Vision, Modeling, and Visualization (2017)
M. Hullin, R. Klein, T. Schultz, A. Yao (Eds.)

Semantic-Aware Image Smoothing

Weihao Li[1,2], Omid Hosseini Jafari[1,2], and Carsten Rother[1,2]

[1] Computer Vision Lab, TU Dresden, Germany
[2] Visual Learning Lab, Heidelberg University, Germany

Abstract

Structure-preserving image smoothing aims to extract semantically meaningful image structure from texture, which is one of the fundamental problems in computer vision and graphics. However, it is still not clear how to define this concept. On the other hand, semantic image labeling has achieved significant progress recently and has been widely used in many computer vision tasks. In this paper, we present an interesting observation, i.e. high-level semantic image labeling information can provide a meaningful structure prior naturally. Based on this observation, we propose a simple and yet effective method, which we term semantic smoothing, by exploiting the semantic information to accomplish semantically structure-preserving image smoothing. We show that our approach outperforms the state-of-the-art approaches in texture removal by considering the semantic information for structure preservation. Also, we apply our approach to three applications: detail enhancement, edge detection, and image segmentation, and we demonstrate the effectiveness of our semantic smoothing method on these problems.

Categories and Subject Descriptors (according to ACM CCS): I.4.3 [Image Processing and Computer Vision]: Enhancement—Smoothing

1. Introduction

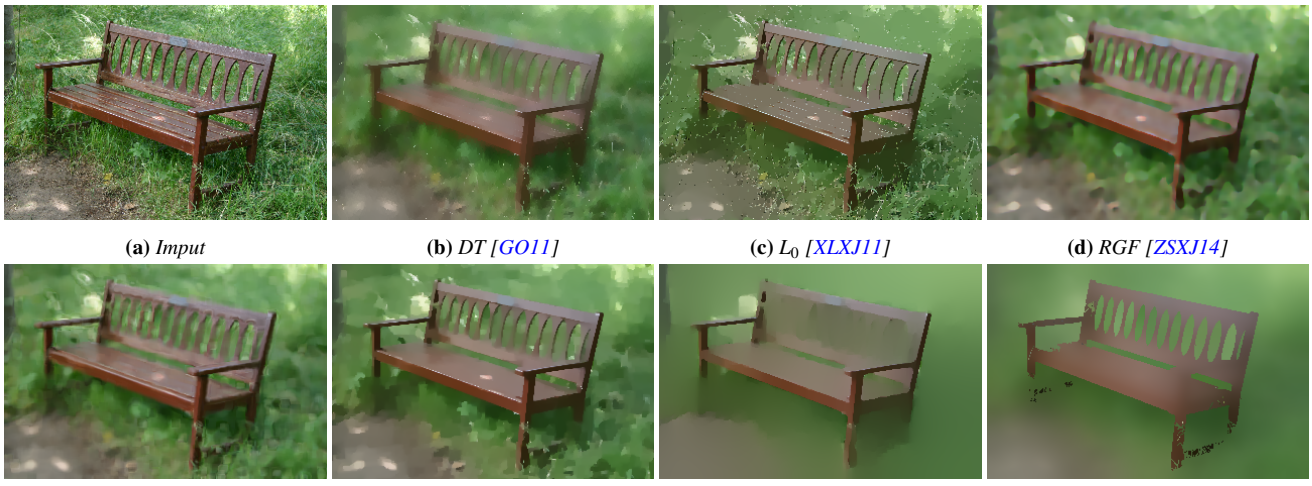
Structure/edge-preserving image smoothing [XYXJ12, KEE13, XRY*15] is one of the fundamental problems in image processing, computational photography, and computer vision. The purpose of image smoothing is to reduce unimportant image texture or noise while preserving semantically meaningful image structures simultaneously [XYXJ12,Yan16]. It has achieved widespread use in various applications, including texture removal, edge extraction, image abstraction, seam carving and tone mapping.

The main challenge of image smoothing is how to obtain and exploit the structural or the edge prior information to distinguish semantically pointless texture or noise from meaningful image structure [XYXJ12, ZSXJ14, Yan16]. The majority of edge-preserving image filters apply low-level feature, i.e. image gradients, as edge prior information, such as bilateral filtering [TM98] and guided filter [HST10]. For structure-preserving image smoothing, relative total variation [XYXJ12], diffusion maps [FFL10], and region covariances [KEE13] measures are used to separate texture from the image structure. Recently, Yang [Yan16] use an edge detector for iterative edge-preserving texture filtering to exploit mid-level vision feature, i.e. structured edges. Although these methods work well for some tasks, it is not clear how to define the meaningful image structure. For example in Figure 1 (b-g), it is difficult for previous approaches to preserve the bench structure when they only consider low-level and mid-level vision features of an image.

In this paper, we present an observation, i.e. high-level semantic information can provide a meaningful structure prior for image smoothing naturally. Recently, semantic labeling has been heavily studied in computer vision community [SWRC06, ZCW*14, LSD15, JGK*17]. Semantic information provides an object-level semantically meaningful structure prior, such as object boundaries, which help to reduce the negative effect of sharp edges inside of objects. Based on this observation, in this paper, we present a simple and yet effective method which exploits semantic labeling information to accomplish texture removal and meaningful structure preservation. We call this new concept *semantic smoothing*. Besides utilizing high-level semantic information, our method also can combine low-level vision features, i.e. image appearance, and mid-level vision information, i.e. image edges.

Our method has two unique properties: meaningful structure preservation and interior detail removal. As shown in Figure 1, input image Figure 1 (a) contains a textured bench in the foreground and a grassland in the background. Current state-of-the-art image smoothing methods cannot successfully separate bench from its texture and preserve its structure as shown in Figure 1 (b)-(g). Our proposed semantic smoothing technique outperforms other approaches by preserving the bench structure effectively as illustrated in Figure 1 (h). To the best of our knowledge, it is the first structure-preserving image smoothing method which exploits high-level semantic segmentation information.

The following sections are organized as follows. The related



(a) *Imput* **(b)** *DT [GO11]* **(c)** L_0 *[XLXJ11]* **(d)** *RGF [ZSXJ14]*

(e) *RG [KEE13]* **(f)** *RTV [XYXJ12]* **(g)** *SF [Yan16]* **(h)** *Ours*

Figure 1: *Semantic Smoothing on MSRC-21 dataset. In this example, image(a) contains a textured bench in a grassland. As a result, it is difficult for the state-of-the-art structure-preserving and edge-preserving smoothing methods to obtain smoothing results with accurate structure (b)-(g). (b) Domain Transform (DT) [GO11], (c) L_0 Smoothing [XLXJ11] ($\lambda = 0.04$), (d) Rolling Guidance Filter (RGF) [ZSXJ14], (e) Region Covariances (RG) [KEE13] (k = 5, σ = 0.2, Model 1), (f) Relative Total Variation (RTV) [XYXJ12] ($\lambda = 0.005$, $\sigma = 3$) and (g) Semantic Filtering (SF) [Yan16]. Our method effectively preserves semantically meaningful structure and smooth out detail and texture. Best viewed in color.*

works are discussed in Section 2. In Section 3 our semantic-aware image smoothing method is described. In Section 4 experimental results and applications are presented.

2. Related Work

We categorize the related work into two aspects: image smoothing and semantic segmentation. First, we discuss edge-preserving and structure-preserving image smoothing methods. Second, we briefly review development progress of semantic segmentation and semantic information in other vision problems.

2.1. Image Smoothing

The image smoothing methods can be separated into two classes: edge-preserving and structure-preserving smoothing. The bilateral filter [TM98] is one of the most popular edge-preserving filtering methods which replaces the intensity value of each pixel in the image with a weighted average of intensity values of its neighboring pixels. In joint bilateral filters [PSA*04, ED04], the range filter is applied to a guidance image from another domain. As edge-preserving image smoothing or filtering methods, we can also mention anisotropic diffusion [PM90], weighted least square [FFLS08], local Laplacian pyramid [PHK11], domain transform [GO11], and semantic filtering [Yan16]. However, it is hard to separate high-contrast textured regions or patterns from the meaningful structures of an image by using these edge-preserving techniques. The structure-preserving image smoothing techniques aim to separate the image structure and texture. One of the most popular structure-preserving image smoothing methods is Xu et al. [XYXJ12], which uses the relative total variation (RTV) measure to decompose structures from textures. They first model a regularization term based on the RTV measure, then solve a global optimization to extract the main structures and to obtain the smoothed image. Zhang *et al.* [ZDXZ15] first segment the input image into superpixels then they build a minimum spanning tree for each superpixel to accelerate image filtering. Shen et al. [SZXJ15] proposes a mutual-structure joint filtering towards preserving common structures of an input and a guidance image. As other structure-preserving image smoothing techniques we can mention total variation [ROF92], local extrema [SSD09], structure adaptive [KD], rolling guidance filter [ZSXJ14], and geodesic [CSRP10]. In contrast, we exploit the semantic segmentation information as a meaningful structure prior for the semantic structure-preserving image smoothing. Recently, several learning-based methods have also been proposed for image filtering [XRY*15, BP16].

2.2. Semantic Segmentation

Semantic segmentation is one of the key problems in image understanding. The goal of semantic segmentation is to label each pixel of the image with the class of its enclosing object. A common pipeline of semantic segmentation is first to train pixel-based classifiers, such as Textonboost [SWRC06] or fully convolutional networks (FCN) [LSD15], then using a probabilistic graphical model, such as CRF [SWRC06, CPK*15, ZJRP*15, LKZ*17], to improve the performance by modeling structured dependencies. With the development of semantic segmentation techniques, other computer vision problems exploit high-level semantic information, such as optical flow [SLSJB16, BLKU16], depth prediction [JGK*17, WSL*15b], depth upsampling [HGY15, SSP*16], object attributes [VRT13, ZCW*14], intrinsic image estimation [VRT13], 3D reconsecration [HZC*13, KLD*14, LSR*12].

However, smoothing image using semantic segmentation information has not been exploited before. In this paper, we propose a novel semantic-aware approach which exploits the semantic information for structure preserving image smoothing.

3. Semantic Smoothing

In this section, we introduce our semantic image smoothing method, which exploits high-level semantic information to achieve semantically meaningful structure preserving smoothing. Given an input image $\mathbf{t}$ and its semantic labeling $\mathbf{s}$, our goal is to compute a new smoothed image $\mathbf{x}$, which is as similar as possible to the input image $\mathbf{t}$ while preserving the semantically meaningful image structure and reducing the texture or noise. We model our *semantic smoothing* as an energy minimization problem. Formally, we define the energy function as a weighted sum of two energy terms

$$E(\mathbf{x}) = E_d(\mathbf{x};\mathbf{t}) + E_r(\mathbf{x};\mathbf{t},\mathbf{s}), \tag{1}$$

where E_d is the data term and E_r is the regularization term.

3.1. Data Term

The purpose of the data term is to minimize the distance between the input image $\mathbf{t}$ and the smoothed image $\mathbf{x}$. Without this data term, there will be a trivial solution where all of the pixels will be assigned to the same color value. We define the data term E_d as

$$E_d(\mathbf{x};\mathbf{t}) = \sum_i (x_i - t_i)^2, \tag{2}$$

where i is the pixel index. With this term, smoothed image $\mathbf{x}$ will be limited within a range around the input image $\mathbf{t}$.

3.2. Regularization Term

The regularization term E_r strive to achieve smoothness by jointly considering the low-level appearance, the mid-level edge, and the high-level semantic information. The regularization term E_r is defined as

$$E_r(\mathbf{x};\mathbf{t},\mathbf{s}) = \sum_i \sum_{j \in \mathcal{N}(i)} W_{i,j}(x_i - x_j)^2, \tag{3}$$

where $\mathcal{N}(i)$ is a set of neighboring (four or eight) pixels around the pixel i and the weight $W_{i,j}$ represents the similarity between the pixel i and the pixel j. Our $W_{i,j}$ consists of three potential functions and is defined as

$$W_{i,j} = \lambda_a w^a_{i,j} + \lambda_e w^e_{i,j} + \lambda_s w^s_{i,j}, \tag{4}$$

where the first factor $w^a_{i,j}$ is the appearance potential which is used to control the low-level information. The second factor $w^e_{i,j}$ is based on the edge detection and is used to control the mid-level information. The last factor $w^s_{i,j}$ is the semantic potential which exploits the high-level semantic information. The weights λ_a, λ_e, and λ_s are used to control the effect of the low-level, the mid-level and the high-level information on the final smoothed output, respectively. These three parts are explained in detail below.

3.2.1. Appearance potential

The appearance potential $w^a_{i,j}$ of the pixel i and the pixel j is defined as

$$w^a_{i,j} = \exp\left(-\frac{||\mathbf{f}_i - \mathbf{f}_j||^2}{\sigma_a}\right), \tag{5}$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ are three-dimensional vectors representing the Lab color values of the pixel i and the pixel j and σ_a is a range parameter.

We use the appearance potential to measure the difference of the low-level vision feature, i.e. color, between the pixel i and the pixel j. In this setting, neighboring pixels of the input image with similar colors are assigned to larger weights and neighboring pixels with different colors are assigned to smaller weights.

3.2.2. Edge potential

The edge potential $w^e_{i,j}$ between the pixel i and the pixel j is defined as

$$w^e_{i,j} = \exp\left(-\frac{\beta^2_{i,j}}{\sigma_e}\right), \tag{6}$$

where $\beta_{i,j} \in [0,1]$ is the boundary strength measure between the pixel i and the pixel j and σ_e is a range parameter.

Recently, Yang [Yan16] uses an edge detector [DZ13] for edge-preserving image filtering. In contrast, we utilize image edges as the mid-level vision cue to help the appearance potential and the semantic potential. In this work, we use the structured edge detector [DZ13] to calculate boundary strength measure $\beta_{i,j}$.

3.2.3. Semantic potential

The semantic potential is the key part of our semantic smoothing. Based on the semantic labeling $\mathbf{s}$, the semantic potential between the pixel i and the pixel j can be written as

$$w^s_{i,j} = \begin{cases} \gamma_{high} & \text{if } s_i = s_j \\ \gamma_{low} & \text{otherwise,} \end{cases} \tag{7}$$

where s_i and s_j present semantic labeling of the pixel i and the pixel j. γ_{high} and γ_{low} are weight parameters and $\gamma_{high} > \gamma_{low}$. When neighboring pixels i and j have the same semantic labeling, we assign a larger weight to encourage these two pixels to have close color values in the output smoothed image. In contrast, when neighboring pixels i and j have different semantic labeling, they are assigned a smaller weight. For each class label, it is possible to set different γ_{high} values to control the different smoothing strength. In this work, for simplicity, we set γ_{high} to 1.0 for all semantic classes and we set γ_{low} to zero.

Semantic information help to reduce the adverse effect of the object's interior sharp edges and texture.

4. Optimization

The objective function in Equation 1 is strictly convex and can be written in a matrix and vector form as

$$E(\mathbf{x}) = (\mathbf{x} - \mathbf{t})^\mathsf{T}(\mathbf{x} - \mathbf{t}) + \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \tag{8}$$

where matrix $\mathbf{A}$ is a Laplacian matrix which is defined as

$$\mathbf{A} = \mathbf{D} - \mathbf{W}, \tag{9}$$

where $\mathbf{W}$ is an adjacency matrix $\{W_{i,j} | j \in \mathcal{N}(i)\}$ and $\mathbf{D}$ is a degree matrix which is defined as

$$D_{i,j} = \begin{cases} \sum_{j \in \mathcal{N}(i)} W_{i,j} & i = j \\ 0 & i \neq j. \end{cases} \tag{10}$$

By setting the gradient of $E(\mathbf{x})$ defined as in Equation 8 to zero, the final smoothing result $\mathbf{x}$ is obtained by solving the linear system based on a large sparse matrix:

$$(\mathbf{I} + \mathbf{A})\mathbf{x} = \mathbf{t} \tag{11}$$

where $\mathbf{I}$ is an identity matrix.

5. Experimental Results and Applications

Our semantic smoothing method can benefit several image editing and manipulation applications due to its special properties, i.e. meaningful structure preservation and interior detail removal.

In this section first, we introduce the datasets which we used in our experiments. Second, we visually compare the texture removal results of our proposed semantic smoothing approach with the state-of-the-art methods. Finally, to show the effect of our approach, we apply it to three applications: detail enhancement, edge detection, and image segmentation.

5.1. Datasets

MSRC-21 dataset [SWRC06] consists of 591 color images with following 21 object classes, such as grass, tree, cow, sheep, water and so forth. Cimpoi [CMV15] also use MSRC-21 dataset for texture recognition and segmentation task. In order to ensure proportional contributions from each class approximately, the dataset is split into 45% training, 10% validation and 45% test images. We use the standard split of the dataset from [SWRC06] to train the textonboost [SWRC06], which incorporates shape, texture, location, and color descriptors. Then, we use the trained textonboost to obtain the semantic segmentation. Lastly, we apply the dense CRF [KK11] to refine the semantic segmentation results and we use this refined version as high-level semantic information input to our smoothing approach.

PASCAL VOC dataset [EVGW*10] consists of one background class and 20 foreground object classes including person, bird, cat, cow, dog and so forth. There are 1464 images for training, 1449 for validating and 1456 for testing, respectively. Recently, the fully convolutional network (FCN) [LSD15] is mainly utilized for estimating the semantic segmentation on PASCAL VOC dataset. Also in this work, we employ the publicly available pre-trained FCN [LSD15] for obtaining the semantic labeling for PASCAL VOC. Then, we use the dense CRF [KK11] to refine the FCN results for using it as the input to our semantic smoothing.

5.2. Texture Removal

Texture removal, which is also called as texture smoothing, aim to separate the meaningful structures from textures. We compare our semantic smoothing results with the state-of-the-art image smoothing techniques, such as Relative Total Variation (RTV) [XYXJ12] and Semantic Filtering (SF) [Yan16]. We use the authors' publicly available implementations. It is difficult to quantitatively evaluate image smoothing methods, therefore similar to the most of the state-of-the-art methods [Yan16, XYXJ12], we present the visual comparison evaluation in Figure 2 and Figure 3.

We visually compare our proposed semantic smoothing technique with [XYXJ12, Yan16] on MSRC-21 dataset (see Figure 2) and PASCAL VOC dataset (see Figure 3). As illustrated in these figures, our semantic-aware image smoothing performs better in terms of preserving meaningful structures and reducing object interior textures. For instance, if we look at the black cow in the first row of Figure 2, there are strong edges inside of the cow's body in other approaches' results, while our approach is able to remove these semantically meaningless edges.

5.3. Applications

5.3.1. Detail Enhancement

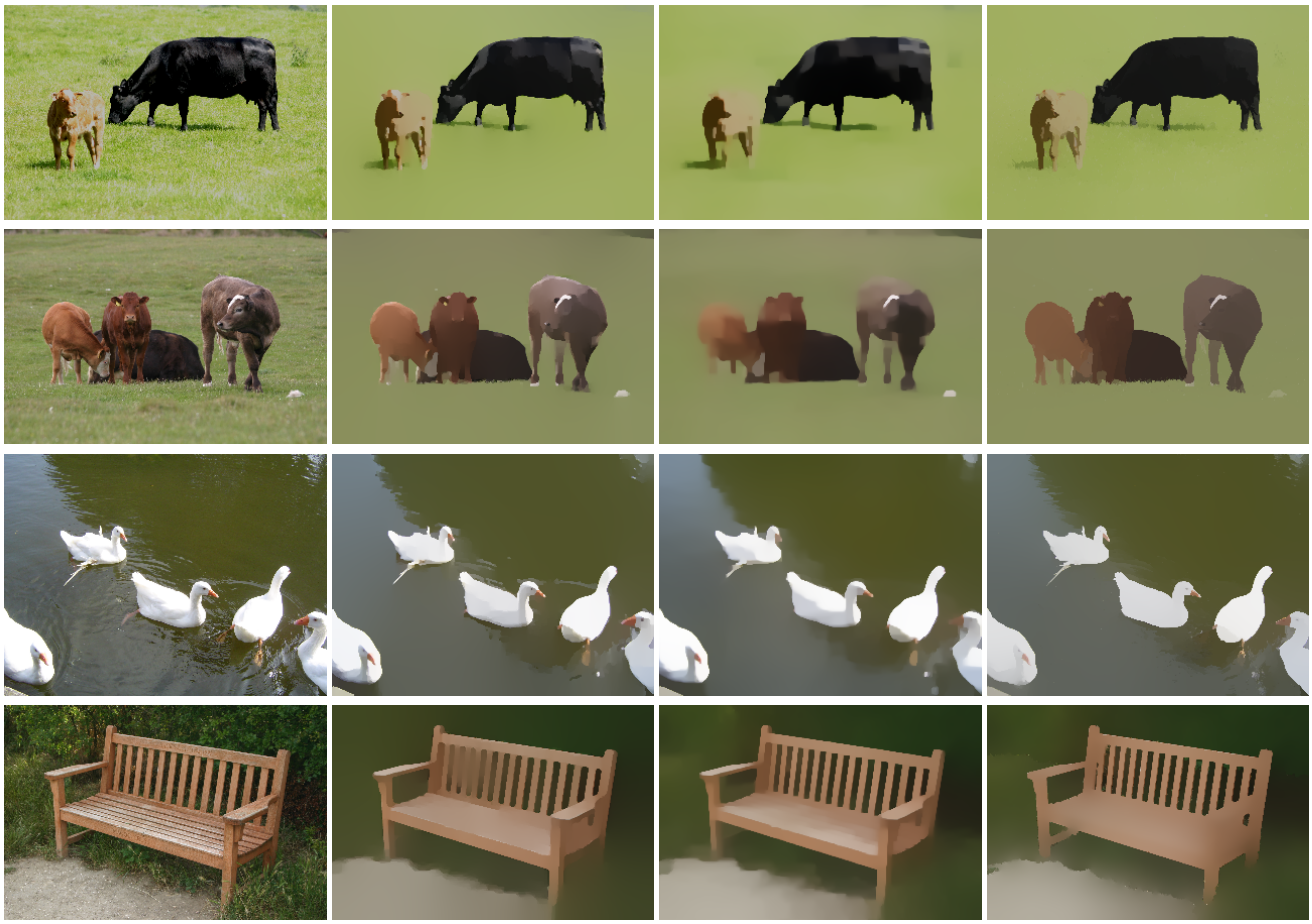
Detail enhancement aims to increase visual appearance of images, which is widely used in image editing. Thanks to the property of structure-preserving image smoothing, i.e. structure-texture decomposition, we can apply our semantic smoothing method to enhance the underlying details or textures of an image. First, we use our semantic smoothing method to decompose the input image into structures and details. Then we add the details back to the input image. That means we augment the contrast in detail components of the input image.

Figure 4 shows two examples. Given two input images Figure 4 (a) and (e) and their semantic smoothing results Figure 4 (b) and (f), we can decompose the texture information Figure 4 (c) and (g) and obtain the detail enhancement results Figure 4 (d) and (h). Since our smoothing method can effectively preserve the object-level structure and remove object interior edges, it can effectively enhance the underlying detail, particularly *interior* texture and edges of objects, without blurring the main structure of objects.

5.3.2. Edge Detection

Edge detection is one of the challenging tasks in computer vision for a long time. The purpose of edge detection is to extract visually salient edges or object boundaries from the input image. Boundary and edge can be used in a broad range of computer vision or graphics tasks, such as semantic segmentation, object recognition, image editing and tone mapping. Our method can be applied to object-level edge extraction thanks to its ability to preserve semantically meaningful structures and remove many unimportant details, such as interior edges of the object especially.

Figure 5 (a) shows an input image in grass texture with a salient foreground, i.e. a cow. Since the texture has high contrast, applying the Canny edge detector [Can86] cannot produce reasonable results directly from the input image, see Figure 5(c). Structured edge detection [DZ13] is a popular edge detection method based on random forests, which can detect salient edges. It achieves better results as demonstrated in Figure 5(e) and thinned edges Figure 5(g), which is obtained by standard non-maximal suppression



(a) *Image* **(b)** *SF [Yan16]* **(c)** *RTV [XYXJ12]* **(d)** *Ours Smoothing*

Figure 2: *Visual comparison of texture removal results on MRSC dataset. (a) input images, (b) Semantic Filtering (SF) [Yan16], (c) Relative Total Variation (RTV) [XYXJ12] and (d) Our semantic smoothing results. Best viewed in color.*

(a) *Imput image* **(b)** *SF [Yan16]* **(c)** *RTV [XYXJ12]* **(d)** *Ours smoothing*

Figure 3: *Visual comparison of texture removal results on PASCAL VOC dataset. (a) input images, (b) Semantic Filtering (SF) [Yan16], (c) Relative Total Variation (RTV) [XYXJ12] and (d) Our semantic smoothing results. Best viewed in color.*

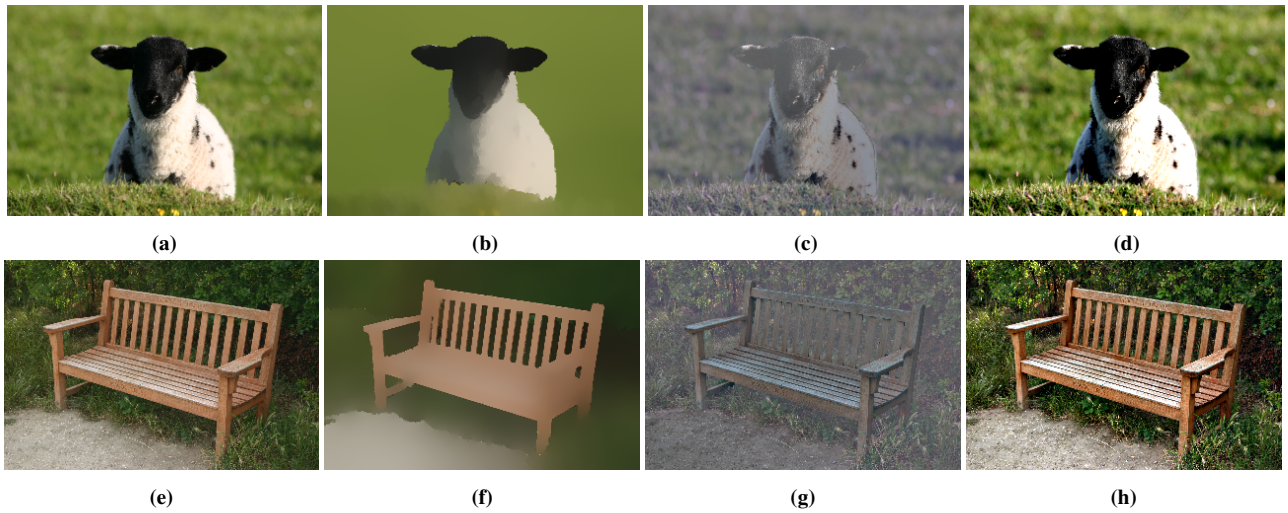


Figure 4: *Detail Enhancement. (a) and (e) are the input images. (b) and (f) are our semantic image smoothing results. (c) and (g) are decomposed texture information outputs. (d) and (h) are the detail enhancement results. Best viewed in color.*

technique. We can see that some of the detected edges come from the textures. In contrast, our method first produces an object-level structure-preserving smoothed image, which removes insignificant details as Figure 5(b). We can improve the result of these edge detection approaches by applying them to our smoothed images Figure 5(b). Figures 5 (d), (f), and (h) illustrate the refined edge detection results of Figures 5 (c), (e), and (g) correspondingly.

5.3.3. Semantic Segmentation

In this section, we show that the smoothed image also can help semantic segmentation. Fully connected conditional random field [KK11], which is also called as dense conditional random field (Dense-CRF), is a very popular tool to refine semantic image segmentation results. We propose to use a modified version of Dense-CRF, which we call Dense-CRF+, where the smoothed images are used to model appearance kernel of Gaussian edge pairwise term instead of the typical RGB color vectors. For the sake of comparison with original Dense-CRF, we use the MSRC-21 dataset, the same data splits and unary potentials as the one used by [KK11].

We choose two standard measures of multi-class segmentation accuracy as [KK11] used, i.e. Overall and Average. Overall is the pixel-wise labeling accuracy, which is computed over the whole image pixels for all classes. Average is the pixel-wise labeling accuracy computed for all classes and the averaged over these classes. The original ground truth labelings of the MSRC-21 dataset are relatively imprecise. There are some regions around objects boundaries left unlabeled. This makes it difficult to evaluate the quantitative performance of semantic segmentation results. Therefore, we evaluated our results on the 94 accurate ground truth labelings provided by [KK11], which is fully annotated at the pixel-level, with accurate labeling around complex boundaries. Table 1 shows the quantitative experimental results. We get the Average accuracy 79.55 and Overall accuracy 88.01. Our method outperforms the original Dense-CRF approach [KK11] on the MSRC-21 dataset. Figure 6 shows some qualitative semantic segmentation results on the MSRC-21 dataset. Our Dense-CRF+ obtains more accurate results than the Dense-CRF, which produces many spatially disjoint object segments. As a future work, it is possible to jointly inference semantic smoothing and segmentation.

Class	Unary	Dense-CRF	Dense-CRF+
Average	76.39	79.37	**79.55**
Overall	83.18	87.78	**88.01**

Table 1: *The quantitative semantic segmentation results on the MSRC-21 dataset.*

6. Conclusion

In this paper, we propose a semantic-aware image smoothing method. Unlike previous image smoothing techniques which use the low-level vision features, such as appearance and gradient, or the mid-level vision features, such as edge or boundary detection, our proposed technique is developed based on the high-level semantic information of the image. Besides exploiting the high-level semantic information, our method also combine the low-level and the mid-level features. Effectiveness of our approach is demonstrated in different applications, including texture removal, detail enhancement, edge detection, and semantic segmentation. The limitation of the semantic smoothing is that it depends on the quality of the semantic segmentation. But with the development of semantic segmentation techniques, particularly using deep learning, we will have enough confidence to believe that using semantic information will be advantageous for image smoothing.In future work, we would like to extend our method by exploiting diverse levels of semantic information, such as instance segmentation [DHS16], object part segmentation [WSL*15a] and material segmentation [BUSB15].

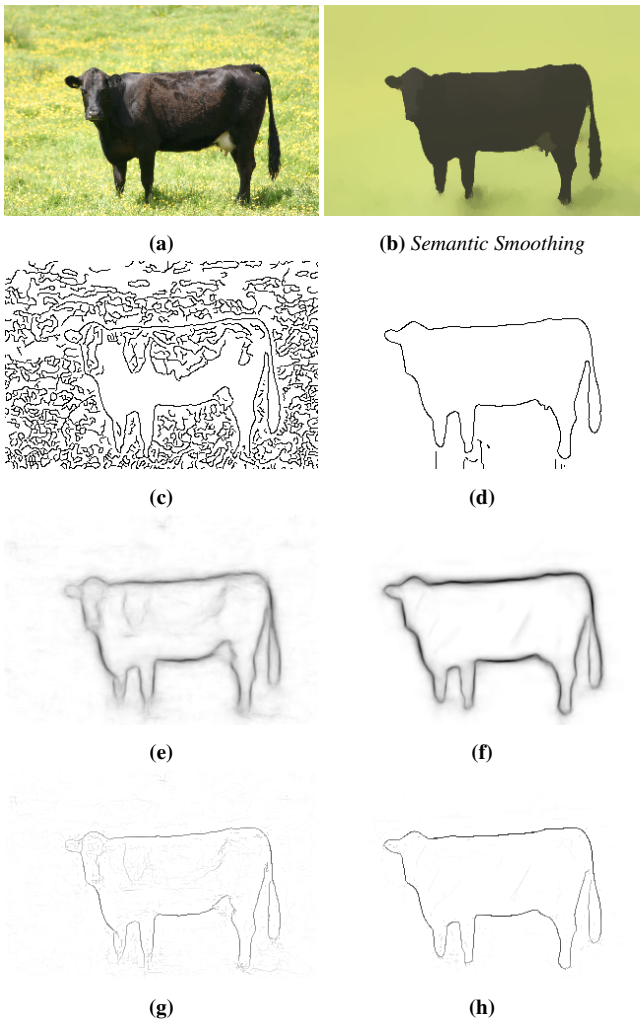


Figure 5: *Edge Detection. (a) Input image, (c) Canny edge detection [Can86] applied to (a), (e) Structure edge detection [DZ13] applied to (a), (g) Non-maximal suppression applied to (e). (b) Our semantic smoothing result, (d) Canny edge detection [Can86] applied to (b), (f) Structure edge detection [DZ13] applied to (b), (h) Non-maximal suppression applied to (f).*

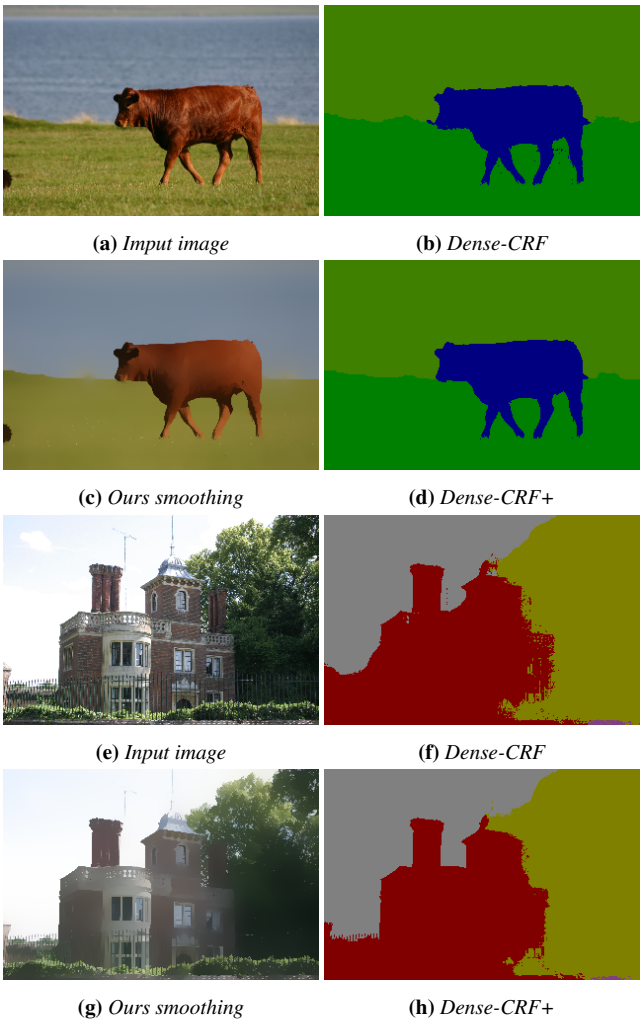


Figure 6: *Semantic segmentation. (a) and (e) are input images. (b) and (f) are Dense-CRF segmentation results. (c) and (g) are our semantic smoothing results. (d) and (h) are Dense-CRF+ segmentation results. Our method predicts segmentations which are localized around object boundaries and are spatially smooth. Best viewed in color.*

References

[BLKU16] BAI M., LUO W., KUNDU K., URTASUN R.: Exploiting semantic information and deep matching for optical flow. In *European Conference on Computer Vision (ECCV)* (2016), Springer, pp. 154–170. 2

[BP16] BARRON J. T., POOLE B.: The fast bilateral solver. In *European Conference on Computer Vision (ECCV)* (2016), Springer, pp. 617–632. 2

[BUSB15] BELL S., UPCHURCH P., SNAVELY N., BALA K.: Material recognition in the wild with the materials in context database. In *Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3479–3487. 6

[Can86] CANNY J.: A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 (1986), 679–698. doi:10.1109/TPAMI.1986.4767851. 4, 7

[CMV15] CIMPOI M., MAJI S., VEDALDI A.: Deep filter banks for texture recognition and segmentation. In *Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3828–3836. 4

[CPK*15] CHEN L.-C., PAPANDREOU G., KOKKINOS I., MURPHY K., YUILLE A. L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR* (2015). 2

[CSRP10] CRIMINISI A., SHARP T., ROTHER C., P'EREZ P.: Geodesic image and video editing. *ACM Trans. Graph. 29*, 5 (Nov. 2010), 134:1–134:15. doi:10.1145/1857907.1857910. 2

[DHS16] DAI J., HE K., SUN J.: Instance-aware semantic segmentation via multi-task network cascades. In *Computer Vision and Pattern Recognition (CVPR)* (2016). 6

[DZ13] DOLLÁR P., ZITNICK C. L.: Structured forests for fast edge detection. In *International Conference on Computer Vision (ICCV)* (2013), pp. 1841–1848. 3, 4, 7

[ED04] EISEMANN E., DURAND F.: Flash photography enhancement via intrinsic relighting. *ACM Trans. Graph. 23*, 3 (Aug. 2004), 673–678. doi:10.1145/1015706.1015778. 2

[EVGW*10] EVERINGHAM M., VAN GOOL L., WILLIAMS C. K., WINN J., ZISSERMAN A.: The pascal visual object classes (voc) challenge. *International Journal of Computer Vision 88*, 2 (2010), 303–338. doi:10.1007/s11263-009-0275-4. 4

[FFL10] FARBMAN Z., FATTAL R., LISCHINSKI D.: Diffusion maps for edge-aware image editing. *ACM Trans. Graph. 29*, 6 (Dec. 2010), 145:1–145:10. doi:10.1145/1882261.1866171. 1

[FFLS08] FARBMAN Z., FATTAL R., LISCHINSKI D., SZELISKI R.: Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Trans. Graph. 27*, 3 (Aug. 2008), 67:1–67:10. doi:10.1145/1360612.1360666. 2

[GO11] GASTAL E. S. L., OLIVEIRA M. M.: Domain transform for edge-aware image and video processing. *ACM Trans. Graph. 30*, 4 (July 2011), 69:1–69:12. doi:10.1145/2010324.1964964. 2

[HGY15] HUANG W., GONG X., YANG M. Y.: Joint object segmentation and depth upsampling. *IEEE Signal Processing Letters 22*, 2 (2015), 192–196. doi:10.1109/LSP.2014.2352715. 2

[HST10] HE K., SUN J., TANG X.: Guided image filtering. In *European Conference on Computer Vision (ECCV)* (2010), Springer, pp. 1–14. 1

[HZC*13] HANE C., ZACH C., COHEN A., ANGST R., POLLEFEYS M.: Joint 3d scene reconstruction and class segmentation. In *Computer Vision and Pattern Recognition (CVPR)* (2013), pp. 97–104. 2

[JGK*17] JAFARI O. H., GROTH O., KIRILLOV A., YANG M. Y., ROTHER C.: Analyzing modular cnn architectures for joint depth prediction and semantic segmentation. In *International Conference on Robotics and Automation (ICRA)* (2017). doi:10.1109/ICRA.2017.7989537. 1, 2

[KD] KYPRIANIDIS J. E., DÖLLNER J.: Image abstraction by structure adaptive filtering. In *TPCG*, pp. 51–58. doi:10.2312/LocalChapterEvents/TPCG/TPCG08/051-058. 2

[KEE13] KARACAN L., ERDEM E., ERDEM A.: Structure-preserving image smoothing via region covariances. *ACM Trans. Graph. 32*, 6 (Nov. 2013), 176:1–176:11. doi:10.1145/2508363.2508403. 1, 2

[KK11] KRÄHENBÜHL P., KOLTUN V.: Efficient inference in fully connected crfs with gaussian edge potentials. In *Neural Information Processing Systems (NIPS)* (Granada, Spain., 2011), pp. 109–117. 4, 6

[KLD*14] KUNDU A., LI Y., DELLAERT F., LI F., REHG J. M.: Joint semantic segmentation and 3d reconstruction from monocular video. In *European Conference on Computer Vision (ECCV)* (2014), Springer, pp. 703–718. 2

[LKZ*17] LARSSON M., KAHL F., ZHENG S., ARNAB A., TORR P. H. S., HARTLEY R. I.: Learning arbitrary potentials in crfs with gradient descent. *CoRR abs/1701.06805* (2017). 2

[LSD15] LONG J., SHELHAMER E., DARRELL T.: Fully convolutional networks for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3431–3440. 1, 2, 4

[LSR*12] LADICKÝ L., STURGESS P., RUSSELL C., SENGUPTA S., BASTANLAR Y., CLOCKSIN W., TORR P. H.: Joint optimization for object class segmentation and dense stereo reconstruction. *International Journal of Computer Vision 100*, 2 (2012), 122–133. doi:10.1007/s11263-011-0489-0. 2

[PHK11] PARIS S., HASINOFF S. W., KAUTZ J.: Local laplacian filters: Edge-aware image processing with a laplacian pyramid. *ACM Trans. Graph. 30*, 4 (July 2011), 68:1–68:12. doi:10.1145/2010324.1964963. 2

[PM90] PERONA P., MALIK J.: Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*, 7 (1990), 629–639. doi:10.1109/34.56205. 2

[PSA*04] PETSCHNIGG G., SZELISKI R., AGRAWALA M., COHEN M., HOPPE H., TOYAMA K.: Digital photography with flash and no-flash image pairs. *ACM Trans. Graph. 23*, 3 (Aug. 2004), 664–672. doi:10.1145/1015706.1015777. 2

[ROF92] RUDIN L. I., OSHER S., FATEMI E.: Nonlinear total variation based noise removal algorithms. *Phys. D 60*, 1-4 (Nov. 1992), 259–268. doi:10.1016/0167-2789(92)90242-F. 2

[SLSJB16] SEVILLA-LARA L., SUN D., JAMPANI V., BLACK M. J.: Optical flow with semantic segmentation and localized layers. In *Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 3889–3898. 2

[SSD09] SUBR K., SOLER C., DURAND F.: Edge-preserving multiscale image decomposition based on local extrema. *ACM Trans. Graph. 28*, 5 (Dec. 2009), 147:1–147:9. doi:10.1145/1618452.1618493. 2

[SSP*16] SCHNEIDER N., SCHNEIDER L., PINGGERA P., FRANKE U., POLLEFEYS M., STILLER C.: Semantically guided depth upsampling. In *German Conference on Pattern Recognition (GCPR)* (2016), Springer, pp. 37–48. doi:10.1007/978-3-319-45886-1_4. 2

[SWRC06] SHOTTON J., WINN J., ROTHER C., CRIMINISI A.: Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European Conference on Computer Vision (ECCV)* (2006), Springer, pp. 1–15. 1, 2, 4

[SZXJ15] SHEN X., ZHOU C., XU L., JIA J.: Mutual-structure for joint filtering. In *International Conference on Computer Vision (ICCV)* (2015), pp. 3406–3414. 2

[TM98] TOMASI C., MANDUCHI R.: Bilateral filtering for gray and color images. In *International Conference on Computer Vision (ICCV)* (1998), IEEE, pp. 839–846. 1, 2

[VRT13] VINEET V., ROTHER C., TORR P.: Higher order priors for joint intrinsic image, objects, and attributes estimation. In *Neural Information Processing Systems (NIPS)* (2013), pp. 557–565. 2

[WSL*15a] WANG P., SHEN X., LIN Z., COHEN S., PRICE B., YUILLE A. L.: Joint object and part segmentation using deep learned potentials. In *International Conference on Computer Vision (ICCV)* (December 2015). 6

[WSL*15b] WANG P., SHEN X., LIN Z., COHEN S., PRICE B., YUILLE A. L.: Towards unified depth and semantic prediction from a single image. In *Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 2800–2809. 2

[XLXJ11] XU L., LU C., XU Y., JIA J.: Image smoothing via l0 gradient minimization. *ACM Trans. Graph. 30*, 6 (Dec. 2011), 174:1–174:12. doi:10.1145/2070781.2024208. 2

[XRY*15] XU L., REN J., YAN Q., LIAO R., JIA J.: Deep edge-aware filters. In *International Conference on Machine Learning (ICML)* (2015), pp. 1669–1678. 1, 2

[XYXJ12] XU L., YAN Q., XIA Y., JIA J.: Structure extraction from texture via relative total variation. *ACM Trans. Graph. 31*, 6 (Nov. 2012), 139:1–139:10. doi:10.1145/2366145.2366158. 1, 2, 4, 5

[Yan16] YANG Q.: Semantic filtering. In *Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4517–4526. 1, 2, 3, 4, 5

[ZCW*14] ZHENG S., CHENG M.-M., WARRELL J., STURGESS P., VINEET V., ROTHER C., TORR P. H.: Dense semantic image segmentation with objects and attributes. In *Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 3214–3221. 1, 2

[ZDXZ15] ZHANG F., DAI L., XIANG S., ZHANG X.: Segment graph based image filtering: Fast structure-preserving smoothing. In *International Conference on Computer Vision (ICCV)* (2015), pp. 361–369. 2

[ZJRP*15] ZHENG S., JAYASUMANA S., ROMERA-PAREDES B., VINEET V., SU Z., DU D., HUANG C., TORR P. H.: Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision (ICCV)* (2015), pp. 1529–1537. 2

[ZSXJ14] ZHANG Q., SHEN X., XU L., JIA J.: Rolling guidance filter. In *European Conference on Computer Vision (ECCV)* (2014), Springer, pp. 815–830. 1, 2