DISPARITY ADAPTED WEIGHTED AGGREGATION FOR LOCAL STEREO

Julia Navarro and Antoni Buades

Universitat de les Illes Balears, Cra. de Valldemossa km 7.5, 07122 Palma, Spain

ABSTRACT

We present a novel aggregation method based on adaptive support weights for local stereo matching. In order to correctly match each block, the adaptive weight distribution should favor pixels sharing the same disparity. State of the art algorithms make this configuration depend only on spatial and color differences, identifying disparity discontinuities with color ones. Compared to these algorithms, we introduce a weight support depending on each tested disparity and not only on the image configuration around the reference pixel. For each tested disparity, we favor pixels in the block matching with smaller cost, which are supposed to be more likely to be correctly represented by this disparity. Besides, we use a multiscale strategy with invalidation criteria to reduce match ambiguity and computational time. Results on the Middlebury stereo benchmark show the performance improvement of the proposed approach in comparison with state of the art.

Index Terms— Stereo, disparity estimation, adaptive support weights, block-matching

1. INTRODUCTION

The goal of stereovision is to estimate the depth of the scene from at least two images taken from different viewpoints. Depth estimation is equivalent to computing the apparent motion of corresponding points in the two images. For an epipolar rectified image pair, all pixels have horizontal motion (which is called disparity) and the problem is then reduced into a 1D correspondence problem.

As comparing each single pixel in both images would be ineffective, local methods compare patches or windows around each pixel. This block-matching (also called area-based) approach assigns for a window around each pixel $\mathbf{p}$ in the first image a matching cost to each candidate window in the secondary image that lies in the same epipolar line. The matching cost of a window is obtained as an aggregation of the cost of each single pixel within the window. The center pixel of the window in the secondary image yielding minimum cost is the match of the pixel $\mathbf{p}$.

The authors were supported by Ministerio de Economía y Competitividad under grant TIN2014-53772-R and CNES research and technology project DAJ/AR/IB/16-10117037.

A known drawback of this approach is the so-called fattening effect. It occurs when the disparity function is not constant within the window. If this happens, the exact window can not be found in the secondary image. As a consequence, the resulting disparity for that pixel is influenced by the different disparities within the window [1]. Therefore, in order to reliably match a block, the depth should vary as less as possible inside the block.

Adaptive support weights algorithms have shown to be competitive [2]. These approaches assign an individual weight between zero and one to each pixel in the window, where values close to one give a strong importance to the pixel in the aggregation process. The idea is that pixels with higher weights should have the same disparity as the reference pixel. The seminal work from Yoon and Kweon [3] proposes a weight based on color similarity and geometric proximity. That is, pixels sharing the same color of the reference pixel are favored, which is equivalent to identify depth discontinuities with color discontinuities. This solution might be effective for scenes with uniform color objects being at different depth planes, however this is not the case for a general scenario with textured objects and slanted surfaces.

In this work, we present an adaptive support weight approach that is able to properly deal with depth discontinuities and slanted surfaces. This new aggregation scheme makes the weight actually depend on the tested disparity and not only on the image configuration around the reference pixel. Apart from color similarity and geometric distance, the proposed weights include a term based on the cost of matching the pixel with the considered disparity. Since the weighted block should compare only pixels having the same depth, for a certain disparity hypothesis these pixels will be the ones with minimum cost.

Besides, following the idea of [4], our algorithm is embedded into a coarse-to-fine strategy in which the search disparity range is restricted at each scale based on the disparity estimated at the previous scale. This strategy reduces the match ambiguity and computational time. At each scale, match validation criteria is used in order to detect and reject incorrect matches.

This paper is organized as follows. In Section 2 we present the state of the art of local methods for disparity estimation. The new block-matching algorithm with adaptive support weights is introduced in Section 3. Section 3.2

and Section 3.3 describe the match validation criteria and the multiscale procedure, respectively. Finally, Section 4 shows the performance of the proposed method by means of a comparison with state of the art approaches.

2. PREVIOUS WORK

Local methods mainly differ in the choice of the matching cost and in the aggregation step. The most common costs are the sum of squared differences, the normalized cross correlation [5], the mutual information [6] and the census transform [7]. While the choice of a different cost might be important in order to be robust to local differences in color, noise, shadows or transparencies, aggregation turns to be the most important in order to avoid the fattening effect. A classification and evaluation of cost aggregation stereo methods is presented in [8].

There exist two different ways of dealing with aggregation: methods directly adapting the shape of the window and adaptive support weights approaches. The first ones try to select the window that best satisfies the condition of constant disparity inside the block. Kanade et al. [9] were the first to address this problem, they used rectangular windows whose shape and size is selected from differences between gray level values. In [10] it was performed correlation with different windows (fixed squared shape and size) containing the reference pixel and taking the one with smallest cost. Patricio et al. [11] excluded neighboring pixels with different gray level to the center one. Hirschmuller et al. [12] divided the correlation window into sub-windows and selected the ones yielding minimum matching cost. Buades et al. [4] tested multiple elongated windows with different orientations and selected the one providing minimum matching cost.

Regarding the use of adaptive support weights, Yoon and Kweon [3] based the aggregation weights of neighboring points on color similarity and geometric distance to the center pixel of the window. Inspired by this approach, Wang et al. [13] introduced this scheme into a dynamical programming framework. Hosni et al. [14] proposed to use geodesic support weights. Kowalczuk et al. [15] used the adaptive matching cost from [3], combined with an iterative disparity refinement. Rhemann et al. [16] used the guided filter [17] with the reference image as guide to filter the cost volume. See [2] for an extensive review of adaptive support weights approaches.

3. STEREO MATCHING USING ADAPTIVE WEIGHTS

Adaptive support weights approaches compute the disparity d at pixel $\mathbf{p}$ as the one that minimizes the general cost C,

$$C(\mathbf{p},d)=\frac{\sum_{\mathbf{q}\in N_\mathbf{p}} w(\mathbf{p},\mathbf{q})\ c(\mathbf{q},d)}{\sum_{\mathbf{q}\in N_\mathbf{p}} w(\mathbf{p},\mathbf{q})}, \tag{1}$$

where $N_\mathbf{p}$ corresponds to a spatial neighborhood of pixel $\mathbf{p}$ and the cost c represents the penalization of matching the pixel $\mathbf{q}$ with $\mathbf{q}+(d,0)$.

Pixels with higher weights should have a similar disparity to the reference pixel in order to fulfill the condition of constant disparity inside the matching window. Since this function is unknown, most methods identify depth discontinuities with color discontinuities and make this weight depend on color differences. Yoon et al. [3] used color similarity (w_c) and spatial distance (w_s) to compute these weights:

$$w(\mathbf{p},\mathbf{q})=w_c(\mathbf{p},\mathbf{q})\cdot w_s(\mathbf{p},\mathbf{q}), \tag{2}$$

$$w_c(\mathbf{p},\mathbf{q})=e^{-\frac{\|I_1(\mathbf{p})-I_2(\mathbf{q})\|^2}{\lambda^2}}, \tag{3}$$

$$w_s(\mathbf{p},\mathbf{q})=e^{-\frac{\|\mathbf{p}-\mathbf{q}\|^2}{\beta^2}}. \tag{4}$$

Inspired by Yoon et al. [3], adaptive weights approaches make this weight depend only on the spatial distance of $\mathbf{p}$ and $\mathbf{q}$ and its color difference. The cost $c(\mathbf{q},d)$ of matching pixels $\mathbf{q}$ and $\mathbf{q}+(d,0)$ reduces to the color difference of these two pixels, what makes the method sensitive to color changes.

We propose to make the aggregation weights $w(\mathbf{p},\mathbf{q})$ depend on the disparity d, particularly on the cost $c(\mathbf{q},d)$. These weights will be denoted as $w(\mathbf{p},\mathbf{q},d)$. In order to make the cost $C(\mathbf{p},d)$ robust to color changes, we use as pixel cost $c(\mathbf{q},d)$ the ZSSD [18] with a squared window of 3×3 pixels. This cost removes the average intensity of the window rendering the comparison independent of the mean intensity.

3.1. Proposed weights

In addition to the spatial proximity and color difference weights, we introduce a new term for the computation of $w(\mathbf{p},\mathbf{q},d)$ based on the cost of assigning to the neighboring pixel $\mathbf{q}$ a disparity d:

$$w_{cost}(\mathbf{q},d)=e^{-\frac{c(\mathbf{q},d)}{\gamma^2}}. \tag{5}$$

We have chosen the cost $c(\mathbf{q},d)$ as the ZSSD [18]:

$$\frac{1}{|B|}\sum_{t\in B}\|I_1(\mathbf{q}+t)-I_2(\hat{\mathbf{q}}+t)-\overline{I_{1|\mathbf{q}+B}}+\overline{I_{2|\hat{\mathbf{q}}+B}}\|^2,$$

where $\hat{\mathbf{q}}=\mathbf{q}+(d,0)$, B denotes the 3×3 window and $\overline{I_{1|\mathbf{q}+B}}$ is the mean of pixel intensities in image I_1 inside the 3×3 window centered at pixel $\mathbf{q}$. The normalization of the patch by its mean permits to make this cost insensitive to changes in illumination and the aggregation on a 3×3 patch makes it more robust than a single pixel difference.

The color similarity weight has also been replaced by a *mean color similarity*, which is the difference of the mean color of the 3×3 patch centered at each pixel:

$$w_{mc}(\mathbf{p},\mathbf{q})=e^{-\frac{\left\|\overline{I_{1|\mathbf{p}+B}}-\overline{I_{1|\mathbf{q}+B}}\right\|^2}{\lambda^2}}. \tag{6}$$

The use of the mean makes the color term more robust to noise and microtexture.

The proposed weights are then obtained by combining the three terms:

$$w(\mathbf{p},\mathbf{q},d) = e^{-\frac{\|\mathbf{p}-\mathbf{q}\|^2}{\beta^2}} e^{-\frac{\left\|\overline{I_{1|\mathbf{p}+B}}-\overline{I_{1|\mathbf{q}+B}}\right\|^2}{\lambda^2}} e^{-\frac{c(\mathbf{q},d)}{\gamma^2}}. \quad (7)$$

In practice we use $\beta = 11$, $\lambda = 6$, $\gamma = 4$ and a 31×31 squared window centered at $\mathbf{p}$. The sub-pixel precision is achieved by directly computing the distances at non integer positions by shifting the entire secondary image.

Algorithm 1 summarizes the proposed method.

Algorithm 1 Stereo matching with adaptive weights.

Require: Stereo pair I_1, I_2, search range $[d_{min}, d_{max}]$, disparity precision $prec$, window size N_p, and λ, β, γ parameters.

Ensure: Disparity map D.

1: $\hat{C}(\mathbf{p}) = \infty$; $D(\mathbf{p}) = 0$; $\forall \mathbf{p} \in I_1$
2: **for** d **from** d_{min} **to** d_{max} **step** $prec$ **do**
3: Compute distance of all 3×3 patches in I_1 with patches in I_2 at disparity d: $c(\mathbf{q},d) = ZSSD(\mathbf{q}, \mathbf{q}+(d,0))$.
4: **for each** pixel $\mathbf{p}$ in I_1 **do**
5: Compute $C(\mathbf{p},d)$ as

$$C(\mathbf{p},d) = \frac{\sum_{\mathbf{q}\in N_\mathbf{p}} w(\mathbf{p},\mathbf{q},d)\ c(\mathbf{q},d)}{\sum_{\mathbf{q}\in N_\mathbf{p}} w(\mathbf{p},\mathbf{q},d)},$$

being $w(\mathbf{p},\mathbf{q},d)$ the proposed in (7).
6: **end for**
7: **if** $C(\mathbf{p},d) < \hat{C}(\mathbf{p})$ **then**
8: $\hat{C}(\mathbf{p}) = C(\mathbf{p},d)$; $D(\mathbf{p}) = d$.
9: **end if**
10: **end for**

3.2. Validation

A match is rejected when the left based disparity and the inverse mapping of the right based estimation differ from more than one pixel. Known as the left-right consistency check [19, 20], it invalidates a match if,

$$d_L(\mathbf{p}) + d_r(\mathbf{p} + d_L(\mathbf{p})) > 1$$

being d_L and d_R the left and right based disparities, respectively.

Ambiguous matches are detected and rejected by using the technique proposed in [4]. This strategy detects non-distinctive pixels by comparing the cost of the best match in the second image and the cost of the best match in the reference image itself. The key idea is that when the latter cost is smaller than the cost of its best match in the second image, probably the patch is part of a repetitive pattern.

Finally, we reject isolated validated regions with an area less of A pixels, being A set in practice to 10 pixels.

3.3. Multiscale strategy

As in [4], our approach is embedded into a coarse-to-fine strategy. This permits to reduce computational cost and match ambiguity. A large search range may lead to mismatches as it increases the possibility of matching with a repetitive pattern.

First of all, we make use of validation criteria to determine the reliability of each computed match. Then, at each scale we locally adapt the search range at each pixel by looking at the minimum and maximum valid disparity values of a neighborhood obtained at the previous coarser scale. Invalid points are assigned the full disparity range.

Each level of the pyramid is obtained by a convolution with a Gaussian kernel with standard deviation $\sigma = 1.2$ and sub-sampling by a factor of two from the initial stereo pair. Disparity computed at coarser scales is up-sampled by bicubic spline interpolation.

4. RESULTS

In this section we illustrate the performance of the proposed method on the Middlebury stereo dataset [21]. A quantitative comparison with different state of the art approaches is reported in Table 1. Values for SGM [22], SNCC [23], IDR [15], Cens5 [12] and TMAP [24] have been obtained from the Middlebury evaluation table [1] (January 2017), while values for CVF [16], ASW [3] and MSMW [4] have been computed using the authors' code with default parameters. The detection of invalid pixels in ASW has been carried out by a left-right consistency check.

The parameters in our method were set for all experiments to $\beta = 11$, $\lambda = 6$, $\gamma = 4$ and a 31×31 squared window centered at $\mathbf{p}$. The sub-pixel precision is set to 1/4 of pixel and the multi-scale procedure uses three scales.

Table 1 displays the error on the Middlebury stereo training dataset [21]. Our algorithm provides competitive results compared to state of the art, achieving the second best average error. The best average error is obtained by IDR which uses the approach ASW by Yoon et al. [3] with a posterior iterative disparity refinement. The average error of IDR is significantly better than ASW while ours is significantly better than ASW. This suggests that more elaborated strategies with the proposed cost could outperform state of the art IDR.

Figure 1 illustrates the visual results for several stereo pairs from the Middlebury training dataset. Despite of the illumination changes present in the MotorcycleE and PianoL pairs, our method is capable to give an accurate result.

Other methods as TMAP, MSMW and ASW fail in many areas and end up having many invalid pixels, in both cases ASW also produces large errors. On the other hand, MSMW and the proposed method are the ones giving better results in the floor of the Playtable pair, while TMAP, ASW and IDR obtain large black areas. By comparing the results between

[1] http://vision.middlebury.edu/stereo/eval3/

Table 1. Evaluation on Middlebury online benchmark [21]. Comparison between our method and SGM [22], SNCC [23], IDR [15], Cens5 [12], TMAP [24], CVF [16], ASW [3] and MSMW [4]. The numbers represent the weighted average for the training dataset over non occluded pixels.

method	resolution	density	bad 0.5	bad 1.0	bad 2.0	bad 4.0	avgErr
SGM	H	84.04	38.5	16.4	7.52	3.92	2.09
SNCC	H	71.29	29.8	12.4	6.03	3.54	2.44
IDR	H	76.10	31.9	11.70	4.58	2.20	1.36
Cens5	H	76.38	35.7	16.6	8.27	4.96	3.14
TMAP	H	69.80	37.3	14.3	5.88	3.50	2.40
CVF	H	70.82	44.52	21.56	9.40	5.90	4.33
ASW	H	73.60	48.43	27.44	14.99	11.19	10.83
MSMW	H	66.92	33.13	16.20	8.28	4.83	3.03
ours	H	71.43	34.35	15.18	6.60	3.38	1.93

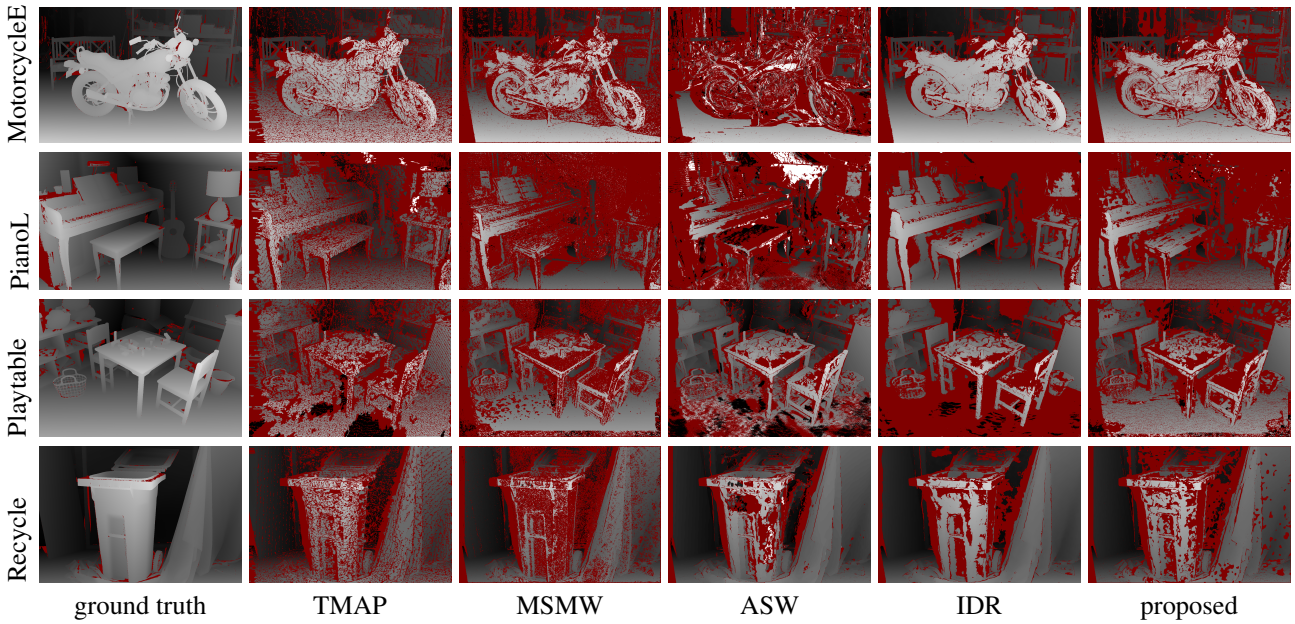

Fig. 1. Visual results for several stereo pairs of the Middlebury training dataset. Comparison between our method and TMAP [24], MSMW [4], ASW [3] and IDR [15].

ASW and IDR, we can again appreciate the significant improvement when applying the refinement proposed in IDR.

5. CONCLUSIONS

This work proposes a novel aggregation weight function for local stereo matching. Apart from the color similarity and spatial distance terms presented by Yoon et al. [3], we include a new term based on the cost of matching the neighboring pixel with the considered disparity. We assume that pixels at the same disparity as the reference pixel will be the ones minimizing the matching cost.

We have shown experimental results on the Middlebury stereo dataset [21], achieving the second best average error. The performance of the proposed cost suggests that more complex procedures with iterative refinement would improve state of the art.

6. REFERENCES

[1] Gwendoline Blanchet, Antoni Buades, Bartomeu Coll, Jean-Michel Morel, and Bernard Rougé, "Fattening free block matching," *Journal of mathematical imaging and vision*, vol. 41, no. 1-2, pp. 109–121, 2011.

[2] Asmaa Hosni, Michael Bleyer, and Margrit Gelautz, "Secrets of adaptive support weight techniques for local stereo matching," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 620–632, 2013.

[3] Kuk-Jin Yoon and In So Kweon, "Adaptive support-weight approach for correspondence search," *IEEE*

Transactions on Pattern Analysis & Machine Intelligence, , no. 4, pp. 650–656, 2006.

[4] Antoni Buades and Gabriele Facciolo, "Reliable multiscale and multiwindow stereo matching," *SIAM Journal on Imaging Sciences*, vol. 8, no. 2, pp. 888–915, 2015.

[5] Marsha J Hannah, "Computer matching of areas in stereo images," Tech. Rep., DTIC Document, 1974.

[6] Paul Viola and William M Wells III, "Alignment by maximization of mutual information," *International journal of computer vision*, vol. 24, no. 2, pp. 137–154, 1997.

[7] Ramin Zabih and John Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Computer VisionECCV'94*, pp. 151–158. Springer, 1994.

[8] Federico Tombari, Stefano Mattoccia, Luigi Di Stefano, and Elisa Addimanda, "Classification and evaluation of cost aggregation methods for stereo correspondence," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[9] Takeo Kanade and Masatoshi Okutomi, "A stereo matching algorithm with an adaptive window: Theory and experiment," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, no. 9, pp. 920–932, 1994.

[10] Andrea Fusiello, Vito Roberto, and Emanuele Trucco, "Efficient stereo with multiple windowing," in *cvpr*. IEEE, 1997, p. 858.

[11] Madan Pørez Patricio, Franois Cabestaing, Olivier Colot, and Pierre Bonnet, "A similarity-based adaptive neighborhood method for correlation-based stereo matching," in *Image Processing, 2004. ICIP'04. 2004 International Conference on*. IEEE, 2004, vol. 2, pp. 1341–1344.

[12] Heiko Hirschmüller, Peter R Innocent, and Jon Garibaldi, "Real-time correlation-based stereo vision with reduced border errors," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 229–246, 2002.

[13] Liang Wang, Miao Liao, Minglun Gong, Ruigang Yang, and David Nister, "High-quality real-time stereo using adaptive cost aggregation and dynamic programming," in *3D Data Processing, Visualization, and Transmission, Third International Symposium on*. IEEE, 2006, pp. 798–805.

[14] Asmaa Hosni, Michael Bleyer, Margrit Gelautz, and Christoph Rhemann, "Local stereo matching using geodesic support weights," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 2093–2096.

[15] Jkedrzej Kowalczuk, Eric T Psota, and Lance C Perez, "Real-time stereo matching on cuda using an iterative refinement method for adaptive support-weight correspondences," *IEEE transactions on circuits and systems for video technology*, vol. 23, no. 1, pp. 94–104, 2013.

[16] Christoph Rhemann, Asmaa Hosni, Michael Bleyer, Carsten Rother, and Margrit Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 3017–3024.

[17] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," in *Computer Vision–ECCV 2010*, pp. 1–14. Springer, 2010.

[18] Heiko Hirschmüller and Daniel Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 9, pp. 1582–1599, 2009.

[19] Pascal Fua, "A parallel stereo algorithm that produces dense depth maps and preserves image features," *Machine vision and applications*, vol. 6, no. 1, pp. 35–49, 1993.

[20] Steven D Cochran and Gérard Medioni, "3-d surface description from binocular stereo," *IEEE transactions on pattern analysis and machine intelligence*, vol. 14, no. 10, pp. 981–994, 1992.

[21] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on Pattern Recognition*. Springer, 2014, pp. 31–42.

[22] Heiko Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 2, pp. 328–341, 2008.

[23] Nils Einecke and Julian Eggert, "A two-stage correlation method for stereoscopic depth estimation," in *Digital Image Computing: Techniques and Applications (DICTA), 2010 International Conference on*. IEEE, 2010, pp. 227–234.

[24] Eric T Psota, Jedrzej Kowalczuk, Mateusz Mittek, and Lance C Perez, "Map disparity estimation using hidden markov trees," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2219–2227.