Combining Semantic Scene Priors and Haze Removal for Single Image Depth Estimation

Ke Wang Enrique Dunn Joseph Tighe Jan-Michael Frahm
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
{kewang,dunn,jtighe,jmf}@cs.unc.edu

Abstract

We consider the problem of estimating the relative depth of a scene from a monocular image. The dark channel prior, used as a statistical observation of haze free images, has been previously leveraged for haze removal and relative depth estimation tasks. However, as a local measure, it fails to account for higher order semantic relationship among scene elements. We propose a dual channel prior used for identifying pixels that are unlikely to comply with the dark channel assumption, leading to erroneous depth estimates. We further leverage semantic segmentation information and patch match label propagation to enforce semantically consistent geometric priors. Experiments illustrate the quantitative and qualitative advantages of our approach when compared to state of the art methods.

1. Introduction

Recovering depth from images is a fundamental problem in computer vision, and has important applications including robotics, surveillance, scene understanding and 3D reconstruction. Most work on visual reconstruction and depth estimation has focused on leveraging multi-view correspondences for depth estimation leaving out the case of single image based depth estimation. The ubiquity of monocular imaging systems in consumer products provides a large collection of archived imagery for which multiple view approaches are not applicable. We consider the problem of estimating relative depth from a single monocular image to infer structural information within the imaged scene.

Recent work on single view depth estimation [6, 11, 15] has focused on extracting image features and geometric information, in order to apply machine learning techniques to approximate a direct mapping from image features to absolute or relative depth. However, since visual appearance is generally insufficient to resolve depth ambiguities, a tremendous burden is posed on the learning algorithm to implicitly reason about the ambiguities. Conversely, the depth perception ability of humans relies on semantic understanding of a scene intuitively. Liu *et al.* [12] incorporated semantic knowledge to unburden the learning algorithm, exploiting geometry priors of each semantic class to simplify the depth prediction model.

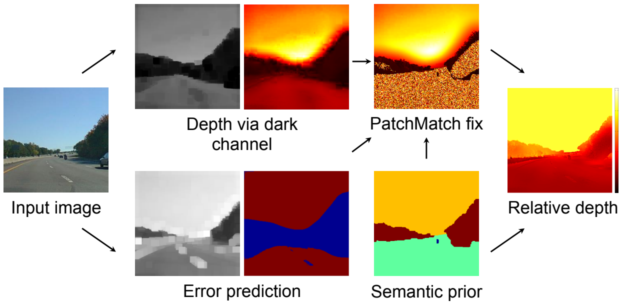


Figure 1. Overview of our proposed approach for non-parametric single view relative depth estimation. Our approach combines PatchMatch depth propagation with semantic priors to achieve semantically congruent results.

We propose a simple monocular depth estimation method, which builds upon previous work based on the dark channel prior [8]. We utilize the dark channel prior to get an initial relative depth estimate. Then we introduce a complementary bright channel to identify potentially erroneous depth estimates attained from the dark channel prior. Additionally, we leverage semantic information to develop a depth propagation framework to correct pixel depth estimates for which the dark channel prior is misleading. Figure 1 depicts an overview of our pipeline. Our method is therefore able to use a single image to achieve semantically congruent depth estimates (See Figure 2 for an example).

2. Related Work

Different approaches have been explored to address the problem of understanding 3D information from monocular images. Our approach is inspired by the relative depth es-

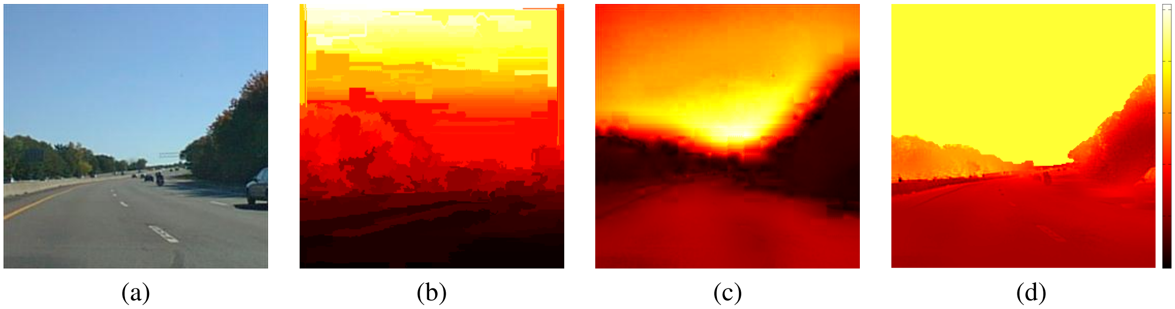


Figure 2. Depth estimation results. Black to yellow gradient depicts increasing scene depth. (a) Input image. (b) Depth estimation by Make3D[15]. (c) Depth estimation by dark channel prior alone[8]. (d) Depth estimation using our approach. Correct depth ordering and detailed depth boundaries are obtained by our approach in image regions not complying with the dark channel prior (e.g. road surface).

timation of patches as a function of haze through the dark channel prior proposed by He *et al.* [8]. In order to estimate the haze it also determines the global atmospheric light as the RGB value of the brightest pixel in the dark channel. To further improve the accuracy of the atmospheric light component Yeh *et al.* [20] introduce the bright channel prior. They leverage the difference between dark channel and bright channel for estimating the improved global atmospheric light. Our approach takes this further by leveraging the difference between dark and bright channel prior to predict unreliable depth estimates caused by the visual appearance of the patches as for example observed for bright scene parts.

Zhou *et al.* [21] built optical models by placing an optical diffuser between the camera lens and the scene, obtaining depth estimates from the modeling of the diffused imaging process. Similarly, Chen *et al.* [3] built self-calibrating cameras through optical reflections. Such optical models require extra imaging devices at imaging time and thus cannot be applied to existing monocular images.

Hoiem *et al.* [11] casted the outdoor scene reconstruction task as a multinomial classification problem. In their work, pixels are classified as *ground*, *sky*, or *vertical*. By “popping up” vertical regions they could build a simple 3D model. However, many commonly found objects cannot be classified into these three classes, *e.g.* buildings with angled surfaces cannot be modeled as *vertical*.

Rather than inferring geometric cues from multiple images of the scene, Saxena *et al.* [15] used a supervised learning approach to directly infer absolute depth of the pixel in the images. They simultaneously collected laser range data and imagery of several scenes. Then a supervised learning framework is applied to predict the depth map as a function of the image using the laser range data as ground truth information. Local features as well as global context are used in their discriminatively trained Markov Random Field (MRF) model [15]. To avoid the costly laser range scanning our method does not rely on a database of known images and their related depth. Moreover our method does not require any prior depth estimates allowing us to use the vast number of freely available labeled image datasets [14, 19].

Unlike other approaches which attempt to map from appearance features to depth directly [15], Liu *et al.* [12] first perform a semantic segmentation of the scene. The semantic information is then used to set the depth of the pixels depending on their classes, for example sky pixels will be far away. In contrast our method relies on the dark and bright channel priors for determining the depth of the image pixels. Hence our method significantly simplifies the depth estimation problem.

Given an image of an object, Hassner and Basri [7] combine the known depths of patches of similar objects to produce a plausible depth estimate. A database of objects representing a single class (e.g. hands, human figures) containing example patches of feasible mappings from the appearance to the depth of each objected is first established. Then by optimizing a global target function representing the likelihood of the candidate depths, Hassner *et al.* could synthesize a depth estimate for the segmented object. In contrast, our proposed method not only estimates the depth of a particular object in the scene but rather estimates the relative depth and geometry of the entire scene.

Oswald *et al.* [13] proposed an algorithmic solution for estimating a three-dimensional model of an object observed in a single image. Based on user input, the algorithm interactively determines the objects silhouette and subsequently computes a silhouette-consistent 3D model, which is precisely the globally minimal surface with user-specified volume of the object of interest. Instead of modeling just a single object from its segmented silhouette, we consider the problem of estimating relative depth of the entire scene from a single image.

Recently Gupta *et al.* [6] proposed to use simple physical scene constraints to obtain relative depth from single images. Their method searches for feasible scene compositions based on the material properties implied by the appearance of the scene elements. This leads to a physically plausible scene layout based on a single image of the scene. While our approach also recognizes scene parts by their appearance our depth is entirely driven by the dark channel prior, which acts as a measurement of the scene depth.

Significant progress has been made to address the problem of single image haze removal recently. For images of outdoor scenes there is a direct relationship between haze density and relative depth of a pixel. The success of these methods lies in using a stronger image prior such as variable scene contrast (Tan *et al.* [17]) or the relationship between surface albedo and transmission medium (Fattal [4]).

3. Method

In this section we present our method for predicting depth from a single image. First, we outline the dark channel prior method in Section 3.1 and discuss its shortcomings for depth estimation. Section 3.2 outlines our proposed method for correcting the errors in the dark channel and Section 3.4 further refines these results by leveraging, alternatively, the geometric segmentation system in [10] and the semantic segmentation system presented in [18].

3.1. Initial depth estimation by dark channel

We wish to predict the depth of the pixels in the scene by leveraging the recent advances in single image haze removal. A hazy image is typically modeled as [4, 8, 17]:

$$\mathbf{I}(x) = \mathbf{J}(x)t(x) + A(1 - t(x)) \tag{1}$$

where $\mathbf{I}(x)$ is the observed intensity of our input image, $\mathbf{J}(x)$ is the scene radiance (i.e. haze free image), A is the global atmospheric light, and $t(x)$ is the medium transmission describing the portion of the light that is not scattered by the atmosphere and reaches the camera. The goal for haze removal is to estimate $\mathbf{J}(x)$ from $\mathbf{I}(x)$, but often $t(x)$ is also predicted. This can be exploited for depth estimation given that under the assumption that the atmospheric light is homogeneous the following holds:

$$t(x) = e^{-\beta d(x)} \tag{2}$$

where $d(x)$ is the depth of the scene point at pixel x and β denotes the atmospheric scattering coefficient. In this work we leverage the dark channel prior of He *et al.* [8] to predict the medium transmission. For a pixel x the dark channel prior ($I^{dark}(x)$) is defined as the darkest channel value over a patch centered around x:

$$I^{dark}(x) = \min_c \left(\min_{\mathbf{y}\in\Omega(\mathbf{x})} I^c(\mathbf{y}) \right) \tag{3}$$

where c is the color channel (either red, green, or blue), $\Omega(\mathbf{x})$ is the patch centered at pixel x, I^c is channel c of the hazy image. Statistical observation shows that except for the sky region, the intensity of the dark channel is low and tends to be zero for haze-free outdoor images. To predict transmission $t(x)$, He *et al.* [8] invert the dark channel of the normalized haze image $I^c(\mathbf{y})/A^c$:

$$t(x) = 1 - w * \min_c \left(\min_{\mathbf{y}\in\Omega(\mathbf{x})} \frac{I^c(\mathbf{y})}{A^c} \right) \tag{4}$$

where w is a constant used for tuning, and A^c is the estimate of the global atmospheric light for channel c. In our experiments we set $w = 0.95$. A is constant for a given image and is estimated as the RGB value of the pixel with the highest dark channel value. Since transmission within a patch is not always constant, the transmission map generated by Equation 4 contains block effects (see Figure 3(b)). Instead of refining the coarse transmission by matting as proposed in [8], we used a guided image filter [9] to achieve similar results at lower computational cost. We leverage the grayscale image as the guidance, and our refined transmission map captures sharp edge discontinuities, as shown in Figure 3(c). Then depth is estimated from the refined transmission map by $d(x) = -\log(t(x))/\beta$.

The dark channel prior method [8] works by assuming that the atmospheric light is of a neutral color and tries to factor this neutral color out by measuring it from the dark channel. Intuitively this works because in most haze free patches there will be at least one pixel of the following type to ensure a low minimum in at least one channel: 1) shadow pixels, where all channels for a pixel are dark, 2) bright color pixels, where 1 or 2 channels for a pixel are dark, 3) dark object pixels, where all channels for a pixel are dark. This fails when there is a patch entirely containing a bright object. See Figure 4 for examples. The white wall in Figure 4(a) and white door in Figure 4(b) introduce bias towards large depth estimates.

3.2. Identifying unreliable estimates

To identify the patches for which the dark channel prior fails to provide a correct depth estimate our method has to identify the potential candidate patches. The candidate patches having neutral bright color can be of two major classes: sky pixels or patches depicting a bright object. For sky regions depicting partial cloud cover, the contrast between predominantly saturated clear sky and high intensity (white) clouds introduces artificial depth variations. Conversely, image regions containing bright objects will tend to yield artificially high values in the dark channel, indicating a large distance to the camera, causing inconsistent depth orderings that are driven by object/region texture instead of scene haze properties. To overcome such ambiguities we propose to flag these unreliable regions through the computation of a dual channel and the implementation of semantically driven reasoning. Next we will detail the computation

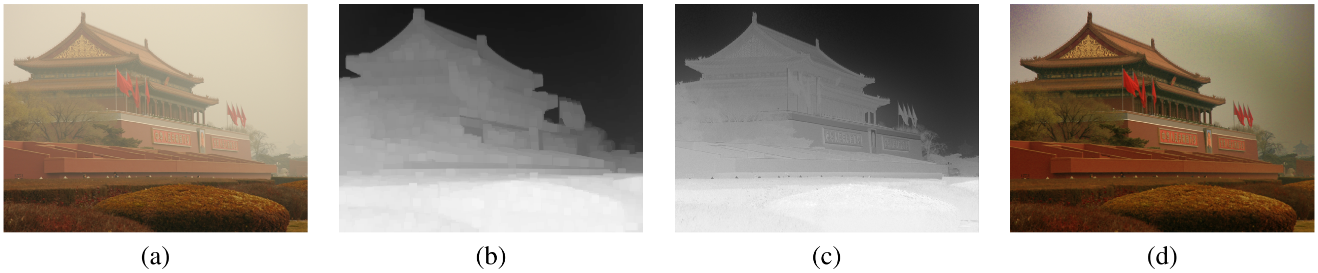


Figure 3. Example of single image haze removal via dark channel prior. (a) Input hazy image $\mathbf{I}(x)$. (b) Estimated coarse transmission map $t(x)$. (c) Refined transmission map after guided image filtering. (d) Final haze-free image $\mathbf{J}(x)$.

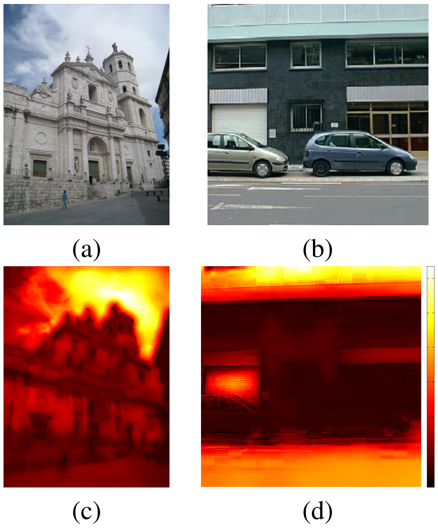


Figure 4. Dark channel prior usually gives wrong depth estimation for bright objects. (a) Input image with white building. (b) Input image with a white door. (c) Depth map for image (a). Depth of the white wall is predicted farther than the building. (d) Depth of the white door is different from the wall it belongs to.

and use of our dual channel prior.

We define the highest channel value within a local patch as the "bright channel":

$$I^{bright}(x) = \max_{c} \left(\max_{\mathbf{y}\in\Omega(\mathbf{x})} I^{c}(\mathbf{y}) \right) \quad (5)$$

The concept of "dual channels" was introduced in Yeh *et al*. [20] for image and video dehazing, where the authors defined these two channels for each pixel to estimate the global atmospheric light A in Equation 1. However, we found that the difference between these two channels may be leveraged to predict unreliable depth estimates.

To find areas where the dark channel prior assumption was unreliable, we take the filtered difference between the highest channel value and the lowest channel value to be below some threshold:

$$unreliable(x) = \mathcal{F}\left(I^{dark}(x) - I^{bright}(x)\right) < \alpha \quad (6)$$

$unreliable(x)$ is true if pixel x is flagged as unreliable. $\mathcal{F}$ is the guided image filter [9]. For all our experiments we empirically set $\alpha = 0.4$. Figure 5 shows an example of this mislabeled prediction. Note that the sky region gets flagged as unreliable, the mitigation strategy for such regions will be described in Section 3.4 where we leverage semantic labels.

3.3. Depth Correction using PatchMatch

For any region that is predicted to be unreliable we need to provide an adjusted depth estimate. We pose this problem as an image completion task and search for matching patches in the input image. We use the PatchMatch technique of Barnes *et al*. [1, 2] to do this efficiently.

For each unreliable depth pixel, we extract the image patch centered around it, and search for visually similar patches centered around reliable depth pixels. Patch similarity is defined via Euclidean distance over pixel values. We initialize each unreliable depth pixel with one reliable pixel position, and transfer the depth accordingly. Then our spatial propagation scheme enables us to find for each pixel labeled as unreliable, an approximate nearest neighbor patch believed to be correct (i.e. not flagged as unreliable). The depth of the matching patch is assigned to the corresponding unreliable pixel (See Figure 6). While such procedure effectively corrects the initially erroneous depth ordering in local neighborhoods, relatively large image regions devoid of reliable pixels (e.g. sky pixels) may be assigned a depth estimate corresponding to arbitrarily distant pixels belonging to separate scene elements. Accordingly, the remaining challenges to be addressed are 1) limiting the scope of depth propagation and 2) enforcing geometric scene priors. We achieve both these goals through the incorporation of semantic image segmentation.

3.4. Leveraging Semantic Segmentation

Semantic segmentation provides higher level pixel associations (i.e. labeled local pixel neighborhoods) that we leverage in our depth propagation framework. We leverage semantic association to determine the scope of our depth propagation, enforce smoothness on our depth estimates

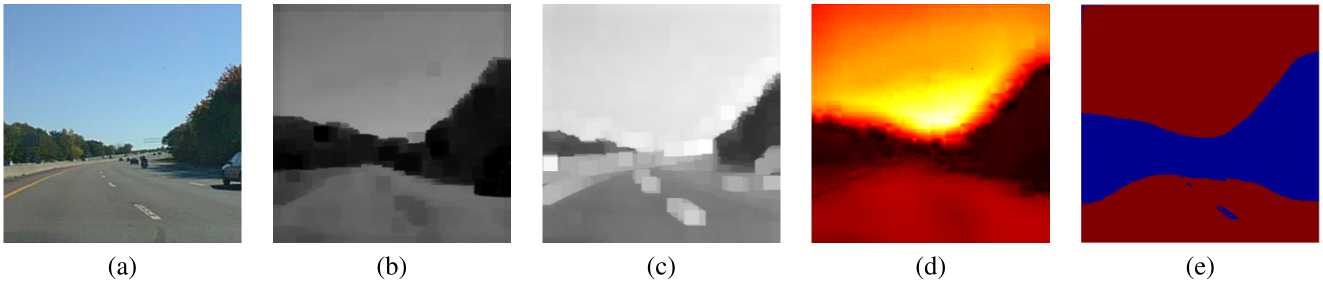


Figure 5. Identifying unreliable estimates. (a) Input image. (b) Dark channel. (c) Bright channel. (d) Depth estimation based on dark channel prior. (e) Unreliable pixels are shown as red. Sky patches and road pixels are predicted as unreliable.

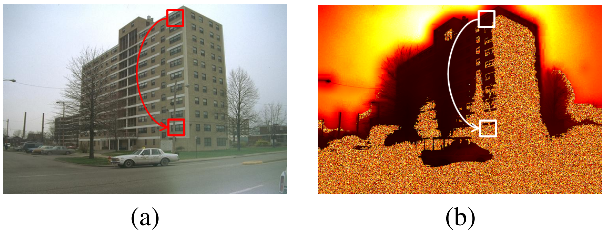


Figure 6. Depth correction using PatchMatch. (a) For unreliable depth estimations, we search for similar patches in correct regions. (b) Depth estimations are transferred accordingly.

within a given semantic region and assign geometric depth priors based on semantic content. We consider the semantic labels generated by approaches of Tighe & Lazebnik [18] and Hoiem *et al*. [10], as shown in Figure 7. The complementary nature of fine grain semantic based object segmentation [18] vs. coarse geometric/structural grouping and labeling [10] is aimed at highlighting the flexibility and generality of our approach.

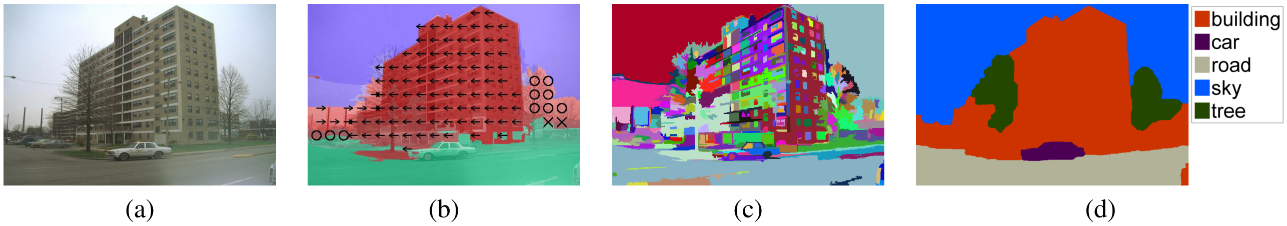
Semantic parsing The use of the system presented in [18] enables accurate fine grain depth boundary estimation from the parsing output. In general, the framework is extensible to include arbitrary number of user specified classes allowing for class specific reasoning regarding scene content and geometry. In this work we leverage the distinction between "stuff" (e.g. ground, sky, grass) and "things" (e.g. cars, signs, bikes). To improve the matching performance, we prefer that matches for patches come from image regions of the same class. For example, if half of the pixels assigned to the "road" label are marked as unreliable, then we only match patches from the reliable road region, rather than any part of the image that was deemed reliable. When a sufficiently large percentage of any given semantic label is flagged as mislabeled (more than 80%), we default back to matching across all un-flagged pixels.

Geometric labeling By extracting the geometric layout of the input image, we can refine depth estimation in a coarse level. The system of [10] labels the geometric layout of the input image into three categories: "horizontal", "vertical" and "sky". Similarly, we find the approximate nearest neighbor patch for each unreliable pixel within the same geometric regions. Coarse level labeling such as [10] can cause the depth map to loose boundaries and details after PatchMatch based depth correction. To attain fine grained boundaries, we further segment the image into superpixels using the efficient graph-based method of Felzenszwalb and Huttenlocher [5]. Then we constrain the nearest neighbor search within the same superpixel, and all candidate patches come from reliable depth pixels. Figure 7 shows the coarse (Figure 7(b)) to fine-grained (Figure 7(c)) segmentation used by our method. If a sufficiently large percentage of any superpixel is flagged as unreliable (over 80%), we loose such constraint to search from all un-flagged pixels.

To mitigate our problem with sky patches we force any pixels determined to be sky by the leveraged labeling [10, 18] to have the original depth predicted by the dark channel prior. Finally, for rich semantic labeled images, we enforce semantic priors between classes, such as "road" or "ground" appears in front of objects it support such as "car" or "building". These relationships between "road", "ground" and "car", "building", are enforced by testing the relative depth ordering against predefined input configurations. If incorrect depth relationship occurs, we uniformly decrease the depth value of the supporting regions until semantic relationships are satisfied. This maintains the depth gradients within the same semantic region, while correcting for depth ordering errors between separate semantic regions. In Figure 8, the side of the building is estimated to be much farther back than the rest of the building; again we correct this by flagging this regions as unreliable.

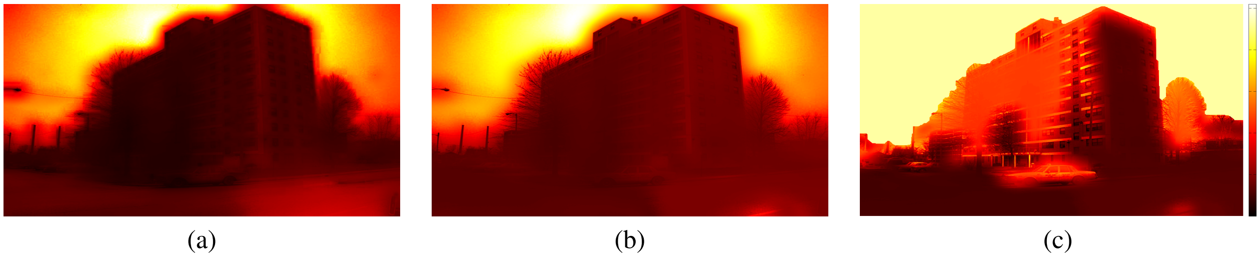
4. Experiments

For evaluation we compare three variations of our proposed method against the *Make3D* system presented in [15] as provided by the authors and our implementation of the *Dark Channel Prior* method [8]. Our first variation is the unconstrained PatchMatch propagation of correct depth estimates directly from the dark channel prior, we will call this variation *PatchMatch*. Our second variation is the use of constrained propagation using the coarse semantic labeling provided by the approach described in Hoiem *et al*. [11], we will refer to this variation as *Geometric Labeling*. Similarly,



(a) (b) (c) (d)

Figure 7. Different segmentation inputs. (a) Input image. (b) Geometric labeling[11]. “Horizontal” is shown in green, “vertical” in red, and “sky” in purple. (c) Superpixel segmentation[5]. (d) Semantic segmentation [18]. Main semantic categories are shown.



(a) (b) (c)

Figure 8. Leveraging semantic information. (a) Dark channel prior based depth estimation. (b) Semantically constrained correction through PatchMatch propagation. Notice road and the supporting building appears at similar depth. (c) Depth after enforcing the semantic priors.

our third and final variation utilizes the fine grain labeling provided by the system described in Tighe and Lazebnik [18], we will deem this variation as *Semantic Labeling* .

We benchmark our method on the Make3D dataset [15, 16], which has 134 test images with ground truth depth maps collected by a laser scanner. In addition, we qualitatively test on the 500 image subset of the LM+SUN dataset used in [18], leveraging the images and semantic segmentation results provided by the authors. In Figure 9, we show examples where our system mitigates errors made by the dark channel prior approach.

For images with colorful objects or objects with enough texture to ensure the dark channel assumption holds, e.g. Figure 9(a), the dark channel prior predicts the correct depth and the only area our system flags as unreliable is the sky, which is subsequently corrected. When the dark channel assumptions are violated, our system successfully flags these regions as shown in Figure 9(b-h). Figure 9 gives multiple examples where our system enforces semantic ordering constraints to migrate errors originated from the use of the dark channel prior. In Figure 9 (b, c, d, & e) the dark channel prior estimates the road to be farther than the cars (b), tree (c, e), or house (d). Our system corrects the depth ordering by semantic order enforcement between road and the supported structures, such as cars, trees and buildings. Thus, we illustrate for a range of outdoor scenes, how we are able to mitigate the bias in depth estimation caused by the dark channel prior.

The results of our quantitative evaluation against ground truth depth maps are presented in Table 1. On the Make3D test set, we compare results using the relative error metric and $\log 10$ error metric, as done originally for *Make3D* [15, 16]. We compute relative error as $e = \frac{|D^* - D|}{D}$ and $\log 10$ error as $e = |\log D^* - \log D|$, where D is the depth map. Since our method predicts the relative depth of the image, we normalize the ground truth depth map from the dataset. Average error on the test dataset can be found in Table 1. On both error metrics, our method outperformed state-of-the-art methods (make3D system[15, 16], dark channel prior[8]).

Table 1. Average error on Make3D dataset. PatchMatch results are acquired based on semantic segmentations. Semantic Prior enforced semantic relationships based on PatchMatch results. Geometric labeling applied PatchMatch based on geometric labelings.

Method	Relative error	$\log 10$ error
Make3D[15, 16]	0.378524	0.522147
Dark channel prior[8]	0.271020	0.490922
PatchMatch[1]	0.261970	0.466927
Semantic Labeling[18]	**0.239856**	0.458029
Geometric Labeling[11]	0.253589	**0.456868**

We also evaluated our method on indoor images as shown in Figure 10. Given that indoors the image formation model of Equation 1 is hardly observable due to the small viewing distances our methods performance degrades significantly.

5. Conclusion

In this paper, we have proposed a simple but effective method based on the dark channel prior to estimate relative depth from single monocular images. Compared with previous works, we summarize our contributions as: 1) We proposed an effective method to predict error-prone areas

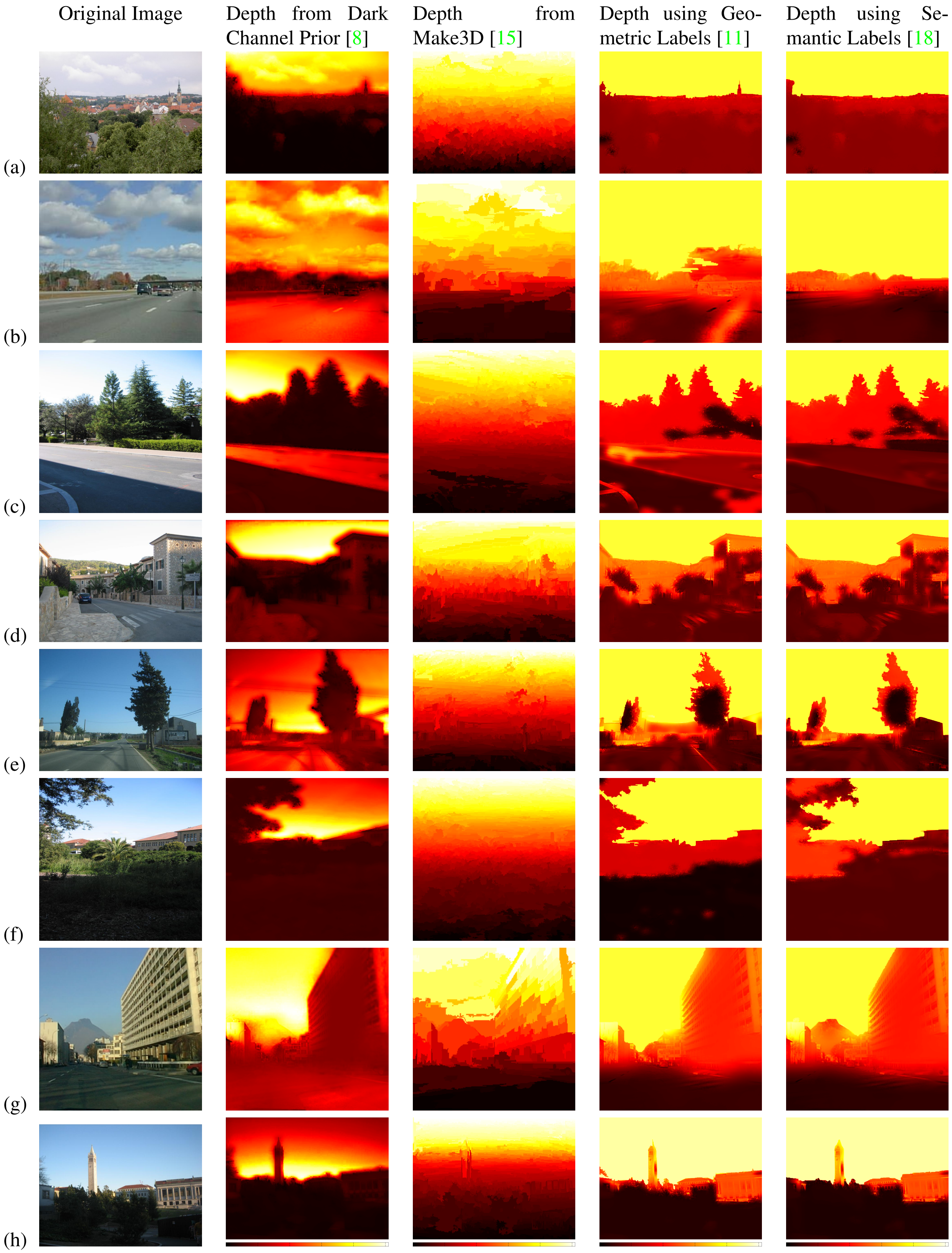
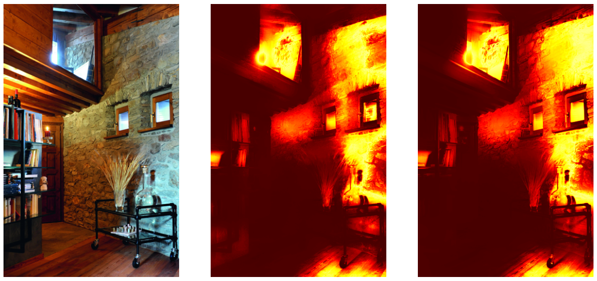

Original Image
Depth from Dark Channel Prior [8]
Depth from Make3D [15]
Depth using Geometric Labels [11]
Depth using Semantic Labels [18]
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)

Figure 9. Comparative results with dark channel prior [8], Make3D [16]. The two rightmost columns depict our estimated relative depth maps using the labelings from [11], [18]. (Best view in color)



Input image Original depth Corrected depth

Figure 10. Our method doesn't work well on indoor scenes.

of the depth estimation by comparing the bright and dark channel priors. 2) We proposed an approach to leverage semantic information to refine the depth estimations.

The proposed framework explored the use of a dual measure to model the suitability of local dark channel prior measurements in the context of depth estimation. We exploited this dual prior within a constrained propagation mechanism to correct likely errors in depth estimation framework proposed by He *et al*. [8]. The constraints on our propagation/correction scheme are provided by semantic labeling attained through state of the art parsing techniques. Accordingly, our framework combines novel prior based depth estimations with semantic reasoning within the context of single view depth estimation.

Our approach is non-parametric, making no assumptions about the underlying model between visual appearance and depth, which is fundamentally different with previous approaches [3, 11, 12, 15, 21]. This makes our algorithm applicable to a wide range of scenes, including highways, coasts, buildings, and forests, as shown in Figure 9. Finally, unlike traditional learning based methods, our approach doesn't require ground truth depth data for training purposes. This makes our methods more applicable to existing images where no ground truth depth data is available.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics*, 28(3):24, 2009. 4, 6

[2] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein. The generalized patchmatch correspondence algorithm. In *ECCV*, pages 29–43. Springer, 2010. 4

[3] Z. Chen, K. Y. K. Wong, Y. Matsushita, X. Zhu, and M. Liu. Self-calibrating depth from refraction. In *ICCV*, pages 635–642, 2011. 2, 8

[4] R. Fattal. Single image dehazing. In *ACM SIGGRAPH 2008 Papers*, SIGGRAPH '08, pages 72:1–72:9, New York, NY, USA, 2008. ACM. 3

[5] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004. 5, 6

[6] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*, 2010. 1, 2

[7] T. Hassner and R. Basri. Example based 3d reconstruction from single 2d images. In *CVPR Workshop, 2006.*, pages 15–15, 2006. 2

[8] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. In *CVPR*, pages 1956–1963. IEEE, 2009. 1, 2, 3, 5, 6, 7, 8

[9] K. He, J. Sun, and X. Tang. Guided image filtering. In *ECCV*, pages 1–14. Springer, 2010. 3, 4

[10] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007. 3, 5

[11] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, volume 1, pages 654–661. IEEE, 2005. 1, 2, 5, 6, 7, 8

[12] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *CVPR*, pages 1253–1260. IEEE, 2010. 1, 2, 8

[13] M. R. Oswald, E. Toppe, and D. Cremers. Fast and globally optimal single view reconstruction of curved objects. In *CVPR*, pages 534–541, 2012. 2

[14] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, May 2008. 2

[15] A. Saxena, S. H. Chung, and A. Ng. Learning depth from single monocular images. *NIPS*, 18:1161, 2006. 1, 2, 5, 6, 7, 8

[16] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 31(5):824–840, 2009. 6, 7

[17] R. T. Tan. Visibility in bad weather from a single image. In *CVPR*, pages 1–8, 2008. 3

[18] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, June 2013. 3, 5, 6, 7

[19] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010. 2

[20] C.-H. Yeh, L.-W. Kang, C.-Y. Lin, and C.-Y. Lin. Efficient image/video dehazing through haze density analysis based on pixel-based dark channel prior. In *Information Security and Intelligence Control (ISIC), 2012 International Conference on*, pages 238–241, 2012. 2, 4

[21] C. Zhou, O. Cossairt, and S. Nayar. Depth from diffusion. In *CVPR*, pages 1110–1117. IEEE, 2010. 2, 8