segmentation, whose goal is to predict lesion segmentation boundaries from dermoscopic lesion images; ii) The DeepGlobe dataset[4]. We choose the validation set of Track 3: Land Cover Classification, whose goal is to predict a pixel-level mask of land cover types (urban, agriculture,

[2] `http://r0k.us/graphics/kodak/`
[3] `https://challenge2018.isic-archive.com`
[4] `http://deepglobe.org`

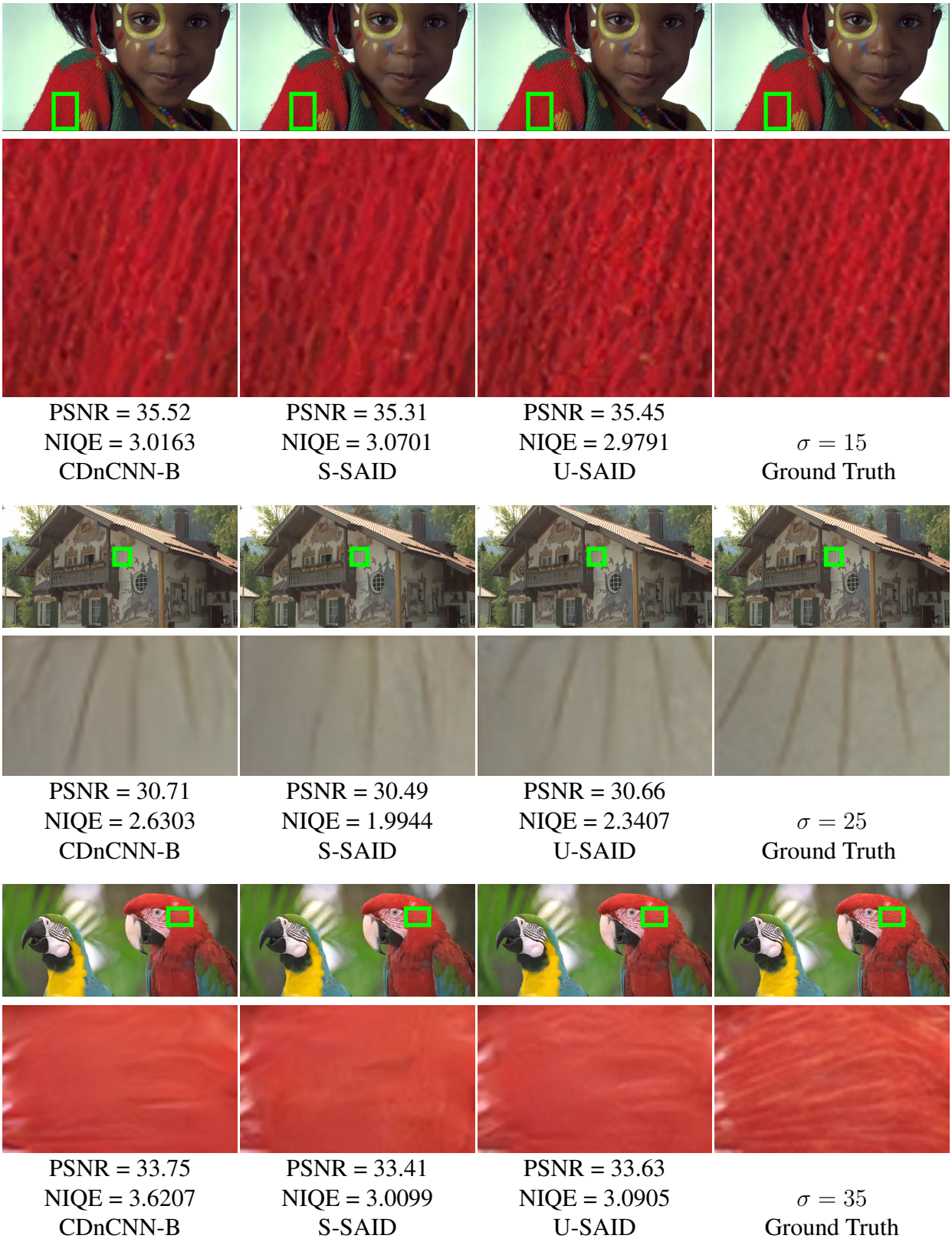


Figure 4.3: More denoised visualizations from Kodak data set by CDnCNN, S-SAID and U-SAID.

Table 4.6: The average Image denoising performance comparison on the Kodak dataset, with noise σ = 15, 25, 35, respectively.

		CDnCNN-B	S-SAID	U-SAID
σ=15	PSNR	**34.75**	34.57	34.62
	SSIM	**0.9242**	0.9217	0.9222
	NIQE	2.7570	2.6288	**2.5690**
σ=25	PSNR	**32.27**	32.07	32.17
	SSIM	**0.8812**	0.8770	0.8790
	NIQE	2.8493	**2.6006**	2.6355
σ=35	PSNR	**30.69**	30.48	30.50
	SSIM	**0.8418**	0.8366	0.8395
	NIQE	2.9753	**2.5619**	2.6687

Figure 4.4: Example image from ISIC 2018 (left: dermoscopic lesion image) and DeepGlobe (right: land satellite image) dataset.

rangeland, forest, water, barren, and unknow) from satellite images. Example images can be found in Fig 4.4.

We add σ = 25 noise to both validation sets, to create unseen testing sets for the trained denoisers. For either denoised validation set, we apply a pyramid scene parsing network (PSPNet) [39], that is pre-trained on the original clean training set. Table 4.7 reports the generalization effects of three denoisers when serving as preprocessing for segmenting unseen noisy datasets: U-SAID performs the best on both datasets, again verifying the benefits of segmentation awareness (that comes "for free" with no knowledge of true segmentation on any dataset). What is noteworthy, while we

Table 4.7: Segmentation results (mIoU) on denoised images of ISIC 2018 and DeepGlobe validation sets.

	noisy	CDnCNN-B	S-SAID	U-SAID
ISIC 2018	0.8061	0.8076	0.8084	**0.8095**
DeepGlobe	0.1309	0.4260	0.4198	**0.4263**

observe in the PASCAL-VOC segmentation experiment that the fully-supervised S-SAID is always superior to the segmentation-unaware CDnCNN-B, it is no longer always the case when applied to unseen datasets of different semantic categories: even CDnCNN-B is able to outperform S-SAID on DeepGlobe. Our hypothesis is that, the full supervision of S-SAID might cause its certain overfitting with PASCAL-VOC object categories. Trained in the unsupervised fashion but still equipped with segmentation awareness, U-SAID is not closely tied with original class semantics on the training set, and might thus generalize better to extracting and preserving semantics from new categories.

4.3.5.3 Denoising for Unseen High-Level Tasks

We now investigate if the segmentation-aware image denoising can also enhance other high-level vision applications, and choose classification and detection as two representative examples. While also listing PSNR and SSIM, we primarily focus on comparing their utility metrics (i.e., accuracy and mAP).

For classification, We choose the challenging CIFAR-100 dataset and add $\sigma = 25$ noise to its validation set. We then pass it through three denoisers, followed by a ResNet-110 classification model, pre-trained on the clean CIFAR-100 training set. As seen from Table 4.8, while U-SAID is second best in terms of both PSNR and SSIM (marginally inferior to CDnCNN-B), it demonstrates a notable boost in terms of both top-1 and top-5 accuracies, with a good margin compared to CDnCNN-B and S-SAID. While S-SAID also outperforms CDnCNN-B in improving classification, U-SAID proves to have even better generalizablity here.

For detection, We choose the MS COCO benchmark [40], and add $\sigma = 15, 25, 35$ noise to its

Table 4.8: Classification results after denoising noisy image inputs (σ = 25) from CIFAR-100.

	noisy	CDnCNN-B	S-SAID	U-SAID
PSNR	20.17	**29.13**	28.94	28.98
SSIM	0.6556	**0.9232**	0.9203	0.9219
Top-1 Acc	11.99	56.86	57.87	**58.16**
Top-5 Acc	29.83	82.64	83.65	**83.70**

Table 4.9: Detection results after denoising noisy MS COCO images.

		noisy	CDnCNN-B	S-SAID	U-SAID
σ = 15	PSNR	24.61	**35.14**	34.92	35.01
	SSIM	0.4796	**0.9440**	0.9410	0.9411
	mAP	0.5110	0.5573	0.5565	**0.5590**
σ = 25	PSNR	20.17	**32.70**	32.48	32.60
	SSIM	0.3233	**0.9137**	0.9095	0.9108
	mAP	0.4401	0.5296	0.5268	**0.5330**
σ = 35	PSNR	17.25	**31.12**	30.89	31.02
	SSIM	0.2383	**0.8861**	0.8803	0.8821
	mAP	0.3663	0.5023	0.4972	**0.5056**

validation set. We evaluate three denoisers in the same way as for the classification experiment, using a pre-trained YOLOv3 detection model [41]. Table 4.9 shows consistent observations as above: U-SAID always leads to the largest improvements in the detection mean average prediction (mAP), and hence has the best task generalizablity among all. Another interesting observation is that S-SAID is not as competitive as CDnCNN-B for the detection task, which we leave for future work to explore.

Both experiments show that the high-level semantics of different tasks are highly transferable for U-SAID, in terms of low-level vision tasks, as in line with [1].

4.3.6 Statistical Significance Study of U-SAID's Improvement

How consistent and statistically meaningful is U-SAID's performance advantage? To answer this, we report the detailed statistics: (1) the p-values of the denoising quality improvement over

Table 4.10: Performance and variance on three different tasks

	CDnCNN-B	S-SAID	U-SAID
PASCAL VOC Segmentation			
mIOU	39.46%	40.19%	**40.35%**
Variance	3.30E-6	3.98E-6	3.15E-6
Cross-set Kodak Denoising			
NIQE	2.87	**2.60**	2.62
Variance	1.74E-4	1.78E-4	6.00E-4
Cross-task CIFAR-100 classification			
top-1 Accuracy	56.89%	57.82%	**58.47%**
top-1 Variance	0.03	0.06	0.02
top-5 Accuracy	82.89%	83.57%	**83.91%**
top-5 Variance	0.02	0.05	0.06

different testing images; and (2) the variance of the performance improvements with different simulated noise patterns, for three representative experiments: PASCAL VOC segmentation (Table 4.5), cross-set KODAK denoising (Table 4.6), and cross-task CIFAR-100 classification (Table 4.8). For each test, we simulated i.i.d. random Gaussian noise ($\sigma = 25$) for each image ten times, and repeat the experiments on them accordingly. Experiment results are shown in Table 4.10.

In the PASCAL VOC segmentation experiment, we performance hypothesis tests to check if U-SAID leads to better segmentation results than CDnCNN-B. Being 95% confident, we obtained p-value $= 1.7305E-9$, which demonstrates the statistical significance of improvement. On the other hand, U-SAID and S-SAID's results do not show significant difference with p-value $= 0.0744 > 0.05$. Without using any segmentation ground truth, our method achieved **statistically similar results** to S-SAID, even under a disadvantageous setting.

For the cross-set Kodak denoising experiment, the NIQE of U-SAID is statistically significantly better than that of CDnCNN-B, with p-value $= 2.6638E-16$. Similarly, S-SAID is better than U-SAID in NIQE with p-value $= 6.7845E-3$.

In CIFAR-100 experiment, for top-1 accuracy, U-SAID yields mean accuracy of 58.47%, which is significantly higher than DnCNN, which has mean = 56.89%, with p-value = 3.6147E-14.

U-SAID has also higher accuracy than S-SAID (mean = 57.82%) with p-value = 1.3486E-6. Similarly for top-5, U-SAID's performance (83.91%) is statistically significant better than DnCNN (82.89%), and S-SAID (83.57%), with p-values of 1.3982E-9 and 4.3994E-3, respectively.

5. CONCLUSION

This paper proposes a segmentation-aware image denoising model that requires no ground-truth segmentation map for training. The proposed U-SAID model leads to comparable performance with its supervised counterpart, in terms of both low-level (denoising) and high-level (segmentation) vision metrics, when trained on and applied to the same noisy dataset (without utilizing extra segmentation information as the latter has to). Furthermore, U-SAID shows remarkable generalizablity to unseen data, semantics, and high-level tasks, all of which endorse it to be a highly robust, effective and general-purpose denoising option.

REFERENCES

[1] D. Liu, B. Wen, X. Liu, Z. Wang, and T. S. Huang, "When image denoising meets high-level vision tasks: A deep learning approach," *arXiv preprint arXiv:1706.04284*, 2017.

[2] B. Cheng, Z. Wang, Z. Zhang, Z. Li, D. Liu, J. Yang, S. Huang, and T. S. Huang, "Robust emotion recognition from low quality and low bit rate video: A deep learning approach," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 65–70, IEEE, 2017.

[3] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, p. 7, 2017.

[4] Z. Fan, L. Sun, X. Ding, Y. Huang, C. Cai, and J. Paisley, "A segmentation-aware deep fusion network for compressed sensing mri," *arXiv preprint arXiv:1804.01210*, 2018.

[5] D. Liu, B. Wen, J. Jiao, X. Liu, Z. Wang, and T. S. Huang, "Connecting image denoising and high-level vision tasks via deep learning," *arXiv preprint arXiv:1809.01826*, 2018.

[6] D. Liu, B. Cheng, Z. Wang, H. Zhang, and T. S. Huang, "Enhance visual recognition under adverse conditions via deep networks," *IEEE Transactions on Image Processing*, 2019.

[7] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[8] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a" completely blind" image quality analyzer.," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.

[9] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Computer Vision, 1998. Sixth International Conference on*, pp. 839–846, IEEE, 1998.

[10] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: nonlinear phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[11] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[12] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, 2006.

[13] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 2272–2279, IEEE, 2009.

[14] B. Wen, S. Ravishankar, and Y. Bresler, "Structured overcomplete sparsifying transform learning with convergence guarantees and applications," *Int. J. Computer Vision*, vol. 114, no. 2, pp. 137–167, 2015.

[15] S. Gu, L. Zhang, W. Zuo, and X. Feng, "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.

[16] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with bm3d?," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2392–2399, IEEE, 2012.

[17] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in neural information processing systems*, pp. 2802–2810, 2016.

[18] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *Pattern recognition (icpr), 2010 20th international conference on*, pp. 2366–2369, IEEE, 2010.

[19] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, Springer, 2016.

[20] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4792–4800, 2016.

[21] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," *arXiv preprint arXiv:1803.03345*, 2018.

[22] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2019.

[23] R. G. VidalMata, S. Banerjee, B. RichardWebster, M. Albright, P. Davalos, S. McCloskey, B. Miller, A. Tambo, S. Ghosh, S. Nagesh, *et al.*, "Bridging the gap between computational photography and visual recognition," *arXiv preprint arXiv:1901.09482*, 2019.

[24] S. Li, I. B. Araujo, W. Ren, Z. Wang, E. K. Tokuda, R. H. Junior, R. Cesar-Junior, J. Zhang, X. Guo, and X. Cao, "Single image deraining: A comprehensive benchmark analysis," *arXiv preprint arXiv:1903.08558*, 2019.

[25] Y. Yuan, W. Yang, W. Ren, J. Liu, W. J. Scheirer, and Z. Wang, "Ug^{2+} track 2: A collective benchmark effort for evaluating and advancing image understanding in poor visibility environments," *arXiv preprint arXiv:1904.04474*, 2019.

[26] A. W. Harley, K. G. Derpanis, and I. Kokkinos, "Segmentation-aware convolutional networks using local attention masks," in *IEEE International Conference on Computer Vision (ICCV)*, vol. 2, p. 7, 2017.

[27] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. J. Belongie, "Learning from noisy large-scale datasets with minimal supervision.," in *CVPR*, pp. 6575–6583, 2017.

[28] Z. Lu, Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao, "Learning from weak and noisy labels for semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 3, pp. 486–500, 2017.

[29] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *arXiv preprint arXiv:1412.6596*, 2014.

[30] W. Li, X. Qian, and J. Ji, "Noise-tolerant deep learning for histopathological image segmentation," in *Image Processing (ICIP), 2017 IEEE International Conference on*, pp. 3075–3079, IEEE, 2017.

[31] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," *arXiv preprint arXiv:1406.2080*, 2014.

[32] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection.," in *CVPR*, vol. 1, p. 4, 2017.

[33] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," *CVPR*, 2018.

[34] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[35] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

[36] J. Xu, L. Zhang, D. Zhang, and X. Feng, "Multi-channel weighted nuclear norm minimization for real color image denoising," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1096–1104, 2017.

[37] K. Dabov, A. Foi, V. Katkovnik, and K. O. Egiazarian, "Color image denoising via sparse 3d collaborative filtering with grouping constraint in luminance-chrominance space.," in *ICIP (1)*, pp. 313–316, 2007.

[38] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pp. 168–172, IEEE, 2018.

[39] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881–2890, 2017.

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[41] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.