MULTISPECTRAL HUMAN CO-SEGMENTATION VIA JOINT CONVOLUTIONAL NEURAL NETWORKS

Sungil Choi *Seungryong Kim* *Kihong Park* *Kwanghoon Sohn*

School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
E-mail: khsohn@yonsei.ac.kr

ABSTRACT

We present a novel human body co-segmentation method for unregistered multispectral, color and thermal, images by leveraging CNNs. The main challenges for that tasks are no-alignment between color and thermal images and an absent of ground truth human segmentation labels. To solve these limitations, our key-insight is to formulate the segmentation network for each modality that solve two sub-tasks, correspondence and classification, in a joint and iterative manner. We formulate the learning framework between multispectral images in a way that training labels for one modality are used to learn the network for the other modality. We estimate dense correspondences between multispectral image pairs using intermediate convolutional activations of CNNs and perform human segmentation for each modality through the conditional random fields (CRF) optimization using unary and pairwise fusion. These two steps are formulated as an iterative framework, enables the network to converge on an optimal solution. Experimental results show that our proposed method outperforms conventional state-of-the-art methods on the VAP benchmark consisting of unregistered multispectral color and thermal images.

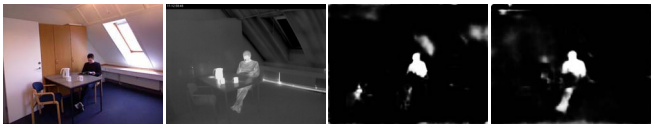
Index Terms— multispectral imaging, weakly supervised learning, human segmentation, visual surveillance, convolutional neural networks

1. INTRODUCTION

Human body segmentation has been one of the most important and fundamental steps for numerous computer vision applications, such as visual surveillance system and action classification [1, 2, 3]. In addition, precise human segmentation benefits the development of white balancing system [4] and object-cutout and paste [5].

Conventionally, a number of approaches have been proposed to segment consistent humans in color images or videos [1, 2]. However, using only color images might induce inherent difficulties in segmenting human body under poor illumination conditions or environments, where the intensities of human body and background are hard to be distinghished. As exemplified in Fig. 1, in a color image, it is hard to distinguish legs of human from the background.

Recently, due to the reduction of the price and their ability to provide visibility in night conditions, thermal sensors have been popularly used for many computer vision tasks as complementary information to supplement the data provided in color images [3, 6, 7]. Especially, the usage of a thermal image combined with a color image could benefit pedestrian detection [6] as well as human segmentation [3] since a thermal image represents the radiated heat of humans in the scene, it enables us to deal with the ambiguity in intensities of human and background. However, combining these multispectral images, such color and thermal image, is not an easy task since they are frequently obtained under different viewpoints and have different attributes.



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

Fig. 1. Importance of joint usage of multispectral, color and thermal, images for human body segmentation: (a) a color image, (b) a thermal image, and estimated human probability maps from (c) the color image, (d) the thermal image. Since multispectral images encode different attributes of a scene, the joint usages of multispectral images can boost the performance of human body segmentation.

Over the past few years, convolutional neural networks (CNNs) have shown revolutionary improvements in performance on various computer vision tasks including segmentation. By leveraging CNNs, human body segmentation performances can be also boosted even for multispectral image cases. However, the key-bottleneck in training CNNs for segmentation is that they require pixel-level haman annotation labels. Collecting accurate pixel-level annotation in large quantities is very labor intensive and somewhat subjective. To overcome these limitations, many approaches have attempted to train segmentation network without pixel-level human annotations rather using weak annotations, such as bounding boxes and image-level labels, which are much easier to collect than pixel-level annotations [8].

In this paper, we propose a novel human body co-segmentation method for unregistered color and thermal images through CNNs. To realize this, our key-insight is to formulate the segmentation network that enables solving two tasks, correspondence and classification, in a joint and iterative manner. To estimate dense correspondences for unregistered color and thermal image, we extract a feature from an intermediate convolutional activation within the segmentation network for each modality. Furthermore, to boost the human segmentation performance through disparate modalities, color and thermal images, we propose two-way fusion techniques, unary fusion and pairwise fusion, in a conditional random field (CRF) model. In an iteration during training networks, intermediate human segmentations from one modality are utilized to learn the network for another modality in a joint manner. By iterating the process of training, human co-segmentation performances for each modality get accurate. In experiments, our proposed method is evaluated on VAP dataset [3] which include unregistrated color and thermal video sequences in an indoor scene, which demonstrates the outstanding performance of our proposed method compared to conventional human segmentation methods.

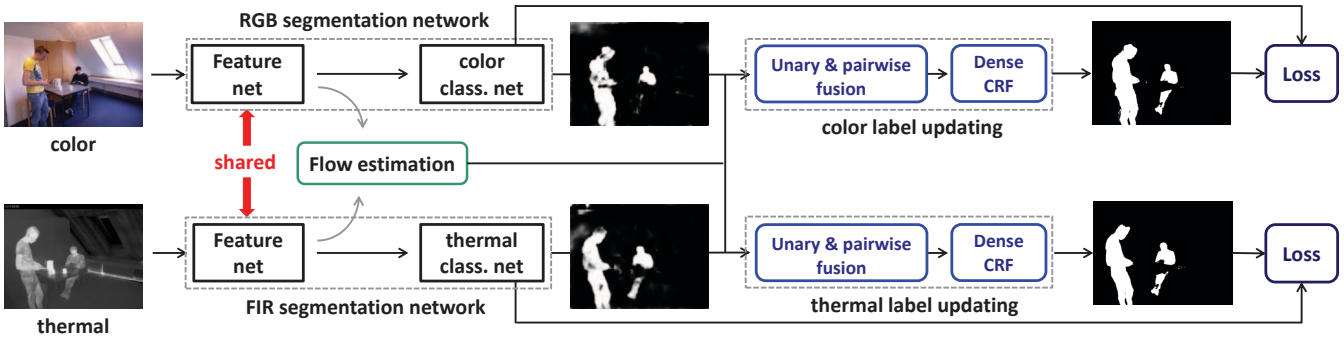


Fig. 2. Overview of our framework, consists of three key-pipelines, color and thermal human segmentation networks, flow estimation, CRF opitimization with the fusion of color and thermal information.

2. PROPOSED METHOD

2.1. Problem Formulation and Overview

Given a color image I_c and a thermal image I_t, our aim is to predict the human segmentation label $S(i) \in \{h, b\}$ for pixel $i = [x_i, y_i]$, where h and b denote a human and background label respectively. Since these multispectral image pairs, i.e., color and thermal images, supplement complementary information to each other, the human segmentation performance through the joint usage of these images can be boosted in comparison to that using an only color image. Recently, most existing methods that exploit CNNs [9, 10, 11] have made a great improvement for the human segmentation task with a pixel-level human annotation in a supervised manner. However, since these multispectral images are frequently taken under different viewpoints in general settings, e.g., surveilance systems, the direct usage of these unregistered multispectral images cannot contribute the performance boosting. Furthermore, a major challenge of human segmentation in multispectral images with CNNs is the lack of ground truth human annotation for training data, where manual annotation is very labor intensive and somewhat subjective.

To alleviate these limitations, we propose a novel human segmentation framework by leveraging CNNs to predict dense human segmentation labels on unregistrated multispectral color and thermal images without ground-truth human annotations. Unlike existing weakly-supervised segmentation methods [8], we formulate the segmentation network to be learned in a way that the resulting segmentation label of one modality can be used to learn other modality segmentation network. To fuse the segmentation labels between multispectral images, we propose two-way fusion techniques, unary and pairwise fusion. Specificially, the score maps from each segmentation network are fused through the max-fusion scheme, and multispectral images are fused to construct the pairwise term in the graphical model. Moreover, to deal with registration problems, we formulate the shared feature extraction layer within the segmentation network to estimate dense correspondences between multispectral images. To boost the performance and convergence, these two tasks, segmentation and correspondence, are formulated synergistically in a joint and iterative manner.

2.2. Network Architecture

Our key-ingradient is to formulate the correspondence and segmentation tasks for unregistered multispectral, color and thermal, images in a joint and iterative manner through CNNs. The overall network of our method consists of two sub-networks, color and thermal human segmentation networks. Each network consists of feature extraction and classification layers, where the feature extraction layers are formulated to be shared across multispectral images to extract consistent features and classification layers are formulated to predict the human segmentation for each modality. As segmentation networks, we adopt the model of fully convolutional networks (FCNs) [9], a variant of VGG-net [12] for classification purpose. The final convolution layer is modified to have 2 classes, i.e., human and background. For correspondence estimation, intermediate convolutional features of last convolutional activations before the pool-3 layer are built through feed-forward processes $\mathcal{F}(I; \mathbf{W}_f)$ where $\mathbf{W}_f$ is the correspondence layer parameters. For segmentation, two independent layers are used to predict the human segmentation labels for each modality through feed-forward processes $\mathcal{F}(\mathcal{F}(I_c; \mathbf{W}_f); \mathbf{W}_c)$ and $\mathcal{F}(\mathcal{F}(I_t; \mathbf{W}_f); \mathbf{W}_t)$ where $\mathbf{W}_c$ and $\mathbf{W}_t$ are segmentation network parameters. Note that overall parameters for color segmentation network are $\mathbf{W} = \{\mathbf{W}_f, \mathbf{W}_c\}$ or vice versa.

2.3. Flow Estimation between Multispectral Images

As their complementary information, the fusion of multispectral images can boost the performance as shown in many applications such as pedestrian detection [6] and segmentation [3]. However, these methods have inherent limitations to be applied on un-registrated multispectral images since they assume that these two images are registrated or taken under same viewpoint. To tackle this problem, in the proposed method, we explicitely estimate two dense correspondence flows, i.e., $f_{c\to t}$ from color I_c to thermal I_t and $f_{t\to c}$ from thermal I_t to color I_c, where $f(i) \in [u_i, v_i]^T$, satisfying $I_t(x_i, y_i) = I_c(x_i + u_i^{c\to t}, y_i + v_i^{c\to t})$ or vice versa. To estimate these correspondeces, intermediate convolutional activations from the network of each modality are first extracted as

$$A_c = \mathcal{F}(I_c; \mathbf{W}_f), \quad A_t = \mathcal{F}(I_t; \mathbf{W}_f). \tag{1}$$

To reduce the outliers, these activations are normalized to have an unit norm such that $A_c / |A_c|$ and $A_t / |A_t|$. They are then used to define the matching evidence in SIFT flow (SF) optimization [13]. Since these activations are a quarter size of an input image, estimated flows through the SF optimization are bilinearly interpolated, and filtered by edge-aware filtering (EAF) [14, 15] with the guidance of each modality image based on the color and flow field consistency assumption.

However, even after the above flow estimation, there might exist errors or outliers on dense flow fields, which might degenerate the human segmentation quality when fusing the modalities. Intuitively the correspondence relation from a color image to a thermal image should be consistent with that from the thermal image to the color image. Thus, in order to remove the outliers on estimated correspondence fields, we adopt a correspondence consistency such that

$$|f_{c\to t}(i + f_{t\to c}(i)) + f_{t\to c}(i)|_1 \leq t, \tag{2}$$

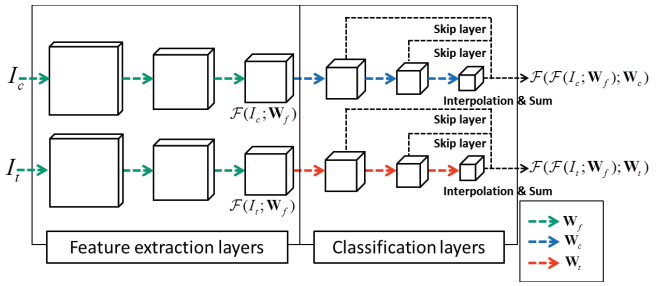


Fig. 3. Network architecture of the segmentation network

where t is a threshold parameter. Inconsistent flows which do not satisfy the condition above are discarded, thus only consistent flows are remained.

2.4. Segmentation Label Updating

We formulate the joint human segmentation for multispectral color and thermal images in a weakly-supervised and iterative manner. During an iteration, the resulting human segmentation results for one modality are used to learn the network for another modality. To fuse the these resulting human segmentation results, we formulate a fully connected conditional random field (CRF) model, which is popularly used to recover detailed structure of object boundaries [16] as a post processing. Unlike conventional methods [10, 11], we incorporate the CRF model to fuse the resulting outputs of two complementary networks and learn the segmentation network of each modality simultaneously.

First of all, initial probabilities for human segmentation can be represented as the convolutional activations from last classification layer of each modality such that

$$P_c = \mathcal{F}(A_c; \mathbf{W}_c), \quad P_t = \mathcal{F}(A_t; \mathbf{W}_t). \tag{3}$$

Then, our energy function in the CRF model can be formulated to optimize the initial human probabilities P_c and P_t in a joint manner. Unlike exising CRF models [11], our CRF model is formulated with a fused probability P_f and a fused guidance image I_f such that

$$E_f(S) = \sum\nolimits_i \phi(P_f(i)) + \sum\nolimits_{i,j} \varphi(I_f(i), I_f(j)) \tag{4}$$

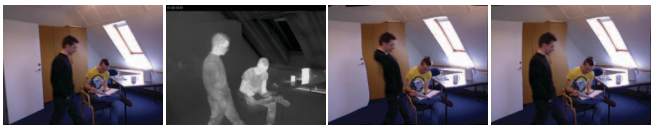
where $\phi(\cdot)$ is an unary data term and $\varphi(\cdot,\cdot)$ is a pairwise regularization term. The unary data term $\phi(P_f(i))$ is computed as $-\log P_f(i)$ and the pairwise regularization term $\varphi(I_f(i), I_f(j))$ is calculated as $k(I_f(i), I_f(j)) \cdot \xi(i,j)$, where $\xi(i,j) = 1$ if $i \neq j$ and zero otherwise. In the following, the fused unary and pairwise terms in our method can be introduced.

2.4.1. *Unary Fusion*

The initial human probabilities P_c and P_t can be used to segment the human in a joint manner. Since each modality encodes different characteristics of the human and has its own distinct advantages at segmenting humans, fusion of probabilities into the unary term enables the method to estimate reliable segmentation labels. Specifically, we warp unregistered score map from thermal image P_t to color domains through a valid flow $f_{t\to c}$ and create a unary term by selecting a larger probability from P_c and $f_{t\to c}(P_t)$ as

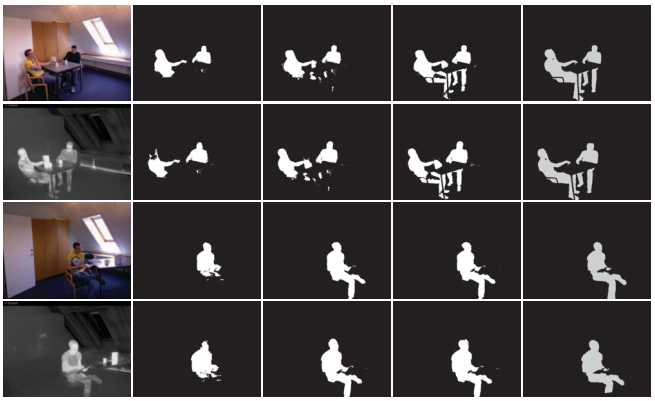
$$\phi(P_f) = -\log(\max(P_c, f_{t\to c}(P_t))). \tag{5}$$

Similarly, for thermal domain, the unary data term are generated by selecting the larger probability from P_t and $f_{c\to t}(P_c)$.



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

Fig. 4. Visualization of warping results for (a) a color image and (b) a thermal image using activation features at (c) iteration 1 and (d) iteration 2. Correspondence estimation performance was enhanced as evolving iterations.



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

Fig. 5. Comparison of qualitative evaluations of human body segmentations as evolving the iteration: (a) input color and thermal images, human segmentation results at (b) iteration 1, (c) iteration 2, and (d) iteration 4 and (e) ground truth. Thanks to joint fusion of multispectral information, human segmentation results the proposed method get improved as evolving the iteration.

2.4.2. *Pairwise Fusion*

The guidance image from only color or thermal might be sensitive to ambiguous segmentation boundaries, i.e. there are similar colors between human body and background across segmentation boundaries. To alleviate this limitation, we propose a pairwise fusion technique in a way that the guidance image I_f is generated using unregistered multimodal image pairs $\{I_c, I_t\}$ and flows $\{f_{t\to c}, f_{c\to t}\}$ between them. For example, in order to refine the human labels in color domain, guidance image I_f is generated by fusing red, green channel of I_c and thermal channel of $f_{t\to c}(I_t)$. With this guidance, the kernel $k(i,j)$ consists of two gaussian kernels depending on the pixel position i and j and pixel intensity $I_f(i)$ and $I_f(j)$:

$$\begin{aligned} k(I_f(i), I_f(j)) = \omega_1 \exp\left(-\frac{|i-j|^2}{2\rho_\alpha^2} - \frac{|I_f(i)-I_f(j)|^2}{2\rho_\beta^2}\right) \\ + \omega_2 \exp\left(-\frac{|i-j|^2}{2\rho_\gamma^2}\right). \end{aligned} \tag{6}$$

This pairwise fusion of multispectral images allows us to have a more informative and richer representation of the scene. Specifically, color modality contributes to extract contour and texture information and thermal modality contributes to extract temperature information. Both of which can be helpful to segment humans. As shown in Fig.1, different modalities have different advantages in human segmentation. When using these two modalities simulataneously, the human segmentation performance can be boosted.

Table 1. Comparison of quantitative evaluations on the VAP benchmark [3]. We measured the mean IOU for human segment labels.

	FCN-CRF [9]	CRF-RNN [10]	Ours (initial)	Ours (final)
Color	49.32	69.24	73.44	**79.68**
Thermal	46.92	64.16	61.23	**69.91**
Average	48.12	66.70	67.34	**74.80**

3. EXPERIMENTAL RESULT

3.1. Experimental Settings

In experiments, we implemented our segmentation networks using the VLFeat MatConvNet library [17]. For convolution layers, the FCN model [9] pretrained on PASCAL-VOC benchmark [18] was used to initialize the network. Note that due to the statistical discrepancy between PASCAL-VOC benchmark [18] and multispectral VAP Trimodal dataset [3], our testing database, an initial parameter in the FCN model [9] cannot provide reliable solutions for our task. Training was performed with stochastic gradient descent (SGD), using a mini-batch of size 20, a learning rate of 0.0001, momentum of 0.9, and a weight decay of 0.0005. For each modality, the CRF energy function was optimized with the fixed associated parameters in 10 iteration, set to $\{\omega_1, \omega_2, \rho_\alpha, \rho_\beta, \rho_\gamma\} = \{5, 3, 50, 10, 3\}$.

The proposed method was evaluated on VAP people segmentation dataset [3] which consists of RGB, thermal, and depth video sequences. Since the dataset consists of a video sequence, our segmentation network might be in overfitting. To remove redundancy, We sampled 520 images for training and 520 images for testing in 5724 annotated frames. We augmented the training data by randomly flipping and cropping the image batch to increase the variation of the training data. While this dataset includes calibration and registration algorithm among modalities, they were not used in our experiment because we consider general settings, i.e., multi-spectral cameras are unregistered.

3.2. Convergence Analysis

We first evaluated our method as evolving the iteration as shown in Fig. 5. To evaluate the convergence of our method, we iteratively performed the whole training process described in sec.2.1. As shown in results, the human segment labels are getting more reliable as the iterations progress. Sepcifically, in first iteration, the network produced miss-classified human areas on some regions that have the ambiguity on the image, especially at leg parts. Since color of the human and the background are very similar, it is hard to distinguish them. As iterations progress, however, sub-networks improve performance by complementing each other. As shown in input images, a thermal image has a larger difference between the leg part and the background than that on color image, thus thermal sub-networks could generate more reliable labels on these regions. This thermal label affects the color label with joint CRF optimization. These improved labels help train the networks for each modality gradually. Furthermore, Fig. 4 shows the dense correspondence estimation results. As expected, dense correspondences between color and thermal images are more reliably estimated as evolving an iteration, which iteratively boost the convergence of our network training.

3.3. Comparison with Other Methods

We evaluated our method in comparison to state-of-the-art human segmentation methods such as FCN-CRF and CRF-RNN[10]. The FCN-CRF combines the pretrained FCN network [9] with CRF post processing [16] for pair comparison. Both of FCN-CRF and CRF-RNN are trained with the supervision of the pixel-level annotations in the PASCAL VOC dataset. For quantitative evaluations, the performance wass measured as pixel intersection-over-union (IOU) [9, 8, 10]. Fig. 6 provides comparison of qualitative evaluations for human segmentation on both color and thermal modality. As expected, the results of FCN-CRF cannot estimate segment labels when the intensities of human and background are similar at each modality. CRF-RNN [10] was robust to this intensity problem by CRF propagation to neighbor pixels but it shows that the larger the problem area, the lower the performance. These problems are due to low performance of single modality estimation. On the other hand, our method segments humans well for both modalities because our networks are updated by synergy of both modality. IOU results of these methods are denoted in Table 1. We achieved 8.1% improvements in average IOU than CRF-RNN [10], which proves that our method definitely overcomes the limitation of single modality based flow estimation by leveraging a modality fusion.

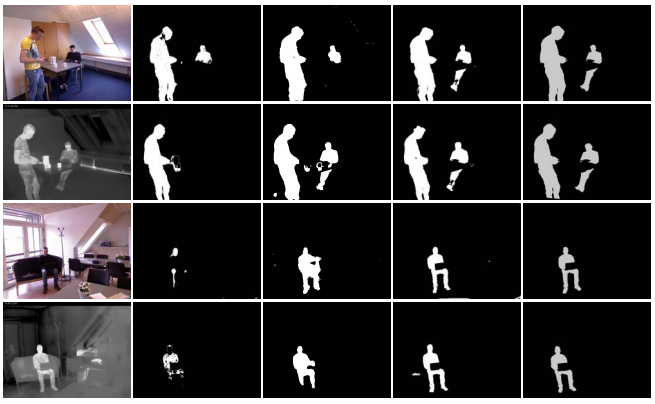


Fig. 6. Comparison of qualitative evaluations of human body segmentations on the VAP benchmark [3]: (a) input color and thermal images, human segmentation results using (b) FCN-CRF [9], (c) CRF-RNN [10], and (d) our method, and (e) ground truth. For a fair comparision, CRF [16] optimization as a post processing scheme is applied to FCN [9].

4. CONCLUSION

We proposed the human body co-segmentation method for unregistered multi-spectral color and thermal image by leveraging CNNs. We formulated the human co-segmentation problem as two sub-problems, dense correspondence and segmentation, in a joint and iterative manner. We formulated the learning framework such that training labels for one modality are gradually updated as an iteration progress to learn the other modality network. To realize that task, dense correspondences are estimated using intermediate convolutional activations of segmentation network, and with this flow field, the human body segmentation labels are fused in a unary and pairwise term within CRF optimization. Experimental results demonstrated that our method provides highly accurate segmentation results in comparison to state-of-the-art methods even without pixel-level human annotations.

5. ACKNOWLEDGEMENTS

This work was supported by Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government(MSIP)(No.2016-0-00197)

6. REFERENCES

[1] V. Vineet and J. Warrell, "Human instance segmentation from video using detector-based conditional random fields," *In Proc. of BMVC*, 2011.

[2] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," *In Proc. of CVPR*, 2003.

[3] C. Palmero, A. Clapés, C. Bahnsen, A. Møgelmose, T. Moeslund, and S. Escalera, "Multi-modal rgb–depth–thermal human body segmentation," *IJCV*, vol. 118, no. 2, pp. 217–239, 2016.

[4] J. Nikkanen, T. Gerasimow, and L. Kong, "Subjective effects of white-balancing errors in digital photography," *Optical Engineering*, vol. 47, no. 11, pp. 113201–113201, 2008.

[5] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video snapcut: robust video object cutout using localized classifiers," *In Proc. of ACM SIGGRAGH*, 2009.

[6] H. Choi, S. Kim, K. Park, and K. Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolution network," *In Proc. of ICPR*, 2016.

[7] S. Hwang, J. Park, N. Kim, Y. Choi, and I. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," *In Proc. of CVPR*, 2015.

[8] G. Papandreou, L. Chen, K. Murphy, and A. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," *In Proc. of CVPR*, 2015.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *In Proc. of CVPR*, 2015.

[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," *In Proc. of CVPR*, 2015.

[11] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv:1412.7062*, 2014.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[13] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. PAMI*, vol. 33, no. 5, pp. 978–994, 2011.

[14] D. Min, S. Choi, J. Lu, B. Ham, K. Sohn, and M. Do, "Fast global image smoothing based on weighted least squares," *IEEE Trans. IP*, vol. 23, no. 12, pp. 5638–5653, 2014.

[15] K. He, J. Sun, and X. Tang, "Guided image filtering," *In Proc. of ECCV*, 2010.

[16] V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *In Proc. of NIPS*, 2011.

[17] A. Vedaldi and Ka. Lenc, "Matconvnet: Convolutional neural networks for matlab," *In Proc. of ACM*, 2015.

[18] M. Everingham, Luc Van G., C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.