STEREO IMAGE DE-FENCING USING SMARTPHONES

Sankaraganesh Jonna*, Sukla Satapathy*, and Rajiv R. Sahay+

Computational Vision Laboratory,
*Department of Computer Science and Engineering, +Department of Electrical Engineering,
*,+Indian Institute of Technology Kharagpur

ABSTRACT

Conventional approaches to image de-fencing have limited themselves to using only image data in adjacent frames of the captured video of an approximately static scene. In this work, we present a method to harness disparity using a stereo pair of fenced images in order to detect fence pixels. Tourists and amateur photographers commonly carry smartphones/phablets which can be used to capture a short video sequence of the fenced scene. We model the formation of the occluded frames in the captured video. Furthermore, we propose an optimization framework to estimate the de-fenced image using the total variation prior to regularize the ill-posed problem.

Index Terms— Image de-fencing, image inpainting, stereo, disparity map, convolutional neural networks, total variation

1. INTRODUCTION

Restoration of images containing occluded objects is a challenging problem and has recently attracted significant attention of researchers. This problem is faced by photographers when they image a scene obstructed by fences e.g. in zoos, museums etc. We observe that image de-fencing is basically an image inpainting [1, 2, 3] problem wherein the fence pixels comprise the region which has to be filled-in. Since fence pixels cover the entire image as shown in Fig. 1 (a), (b) manual segmentation is too cumbersome. In several fence removal methods [4, 5, 6, 7, 8, 9] a video is captured by panning a camera relative to the scene. Subsequently, these methods use image data in the frames of the video to segment the fences. However, in this work we exploit the disparity of pixels in two adjacent frames to localize fence pixels. Note that generally fences are closer to the panning camera and therefore undergo greater motion parallax due to relative motion.

The main challenge is robust identification of the fences/occlusion in bad illumination, heavy clutter in the background, noise etc. In this work, we propose a depth-based technique to estimate the spatial location of fence pixels. Depth maps can be obtained using laser scanners or the inexpensive Kinect sensor. However, laser scanners are slow and expensive and
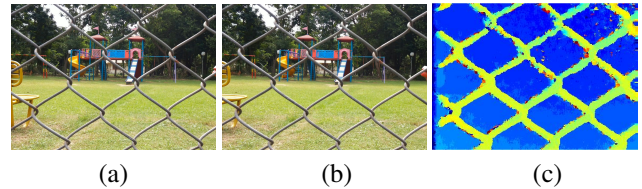

Fig. 1. (a), (b) Left and right images captured using a mobile phone. (c) Stereo disparity map obtained by proposed algorithm.

Kinect is limited to capturing depth maps of indoor scenes, respectively. These days smartphones have become ubiquitous and come equipped with sophisticated cameras. Hence, the depth map for estimating fence pixel locations can be obtained using stereo.

In this work, we use only two frames from the captured video using a smartphone and consider them as a stereo-pair to compute the disparity map and subsequently for information fusion. We assume that the objects occluded in left frame are visible in the right image. Zbontar et al. [10] proposed a supervised learning approach to estimate disparity maps by training a convolutional neural network to compare image patches. We employed pre-trained "fast-CNN" from [10] for estimating a raw disparity map in our work which we refine using morphological filtering and image matting [11]. In Figs. 1 (a), (b), we show left and right occluded frames captured using a mobile camera. The stereo disparity map computed using the proposed algorithm is shown in Fig. 1 (c). In the second step of our algorithm we use the estimated fence pixels to obtain optical flow. Finally, the de-fenced image is reconstructed with total variation (TV) [12] as the regularizer using the split Bregman [13] optimization framework.

2. RELATED WORK

Several works have appeared in the literature [4, 5, 6, 7, 8] which use only image data for fence detection. A computational model for periodic pattern perception in [14] is based on the theory of frieze and wallpaper groups. Another effective fence mask detection algorithm based on detection of

regular textures is proposed by Park et al. in [15]. The work in [6] addressed the video de-fencing problem using visual parallax as a cue for soft fence detection distinguishing fences from the background pixels. The occluding foreground along with the restored background are obtained using visual parallax in [7]. Yi et al. [8] proposed a bottom-up framework for fence detection by clustering pixels into coherent groups using color and motion features through graph-cut based optimization. In our previous works [9, 16] we have proposed a supervised learning approach for automatic identification of occlusions/fences using only image data.

In this work, we propose a depth-based technique to detect the fence mask from the disparity map obtained from a stereo image pair. The works of Wang et al. [17] and Jonna et al. [18] are most closely related to our method. The algorithm in [17] leveraged advantages of stereo image pair for joint completion of missing texture and depth. The limitation of [17] is that the region to be inpainted has to be specified manually. The technique of [18] proposed a multimodal approach for image de-fencing where the fence masks are extracted automatically from depth maps corresponding to the color images obtained using Kinect sensor. One limitation of the Kinect sensor is that it provides the depth map of a scene in a constrained laboratory environment and only works accurately if the scene is within a particular distance from sensor (viewing distance range is 1.2m to 3.5m). However, most fenced images are captured in outdoor settings such as zoos, play grounds, museums etc. The applications of real-world image de-fencing are generally in places, where one can not use the Kinect sensor. Hence, the depth map for estimating fence pixel locations has to be obtained using stereo.

Recently, Zbontar et al. [10] proposed a supervised learning algorithm for estimating disparity map by training convolutional neural network (CNN) architectures [19, 20]. The raw disparity maps obtained using CNNs in [10] are refined by a series of post-processing steps: cost aggregation, semiglobal matching, left-right image consistency check and filtering. Unlike, the algorithms in [6, 7, 8], we use only two frames for both fence segmentation and information fusion. As we are using smartphones for capturing the left and right stereo images we have chosen a learning-based method [10] for disparity estimation wherein we don't require any a priori camera calibration.

3. METHODOLOGY

3.1. Problem formulation

The degradation model for the frames of the captured video is

$$\mathbf{y}_m^{obs} = \mathbf{O}_m \mathbf{y}_m = \mathbf{O}_m \mathbf{W}_m \mathbf{x} + \mathbf{n}_m \tag{1}$$

where $\mathbf{y}_m^{obs}$ is the mth observation wherein the occluded pixels have been excluded using $\mathbf{O}_m$, $\mathbf{y}_m$, $m = 1, 2$ are the left and right images comprising the stereo pair, $\mathbf{O}_m$ is the fence mask

corresponding to m^{th} image, $\mathbf{W}_m$ is the warp matrix, $\mathbf{x}$ is the de-fenced image and $\mathbf{n}_m$ is the Gaussian noise.

3.2. Detection of fence using stereo

In the stereo problem, given a pair of images $\mathbf{y}_m$ $(m = 1, 2)$ we want to compute the disparity map $\mathbf{D}$. Since fences are closer to the camera, we have exploited disparity/depth cue for the fence pattern segmentation. Following the taxonomy of Scharstein et al. [21], a stereo algorithm consists of four steps: matching cost computation, cost aggregation, optmization and disparity refinement. The authors in [10] proposed two supervised learning based convolutional neural network architectures for matching cost computation followed by a set of post processing operations. Note that they trained two CNNs 'fast' and 'accurate' on pairs of small images patches wherein the true disparity values are known. The advantage of the method in [10] is that it does not require a priori camera calibration. Therefore, we employed the 'fast' pre-trained CNN model in [10] which is trained on 17 million example images extracted from training dataset of KITTI 2015 [22] to generate the raw disparity map.

Fig. 2. Architecture of the pre-trained 'fast' CNN [10].

The architecture of the 'fast' CNN trained for similarity measure on small image patches consists of two sub-networks shown in Fig. 2. Each sub-network is made up of convolutional layers followed by ReLU layers except the last convolutional layer. Extracted features from each one of the sub-network are used to compute cosine similarity. The negative value of the cosine similarity measure is treated as the matching cost. To obtain the final fence mask, we use a matting technique [11].
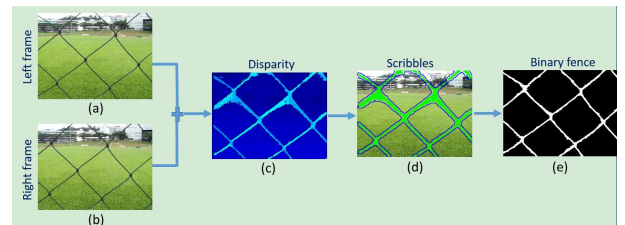
Fig. 3. Flowchart of the proposed fence segmentation algorithm.

The overall workflow of the proposed fence segmentation algorithm is shown in Fig. 3. In Figs. 3 (a), (b), we show the

left and right stereo images taken from a video clip, respectively. Estimated raw disparity map using pre-trained CNN [10] of Fig. 2 is shown in Fig. 3 (c). The output in Fig. 3 (c) is dilated and then canny edge operator is applied. In Fig. 3 (d) we show the response obtained by Canny edge detection algorithm with blue color which are treated as background scribbles. Similarly, foreground scribbles (green color) are obtained by erosion operation on the image in Fig. 3 (c). The combination of both foreground and background scribbles is shown in Fig. 3 (d). We fed these automatically generated scribbles to the method of [11] and obtain the refined binary fence mask shown in Fig. 3 (e) which is generated by thresholding the alpha map obtained from [11].

3.3. Estimation of optical flow

The two stereo images in our work have been extracted from a video captured using a smartphone with approximately horizontal motion. To fill-in the occluded information in the reference left frame using visible data from the right view, we need pixel correspondence between them. Optical flow can be estimated accurately between visible areas of two images, however, we observed some erroneous values near the occlusion boundaries. To avoid these errors, we blurred the fences in the stereo images with a Gaussian kernel ($\sigma = 2$) before applying the method in [23] to estimate optical flow. Note that now we already know the spatial locations of the fence pixels which are estimated in the previous section.

3.4. Optimization using TV

As we model image de-fencing as an inverse ill-posed problem, total variation of the de-fenced image can be posed as a regularization constraint. Since TV prior is non-smooth split Bregman technique [13] is used to solve the resulting minimization problem to reconstruct the fence-free image. The de-fenced image is the solution of the following constrained optimization problem

$$\arg\min_{\mathbf{x}} \frac{1}{2} \sum_{m=1}^{p} \parallel \mathbf{y}_m - \mathbf{O}_m \mathbf{W}_m \mathbf{x} \parallel_2^2 + \mu \parallel \mathbf{d} \parallel_1 \; s.t. \; \mathbf{d} = \nabla \mathbf{x}$$

where p is the number of frames chosen from the video and μ is the regularization parameter. The above optimization framework is a combination of both $l1$ and $l2$ terms and hence difficult to solve. We employ the split Bregman iterative framework described in [13] to solve the above problem. We use an alternative unconstrained formulation as

$$\arg\min_{\mathbf{x}} \frac{1}{2} \sum_{m=1}^{p} \parallel \mathbf{y}_m - \mathbf{O}_m \mathbf{W}_m \mathbf{x} \parallel_2^2 + \mu \parallel \mathbf{d} \parallel_1$$
$$+ \frac{\lambda}{2} \parallel \mathbf{d} - \nabla \mathbf{x} \parallel_2^2 \quad (2)$$
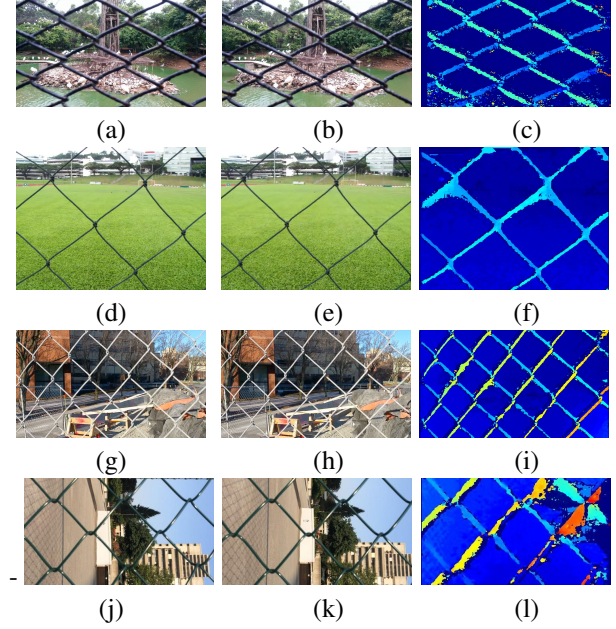

Fig. 4. Row 1: First and second columns show the left and right stereo frames captured by us using a smartphone. Rows 2 to 4: First and second columns depict stereo pairs from video clips of [6, 7, 8], respectively. Third column: estimated disparity maps using the proposed algorithm.

where λ is the shrinkage parameter. The iterates to solve the above equation are as

$$[\mathbf{x}^{k+1}, \mathbf{d}^{k+1}] = \arg\min_{\mathbf{x},\mathbf{d}} \frac{1}{2} \sum_{m=1}^{p} \parallel \mathbf{y}_m - \mathbf{O}_m \mathbf{W}_m \mathbf{x}^k \parallel_2^2$$
$$\quad (3)$$
$$+ \mu \parallel \mathbf{d}^k \parallel_1 + \frac{\lambda}{2} \parallel \mathbf{d}^k - \nabla \mathbf{x}^k + \mathbf{b}^k \parallel_2^2$$

$$\mathbf{b}^{k+1} = \nabla \mathbf{x}^{k+1} + \mathbf{b}^k - \mathbf{d}^{k+1}$$

We can now split the above problem into two sub-problems as

Sub Problem 1:

$$[\mathbf{x}^{k+1}] = \arg\min_{\mathbf{x}} \frac{1}{2} \sum_{m=1}^{p} \parallel \mathbf{y}_m - \mathbf{O}_m \mathbf{W}_m \mathbf{x}^k \parallel_2^2$$
$$+ \frac{\lambda}{2} \parallel \mathbf{d}^k - \nabla \mathbf{x}^k + \mathbf{b}^k \parallel_2^2$$

This sub-problem is solved by a steepest descent method.

Sub Problem 2:

$$[\mathbf{d}^{k+1}] = \arg\min_{\mathbf{d}} \mu \parallel \mathbf{d}^k \parallel_1 + \frac{\lambda}{2} \parallel \mathbf{d}^k - \nabla \mathbf{x}^{k+1} + \mathbf{b}^k \parallel_2^2$$

The above sub-problem can be solved by applying the shrinkage operator as follows

$$\mathbf{d}^{k+1} = shrink(\nabla \mathbf{x}^{k+1} + \mathbf{b}^k, \frac{\lambda}{\mu})$$

$$\mathbf{d}^{k+1} = \frac{\nabla \mathbf{x}^{k+1} + \mathbf{b}^k}{\mid \nabla \mathbf{x}^{k+1} + \mathbf{b}^k \mid} * max(\mid \nabla \mathbf{x}^{k+1} + \mathbf{b}^k \mid -\frac{\lambda}{\mu}, 0)$$

The variable $\mathbf{b}$ can be updated as $\mathbf{b}^{k+1} = \nabla \mathbf{x}^{k+1} + \mathbf{b}^k - \mathbf{d}^{k+1}$. To obtain the best estimate of the de-fenced image we tuned the parameters μ, λ.

4. EXPERIMENTAL RESULTS

To validate the robustness of the proposed method we have reported both fence detection and image de-fencing results using videos captured by us with a smartphone as well as video sequences reported in [6, 7, 8]. We have generated all the results using a 3.4 GHz Intel Core i7 processor with 16 GB of RAM. The execution time of our non-optimized MATLAB implementation is of the order of few tens of seconds.

Initially, we report the fence segmentation results. The images shown in Figs. 4 (a) and (b) are taken from a short video clip captured by us using a smartphone camera. In Fig. 4 (c), we show the raw disparity map obtained using the pretrained CNN [10] of Fig. 2 which is post-processed by morphological operations and learning-based matting method of [11] as outlined in section 3.2. Next, we conduct an experiment on a dataset taken from [6]. In Figs. 4 (d) and (e), we show the stereo pair and the computed disparity map is depicted in Fig. 4 (f). We conducted one more experiment using a video reported in [7]. In Figs. 4 (g) and (h) we show two frames taken from their video sequence and the corresponding disparity map estimated using our algorithm is shown in Fig. 4 (i). Finally, we used two frames from a video sequence reported in [8] which are shown in Figs. 4 (j) and (k). In Fig. 4 (l) we show the obtained disparity map using our approach.

Now we show the performance of the proposed algorithm for removing fences. In Fig. 5 (a) and (b) we show two fenced images taken from a video sequence captured by us using a smartphone. The de-fenced image is shown in Fig. 5 (c). Note that there are no residual traces of fences in the reconstructed output wherein all details have been preserved. We now provide comparisons with state-of-the-art image/video de-fencing algorithms using video sequences reported from their works [6, 7, 8]. In Fig. 5 (d), we show the reference frame taken from a video sequence in [6]. The de-fenced images obtained using the method of [6] and the proposed algorithm are shown in Figs. 5 (e) and (f), respectively. Note that de-fenced image obtained using [6] is blurred whereas the proposed algorithm preserves sharp details. This is evident in the close-up insets of Figs. 5 (e) and (f). Next, a frame taken from a video sequence reported in [7] is shown in Fig. 5 (g). In Figs. 5 (h) and (i), we show the de-fenced results obtained by [7] and the proposed algorithm, respectively. Interestingly, using only 2 frames the proposed algorithm achieves comparable results to [7] which in contrast used 5 frames. Finally, we perform an experiment on a recent video sequence reported in [8]. A frame taken from the video is shown in Fig. 5 (j). The inpained image obtained using [8] is shown
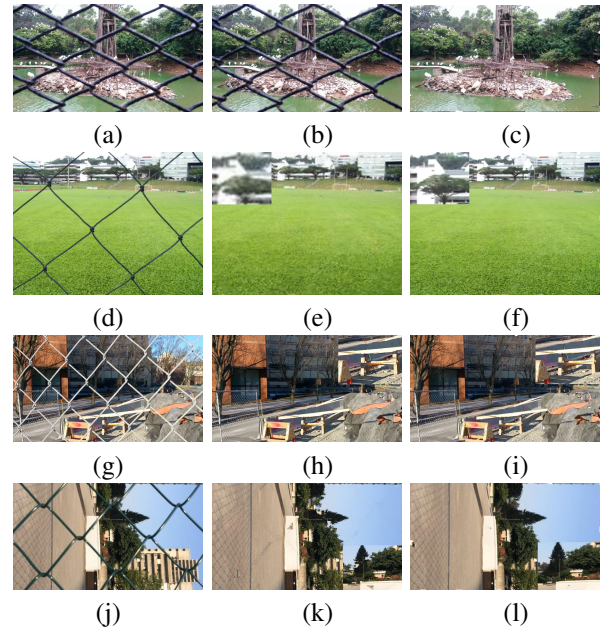

Fig. 5. (a), (b) Two fenced observations from a video captured by us using a smartphone. (c) De-fenced image obtained by the proposed technique. Rows 2 to 4: Comparisons with the existing image/video de-fencing algorithms. **First column**: Sample reference images from video sequences [6, 7, 8], respectively. **Second column**: De-fenced images obtained using state-of-the-art image de-fencing methods [6, 7, 8], respectively. **Third column**: De-fenced results obtained using the proposed algorithm.

in Fig. 5 (k). The corresponding de-fenced image obtained using the proposed algorithm is shown in Fig. 5 (l). Observe that the result in Fig. 5 (k) contains some artifacts whereas the obtained de-fenced image in Fig. 5 (l) is reconstructed well. The same can be observed in insets of both the Figs. 5 (k) and (l), respectively. Note that the algorithm in [8] employed inpainting technique of [2] for image de-fencing.

5. CONCLUSIONS

In this work, we present a novel algorithm for fence segmentation and removal using a stereo-pair of fenced images captured with a smartphone. Initially, we harnessed the disparity cue for robust fence segmentation. We computed the raw disparity map corresponding to a pair of images using pretrained CNNs [10] followed by a series of post-processing steps for obtaining an accurate fence mask. Subsequently, an optimization framework is formulated and the ill-posed inverse problem of image de-fencing is solved using split Bregman iterative procedure assuming total variation of the defenced image as the regularization constraint. We compare the performance of the proposed algorithm with several state-of-the-art works for image de-fencing.

6. REFERENCES

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. ACM SIGGRAPH*, 2000, pp. 417–424.

[2] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, 2004.

[3] C. Guillemot and O. Le Meur, "Image inpainting : Overview and recent advances," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 127–144, 2014.

[4] Y. Liuy, T. Belkina, J. H. Hays, and R. Lublinerman, "Image de-fencing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[5] V. S. Khasare, R. R. Sahay, and M. S. Kankanhalli, "Seeing through the fence: Image de-fencing using a video sequence," in *Proc. IEEE Int. Conf. Image Process.*, 2013, pp. 1351–1355.

[6] Y. Mu, W. Liu, and S. Yan, "Video de-fencing," *IEEE Trans. Circts. Sys. Vid. Tech.*, vol. 24, no. 7, pp. 1111–1121, 2014.

[7] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman, "A computational approach for obstruction-free photography," *ACM Trans. Graph.*, vol. 34, no. 4, p. 79, 2015.

[8] R. Yi, J. Wang, and P. Tan, "Automatic fence segmentation in videos of dynamic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 705–713.

[9] S. Jonna, K. K. Nakka, V. S. Khasare, R. R. Sahay, and M. S. Kankanhalli, "Detection and removal of fence occlusions in an image using a video of the static/dynamic scene," *J. Opt. Soc. Am. A*, vol. 33, no. 10, pp. 1917–1930, 2016.

[10] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, pp. 1–32, 2016.

[11] Y. Zheng and C. Kambhamettu, "Learning based digital matting," in *Proc. Int. Conf. Comput. Vis.*, Sept 2009, pp. 889–896.

[12] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, no. 1, pp. 259–268, 1992.

[13] T. Goldstein and S. Osher, "The split bregman method for *l*1-regularized problems," *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 323–343, 2009.

[14] Y. Liu, R. T. Collins, and Y. Tsin, "A computational model for periodic pattern perception based on frieze and wallpaper groups," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 354–371, 2004.

[15] M. Park, K. Brocklehurst, R. T. Collins, and Y. Liu, "Deformed lattice detection in real-world images using mean-shift belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1804–1816, 2009.

[16] S. Jonna, K. K. Nakka, and R. R. Sahay, "My camera can see through fences: A deep learning approach for image de-fencing," in *Asian Conference on Pattern Recognition*, 2015, pp. 261–265.

[17] L. Wang, H. Jin, R. Yang, and M. Gong, "Stereoscopic inpainting: Joint color and depth completion from stereo images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[18] S. Jonna, V. S. Voleti, R. R. Sahay, and M. S. Kankanhalli, "A multimodal approach for image de-fencing and depth inpainting," in *Proc. Int. Conf. Advances in Pattern Recognition*, 2015, pp. 1–6.

[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[21] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1, pp. 7–42, 2002.

[22] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.

[23] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, 2011.