Task-specific Scene Structure Representations*

Seunghyun Shin, Jisu Shin, and Hae-Gon Jeon

AI Graduate School, GIST, South Korea
{seunghyuns98,jsshin98}@gm.gist.ac.kr, haegonj@gist.ac.kr

Abstract. Understanding the informative structures of scenes is essential for low-level vision tasks. Unfortunately, it is difficult to obtain a concrete visual definition of the informative structures because influences of visual features are task-specific. In this paper, we propose a single general neural network architecture for extracting task-specific structure guidance for scenes. To do this, we unfold the traditional graph-partitioning problem into a learnable network, named *Scene Structure Guidance Network (SSGNet)*, to represent the task-specific informative structures. In addition, our SSGNet is light-weight ($\sim$ 55K parameters), and can be used as a plug-and-play module for off-the-shelf architectures. Our main contribution is to show that such a simple network can achieve state-of-the-art results for several low-level vision applications including joint upsampling and image denoising even on unseen datasets, compared to existing methods which use structural embedding frameworks.

Keywords: Low-level Vision · Structure Guidance · Unsupervised

1 Introduction

Methods for estimating scene structures have attracted wide research attention for the past several decades. Clearly, the goodness of scene structures depends on the target applications, and is defined by either training data or objective functions. Nevertheless, the question of how to effectively exploit structure guidance information remains unanswered.

In this paper, we propose a *Scene Structure Guidance Network (SSGNet)*, a single general neural network architecture for extracting task-specific structural features of scenes. Our SSGNet is lightweight in both size and computation, and is a plug-and-play module that can be applied to any baseline low-level vision architectures. The SSGNet computes a set of parameterized eigenvector maps, whose combination is selectively determined in favor of the target domain. To achieve this, we introduce two effective losses: (1) *Eigen loss*, motivated by the traditional graph partitioning problem [21], forms a basis set of scene structures based on weight graphs on an image grid. (2) *Spatial loss* enforces the sparsity of each eigenvector for diverse representations of scene structures. We note that, without any supervision, our SSGNet can successfully learn to generate

* This paper is the short version of AAAI'23 and is NEVER considered an official publication.

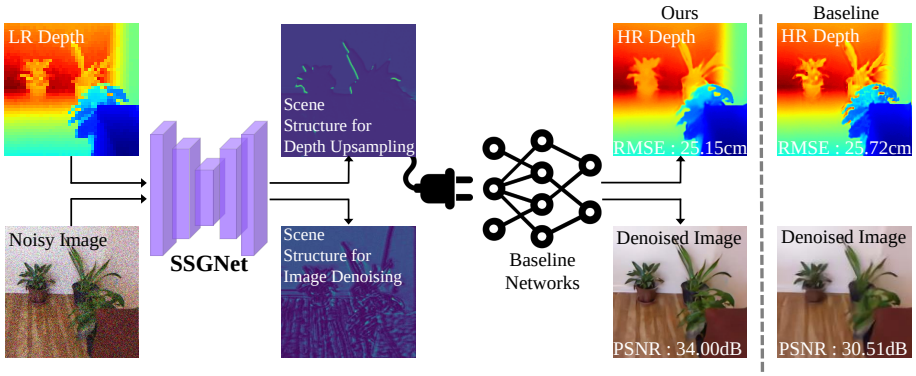


Fig. 1. Our SSGNet is a lightweight architecture and can be applied as a plug-and-play module to improve the performance of baseline networks for low-level vision tasks.

task-specific and informative structural information as shown in Figure 1. To demonstrate the wide applicability of our SSGNet, we conduct extensive experiments on several low-level vision applications, including joint upsampling and image denoising, and achieve state-of-the-art results, even in cross-dataset generalization.

2 Related Work

2.1 Low-level vision tasks

The goal of low-level vision tasks such as denoising, super-resolution, deblurring and inpainting is to recover a sharp latent image from an input image that has been degraded by the inherent limitations of the acquisition systems. In the past decade, there have been significant improvements in low-level vision tasks, and recently deep learning-based techniques have especially proven to be powerful systems.

With the help of inductive bias [4], convolutional neural networks (CNNs) with a pixel-wise photo consistency loss [16] are adopted. To mitigate the issue on inter-pixel consistency on CNNs, generative adversarial networks (GANs) [8]-based methods are proposed to produce visually pleasing results with perceptual losses [14] based on high-level semantic features. Nowadays, a vision transformer (ViT) [6] has been used to capture both local and global image information by leveraging the ability to model long-range context.

Such approaches have shown good progress with structural details. For regularization, adding robust penalties to objective functions [23] suppresses high-frequency components, and hence the results usually provide a smooth plausible reconstruction. However, those constraints often suffer from severe overfitting to noisy labels and are sensitive to hyperparameters, which leads to a lack of model generality.

2.2 Structural information

Extensive studies on low-level vision have verified the feasibility and necessity of the image prior including image edges and gradients. One of the representative works involves joint image filters which leverage a guidance image as a prior and transfer its structural details to a target image for edge-preserved smoothing [24, 11, 28].

Such structure information can be defined in practice, depending on the tasks. Both super-resolution [20, 22, 26, 7] and image denoising [17], which utilize a patch similarity, generate gradient maps to reconstruct high frequency details or suppress image noises. Works in [9, 13] infer object boundaries to refine initial predictions in visual perception tasks, including depth estimation/completion. Also, image inpainting [19, 27, 10, 2], filling in missing parts of corrupted scenes, adopt edge maps from traditional method like Canny edge detector [1] to hallucinate their own scene structures.

In spite of promising results from the state-of-the-art methods learning meaningful details for each task, they require a high modeling capacity with numerous parameters and ground-truth structure maps for training. In contrast, our SSGNet, a very small network generating scene structures without any supervision, has advantages for various low-level vision tasks, simply by embedding as an additional module.

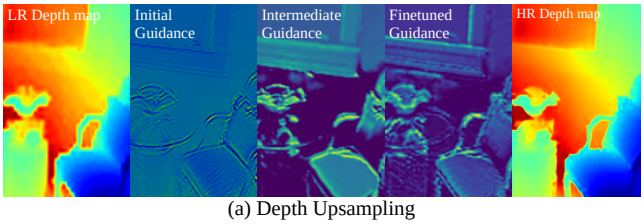
3 Methodology

Motivation Spectral graph theory proves that the eigenvectors of the graph Laplacian yield minimum-energy graph partitions, and each smallest eigenvector partitions the graph into soft-segments based on its adjacent matrix.

In the image domain, a reference pixel and its similarity to neighboring pixels can be interpreted as a node and edges in a graph, respectively. In general, affinity is defined by appearance similarities (the absolute of intensity differences). With this motivation, images can be decomposed into soft image clusters from a precomputed affinity matrix. In addition, scene configurations in images can be described as a set of eigenvectors whose smallest eigenvalues indicate connected components on the affinity matrix.

Scene Structure Guidance Network In this work, our goal is to train the proposed network, SSGNet, without any supervision because it is infeasible to define a unique objective function for a task-specific structure guidance. To accomplish this, we devise a learnable and parametric way of efficiently representing scene structures. The output of our SSGNet is associated with learnable weights that will be finetuned in accordance with an objective function of target applications. To optimize SSGNet in an unsupervised manner, we define a loss function $\mathcal{L}_{ssg}$ which is a linear combination of two loss terms as follows:

Eigen Loss The main objective of SSGNet is to obtain a set of smallest eigenvectors $\mathbf{Y}$ of the graph Laplacian $\mathbf{L}$. Since an image is segmented based on a constructed affinity matrix in spectral graph theory, the form of the matrix depends on the pixel-level similarity encoding. In this work, we adopt the sparse



(a) Depth Upsampling

(b) Image Denoising

Fig. 2. Examples of task-specific scene structures: intial, intermediate and final results from SSGNet for (a) joint depth upsampling and (b) image denoising.

KNN-matting matrix [3], which collects nonlocal neighborhoods j of a pixel i by the k-nearest neighbor algorithm (KNN). Using the sparse KNN-matting matrix, we can take account of both spatial distance and color information with less computational cost than a traditional similarity matrix. The graph Laplacian $\mathbf{L}$ is finally obtained by $\mathbf{L} = \mathbf{D} - \mathbf{W}$.

We can finally obtain a set of eigenvectors $\mathbf{Y}$ by minimizing the quadratic form of $\mathbf{L}$, $\mathcal{L}_{eigen}$, as below:

$$\mathcal{L}_{eigen} = \sum_{k} \mathbf{Y}_k^T \mathbf{L} \mathbf{Y}_k. \tag{1}$$

Spatial Loss Our spatial loss $\mathcal{L}_{spatial}$ considers the sparsity of each eigenvector to enforce diverse representations of scene structure, defined as below:

$$\mathcal{L}_{spatial} = \sum_{k} (|\mathbf{Y}_k|^\gamma + |1 - \mathbf{Y}_k|^\gamma) - 1, \tag{2}$$

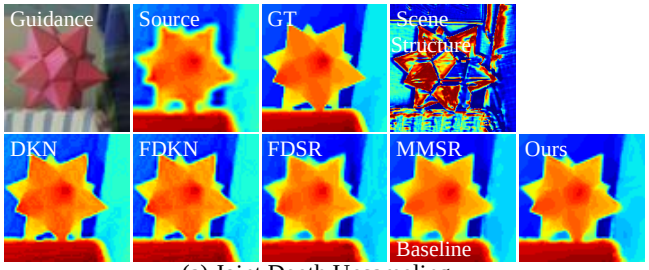
where $|\cdot|$ indicates an absolute value, and the hyperparameter γ is set to 0.9 in our implementation. With the $\mathcal{L}_{spatial}$ and the softmax operation together, it makes each pixel across the eigenvectors have different value due to the sparsity penalty, which produces diverse feature representations of image structures.

In total, the final loss function for SSGNet is defined as:

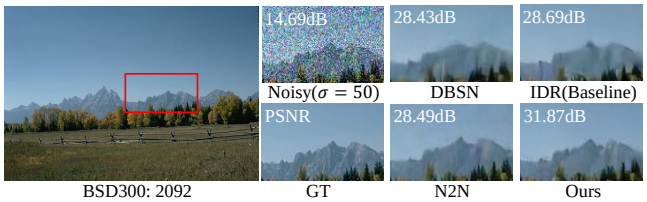
$$\mathcal{L}_{ssg} = \mathcal{L}_{eigen} + \lambda \mathcal{L}_{spatial} \tag{3}$$

where λ is the hyper-parameter, and is empirically set to 40.

Our SSGNet is pretrained on a single dataset and can be embedded in various baseline networks after passing through an additional single convolution layer which acts as an attention module. In favor of the target domain on each task,



(a) Joint Depth Upsampling



(b) Image Denoising

Fig. 3. Comparison results on (a) joint depth upsampling and (b) image denoising.

Dataset	Scale		Supervised			Self-Supervised	
			DKN[15]	FDKN[15]	FDSR[12]	MMSR[5]	Ours
2014	×4	RMSE	2.878	2.593	3.217	<u>1.953</u>	**1.819**
		MAE	0.739	0.659	0.595	<u>0.573</u>	**0.451**
	×8	RMSE	3.642	3.510	3.606	<u>2.765</u>	**2.714**
		MAE	0.775	0.871	0.885	<u>0.785</u>	**0.675**

Table 1. Quantitative results on joint depth upsampling tasks. The best and the second best results are marked as **bold** and <u>underlined</u>, respectively. (unit:cm)

this layer produces adaptive structural information of input scenes by linearly combining the set of eigenvectors. In Figure 2, we visualize how the eigenvectors from SSGNet change differently at each iteration during finetuning on each task. We claim that it is possible for our SSGNet to capture informative and task-specific structures through gradient updates from backpropagation.

4 Experiments

We conduct a variety of experiments on low-level vision tasks, including self-supervised joint depth upsampling and unsupervised single image denoising, to demonstrate the effectiveness of our SSGNet. Prior to the evaluations, we train our SSGNet on a well-known NYUv2 dataset. With the pre-trained weight of SSGNet, we embed it to the baseline networks, which are existing CNN architectures, and finetune on each task.

As shown in Table 1, MMSR [5] with our SSGNet embedded achieves the best performance over the comparison methods. In addition, as shown in Table 2, IDR [29] with our SSGNet embedded achieves the best performance among all the competitive methods regardless of the noise levels. Figure 3 also shows some

Method		Kodak		BSD300		BSD68	
		$\sigma = 25$	$\sigma = 50$	$\sigma = 25$	$\sigma = 50$	$\sigma = 25$	$\sigma = 50$
DBSN[25]	PSNR	32.07	28.81	31.12	27.87	28.81	25.95
	SSIM	0.875	0.783	0.881	0.782	0.818	0.703
N2N[18]	PSNR	**32.39**	29.23	31.39	28.17	29.15	26.23
	SSIM	**0.886**	0.803	0.889	0.799	0.831	0.725
IDR[29]	PSNR	32.36	29.27	31.48	28.25	29.20	26.25
	SSIM	0.884	0.803	0.890	0.802	**0.835**	0.726
Ours	PSNR	**32.39**	**29.34**	**31.52**	**28.33**	**29.25**	**26.36**
	SSIM	0.885	**0.806**	**0.891**	**0.805**	**0.835**	**0.731**

Table 2. Quantitative results on single image denoising.

example results. We highlight that the result demonstrates the strong generalization capabilities of our SSGNet on unseen data again.

5 Conclusion

In this paper, we present a single general network for representing task-specific scene structures. We cast the problem of the acquisition of informative scene structures as a traditional graph partitioning problem on the image domain, and solve it using a lightweight CNN framework without any supervision, *Scene Structure Guidance Network (SSGNet)*. Our SSGNet computes coefficients of a set of eigenvectors, enabling to efficiently produce diverse feature representations of a scene with our proposed two loss terms, the eigen loss and the spatial loss. We show the promising performance gains for both the joint depth upsampling and image denoising, even with the good cross-dataset generalization capability.
Remark This paper is a re-publishing of the paper which has been published in AAAI2023 by request of the IW-FCV2023 program committee to share the research results.

References

1. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **8**(6), 679–698 (1986)
2. Cao, C., Fu, Y.: Learning a sketch tensor space for image inpainting of man-made scenes. In: Proceedings of International Conference on Computer Vision (ICCV) (2021)
3. Chen, Q., Li, D., Tang, C.K.: Knn matting. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **35**(9), 2175–2188 (2013)
4. Cohen, N., Shashua, A.: Inductive bias of deep convolutional networks through pooling geometry. In: International Conference on Learning Representations (ICLR) (2017)
5. Dong, X., Yokoya, N., Wang, L., Uezato, T.: Learning mutual modulation for self-supervised cross-modal super-resolution. In: Proceedings of European Conference on Computer Vision (ECCV) (2022)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
7. Fang, F., Li, J., Zeng, T.: Soft-edge assisted network for single image super-resolution. IEEE Transactions on Image Processing (TIP) **29**, 4656–4668 (2020)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the Neural Information Processing Systems (NeurIPS) (2014)
9. Gu, S., Zuo, W., Guo, S., Chen, Y., Chen, C., Zhang, L.: Learning dynamic guidance for depth image enhancement. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
10. Guo, X., Yang, H., Huang, D.: Image inpainting via conditional texture and structure dual generation. In: Proceedings of International Conference on Computer Vision (ICCV) (2021)
11. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **35**(6), 1397–1409 (2012)
12. He, L., Zhu, H., Li, F., Bai, H., Cong, R., Zhang, C., Lin, C., Liu, M., Zhao, Y.: Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
13. Jin, L., Xu, Y., Zheng, J., Zhang, J., Tang, R., Xu, S., Yu, J., Gao, S.: Geometric structure based and regularized depth estimation from 360 indoor imagery. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
14. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of European Conference on Computer Vision (ECCV) (2016)
15. Kim, B., Ponce, J., Ham, B.: Deformable kernel networks for joint image filtering. International Journal on Computer Vision (IJCV) **129**(2), 579–600 (2021)
16. Li, Y., Huang, J.B., Ahuja, N., Yang, M.H.: Deep joint image filtering. In: Proceedings of European Conference on Computer Vision (ECCV) (2016)
17. Liu, Y., Anwar, S., Zheng, L., Tian, Q.: Gradnet image denoising. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) (2020)

18. Moran, N., Schmidt, D., Zhong, Y., Coady, P.: Noisier2noise: Learning to denoise from unpaired noisy data. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
19. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: Proceedings of International Conference on Computer Vision Workshop (ICCVW) (2019)
20. Pickup, L., Roberts, S.J., Zisserman, A.: A sampled texture prior for image super-resolution. In: Proceedings of the Neural Information Processing Systems (NeurIPS) (2003)
21. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **22**(8), 888–905 (2000)
22. Sun, J., Xu, Z., Shum, H.Y.: Image super-resolution using gradient profile prior. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
23. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
24. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Proceedings of International Conference on Computer Vision (ICCV) (1998)
25. Wu, X., Liu, M., Cao, Y., Ren, D., Zuo, W.: Unpaired learning of deep image denoising. In: Proceedings of European Conference on Computer Vision (ECCV) (2020)
26. Xie, J., Feris, R.S., Sun, M.T.: Edge-guided single depth image super resolution. IEEE Transactions on Image Processing (TIP) **25**(1), 428–438 (2015)
27. Yang, J., Qi, Z., Shi, Y.: Learning to incorporate structure knowledge for image inpainting. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2020)
28. Zhang, Q., Shen, X., Xu, L., Jia, J.: Rolling guidance filter. In: Proceedings of European Conference on Computer Vision (ECCV) (2014)
29. Zhang, Y., Li, D., Law, K.L., Wang, X., Qin, H., Li, H.: Idr: Self-supervised image denoising via iterative data refinement. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)