Task-Specific Scene Structure Representations

Jisu Shin*, Seunghyun Shin*and Hae-Gon Jeon†

AI Graduate School, GIST, South Korea
{jsshin98, seunghyuns98}@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 first analyze traditional spectral clustering methods, which computes a set of eigenvectors to model a segmented graph forming small compact structures on image domains. We then unfold the traditional graph-partitioning problem into a learnable network, named *Scene Structure Guidance Network (SSGNet)*, to represent the task-specific informative structures. The SSGNet yields a set of coefficients of eigenvectors that produces explicit feature representations of image structures. In addition, our SSGNet is light-weight (56K parameters), and can be used as a plug-and-play module for off-the-shelf architectures. We optimize the SSGNet without any supervision by proposing two novel training losses that enforce task-specific scene structure generation during training. 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. We also demonstrate that our SSGNet generalizes well on unseen datasets, compared to existing methods which use structural embedding frameworks. Our source codes are available at https://github.com/jsshin98/SSGNet.

1 Introduction

Methods for estimating scene structures have attracted wide research attention for the past several decades. As an example, texture representations based on image edges have been extensively studied with impressive performance on low-level vision tasks, *i.e.* image denoising (Tomasi and Manduchi 1998), deblurring (Krishnan and Fergus 2009; Levin et al. 2007), super-resolution (Tai et al. 2010) and inpainting (Nazeri et al. 2019; Yang, Qi, and Shi 2020; Guo, Yang, and Huang 2021). Another aspect of scene structures involves inferring robust object boundaries to quantify uncertainty and refine initial predictions in visual perception tasks including joint filtering (He, Sun, and Tang 2012; Guo et al. 2018; Li et al. 2016) and depth completion (Eldesokey et al. 2020). Clearly, the goodness of scene structures depends on the target applications, and is defined by either training data or objective functions.

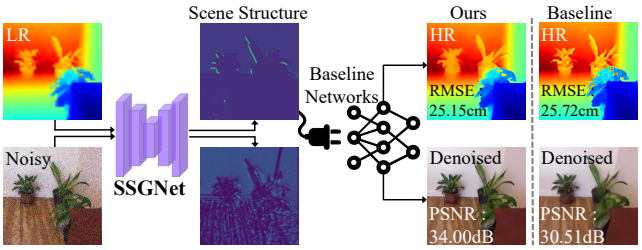


Figure 1: Our SSGNet is a lightweight architecture and can be applied as a plug-and-play module to improve the performance of baseline networks for low-level vision tasks.

More recent approaches of extracting informative scene structures have focused on capturing task-specific features from various learning frameworks. One interesting work for joint filtering in (de Lutio et al. 2022) builds graph nodes on learned features from a guidance image to encode semantic information, and represents scene structures by segmenting the graph edges based on objective functions. However, they have heavy computational burdens and are not implemented as an end-to-end architecture. To formulate an end-to-end architecture, edge priors, directly obtained from conventional edge detection (Irwin et al. 1968; Canny 1986), are used as a guide. Typically, image edges (or gradients) represent high frequency features and can be forced to generate fine details in the prediction results (Fang, Li, and Zeng 2020). Nevertheless, the question of how to effectively exploit structure guidance information remains unanswered. Tremendous efforts have been made to only generate a single purpose scene structure with each different architecture.

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

*These authors contributed equally.
†Corresponding author

achieve this, we introduce two effective losses: (1) *Eigen loss*, motivated by the traditional graph partitioning problem (Shi and Malik 2000), 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 task-specific and informative structural information as shown in Fig.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

Our work is closely related to scene structure embedding for low-level vision tasks.

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 (*i.e.* sensor size, depth of field or light efficiency). 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 (Cohen and Shashua 2017), convolutional neural networks (CNNs) with a pixel-wise photo consistency loss (Li et al. 2016; Zhang and Sabuncu 2018; Zhong et al. 2021) are adopted. To mitigate the issue on inter-pixel consistency on CNNs, generative adversarial networks (GANs) (Goodfellow et al. 2014; Zhu et al. 2017; Karras, Laine, and Aila 2019; Liu et al. 2021a; Wang et al. 2021)-based methods are proposed to produce visually pleasing results with perceptual losses (Johnson, Alahi, and Fei-Fei 2016; Fuoli, Van Gool, and Timofte 2021; Suvorov et al. 2022) based on high-level semantic features. Nowadays, a vision transformer (ViT) (Dosovitskiy et al. 2021; Liu et al. 2021b; Caron et al. 2021; Chen et al. 2021) 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 (Tibshirani 1996; Xu et al. 2010; Loshchilov and Hutter 2019; de Lutio et al. 2022) 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 (Tomasi and Manduchi 1998; He, Sun, and Tang 2012; Zhang et al. 2014).

Such structure information can be defined in practice, depending on the tasks. Both super-resolution (Pickup, Roberts, and Zisserman 2003; Sun, Xu, and Shum 2008; Xie, Feris, and Sun 2015; Fang, Li, and Zeng 2020) and image denoising (Liu et al. 2020), which utilize a patch similarity, generate gradient maps to reconstruct high frequency details or suppress image noises. Works in (Gu et al. 2017; Jin et al. 2020) infer object boundaries to refine initial predictions in visual perception tasks, including depth estimation/completion. Also, image inpainting (Nazeri et al. 2019; Yang, Qi, and Shi 2020; Guo, Yang, and Huang 2021; Cao and Fu 2021), filling in missing parts of corrupted scenes, adopt edge maps from traditional method like Canny edge detector (Canny 1986) 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

Motivated by spectral graph theory (Shi and Malik 2000; Levin, Rav-Acha, and Lischinski 2008; Levin, Lischinski, and Weiss 2007), a set of basis represents scene configurations as a linear combination of the basis. Such parameterization provides a restrictive solution space to accommodate semantic entities like textures and object boundaries. Following the works in (Tang and Tan 2019; Bloesch et al. 2018), we begin with an introduction to spectral methods, and then parameterize scene structures which can be used as guidance for various vision tasks.

3.1 Motivation

Let us set a weighted undirected graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ in an arbitrary feature space with a set of nodes $\mathbf{V}$, and a set of edges $\mathbf{E}$, whose weight can be represented as an $N \times N$ non-negative adjacency matrix $\mathbf{W} = \{w(i,j) : (i,j) \in \mathbf{E}\}$ where i, j denote graph nodes. The Laplacian matrix L of this graph is then obtained by $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with the row-wise sum of $\mathbf{W}$ on its diagonal. Since the Laplacian matrix is a positive semidefinite matrix, for every N dimensional vector y from the matrix $\mathbf{Y}$ which consists of a set of vectors, it holds that

$$y^T \mathbf{L} y = \sum_{(i,j)\in\mathbf{E}} w(i,j)\{y(i) - y(j)\}^2 \geq 0. \qquad (1)$$

To minimize the Eq.(1), the indicator vector y should take similar values for nodes i and j. When the adjacent value $w(i,j)$ is high, the two nodes are more tightly coupled.

Spectral graph theory in (Fiedler 1973; Shi and Malik 2000) proves that the eigenvectors of the graph Laplacian yield minimum-energy graph partitions, and each smallest

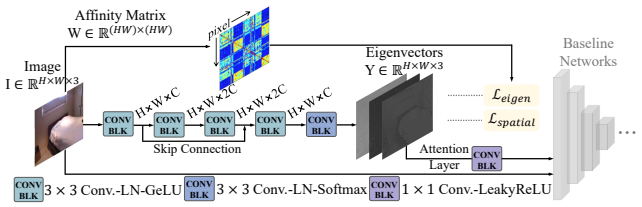


Figure 2: An overview of SSGNet. LN, GeLU, and LeakyReLU denote the layer normalization, GeLU activation, and LeakyReLU activation, respectively. The eigenvectors are integrated via the attention layer, and then embedded to any baseline network.

eigenvector, like the indicator vector y, 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 (Boykov, Veksler, and Zabih 2001), respectively. In general, affinity is defined by appearance similarities (*i.e.* the absolute of intensity differences). With this motivation, images can be decomposed into soft image clusters from a pre-computed 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.

3.2 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.

Given single color images $\mathbf{I} \in \mathbb{R}^{h\times w\times 3}$, our SSGNet σ yields a set of eigenvectors $\mathbf{Y} \in \mathbb{R}^{h\times w\times n}$, where n denotes the number of eigenvectors and is empirically set to 3:

$$\mathbf{Y} = \sigma(\mathbf{I}). \tag{2}$$

As illustrated in Fig.2, SSGNet takes a simple encoder-decoder architecture ($\sim$ 56K), consisting of two 3×3 convolutional layers and three 3×3 deconvolutional layers with layer normalizations (Ba, Kiros, and Hinton 2016) and gelu activations (Hendrycks and Gimpel 2016) after each layer except for the last softmax layer. 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 $\mathbf{L}$, inspired by the spectral graph theory (Fiedler 1973; Shi and Malik 2000; Levin, Rav-Acha, and Lischinski 2008).

To generate the graph Laplacian $\mathbf{L}$, we trace back all the way down to some traditional similarity matrix methods. 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 (Levin, Rav-Acha, and Lischinski 2008; Levin, Lischinski, and Weiss 2007; Chen, Li, and Tang 2013). In this work, we adopt the sparse KNN-matting matrix (Chen, Li, and Tang 2013). To be specific, we first collect nonlocal neighborhoods j of a pixel i by the k-nearest neighbor algorithm (KNN) (Cover and Hart 1967). Then, we define the feature vector $\varphi(i)$ at a given pixel i as follows:

$$\varphi(i) = (r, g, b, d_x, d_y)_i, \tag{3}$$

where (r, g, b) denotes each color channel, and (d_x, d_y) is a weighted spatial coordinate for the x- and y-axes. We follow the KNN kernel function KNN(i) to construct the sparse affinity matrix $\mathbf{W}$ based on feature vectors φ:

$$\mathbf{W}(i,j) = \begin{cases} 1 - \parallel \varphi(i) - \varphi(j) \parallel, & j \in \text{KNN}(i) \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $j \in$ KNN(i) are the k-nearest neighbors of i based on the distance defined by φ. 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 $\mathbf{L} = \mathbf{D} - \mathbf{W}$ as the same manner, described in Sec.3.1.

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

$$\mathcal{L}_{eigen} = \sum_k \mathbf{Y}_k^T \mathbf{L} \mathbf{Y}_k. \tag{5}$$

However, we observe that SSGNet sometimes produces undesirable results during the training phase because of the degenerate case, where the rank of $\mathbf{Y}$ may be lower, and needs an additional loss term to regularize it.

Spatial Loss Since our SSGNet uses a softmax function in the last layer to prevent the eigenvectors from converging to zero vectors, we only need to handle the degenerate case, where all eigenvectors have the same value. Our spatial loss $\mathcal{L}_{spatial}$ considers the sparsity of each eigenvector to enforce diverse representations of scene structure, defined as below:

$$\mathcal{L}_{spatial} = \sum_k (|\mathbf{Y}_k|^\gamma + |1 - \mathbf{Y}_k|^\gamma) - 1, \tag{6}$$

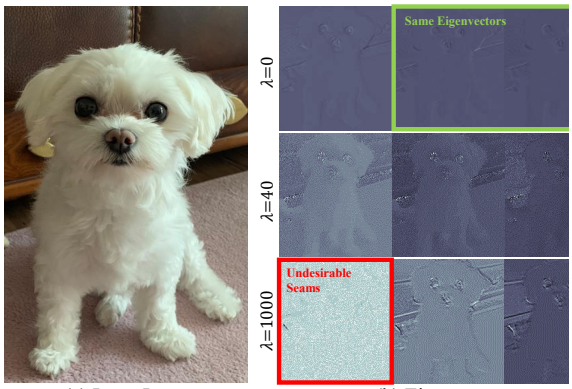
where $|\cdot|$ indicates an absolute value, and the hyperparameter γ is set to 0.9 in our implementation. We can intuitively figure out that $\mathcal{L}_{spatial}$ has a minimum value when $\mathbf{Y}_k$ is either 0 or 1 for each pixel. With the $\mathcal{L}_{spatial}$ and the softmax operation together, we show that if a pixel of one eigenvector converges near to 1, the pixel of other eigenvectors should go to 0. This 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:

$$\mathcal{L}_{ssg} = \mathcal{L}_{eigen} + \lambda \mathcal{L}_{spatial} \tag{7}$$

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, this layer produces adaptive structural information of input scenes by linearly combining the set of eigenvectors.



(a) Input Image (b) Eigenvectors

Figure 3: Visualization of the sets of eigenvectors according to λ = 0, 40, and 1000.

3.3 Analysis

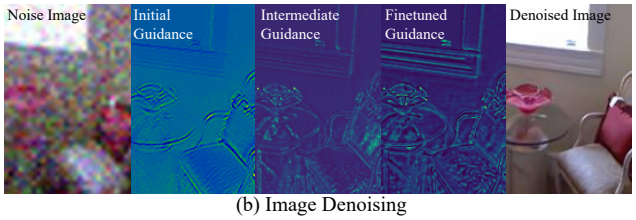
To the best of our knowledge, the SSGNet is the first to unfold the eigen-decomposition problem into a learnable network. To validate its effectiveness, we provide a series of analyses on SSGNet.

First, we analyze our loss function in Eq.(7) by tuning the hyper-parameter λ used as a balancing term between $\mathcal{L}_{spatial}$ and $\mathcal{L}_{eigen}$. In our experiment, the best performance is obtained with $\lambda = 40$. In Fig.3, we show the visualization results for three different λ values, including λ = 0, 40, and 1000. When λ is set to 0, $\mathcal{L}_{spatial}$ is not forced enough to give a sparsity penalty across eigenvectors, which leads to the degenerate case. Otherwise, if λ is set to 1000, the image is not well-segmented because the overwhelming majority of $\mathcal{L}_{spatial}$ causes undesirable seams on the image. From this, we can see that the absence of either one leads to an undesirable situation, which emphasizes the role of each of the two terms in our loss functions.

Next, we demonstrate that our SSGNet yields task-specific structural guidance features. As we highlighted, the SSGNet can be embedded in baseline networks. When the pretrained SSGNet is attached to baseline networks, the network parameters on SSGNet are finetuned to produce guidance features suitable for each task as the training proceeds. In Fig.4, we visualize how the eigenvectors from SSGNet change at each iteration during finetuning, including joint depth upsampling (Dong et al. 2022) and single image denoising (Zhang et al. 2022).

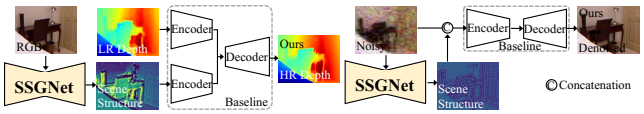
The joint depth upsampling needs accurate object boundaries as a prior (Li et al. 2014). For obvious reasons, an objective function in the joint depth upsampling encourages a greater focus on reconstructing object boundaries. As shown in Fig.4(a), our SSGNet generates attentive features on them during fine-tuning. In addition, for image denoising, it is essential to preserve fine detailed textures. In Fig.4(b), with the meaningful scene structures from our SSGNet, the plausible result is inferred as well. We claim that it is possible for our SSGNet to capture informative and task-specific structures through gradient updates from backpropagation (LeCun et al. 1989). We will describe the experimental details and

(a) Depth Upsampling



(b) Image Denoising

Figure 4: Examples of task-specific scene structures: intial, intermediate and final results from SSGNet for (a) joint depth upsampling and (b) image denoising.



(a) Depth Upsampling (b) Image Denoising

Figure 5: Illustrations of SSGNet for low-level vision tasks. The yellow colored networks indicate our SSGNet that outputs informative task-specific structure guidances.

SSGNet's quantitative benefits on each task in Sec.4.

3.4 Training Scheme

We implement the proposed framework using a public Pytorch (Paszke et al. 2019), and utilize the Adam (Kingma and Ba 2014) optimizer with $\beta_1 = 0.9$ and, $\beta_2 = 0.999$. The learning rate and the batch size are set to 0.0001 and 4 on SSGNet, respectively. We train the proposed framework on images with a 256×256 resolution. Since the proposed framework consists of fully convolutional layers, images with higher resolutions than that used in the training phase are available in inference. The training on SSGNet took about 40 hours on two NVIDIA Tesla v100 GPUs.

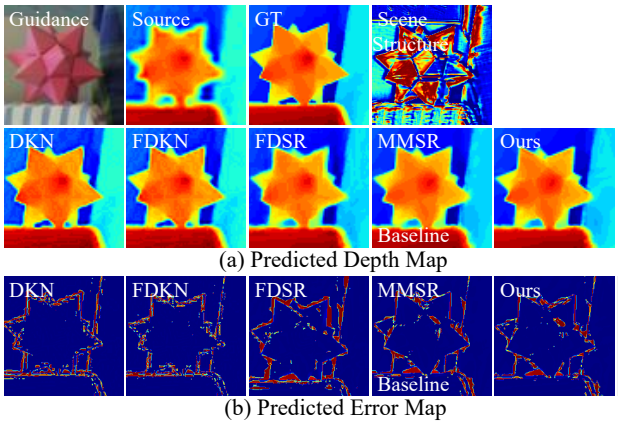
4 Experiments

We conduct a variety of experiments on low-level vision tasks, including self-supervised joint depth upsampling (Sec.4.1) and unsupervised single image denoising (Sec.4.2), to demonstrate the effectiveness of our SSGNet. Moreover, we provide an extensive ablation study (Sec.4.3) to precisely describe the effects of each component in SSGNet. Note that the higher resolution version of experimental results is reported in our supplementary material.

Baselines with SSGNet In this section, our goal is to validate a wide applicability of SSGNet. To do this, we incorporate SSGNet into existing CNN architectures for the joint depth upsampling and the unsupervised image denoising by simply embedding scene structures from ours to the models.

Dataset	Scale	Supervised						Self-Supervised					
		DKN		FDKN		FDSR		P2P		MMSR		Ours	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
2005	×4	1.103	0.275	0.964	0.222	0.886	0.211	1.288	0.273	0.708	0.239	**0.612**	**0.188**
	×8	1.182	0.288	1.629	0.339	1.043	0.333	1.177	0.280	1.043	0.319	**0.830**	**0.245**
2006	×4	1.623	0.297	1.337	0.222	1.198	**0.198**	2.604	0.413	0.555	0.232	**0.504**	0.201
	×8	1.790	0.307	1.883	0.305	1.170	0.267	2.684	0.300	0.723	0.261	**0.648**	**0.225**
2014	×4	2.878	0.739	2.593	0.659	3.217	0.595	4.019	0.822	1.953	0.573	**1.819**	**0.451**
	×8	3.642	0.775	3.510	0.871	3.606	0.885	3.894	0.920	2.765	0.785	**2.714**	**0.675**

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



(a) Predicted Depth Map

(b) Predicted Error Map

Figure 6: Comparison results on the joint depth upsampling with a resolution factor of 8 on the Middlebury 2005 dataset. We visualize the predictions and their corresponding error maps of competitive methods and ours.

Prior to the evaluations, we train our SSGNet on a well-known NYUv2 dataset (Silberman and Fergus 2011), consisting of 1,000 training images and 449 test images. With the pre-trained weight of SSGNet, we embed it to the baseline networks and finetune on each task. As mentioned above, we do not need any supervision for training SSGNet. We note that NYUv2 dataset is not used for evaluations, to validate the zero-shot generalization across various datasets.

4.1 Joint Depth Upsampling

Joint depth upsampling leverages the explicit structure detail of the input image as a guidance and transfers it to the target low-resolution depth map for enhancing spatial resolution. With this application, we demonstrate the synergy of the structure details from clean input images and the proposed learnable scene structure from SSGNet.

For this experiment, we choose MMSR (Dong et al. 2022) as a baseline depth upsampling network. MMSR introduces a mutual modulation strategy with the cross-domain adaptive filtering and adopts a cycle consistency loss to train the model in a fully self-supervised manner. Instead of directly using the input image as the guidance, we employ the structure guidance from the pretrained SSGNet in Fig.5(a), and follow the training scheme of MMSR for fair comparisons such that all the supervised methods are trained on NYUv2 dataset.

We also follow the evaluation protocol described in (Dong et al. 2022) to quantitatively measure the root mean square error (RMSE) and the mean absolute error (MAE). To be specific, we use the Middlebury stereo dataset 2005 (Scharstein and Pal 2007), 2006 (Hirschmuller and Scharstein 2007), and 2014 (Scharstein et al. 2014)[1], and augment them, which provides 40, 72, and 308 image-depth pairs, respectively, using a public code[2].

We compare with various state-of-the-art models, including supervised, DKN (Kim, Ponce, and Ham 2021), FDKN (Kim, Ponce, and Ham 2021) and FDSR (He et al. 2021), and self-supervised manners, P2P (Lutio et al. 2019) and MMSR (Dong et al. 2022). As shown in Tab.1, MMSR with our SSGNet embedded achieves the best performance in almost datasets over the comparison methods. Our SSGNet brings the performance gain over the second best method is about 10.4% and 11.8% with respect to RMSE and MAE, respectively. It is also noticeable that the scene structure contributes to reducing the errors in the star-like object boundary and the inside surface, visualized in Fig.6. We highlight that the result demonstrates the strong generalization capabilities of our SSGNet on unseen data again.

4.2 Image Denoising

We treat single image denoising to check the effectiveness of our SSGNet if the scene structure in the input image is corrupted by noise. For this experiment, we use IDR (Zhang et al. 2022) as a baseline image denoising network. IDR suppresses the image noise in a self-supervised manner by proposing an iterative data refinement scheme. The key of IDR is to reduce a data bias between synthetic-real noisy images and ideal noisy-clean images. To embed the scene structure to IDR, we simply concatenate it from our pretrained SSGNet with the noisy input image in Fig.5(b). As the rounds go on iteratively, our SSGNet focuses more on texture information of input scenes by ignoring the image noise, as already displayed in Fig.4.

To validate the applicability to the image denoising task as well, we compare our results with various state-of-the-art self-supervised models, including BM3D (Mäkinen,

[1]Since Middlebury 2003 provides neither depth maps nor camera parameters, we could not use it in this evaluation.

[2]Downloaded from https://rb.gy/bxyqgi

Method	Kodak				BSD300				BSD68			
	$\sigma = 25$		$\sigma = 50$		$\sigma = 25$		$\sigma = 50$		$\sigma = 25$		$\sigma = 50$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
BM3D	31.88	0.869	28.64	0.772	30.47	0.863	27.14	0.745	28.55	0.782	25.59	0.670
N2V	31.63	0.869	28.57	0.776	30.72	0.874	27.60	0.775	27.64	0.781	25.46	0.681
Nr2n	31.96	0.869	28.73	0.770	29.57	0.815	26.18	0.684	N/A	N/A	N/A	N/A
DBSN	32.07	0.875	28.81	0.783	31.12	0.881	27.87	0.782	28.81	0.818	25.95	0.703
N2N	**32.39**	**0.886**	29.23	0.803	31.39	0.889	28.17	0.799	29.15	0.831	26.23	0.725
IDR	32.36	0.884	29.27	0.803	31.48	0.890	28.25	0.802	29.20	**0.835**	26.25	0.726
Ours	**32.39**	0.885	**29.34**	**0.806**	**31.52**	**0.891**	**28.33**	**0.805**	**29.25**	**0.835**	**26.36**	**0.731**

Table 2: Quantitative results on single image denoising.

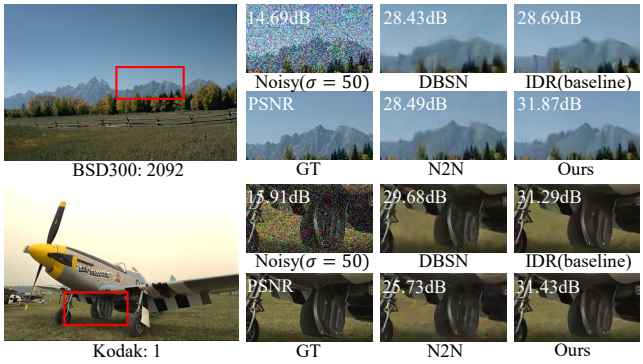


Figure 7: Examples of the single image denoising. For the noisy level $\sigma = 50$, we visualize the results from IDR+SSGNet as well as the state-of-the-art methods.

Azzari, and Foi 2019) N2V (Krull, Buchholz, and Jug 2019), Nr2n (Moran et al. 2020) DBSN (Wu et al. 2020), N2N (Lehtinen et al. 2018), and IDR (Zhang et al. 2022). For the evaluation, we strictly follow the experimental setup in (Zhang et al. 2022). We quantitatively measure PSNR and SSIM on Kodak (Kodak 1993), BSD300 (Movahedi and Elder 2010) and BSD68 (Martin et al. 2001) datasets for the zero-shot generalization. The models are trained on Gaussian noise with the continuous noise level $\sigma = [0, 50]$ and tested on $\sigma = 25$ and 50.

As shown in Tab.2, IDR with our SSGNet embedded achieves the best performance among all the competitive methods regardless of the noise levels. We emphasize that the performance gain by our SSGNet is about 0.58dB on average. Considering the performance difference between the second and the third best methods is about 0.26dB achieved by the paradigm shift from a statistical reasoning of image restoration (Lehtinen et al. 2018) to the iterative refinement (Zhang et al. 2022), SSGNet makes meaningful contribution. Fig.7 shows some example results. With the powerful capability of IDR on the noise suppression, our SSGNet preserves the scene texture of the objects well.

4.3 Ablation Study

An extensive ablation study is conducted to examine the effect of each component on SSGNet: the hyper-parameter λ in our loss function and the number of eigenvectors. We additionally test alternative scene structures computed from Canny Edge (Canny 1986) with different thresholds. For this ablation study, we measure RMSE and MAE on the Middlebury 2005 dataset for the joint depth upsampling ($\times 8$), and PSNR and SSIM on the Kodak dataset for the single image denoising ($\sigma = 50$), whose results and examples are reported in Tab.3 and Fig.8, respectively.

Choice of Hyper-parameter λ Since our loss function requires the selection of a hyper-parameter λ, it is important to study the sensitivity of the performances to the choice of λ. We carry out this experiment for six different values: 0.001, 0.1, 1, 100 and 1000 as well as 40 in our setting.

As a result, SSGNet's performance is insensitive to the choice of λ. In the joint depth upsampling, the performance difference according to λ is very marginal in that RMSE and MAE are at most 0.02cm and 0.001cm off the optimal values. In contrast, the performance gain for the image denoising when using $\lambda = 40$ is relatively large. Compared to $\lambda = 1000$ which shows the second best performance, the improvement from 0.08dB in PSNR when using $\lambda = 40$ brings more benefits for the comparisons with the state-of-the-art methods. In total, we find the optimal trade-off between these two tasks.

The Number of Eigenvectors The number of eigenvectors to represent scene structures is closely related to the number of learnable parameters in SSGNet. It is important for us to determine the optimal trade-off parameter in consideration of both the minimum number and the performances on these two tasks.

We investigate the performances of SSGNet with two, five, seven and ten as well as three eigenvectors. Interestingly, we observe the similar phenomenon just as above. The performance degradation on the joint depth upsampling is very small (about 0.02cm in RMSE and 0.003 in MAE), and the performance gain by 0.07dB in PSNR over the second best value on the image denoising is achieved. For the same reason, we set the number of eigenvectors to 3.

Comparison with Hand-crafted Structure Prior Hand-crafted edge detection is widely used for representing scene structures, even in recent models for low-level vision *i.e.* inpainting (Guo, Yang, and Huang 2021; Dong, Cao, and Fu 2022) and super-resolution (Nazeri, Thasarathan, and Ebrahimi 2019). We employ one of the representative hand-crafted edge map detection methods, Canny edge.

Task	Depth Upsampling		Denoising	
	RMSE	MAE	PSNR	SSIM
Ours	0.83	0.245	29.34	0.806
Hyper-parameter λ				
0.001	0.84	0.246	29.17	0.801
0.1	0.83	0.245	29.15	0.800
1	0.82	0.244	29.13	0.800
100	0.81	0.244	29.16	0.801
1000	0.84	0.248	29.26	0.803
# of eigenvectors				
2	0.84	0.272	29.15	0.800
5	0.81	0.242	29.16	0.800
7	0.82	0.242	29.17	0.800
10	0.82	0.244	29.27	0.804
Canny Edge				
ψ={0.6, 0.9, 1.4}	0.90	0.298	24.86	0.510
ψ={1.0, 2.0, 3.0}	0.90	0.282	24.37	0.489

Table 3: Ablation study for the effects of each component of SSGNet. We use the Middlebury 2005 with $\times 8$ for the joint depth upsampling, and the Kodak with a noise level σ=50 for the single image denoising.

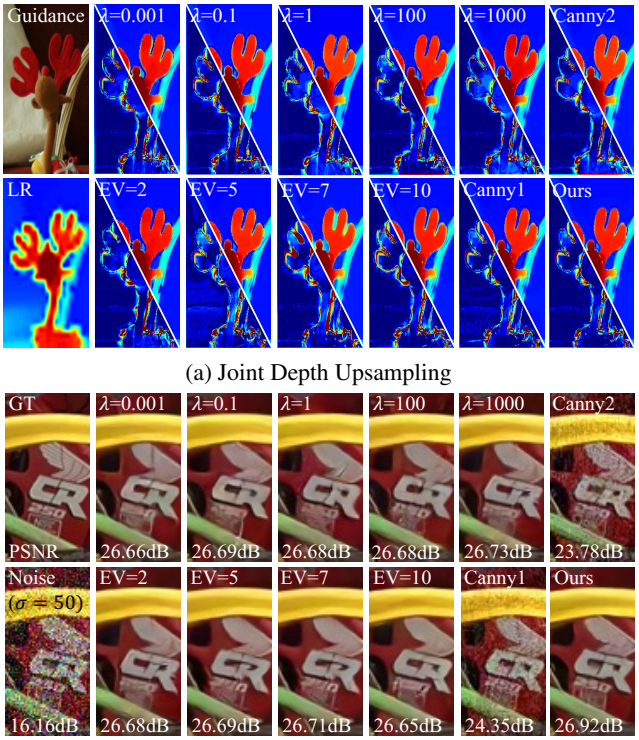
For fair comparison, we generate a set of edge maps with various thresholds for image gradients ψ, and embed it into the baseline networks. We note that the edge maps are used as input of our attention layer for the networks to selectively choose informative scene structures during the training phase. Here, we set two types of thresholds ψ to {0.6, 1.5, 2.4} and {1.0, 2.0, 3.0} that we manually find the best settings to configure image textures and object boundaries.

As shown in Tab.3, the interesting fact is that the performance drop of the joint depth upsampling is not huge when using the hand-crafted edge maps (within 0.07cm in RMSE and 0.037cm in MAE). On the other hand, there is the large performance gap between ours and the Canny edge maps (about 4dB in PSNR and 0.3 in SSIM).

Two possible reasons why the Canny edge fails to generate task-specific scene representations are: (1) The edge maps are not affected by back-propagation in training phase. (2) Based on the experimental results for the image denoising, the Canny edge is sensitive to image noise, which may corrupt the estimated scene structures and eventually not work as a prior. On the other hand, as displayed in Fig.8, our SSGNet returns the sharpest images, enabling the contents to be read. We thus argue that this experiment demonstrates the efficacy of our learnable structure guidance.

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



(a) Joint Depth Upsampling

(b) Image Denoising

Figure 8: Qualitative comparison for different settings of SSGNet. EV denotes the number of eigenvectors, and Canny1 and Canny2 mean edge threshold settings such as ψ = {0.6, 0.9, 1.4} and ψ = {1.0, 2.0, 3.0}, respectively. For the joint depth upsampling, we display the reconstruction results and the error maps, together.

of a scene with a small number of learnable parameters. With our proposed two loss terms, the eigen loss and the spatial loss, SSGNet is first initialized to parameterize the scene structures. The SSGNet is then embedded into the baseline networks and the parameters are fine-tuned to learn task-specific guidance features as the training proceeds. Lastly, we show the promising performance gains for both the joint depth upsampling and image denoising, even with the good cross-dataset generalization capability.

Discussion Although our SSGNet achieves the state-of-the-art results for the tasks with the simple embedding approach across the baseline networks, there are still rooms for improvements.

To suggest our future directions, we conduct additional small experiments for image super-resolution and unguided depth completion tasks which is a dense depth prediction from a sparse input depth without any guidance image. In these experiments, the super-resolution only use downsampled input images to extract scene structures, and the depth completion relies on the pretrained weight of SSGNet to represent scene configurations. We choose SeaNet (Fang, Li, and Zeng 2020), a CNN architecture equipped with a separate scene texture estimation branch, and pNCNN (Eldesokey

Scale	×2		×3		×4	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SeaNet	38.08	0.9609	34.55	0.9282	32.33	0.8981
SeaNet+	38.15	0.9611	34.65	**0.9290**	32.44	0.8981
Ours	**38.18**	**0.9612**	**34.68**	**0.9290**	**32.51**	**0.8983**

Scale	KITTI		NYU	
	RMSE	MAE	RMSE	MAE
pNCNN	1013.08	**251.53**	0.058	0.144
Ours	**1009.51**	256.06	**0.056**	**0.138**

Table 4: Additional experiments on Set5 (Bevilacqua et al. 2012) for image super-resolution with ×2, ×3 and ×4, and NYUv2 (Silberman et al. 2012) and KITTI (Uhrig et al. 2017) datasets for unguided depth completion. (unit:cm)

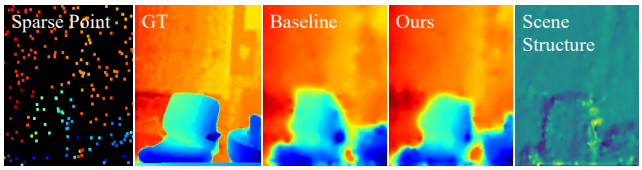


Figure 9: Comparison results on unguided depth completion on the NYUv2 dataset with pNCNN. Even no a guidance image, our SSGNet establishes the scene structure well.

et al. 2020), a lightweight probabilistic CNN ($\sim$ 670K) to refine initial dense depth predictions based on a statistical uncertainty measure, for the image super-resolution and the unguided depth completion, respectively.

Tab.4 reports that we obtain the performance gains with our SSGNet over the baseline models. Particularly, the synergy between our SSGNet and pNCNN is noticeable in Fig.9. Unfortunately, the baseline models with our SSGNet do not reach the quality of huge size models, a ViT-based image super-resolution (Liang et al. 2021) and a GAN-based unguided depth completion (Lu et al. 2020).

One of the future directions is to devise the best incorporation scheme of our SSGNet in that their structures are too tricky to intuitively embed it. Another is that a joint multi-modality training from heterogeneous data is expected to represent more informative scene structures and to extend the applicability of SSGNet.

Acknowledgements

This research was partially supported by 'Project for Science and Technology Opens the Future of the Region' program through the INNOPOLIS FOUNDATION funded by Ministry of Science and ICT (Project Number: 2022-DD-UP-0312), GIST-MIT Research Collaboration funded by the GIST, the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program in part (P0019797), the National Research Foundation of Korea (NRF) (No.2020R1C1C1012635) grant funded by the Korea government (MSIT), Vehicles AI Convergence Research & Development Program through the National IT Industry Promotion Agency of Korea (NIPA) funded by the Ministry of Science and ICT (No.S1602-20-1001), and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST), No.2021-0-02068, Artificial Intelligence Innovation Hub)

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bevilacqua, M.; Roumy, A.; Guillemot, C.; and Alberi-Morel, M. L. 2012. Low-complexity single-image super-resolution based on non-negative neighbor embedding. In *Proceedings of British Machine Vision Conference (BMVC)*.

Bloesch, M.; Czarnowski, J.; Clark, R.; Leutenegger, S.; and Davison, A. J. 2018. CodeSLAM—learning a compact, optimisable representation for dense visual SLAM. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Boykov, Y.; Veksler, O.; and Zabih, R. 2001. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(11): 1222–1239.

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 8(6): 679–698.

Cao, C.; and Fu, Y. 2021. Learning a sketch tensor space for image inpainting of man-made scenes. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-trained image processing transformer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Q.; Li, D.; and Tang, C.-K. 2013. KNN matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(9): 2175–2188.

Cohen, N.; and Shashua, A. 2017. Inductive bias of deep convolutional networks through pooling geometry. In *International Conference on Learning Representations (ICLR)*.

Cover, T.; and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1): 21–27.

de Lutio, R.; Becker, A.; D'Aronco, S.; Russo, S.; Wegner, J. D.; and Schindler, K. 2022. Learning Graph Regularisation for Guided Super-Resolution. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dong, Q.; Cao, C.; and Fu, Y. 2022. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dong, X.; Yokoya, N.; Wang, L.; and Uezato, T. 2022. Learning Mutual Modulation for Self-Supervised Cross-Modal Super-Resolution. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Eldesokey, A.; Felsberg, M.; Holmquist, K.; and Persson, M. 2020. Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fang, F.; Li, J.; and Zeng, T. 2020. Soft-edge assisted network for single image super-resolution. *IEEE Transactions on Image Processing (TIP)*, 29: 4656–4668.

Fiedler, M. 1973. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2): 298–305.

Fuoli, D.; Van Gool, L.; and Timofte, R. 2021. Fourier space losses for efficient perceptual image super-resolution. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

Gu, S.; Zuo, W.; Guo, S.; Chen, Y.; Chen, C.; and Zhang, L. 2017. Learning dynamic guidance for depth image enhancement. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guo, X.; Li, Y.; Ma, J.; and Ling, H. 2018. Mutually guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(3): 694–707.

Guo, X.; Yang, H.; and Huang, D. 2021. Image Inpainting via Conditional Texture and Structure Dual Generation. In *Proceedings of International Conference on Computer Vision (ICCV)*.

He, K.; Sun, J.; and Tang, X. 2012. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(6): 1397–1409.

He, L.; Zhu, H.; Li, F.; Bai, H.; Cong, R.; Zhang, C.; Lin, C.; Liu, M.; and Zhao, Y. 2021. 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)*.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hirschmuller, H.; and Scharstein, D. 2007. Evaluation of cost functions for stereo matching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Irwin, F.; et al. 1968. An isotropic 3x3 image gradient operator. *Presentation at Stanford AI Project*, 2014(02).

Jin, L.; Xu, Y.; Zheng, J.; Zhang, J.; Tang, R.; Xu, S.; Yu, J.; and Gao, S. 2020. Geometric structure based and regularized depth estimation from 360 indoor imagery. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kim, B.; Ponce, J.; and Ham, B. 2021. Deformable kernel networks for joint image filtering. *International Journal on Computer Vision (IJCV)*, 129(2): 579–600.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Kodak, E. 1993. Kodak lossless true color image suite (PhotoCD PCD0992). *Journal of Signal and Information Processing*, 6.

Krishnan, D.; and Fergus, R. 2009. Fast image deconvolution using hyper-Laplacian priors. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

Krull, A.; Buchholz, T.-O.; and Jug, F. 2019. Noise2void-learning denoising from single noisy images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4): 541–551.

Lehtinen, J.; Munkberg, J.; Hasselgren, J.; Laine, S.; Karras, T.; Aittala, M.; and Aila, T. 2018. Noise2Noise: Learning image restoration without clean data. *Proceedings of the International Conference on Machine Learning (ICML)*.

Levin, A.; Fergus, R.; Durand, F.; and Freeman, W. T. 2007. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)*, 26(3): 70–es.

Levin, A.; Lischinski, D.; and Weiss, Y. 2007. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(2): 228–242.

Levin, A.; Rav-Acha, A.; and Lischinski, D. 2008. Spectral matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(10): 1699–1712.

Li, J.; Lu, Z.; Zeng, G.; Gan, R.; and Zha, H. 2014. Similarity-aware patchwork assembly for depth image super-resolution. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Y.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2016. Deep joint image filtering. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Liu, H.; Wan, Z.; Huang, W.; Song, Y.; Han, X.; and Liao, J. 2021a. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Y.; Anwar, S.; Zheng, L.; and Tian, Q. 2020. Gradnet image denoising. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Lu, K.; Barnes, N.; Anwar, S.; and Zheng, L. 2020. From depth what can you see? Depth completion via auxiliary image reconstruction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lutio, R. d.; D'aronco, S.; Wegner, J. D.; and Schindler, K. 2019. Guided super-resolution as pixel-to-pixel transformation. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Mäkinen, Y.; Azzari, L.; and Foi, A. 2019. Exact transform-domain noise variance for collaborative filtering of stationary correlated noise. In *Proceedings of International Conference on Image Processing (ICIP)*.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Moran, N.; Schmidt, D.; Zhong, Y.; and Coady, P. 2020. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Movahedi, V.; and Elder, J. H. 2010. Design and perceptual validation of performance measures for salient object segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*.

Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; and Ebrahimi, M. 2019. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of International Conference on Computer Vision Workshop (ICCVW)*.

Nazeri, K.; Thasarathan, H.; and Ebrahimi, M. 2019. Edge-informed single image super-resolution. In *Proceedings of International Conference on Computer Vision Workshop (ICCVW)*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

Pickup, L.; Roberts, S. J.; and Zisserman, A. 2003. A sampled texture prior for image super-resolution. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; and Westling, P. 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition (GCPR)*, 31–42. Springer.

Scharstein, D.; and Pal, C. 2007. Learning conditional random fields for stereo. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shi, J.; and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 22(8): 888–905.

Silberman, N.; and Fergus, R. 2011. Indoor scene segmentation using a structured light sensor. In *Proceedings of International Conference on Computer Vision Workshop (ICCVW)*.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Sun, J.; Xu, Z.; and Shum, H.-Y. 2008. Image super-resolution using gradient profile prior. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Tai, Y.-W.; Liu, S.; Brown, M. S.; and Lin, S. 2010. Super resolution using edge prior and single image detail synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tang, C.; and Tan, P. 2019. Ba-net: Dense bundle adjustment network. In *International Conference on Learning Representations (ICLR)*.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288.

Tomasi, C.; and Manduchi, R. 1998. Bilateral filtering for gray and color images. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Uhrig, J.; Schneider, N.; Schneider, L.; Franke, U.; Brox, T.; and Geiger, A. 2017. Sparsity Invariant CNNs. In *International Conference on 3D Vision (3DV)*.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Wu, X.; Liu, M.; Cao, Y.; Ren, D.; and Zuo, W. 2020. Unpaired learning of deep image denoising. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Xie, J.; Feris, R. S.; and Sun, M.-T. 2015. Edge-guided single depth image super resolution. *IEEE Transactions on Image Processing (TIP)*, 25(1): 428–438.

Xu, Z.; Zhang, H.; Wang, Y.; Chang, X.; and Liang, Y. 2010. L1/2 regularization. *Science China Information Sciences*, 53(6): 1159–1169.

Yang, J.; Qi, Z.; and Shi, Y. 2020. Learning to incorporate structure knowledge for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Zhang, Q.; Shen, X.; Xu, L.; and Jia, J. 2014. Rolling guidance filter. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Zhang, Y.; Li, D.; Law, K. L.; Wang, X.; Qin, H.; and Li, H. 2022. IDR: Self-Supervised Image Denoising via Iterative Data Refinement. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

Zhong, Y.; Yuan, B.; Wu, H.; Yuan, Z.; Peng, J.; and Wang, Y.-X. 2021. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of International Conference on Computer Vision (ICCV)*.