Signal Processing: Image Communication 90 (2021) 116048

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Single-image depth estimation by refined segmentation and consistency reconstruction

Huajun Liu [a,*], Dian Lei [a], Qing Zhu [b], Haigang Sui [c], Huanran Zhang [a,*], Ziyan Wang [a]

[a] School of Computer Science, Wuhan University, China
[b] Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, China
[c] State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, China

ARTICLE INFO

Keywords:
Depth estimation
Image segmentation
Consistency reconstruction
Single image

ABSTRACT

Recent years have witnessed tremendous success of single-image depth estimation. However, most of the existing approaches merely use scene descriptions of a whole image to retrieve its candidates, which may end up with undesirable depth supports for local regions. In this paper, we propose a segmentation method for single-image depth estimation based on data-driven framework. First, a per-pixel boundary spreading method is presented to improve the image segmentation and provide local regions for image retrieval. Second, a local-region image retrieval is conducted to provide a powerful support for the depth estimation of each segmented part. Third, a scene similarity matrix is constructed and combined with the initial depth prior to establish the correlations across different regions for a consistent depth optimization. Experiments show that applying our method to classic data-driven methods can improve the performance of depth estimation. Besides, our results also manifest clearer depth boundaries in some local regions than the state-of-the-art methods based on deep learning framework.

1. Introduction

Depth estimation is the process of predicting the depth map of a scene. The depth information is quite significant for understanding geometric relationship in the scene, which can provide proper representations for objects and environments and make contributions to various fields in computer vision, such as 3D reconstruction [1–4], scene understanding [5,6], human pose estimation [7–9], *etc*. More recently, depth sensors have been used to simultaneously capture color images and their corresponding depth map. However, for RGB images without any depth information, it seems unpractical and labor-intensive to re-capture their depth information using the depth sensors. In this case, it is still significant and fundamental to recover the depth information for those existing RGB images. Some approaches estimate depth maps from stereo images [10] or motion sequences [11–13], which can provide sophisticated geometric information. In contrast, it is much more challenging to estimate depths from a single image [1,14–17] without using any auxiliary cues like stereo correspondences and temporal information, which has become an active area within computer vision field.

To tackle the challenge of single-image depth estimation, various method are proposed in recent years. Data-driven has been a popular way of depth estimation for single image [18–20], which makes use of similar geometric characteristics in a large scale RGBD database and retrieves image candidates from the database based on appearance similarity. By densely aligning them, the depth information of the input image can be approximately estimated by global optimization [19] or Poisson reconstruction [20]. However, it is not easy to accurately estimate the depth of local regions if we only retrieve candidate images for a whole image, which would lead to unsatisfactory depth estimation results and ambiguous depth boundaries. More recently, with the rapid development of deep learning technology, much attention has been attracted to leveraging convolutional neural networks (CNNs) for single-image depth estimation [14,17,21–26]. Generally, the CNN-based methods focus on extracting abstract features of the input data and then predicting the dense pixel-wise depth map, which, to some extent, may neglect some detail in local areas like object boundaries. Hence, it has been a key issue for these previous methods to achieve an accurate depth estimation for local regions.

Motivated by this unsolved issue, we move a step further in the data-driven category and propose a segmentation strategy for single-image depth estimation, which can provide more powerful supports for local regions and help to produce clear and sharp depth boundary for the results. The first stage is per-pixel boundary spreading, which is proposed to segment the image for the second stage, with the aim to obtain more

* Corresponding authors.
E-mail addresses: huajunliu@whu.edu.cn (H. Liu), dian_lei@whu.edu.cn (D. Lei), zhuq66@263.net (Q. Zhu), haigang_sui@263.net (H. Sui), huanranZ@whu.edu.cn (H. Zhang), zenobia@whu.edu.cn (Z. Wang).

https://doi.org/10.1016/j.image.2020.116048
Received 18 November 2019; Received in revised form 14 September 2020; Accepted 25 October 2020
Available online 31 October 2020

suitable candidates for local regions. Next, image retrieval is conducted to leverage similar scene geometric information in the database, which follows the framework of data-driven approaches. The third stage relies on a newly-introduced consistency constraint to establish associations for segmented scenes for cross-region depth optimization, which can eliminate the incorrect depth results caused by separated estimations of the segmented regions. Experimental results validate that the proposed approach and constraint can be freely applied to the existing data-driven clues to effectively improve the accuracy of depth estimation. As for the state-of-the-art CNN-based methods, our method also performs better in local detail like depth boundaries.

In essence, the contributions of our paper are listed as follows:
(i) a segmentation method is proposed to estimate an accurate depth map for single image, which can be applied to the existing data-driven frameworks, and may provide similar ideas to the neural-network-based approaches;
(ii) a per-pixel boundary spreading approach is proposed to segment the image into parts of similar characteristics, which is designed for local-region image retrieval and better depth estimation;
(iii) a consistency constraint for cross-region depth optimization is introduced to establish associations for segmented scenes and alleviate the depth inconsistency caused by segmentation.

2. Related work

Since the proposed method involves three aspects: depth estimation, image segmentation and consistency alignment, the related work is discussed respectively.

2.1. Depth estimation

A widely-used approach to recover the image depth map is through exploiting video sequences [11–13,27–29], which can provide rich temporal information for understanding the scene structures. Zhang et al. devised a powerful framework of bundle optimization, using photo-consistency and geometric coherence constraints to associate different views and recover the depth map for static scenes merely by a single moving camera [27]. Yang et al. extended the approach to deal with depth estimation of dynamic scenes with the optical flow introduced [28]. However, both of the methods use very few neighboring frames with relatively short baselines for depth optimization. Later, Zhang et al. proposed to recover depth map from a trinocular video sequence by detecting dynamic regions, which could automatically prolong the baselines and improve the accuracy of the depth estimation for both static and dynamic regions [29]. The above methods all work on recovering depth according to multi-view geometric constraints of the same scene.

Another research direction followed by many methods is to estimate depth by utilizing user annotations and semantic labels [30,31], as depth information and semantic information are both naturally associated with scene structure clues. Russell et al. modeled the single-image depth estimation by making use of the geometric class information (ground, standing, attached) and edge relations (support, occlusion, attachment) in human-labeled regions [30]. Liu et al. integrated the geometric and depth priors with the semantic class to improve the quality of depth estimation [31]. However, these two methods require manual collection and annotation of the pixel-wise semantic labels, which can be really tedious and labor-intensive.

What is more, various attempts have been made for single-image depth estimation, one of which is leveraging data-driven approaches, where the depth maps of candidates are retrieved from the existing RGBD databases [18–20,32]. As a milestone, Konrad et al. used the histograms of gradients (HOG) descriptor [33] to retrieve a couple of color images that resemble the input image most, where the depth result is fused simply by computing the median depth value and the depth boundary is refined by a bilateral filter [18]. Karsch et al. proposed a depth transfer algorithm [19], which densely warps candidate images to the input image through a pixel-level dense alignment like SIFT Flow [34] and adopts a global form of depth fusion to estimate final depth map. Targeting at improving the speed and quality of image retrieval, Herrera et al. proposed to obtain depth priors by clustering similar images in the database [32]. To alleviate the dependency on similar databases, Choi et al. proposed a gradient-domain framework to estimate single image depth [20], where depth gradients are transferred as reconstruction cues instead of depth values and integrated with Poisson reconstruction. However, all of the above data-driven methods attempt to retrieve candidates using scene descriptors for the whole image, which may lead to an unsatisfactory retrieval for local scenes and produce final depth maps with vague boundaries. For instance, the depth boundary of the depth map estimated by [19] may not be clear and sharp, while [18,20,32] require filters to post-process and refine the depth boundary.

Another research direction of single-image depth estimation lies on the CNN-based deep learning methods. Eigen et al. [14] proposed a combination of a coarse network to predict a global depth distribution and a fine network to refine the depth map. Later, they [21] extended this work to perform surface normal estimation, semantic label estimation and depth estimation at the same time. Many CNN-based techniques for depth estimation are combined with CRF models [15,22,23]. Laina et al. [24] designed a depth estimation network based on the ResNet architecture [35], which adopted an up-projection structure to improve the resolution of an estimated depth map. Fu et al. [36] proposed the deep ordinal regression network (DORN), which transformed the depth regression task into a classification task and yielded the state-of-the-art depth estimation performance. More recently, Lee et al. [26] proposed an idea of relative depth and present an estimation method for relative depth maps. Undoubtedly, CNN-based methods are able to achieve optimal depth maps in global review, as they focus on extracting high-level features of the input data and predicting pixel-wise depth maps. But to some extent, these methods may also miss some detail in local depth boundaries.

2.2. Image segmentation

In order to enhance the supports for the local regions and obtain clear depth boundaries, we propose to segment the input image and retrieve candidates for the segmented parts separately. To the best of our knowledge, classic methods for image segmentation include self-adaptive segmentation approaches [37–39] and the handcrafted segmentation approaches [40,41]. The former ones can produce relatively desirable boundary adherence by over-segmentation, but it excessively cuts off the depth relations for cross-regions, which may lead to intractable depth errors. And the latter ones may easily result in rough segmentation boundary, thus they may require further boundary processing for roughly-segmented images. Different from previous methods that rely too much on post-processing the depth boundary with joint filters [18,20,32], we arrange a segmentation stage to enhance the supports for local scenes and acquire sharp depth boundary by our proposed per-pixel boundary spreading.

2.3. Consistency alignment

If we retrieve candidates and predict depths right after the first image segmentation stage, the depth relations will be incorrect due to the separate retrievals for different regions. In other research fields of computer vision, various solutions have been introduced to handle similar issues. For instance, in image stitching, Zaragoza et al. leveraged global-matched features to ensure the smoothness across partitioned grids [42]. With respect to image super-resolution, Sun et al. employed global color and texture gradient of the low-resolution image to compensate the possible artifacts along the salient edges of divided

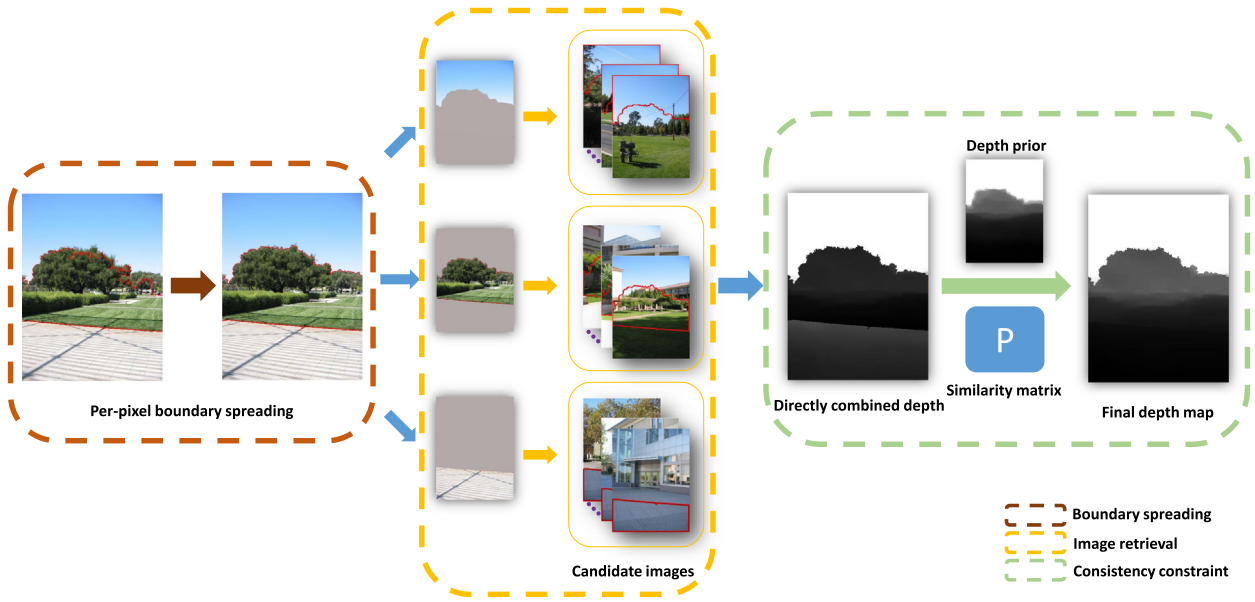

Fig. 1. The pipeline of our method for single image depth extraction. The first stage, *per-pixel boundary spreading*, is to obtain clear depth boundary while ensures a plausible segmentation of local regions. The second stage, *image retrieval for local segments*, is to retrieve candidates for each segmented region according to the data-driven framework. The third stage, *depth estimation with consistency constraint*, is to construct a scene similarity matrix and combine it with the depth prior to establish associations of local regions and obtain the final depth map. The *depth prior* is the depth map produced by the $E_o(D)$ term in Section 3.3, which is the original depth model term proposed in each data-driven baseline. The *directly combined depth* is the depth map directly produced by using $E_o(D)$ on each segment. What is more, the *final depth map* is predicted by combining the above two maps together with our proposed consistency term $E_c(D)$ in Section 3.3.

segments [43]. As for shadow removal work, Wu et al. devised a spatial smoothing approach to exploit the consistency across neighboring edge patches to maintain global edge contours and remove isolated false detection [44]. Inspired by the above methods, we intend to establish global depth relations across the segmented regions, with the aim to improve the accuracy of depth estimation.

Therefore, based on data-driven approaches, we conduct an image segmentation step for a better retrieval, which alleviates the dependency on the scene descriptions of a whole image. Then we establish the relations across different regions to fix the depth errors between the separated regions. In this way, the depth estimation accuracy for single image can be significantly improved when applied to conventional data-driven methods. Besides, compared to the existing deep learning methods, the local detail in depth maps predicted by our method are more clear.

3. Proposed approach

Our method targets at the depth estimation for single image with a newly-introduced segmentation strategy and a novel consistency constraint. The whole pipeline of our method is shown in Fig. 1. The first stage, per-pixel boundary spreading, is designed to obtain clear depth boundary while ensures a plausible segmentation of local regions. In the second stage, image retrieval is conducted for each segmented region according to the data-driven framework. Lastly, in the third stage, a scene similarity matrix is constructed and combined with the depth prior to establish associations of different regions, which can fix the depth inconsistency caused by segmentation. The combination of these three stages can contribute to a better result of depth estimation with clear local detail.

3.1. Per-pixel boundary spreading

In this stage, we resort to SLIC [40] for an initial segmentation of the input image, which is a classic handcrafted segmentation algorithm. However, though SLIC can produce better boundary adherences by segmenting an image into more parts, it would be hard to construct the relations of different regions and result in inaccurate depth prediction

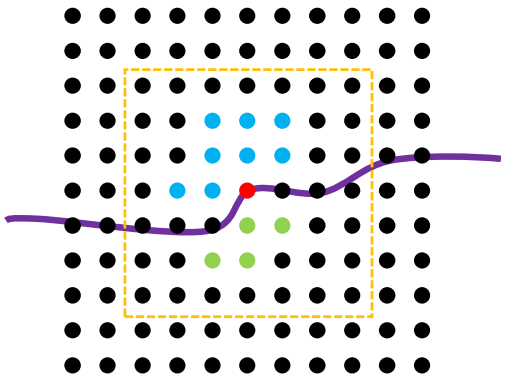

Fig. 2. The neighborhood $N(C_i)$ of the pixel C_i. The purple curve represents the boundary. The red point is the current boundary pixel C_i and the dotted square is its searching range. The blue points are denoted as $\{D_p\}$ and green points are denoted as $\{D_q\}$. In this situation, n_q is smaller than n_p, so the boundary should spread from the blue pixels to the green ones.

(see the experiment in Fig. 4(d–f)). Conversely, segmenting an image into fewer parts may result in rough segmentation boundary(see the experiment in Fig. 3(a)), which makes it necessary to conduct further processing for roughly-segmented images. To alleviate the reliance on post processing like joint filters [18,20,32], we arrange a per-pixel boundary spreading stage to enhance the supports for local scenes and acquire clear depth boundaries.

The first step of boundary spreading is to assign two different labels for pixels respectively on two adjacent regions divided by the initial boundary obtained by SLIC. We adopt HSV color space instead of RGB space to describe the pixels, because the color similarity can be better-distinguished by the H (hue) and S (saturation) channels in HSV space. Therefore, for all the pixels in the neighboring region $N(C_i)$ of a boundary pixel C_i, we translate them to HSV color space, in order to measure the color similarities and distinguish the incorrectly-labeled pixels.

Next, the spreading directions of the current boundaries need to be determined. After translation to HSV space, the colors of boundary pixel C_i and its neighboring pixel D_j ($D_j \in N(C_i)$) are denoted as I_{C_i} and I_{D_j} respectively. If $\left\|I_{D_j} - I_{C_i}\right\|_2^2 < T$ is satisfied, we would consider D_j and C_i similar to each other. In this way, all the similar pixels on two sides of the boundary will be classified as $\{D_p\}$ and $\{D_q\}$ ($D_p, D_q \in N(C_i)$), whose numbers are n_p and n_q respectively. As illustrated in Fig. 2, the spreading direction of the segmentation boundary is determined by the comparison of the pixel numbers n_p and n_q. Namely, if n_q is smaller than n_p, $\{D_q\}$ would be recognized as incorrectly-labeled pixels caused by the initial segmentation, and vice versa. Accordingly, the incorrectly-labeled pixels $\{D_j^m\}$ can be expressed as:

$$D_j^m = \begin{cases} D_q & , \quad n_q < n_p \\ D_p & , \quad n_p < n_q \end{cases} \tag{1}$$

These incorrect labels of those pixels need to be modified, which, can be vividly described as a process of boundary spreading. A new temporary boundary will be generated when all the boundary pixels finish the spreading, during which the number of pixels whose labels has been modified is counted. The spreading process would keep iterating on the temporary boundary pixels until the number of modified pixels is below a threshold P. Finally the spreading for pixels on one boundary is done and the final boundary is fixed.

However, the pixels on temporary boundary of each iteration tend to move towards the spreading direction. Thus, if the comparisons are only performed on each temporary boundary, the spreading would not come to an end as it is always updating. Considering this, we introduce an effective color consistency constraint provided by the initial boundaries, which emphasize the consistency of the whole spreading process. For a temporary boundary pixel C_i', we use the color description $I'_{C_i'}$ to measure the similarity with the adjacent pixels. The dynamically changed $I'_{C_i'}$ can be expressed as:

$$\mathrm{I}'_{C_i'} = w_1 \cdot I_{C_i'} + (1 - w_1) \cdot I_{C_i} \tag{2}$$

where I_{C_i} and $I_{C_i'}$ represent the colors of initial boundary pixel C_i and temporary boundary pixel C_i', respectively. Note that we use a variable for each C_i' to record its initial boundary pixel and build the correlation between them. Besides, w_1 represents the weight of $I_{C_i'}$, which is expressed as $w_1 = e^{-\left\|I_{C_i'} - I_{C_i}\right\|^2/\sigma}$ and is used to maintain the colors consistency of the pixels.

With the above spreading process done, there may still exist some *empty holes* in the segmented image. Thus we ought to deal with them to eliminate the negative impacts on the following depth estimation. We tend to consider a region as an empty hole if the pixel number in this region is below a default threshold R. The label of the pixels in this region would be modified to the label of its surrounding region, then the incorrect boundaries would be eliminated and the empty holes would be removed. Up to now, the image segmentation is done.

3.2. Image retrieval for local segments

As shown in Fig. 1, the second stage, image retrieval, is conducted with the input images that have been segmented using our boundary spreading process. We adopt similar retrieval approaches which have been used in previous data-driven methods [18–20]. Regarding an input image as a reference, scene descriptors are used to calculate the feature similarities between candidates in database and the input reference image. We can adopt descriptors such as GIST and PHOG, which have been used in [19,20]. But different from the retrieving process of these methods, in our work, the image retrieval is conducted for each corresponding segment, thus ensuring a relatively powerful support for elaborate depth estimation for local regions. Likewise, the third stage of our method is also conducted in this way.

To be specific, our image retrieval stage is conducted as follows. First, we process all the training images by partitioning the whole scene into separated parts, according to the segmentation results of the former stage. Then, with respect to local region retrieval, we focus on the similarity of the candidate in the exact same region as in the reference. That is to say, for each independent segment, multiple images in the database would be retrieved as its own candidates.

3.3. Depth estimation with consistency constraint

If the depth of each non-overlapping region is estimated separately in the image retrieval stage, the associations across all the segments will not be consistent enough. To this end, a scene similarity matrix is constructed and combined with depth prior in this stage, which works as a consistency constraint term for adjacent regions.

The consistency constraint proposed by us can be freely applied to any data-driven depth model, no matter what depth clue it adopts. In our paper, we illustrate our method based on two classic frameworks, [19] using depth value clue and [20] using depth gradient clue. The full energy function for depth estimation $E(D)$ is:

$$E(D) = E_o(D) + \lambda E_c(D) \tag{3}$$

where $E_o(D)$ is the initial depth model term used in corresponding data-driven method, and $E_c(D)$ is our newly-proposed consistency term which helps to rebuild the associations between the segmented regions. Besides, λ is the parameter for the term $E_c(D)$, which is set as $\lambda = 1$ in our experiments. Since the function requires an unconstrained, non-linear optimization, we adopt iteratively reweighted least squares(IRLS) to minimize our objective function. IRLS is a desirable solution for this kind of optimization, which leverages a linear function of the parameters to approximates the objective and minimizes the squared residual repetitively until convergence.

As for $E_c(D)$, we first construct a scene similarity matrix P to establish the correlations between the candidates retrieved for segmented regions (local) and the whole image (global). Then we construct the consistency constraint term using P as well as the depth prior.

Assuming the number of local images and global images are M and N respectively, the size of the similarity matrix P is $M \times N$. Regarding the input image as a reference, feature distances between the candidates and the reference can be calculated by scene descriptors used in [19,20]. To be specific, [19] adopts GIST feature vectors (e.g. denoted as G_1 and G_2) to determine the scene distance between two images, which is defined as $d = \|G_1 - G_2\|$. [20] measures the distance between two images by the PHOG descriptors (e.g. denoted as F_1 and F_2), which is defined as $d = \|F_1 - F_2\|_2^2$. Then the P_{lk}, which represents the relation for the lth local image and the kth global image, can be expressed as:

$$P_{lk} = 1 - \left(\left(\frac{d_{loc}^l}{d_{glo}^k} - d_{\min} \right) \Big/ \left(d_{\max} - d_{\min} \right) \right), \tag{4}$$

where d_{loc}^l represents the distance between the lth local image and corresponding part of the reference image, and d_{glo}^k represents the distance between the kth global image and the reference image. Besides, $d_{\max} = \max\left(d_{loc}^l / d_{glo}^k\right)$, $d_{\min} = \min\left(d_{loc}^l / d_{glo}^k\right)$, $l = 1 \ldots M, k = 1 \ldots N$.

[19,20] adopt different depth clues, which are depth value and depth gradient, respectively. Accordingly, for each clue, we propose corresponding $E_c(D)$ term to reconstruct the depth consistency, which are illustrated as follows.

Situation1. Depth value clue.

The original depth model $E_o(D)$ [19] is:

$$E_o(D) = \sum_{i \in pixels} E_t\left(D_i\right) + \alpha E_s\left(D_i\right) + \beta E_p\left(D_i\right). \tag{5}$$

For a single image, this objective in [19] contains three terms: a data term E_t to measure the distance from the inferred depth map D to each of the candidates, a smoothness term E_s to maintain the

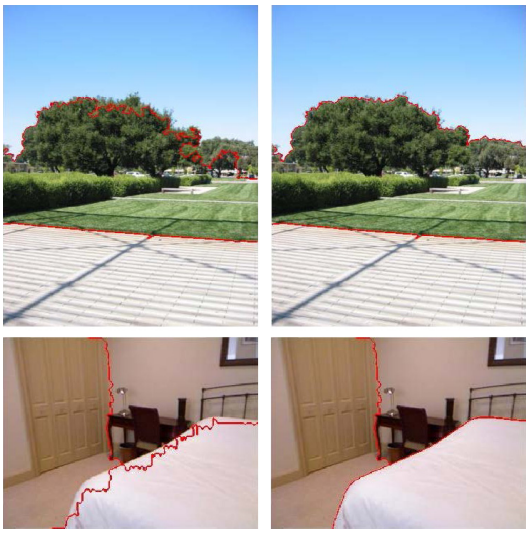

Fig. 3. The examples of the rough segmentation using SLIC and our per-pixel boundary spreading method: (a) the segmentation results of SLIC; (b) the segmentation results of our method.

correlation between texture and depth, and prior term E_p to compute the distance between D and the average depth in the database. In [19], the parameters are set as α=10 and β=0.5.

The consistency constraint term $E_c(D)$ is denoted as:

$$E_c(D) = \sum_{i \in pixels} \phi\left(D_i - \sum_{k=1}^{N}\sum_{l=1}^{M} P_{lk}\omega_i^k D_i'^k\right), \tag{6}$$

where the distance is measured by ϕ, a robust error norm (we use an approximation to L1 norm, $\phi(x) = \sqrt{x^2+\varepsilon}$, with $\varepsilon = 10^{-4}$). $D_i'^k$ is the aligned depth of the kth candidate at pixel i from the global images, and ω_i^k is the accuracy confidence measure of $D_i'^k$ [19]. P_{lk} represents the similarity for the lth local image and the kth global image.

Situation2. Depth gradient clue.

The original depth model $E_o(D)$ [20] is:

$$E_o(D) = \sum_{i \in pixels} E_t\left(D_i\right) = \sum_{i \in pixels} \phi\left(D_i - D_i^l\right), \tag{7}$$

where E_t term represents the distance between the inferred depth map D and each candidate, which is measured by ϕ. D_i^l is the reconstructed depth at pixel i from the lth local image, which determined by a weighted median calculation.

The consistency constraint term $E_c(D)$ is denoted as:

$$E_c(D) = \sum_{i \in pixels} \phi\left(D_k - P_{lk} D_i^k\right), \tag{8}$$

where D_i^k is the reconstructed depth at pixel i from the kth global image, which is determined by a weighted median calculation, and P_{lk} is the similarity between the lth local image in Eq. (7) and the kth global image.

For these two frameworks that our method refers to, we make qualitative and quantitative comparisons in the next section, which are shown in Figs. 6, 7, Tables 1 and 2.

4. Experiments

4.1. Datasets and evaluation metrics

We set training databases and test images using two public RGB-D datasets, i.e. *Make3D* [45] and *NYU* [3]. The baselines compared with our method include data-driven methods [19,20] and deep learning methods [24,26,36]. For explicit demonstrations, we use the same color range for each comparison respectively.

Table 1
The metric comparisons with data-driven methods on Make3D dataset.

Method	REL	$\log_{10}$	RMS
DT	0.361	0.148	15.1
ours + DT clue	0.346	0.138	13.8
DA	0.428	0.176	18.9
ours + DA clue	0.373	0.157	16.7

Table 2
The metric comparisons with data-driven methods on NYUv2 dataset.

Method	REL	$\log_{10}$	RMS
DT	0.374	0.134	1.20
ours + DT clue	0.353	0.126	1.04
DA	0.477	0.156	1.46
ours + DA clue	0.434	0.142	1.32

Make3D database has been used in [19,20]. It contains 534 outdoor images with their depth information, which are captured by a laser range finder from outdoor. There are 400 training images and 134 test images in this database. The resolution of the color images are 1704 × 2272 pixels, and the resolution of the corresponding depth images are 550 × 305 pixels. For consistency of evaluation results, all color and depth images are resized to 460 × 345 pixels by bilinear interpolation.

NYUv2 database has been used in [20,24,26,36]. It is composed by 1449 color images with their depth information, which is captured by the Kinect sensor in indoor scenes. The training and test dataset are not clarified in this database, so a Leave-One-Out approach is adopted for testing in data-driven methods. The resolutions of both color images and depth images are 640 × 480 pixels. For an explicit comparison with the previous work, all of them have been resized to 320 × 240 pixels.

The similarity scores between an estimated depth map and the corresponding ground truth are calculated based on existing evaluation metrics listed as follows:

- Relative (REL) error: $\frac{|D-D^*|}{D^*}$;
- $\log_{10}$ error: $|\log_{10}(D) - \log_{10}(D^*)|$;
- Root mean squared (RMS) error: $\sqrt{\sum_{i=1}^{N}\left(D_i - D_i^*\right)^2 / N}$;

where D is the estimated depth and D^* is the ground truth depth.

4.2. Implementation detail

We demonstrate the performance of our proposed method in comparison with two conventional methods, Depth Transfer (DT) algorithm [19] and Depth Analogy (DA) algorithm [20] qualitatively and quantitatively, both of which are data-driven approaches using a large scale of RGB-D database. We obtain the depth estimation results of DT based on available source code provided by the author. The results of DA and ours are based on our own MATLAB program. Considering the trade-off between accuracy and runtime efficiency, the number K of the candidates to be retrieved is set to 7 throughout our paper. For boundary spreading, the searching region of each boundary pixel is set as 20 × 20. In addition, we set the distance threshold of color similarity T as 0.03 and the pixel number threshold of empty holes R as 1000. The threshold P to control the iteration of boundary spreading is set as 15.

As for the implementation of deep learning baselines [24,26,36], we evaluate all the models on the widely-used dataset NYUv2 for comparison. For [24,36], the depth map results are generated using the

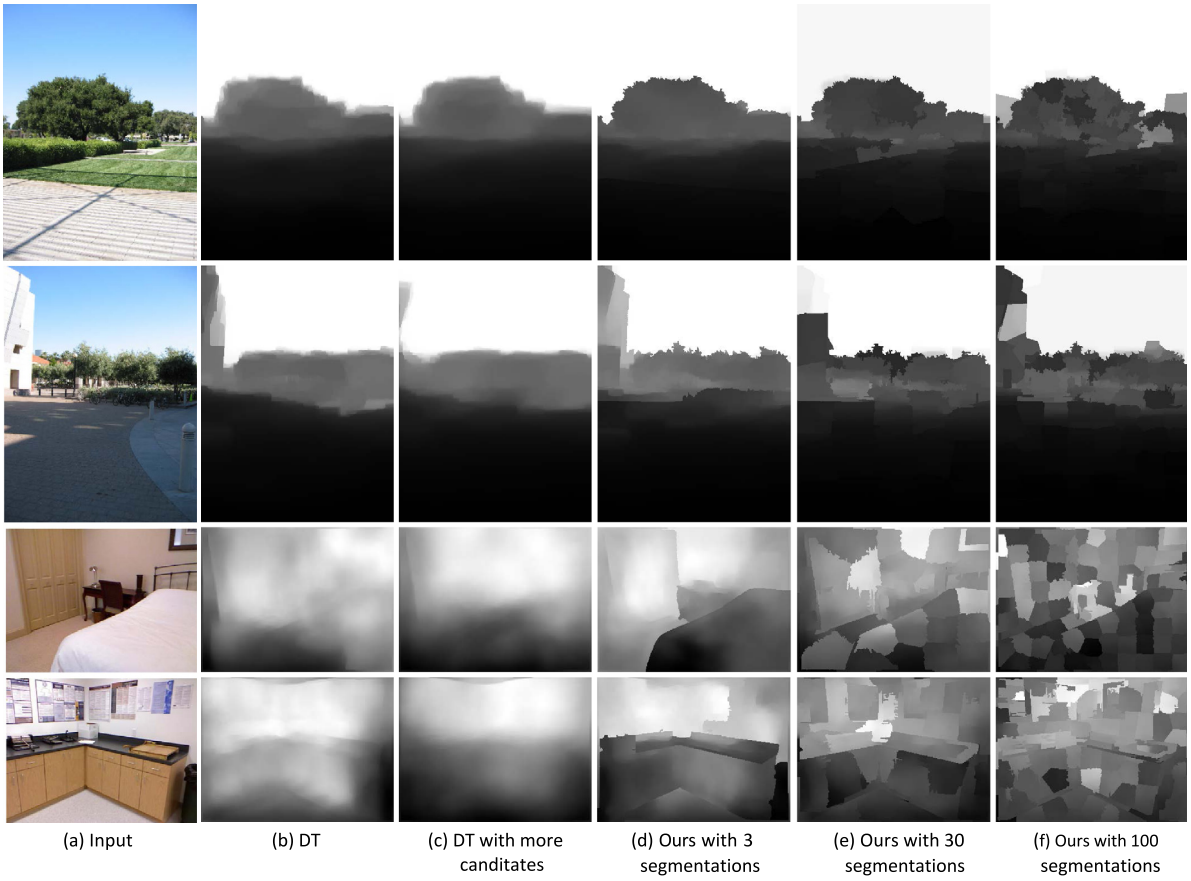

Fig. 4. The depth results using different segmentations and candidate images. (a) input image; (b–c) the depth results of DT with 7 and 21 candidate images respectively; (d–f) the depth results of our method with the image segmented into 3, 30 and 100 regions respectively.

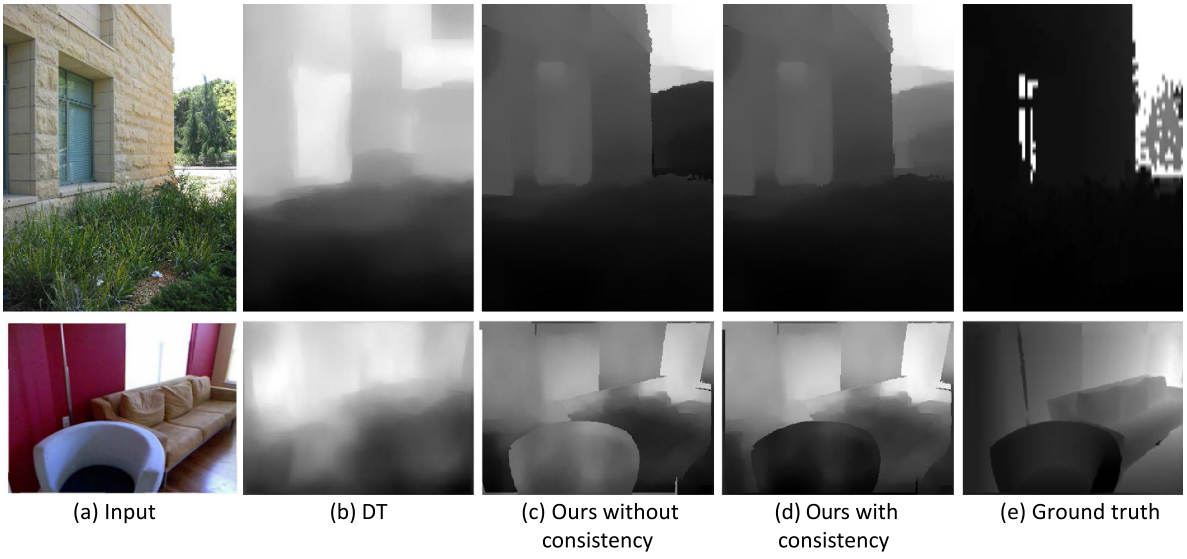

Fig. 5. The comparisons of results with and without consistency constraint. (a) input image; (b) the depth results of DT; (c) our method without consistency; (d) our method with consistency; (e) the ground truth.

source code provided by the respective authors, and the performance scores are excerpted from the respective papers. For [26], the model are trained 140 epochs and evaluated according to the scheme presented in the paper. All the deep learning methods are executed on a single NVIDIA RTX 2080Ti GPU.

4.3. Experimental results

In this section, we first make explicit analysis of our proposed method. Then, we evaluate the qualitative and quantitative results of our method as well as the state-of-the-art baselines.

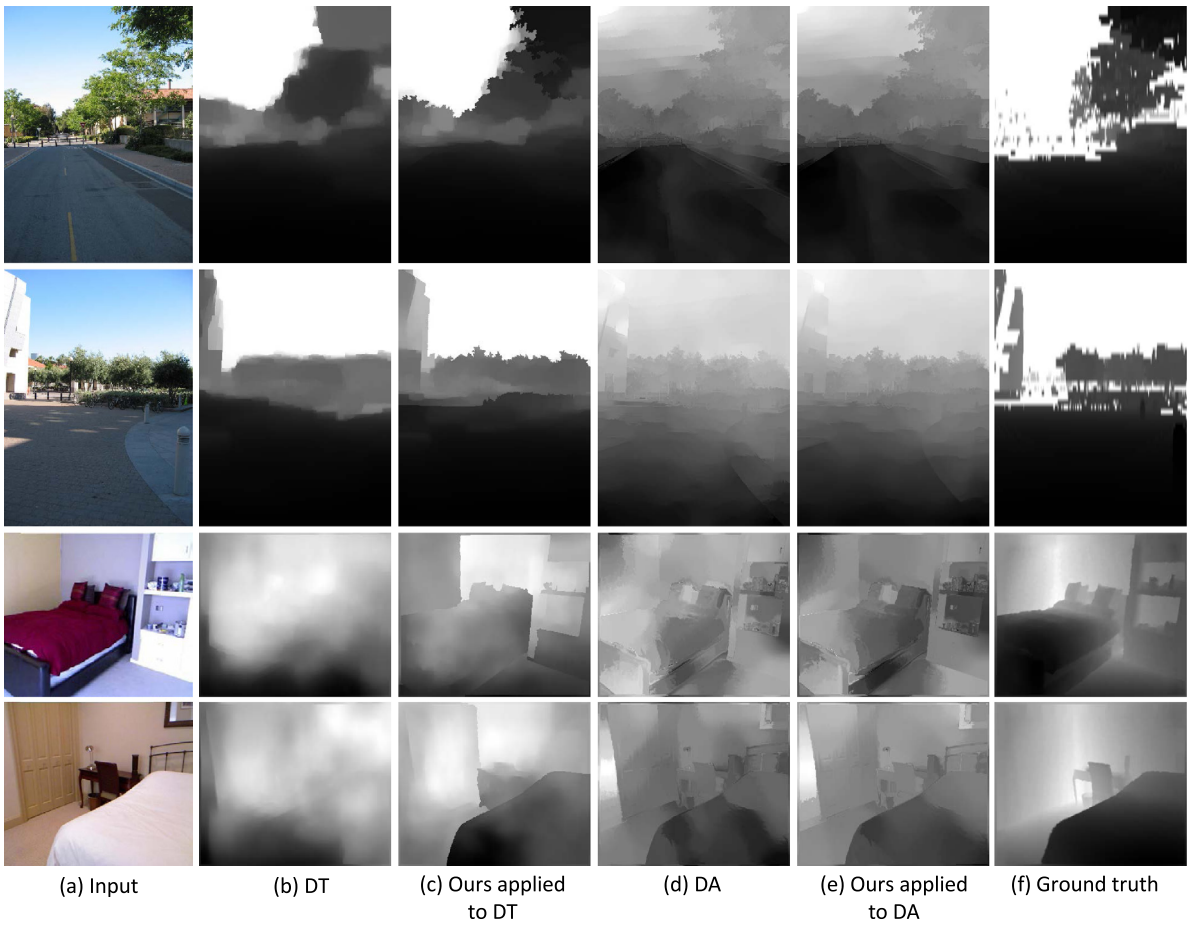

Fig. 6. Experimental results using Make3D dataset and NYUv2 dataset. (a) input image; (b)(d) the depth results of DT and DA; (c)(e) the depth results of our method applied to them; (f) the ground truth.

4.3.1. Analysis of our approach

The parameter setting of segmentation. To determine a proper parameter for the first segmentation step, we conduct experiments using different variations of segmentation numbers. Fig. 4(d–f) show the estimation results with the number set as 3, 30 and 100, respectively. We can observe that the depth map tends to be more scattered with the increase of the segmentation number, which indicates that a larger number of segmentation will excessively cut off the depth relations across regions, making it difficult to maintain depth consistency. Hence, we set the segmentation number as 3 in our experiments. In addition, this experiment could also validate that the over-segmentation of SLIC would not result in desirable depth estimation.

Comparisons with SLIC. With the segmentation number set as 3, we compare the segmentation results of using the proposed per-pixel boundary spreading method with the results of SLIC, as shown in Fig. 3. We can notice that the segmentation boundary obtained by SLIC is quite vague and cannot separate objects well, while our method can deal with pixels more elaborately and obtains a much more satisfying boundary.

The effect brought by image segmentation. As we illustrated in Section 3, the number of retrieved candidates will be larger if the segmentation step is conducted. To this end, we present an experiment to expound that the final effect of depth estimation is due to the powerful local-region support by segmentation, not the increased number of candidate images. We collect all the retrieved images in our method and make DT use the exact same number of candidates in the corresponding scene to estimate the depth map. The comparisons are shown in Fig. 4(c, d). It can be noticed that compared to our result, DT do not achieve desirable depth map even when using more candidates, which indicates the effect of the first image segmentation step.

The effect of consistency constraint. The essence of our presented consistency constraint is also evaluated and demonstrated. We show the comparison results with and without the consistency constraint in Fig. 5. The results without the consistency constraint are obtained by directly combining the depths of the segmented regions. We can see in Fig. 5(c) that there exist incorrect depth results. By contrast, with the consistency constraint used, the depth relation is much more smoothed and a convincing depth map can be obtained, as shown in Fig. 5(d).

4.3.2. Comparisons with classic data-driven methods

Fig. 6 shows the depth results of two classic models and our method applied to them on both datasets. It is clear that applying our method to both models is able to produce better depth map than using DT or DA alone. The depth boundaries of DT are over-smoothed across different local objects, and are quite vague in local regions due to the global scene description. As for the results of DA, there also exist inaccurate depths in some local scenes. Meanwhile, in reference to the ground truth, our approach manifest more comparable and accurate local depths than both models. More results of our method applied to DA and DT using Make3D dataset and NYUv2 dataset are shown in Fig. 7.

Moreover, Table 1 shows the quantitative results of the two traditional models and our method applied to them, which is conducted on Make3D dataset, while Table 2 shows the comparison results on NYUv2 dataset. It can be observed that our proposed method outperforms DT and DA in all evaluation metrics, indicating a higher quality of depth estimation by our method.

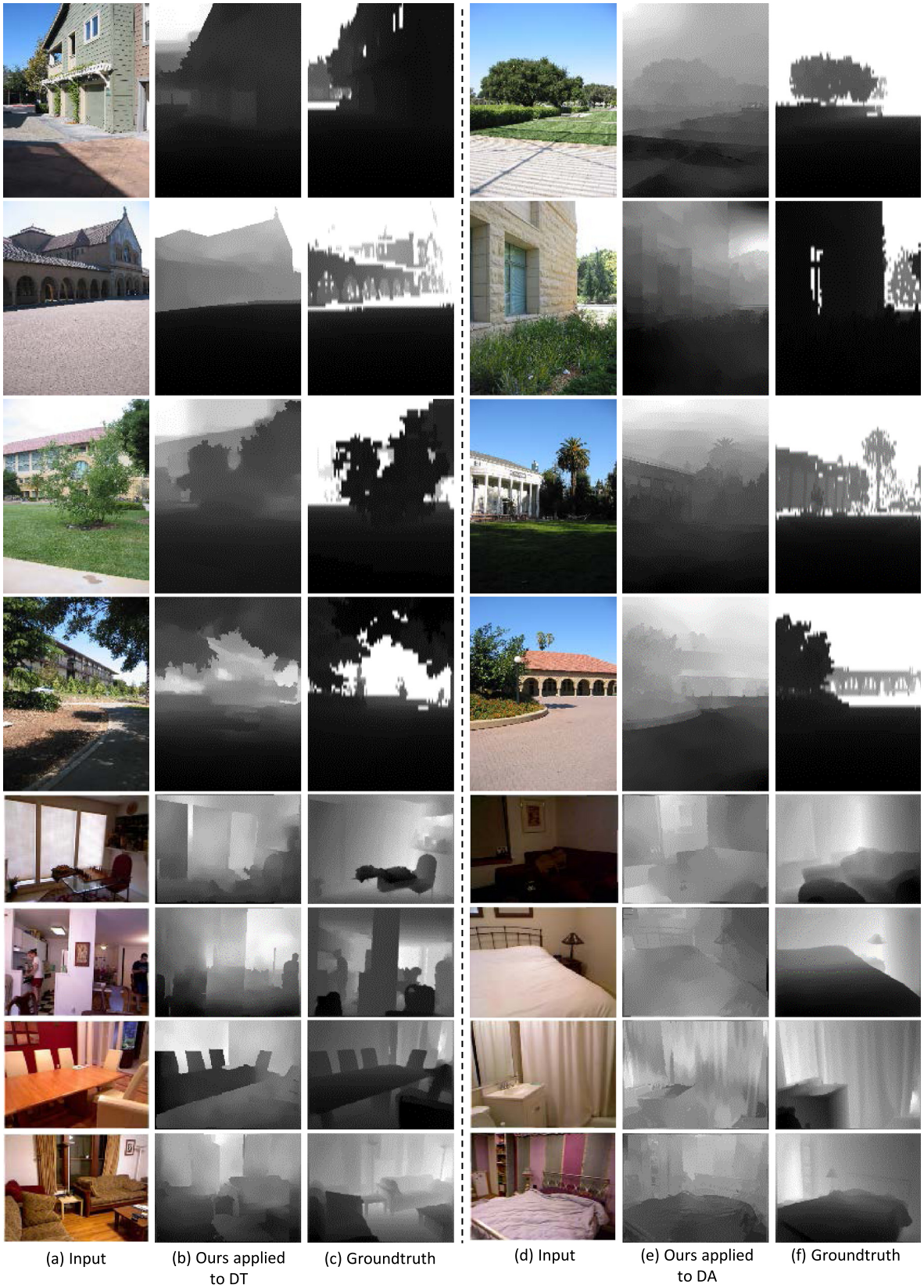

Fig. 7. More results of our method applied to DA and DT using the Make3D dataset and the NYUv2 dataset. (a) (d) the input image; (b) the depth results of our method applied to DT; (e) the depth results of our method applied to DA; (c) (f) the ground truth.

4.3.3. *Comparisons with deep learning methods*

As we can see in Table 3, the quantitative performances of the state-of-the-art CNN-based methods are quite satisfying in global view. Fig. 8 shows the qualitative comparisons with CNN-based models for depth estimation on NYUv2 dataset. In general, depth maps predicted by the deep learning methods are more desirable than our method, as they

Fig. 8. Comparisons of the state-of-the-art deep learning methods and our method applied to DT on NYUv2 dataset.

Table 3
The metric comparisons with deep learning methods on NYUv2 dataset.

Method	REL	$\log_{10}$	RMS
ours + DT clue	0.353	0.126	1.04
Laina et al. [24]	0.127	0.055	0.573
Fu et al. [36]	0.115	0.051	0.509
Lee et al. [26]	0.135	0.055	0.539

extract high-level features of the scene by learning a large number of parameters for the deep neural networks. However, we can discover that it is not easy for them to acquire clear and sharp boundaries where there are depth saltations, e.g. the bed in the first row, the chairs in the second and last row, the lamp in the third row, and the persons in the fourth row. As for our results, some clear and sharp depth boundaries can be restored, such as the depth detail around the chair edges and the person areas. Thus, deep learning methods may have some inherent flaws due to the lost information in abstract features, which also inspires us to combine our work with CNN models to achieve higher quality of depth estimation.

5. Discussions

As we aim to improve the depth estimation by a segmentation strategy, the run-time of retrieving the candidates would be longer depending on the segmentation number. Moreover, as a data-driven approach, our method leverages descriptors like GIST, HOG or PHOG, which has to retrieve candidates in the corresponding regions of two images. In this way, our segmentation approach tends to retrieve the candidates for segmented regions at similar location of the other image, and the depth information of other parts cannot be leveraged. For example, if a tree on the right side was segmented, but the database only contains an image with a quite similar tree on the left side, it would not be retrieved by our method. This is the limitation we will work on in the future.

What is more, as we stated above, there might exist some inherent drawbacks in the prevalent deep learning methods, due to the general process of extracting abstract features and upsampling to make dense predictions. We are quite motivated by this actuality, and we are willing to extend our work in the future and combine it with neural networks to achieve higher quality of depth estimation.

6. Conclusion

In this paper, we propose a data-driven approach of estimating the depth of single image using a segmentation strategy, where a per-pixel boundary spreading approach is presented to provide better supports for local regions. Besides, a similarity matrix is constructed and combined with the initial depth prior to associate the segmented regions. Experiments have shown that our method can be applied to data-driven frameworks to obtain better depth information and boundary results.

CRediT authorship contribution statement

Huajun Liu: Conceptualization, Methodology, Writing - original draft, Project administration, Funding acquisition. **Dian Lei:** Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Qing Zhu:** Supervision, Funding acquisition. **Haigang Sui:** Writing - review & editing, Funding acquisition. **Huanran Zhang:** Software, Validation, Visualization. **Ziyan Wang:** Data curation, Validation, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) (41771427, 41631174 and 41771457).

References

[1] D. Hoiem, A.A. Efros, M. Hebert, Automatic photo pop-up, in: ACM SIGGRAPH 2005 Papers, 2005, pp. 577–584.
[2] A. Saxena, M. Sun, A.Y. Ng, Learning 3-D Scene Structure from a Single Still Image, 2007, pp. 1–8.
[3] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGBD images, in: European Conference on Computer Vision, Springer, 2012, pp. 746–760.
[4] R. Mur-Artal, J.D. Tardós, Orb-slam2: An open-source slam system for monocular, stereo, and RGB-D cameras, IEEE Trans. Robot. 33 (2017) 1255–1262.
[5] X. Ren, L. Bo, D. Fox, RGB-(D) Scene Labeling: Features and Algorithms, 2012, pp. 2759–2766.
[6] S. Song, S.P. Lichtenberg, J. Xiao, Sun RGB-D: A RGB-D scene understanding benchmark suite, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 567–576.
[7] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M.J. Finocchio, R. Moore, A.A. Kipman, A. Blake, Real-Time Human Pose Recognition in Parts from Single Depth Images, 2011, pp. 1297–1304.
[8] J. Taylor, J. Shotton, T. Sharp, A. Fitzgibbon, The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 103–110.
[9] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, et al., Efficient human pose estimation from single depth images, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2012) 2821–2840.
[10] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, Int. J. Comput. Vis. 47 (2002) 7–42.
[11] G.J. Brostow, J. Shotton, J. Fauqueur, R. Cipolla, Segmentation and recognition using structure from motion point clouds, in: European Conference on Computer Vision, Springer, 2008, pp. 44–57.
[12] A. Flint, D. Murray, I. Reid, Manhattan scene understanding using monocular, stereo, and 3d features, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2228–2235.
[13] K. Yamaguchi, D. McAllester, R. Urtasun, Efficient joint segmentation, occlusion labeling, stereo and flow estimation, in: European Conference on Computer Vision, Springer, 2014, pp. 756–771.
[14] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, in: Advances in Neural Information Processing Systems, 2014, pp. 2366–2374.
[15] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, M. He, Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1119–1127.
[16] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2650–2658.
[17] A. Chakrabarti, J. Shao, G. Shakhnarovich, Depth from a single image by harmonizing overcomplete local network predictions, in: Advances in Neural Information Processing Systems, 2016, pp. 2658–2666.
[18] J. Konrad, M. Wang, P. Ishwar, C. Wu, D. Mukherjee, Learning-based, automatic 2D-to-3D image and video conversion, IEEE Trans. Image Process. 22 (2013) 3485–3496.
[19] K. Karsch, C. Liu, S.B. Kang, Depth transfer: Depth extraction from video using non-parametric sampling, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 2144–2158.
[20] S. Choi, D. Min, B. Ham, Y. Kim, C. Oh, K. Sohn, Depth analogy: Data-driven approach for single image depth estimation using gradient samples, IEEE Trans. Image Process. 24 (2015) 5953–5966.
[21] D. Eigen, R. Fergus, Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture, 2014.
[22] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, A. Yuille, Towards unified depth and semantic prediction from a single image, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 2800–2809.
[23] F. Liu, C. Shen, G. Lin, I. Reid, Learning depth from single monocular images using deep convolutional neural fields, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2015) 2024–2039.
[24] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab, Deeper depth prediction with fully convolutional residual networks, in: 3D Vision, 3DV, 2016 Fourth International Conference on, IEEE, 2016, pp. 239–248.
[25] J.H. Lee, M. Heo, K.R. Kim, C.S. Kim, Single-image depth estimation based on fourier domain analysis, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2018.
[26] J.H. Lee, C.S. Kim, Monocular depth estimation using relative depth maps, in: The IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019.
[27] G. Zhang, J. Jia, T.T. Wong, H. Bao, Consistent depth maps recovery from a video sequence, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 974–988.
[28] M. Yang, X. Cao, Q. Dai, Multiview video depth estimation with spatial-temporal consistency, in: BMVC, Citeseer, 2010, pp. 1–11.
[29] W. Yang, G. Zhang, H. Bao, J. Kim, H.Y. Lee, Consistent depth maps recovery from a trinocular video sequence, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1466–1473.
[30] B.C. Russell, A. Torralba, Building a database of 3D scenes from user annotations, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 2711–2718.
[31] B. Liu, S. Gould, D. Koller, Single image depth estimation from predicted semantic labels, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 1253–1260.
[32] J.L. Herrera, C.R. del Blanco, N. García, Automatic depth extraction from 2D images using a cluster-based learning framework, IEEE Trans. Image Process. 27 (2018) 3288–3299.
[33] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: International Conference on Computer Vision & Pattern Recognition, CVPR'05, IEEE Computer Society, 2005, pp. 886–893.
[34] C. Liu, J. Yuen, A. Torralba, Sift flow: Dense correspondence across scenes and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2010) 978–994.
[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.
[36] H. Fu, M. Gong, C. Wang, K. Batmanghelich, D. Tao, Deep ordinal regression network for monocular depth estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2002–2011.
[37] A. Levinshtein, A. Stere, K.N. Kutulakos, D.J. Fleet, S.J. Dickinson, K. Siddiqi, Turbopixels: Fast superpixels using geometric flows, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 2290–2297.
[38] A. Vedaldi, S. Soatto, Quick shift and kernel methods for mode seeking, in: European Conference on Computer Vision, Springer, 2008, pp. 705–718.
[39] O. Veksler, Y. Boykov, P. Mehrani, Superpixels and supervoxels in an energy optimization framework, in: European Conference on Computer Vision, Springer, 2010, pp. 211–224.
[40] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 2274–2282.
[41] Y.J. Liu, C.C. Yu, M.J. Yu, Y. He, Manifold slic: A fast method to compute content-sensitive superpixels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 651–659.
[42] J. Zaragoza, T.J. Chin, M.S. Brown, D. Suter, As-projective-as-possible image stitching with moving dlt, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2339–2346.
[43] J. Sun, J. Zhu, M.F. Tappen, Context-constrained hallucination for image super-resolution, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 231–238.
[44] Q. Wu, W. Zhang, B.V. Kumar, Strong shadow removal via patch-based shadow edge detection, in: 2012 IEEE International Conference on Robotics and Automation, IEEE, 2012, pp. 2177–2182.
[45] A. Saxena, M. Sun, A.Y. Ng, Make3D: Learning 3D scene structure from a single still image, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2008) 824–840.