On guiding video object segmentation

Diego Ortego
Insight Centre for Data Analytics
Dublin City University (DCU)
Dublin (Ireland)
diego.ortego@insight-centre.org

Kevin McGuinness
Insight Centre for Data Analytics
Dublin City University (DCU)
Dublin (Ireland)
kevin.mcguinness@insight-centre.org

Juan C. SanMiguel
VPULab
Universidad Autónoma de Madrid (UAM)
Madrid (Spain)
juancarlos.sanmiguel@uam.es

Eric Arazo
Insight Centre for Data Analytics
Dublin City University (DCU)
Dublin (Ireland)
eric.arazo@insight-centre.org

José M. Martínez
VPULab
Universidad Autónoma de Madrid (UAM)
Madrid (Spain)
josem.martinez@uam.es

Noel E. O'Connor
Insight Centre for Data Analytics
Dublin City University (DCU)
Dublin (Ireland)
noel.oconnor@insight-centre.org

Abstract—This paper presents a novel approach for segmenting moving objects in unconstrained environments using guided convolutional neural networks. This guiding process relies on foreground masks from independent algorithms (i.e. state-of-the-art algorithms) to implement an attention mechanism that incorporates the spatial location of foreground and background to compute their separated representations. Our approach initially extracts two kinds of features for each frame using colour and optical flow information. Such features are combined following a multiplicative scheme to benefit from their complementarity. These unified colour and motion features are later processed to obtain the separated foreground and background representations. Then, both independent representations are concatenated and decoded to perform foreground segmentation. Experiments conducted on the challenging DAVIS 2016 dataset demonstrate that our guided representations not only outperform non-guided, but also recent and top-performing video object segmentation algorithms.

Index Terms—Video object segmentation, foreground segmentation, attention, deep learning.

I. INTRODUCTION

SEGMENTING an image into regions is key for identifying objects of interest. For example, image segmentation [1] clusters image pixels with common properties (e.g. colour, textures, etc), while semantic segmentation [2] categorises each pixel into a set of predefined classes. Foreground segmentation is a particular case of semantic segmentation with two categories: foreground and background. The former contains the objects of interest in an image, which may correspond to salient objects [3] [4], generic objects [5] [6], moving objects [7], spatio-temporal relevant patterns [8] or even weak labels [9]. The latter consists on the non-relevant data, being usually the static scene objects. Moreover, foreground segmentation in unconstrained environments is known as video object segmentation (VOS) and faces many challenges related to camera motion, shape deformations, occlusions or motion blur [10].

VOS has become an active research area, as demonstrated by the widespread use of the DAVIS benchmark [11]. Existing algorithms are supervised, semi-supervised and unsupervised. Supervised approaches employ frame-by-frame human intervention [12] whereas semi-supervised ones only require initialization (e.g. annotations of the objects to segment in the first frame) [13] [14]. Conversely, unsupervised VOS does not involve human intervention, requiring the automatic detection of relevant moving objects [15].

Currently deep learning is significantly advancing computer vision performance [16] such as for VOS, where state-of-the-art approaches employ convolutional neural networks [7] [17]. Performance improvement can be achieved by increasing the complexity of the network [18] [19], but also by learning better models without requiring new architectures. For instance, loss function variations [20], transfer learning [21], data augmentation [22] or applying spatial attention [23] are techniques widely explored. In particular, spatial attention can be used to highlight activations in feature maps of the network, thus enabling training of more accurate models. In [23], attention is extracted from convolutional features from a semantic segmentation network to promote those activations that do not respond to several classes, but to only one. Also, [24] generates attention maps responding to visual patterns in different scales and levels from convolutional features, thus capturing different semantic patterns such as body parts, objects, or background that improve pedestrian attribute recognition. Furthermore, [25] uses three attentions maps (general, residual and channel attentions) extracted from convolutional features to weight the cross-correlation of a fully-convolutional siamese tracker that adapts the offline learned model to the online tracked targets. Additionally, content-based image retrieval [26] can be enhanced by using visual attention models to weight the contribution of the activations from different spatial regions.

This paper proposes a novel approach to improving the performance of an unsupervised VOS algorithm by using an independent foreground segmentation (i.e. the mask from an existing algorithm in the literature) to guide the segmentation process by focusing on relevant activations. First, given a video frame and its associated optical flow, two networks compute appearance and motion feature maps. Second, both feature
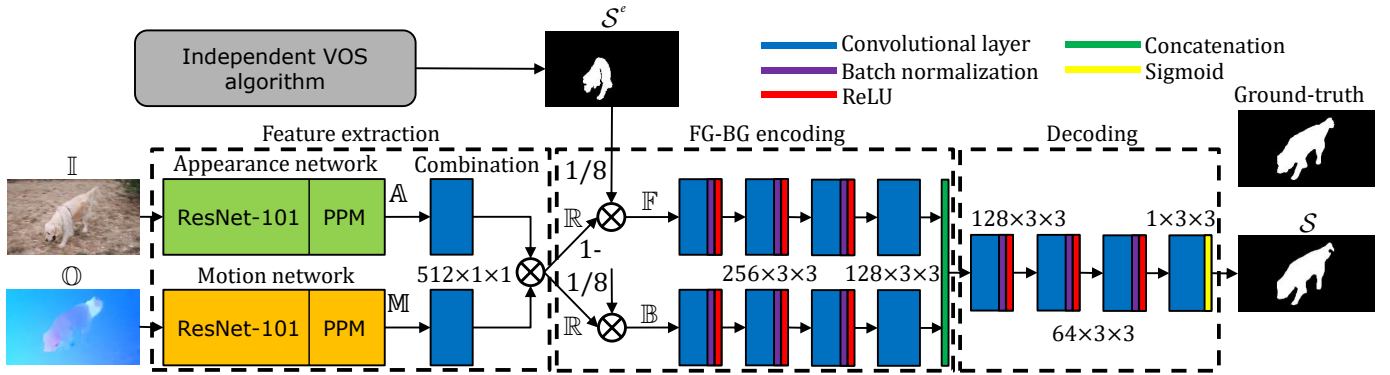
Fig. 1. Overview of the proposed architecture. Video object segmentation is performed making use of appearance and motion convolutional representations computed using features extracted after the Pyramid Pooling Module (PPM) of PSPNet [2]. Then, such representations are decoded with the guide of an attention mechanism provided by the foreground segmentation mask $\mathcal{S}^e$ computed using an independent algorithm. Key: FG (foreground). BG (background).

maps are unified to exploit their complementarity. Third, the foreground mask of an independent algorithm is used to encode foreground and background information. Finally, both representations are concatenated and decoded to produce a foreground mask. We validate the proposed approach in the recent DAVIS 2016 dataset [11], demonstrating the improvements achieved by our approach, whose novelty lies in the use of an independent foreground mask to separate foreground and background representations and learn a better foreground segmentation model.

The remainder of this paper is organized as follows: Section II overviews the proposed approach whereas Section III and IV describe the proposed algorithm and the experiments performed. Finally, Section V concludes this paper.

II. ALGORITHM OVERVIEW

VOS can be formulated as a pixel-wise labelling of each video frame $\mathbb{I}$ as either foreground (1) or background (0), thus generating a foreground segmentation mask $\mathcal{S}$. We propose an approach based on convolutional neural networks (CNNs) that uses a frame $\mathbb{I}$ together with its corresponding optical flow $\mathbb{O}$ and the foreground mask of an independent algorithm $\mathcal{S}^e$ (from now on, the independent foreground mask), to compute a foreground segmentation $\mathcal{S}$ (see Figure 1). Our CNN-VOS approach starts with an appearance network computing a convolutional feature map $\mathbb{A}$ from a video frame $\mathbb{I}$ and a motion network generating a convolutional feature map $\mathbb{M}$ from the optical flow $\mathbb{O}$ associated to $\mathbb{I}$. Both feature maps are then combined by element-wise multiplication to obtain a unique representation $\mathbb{R}$ comprising both appearance and motion information. Subsequently, the independent foreground mask $\mathcal{S}^e$ is used to weight $\mathbb{R}$ and separately encode foreground $\mathbb{F}$ and background $\mathbb{B}$ information based on the previously computed feature maps, which are processed by two independent networks. Finally, the output of the two sub-networks is concatenated and decoded with a convolutional network to produce the foreground mask $\mathcal{S}$.

III. ALGORITHM DESCRIPTION

A. Appearance network

The appearance network learns a model $\mathbb{A}$ from spatial distribution of the colour information of a video frame $\mathbb{I}$. In particular, we use PSPNet [2], a fully-convolutional neural network for semantic segmentation which relies on ResNet [19] followed by a hierarchical context module (Pyramid Pooling Module) that harvest information from different scales. This module processes different downsamplings of ResNet features and later upsamples them for concatenating all features to form an improved representation that includes the original ResNet features. In particular, we use ResNet-101 to obtain $\mathbb{A}$ by extracting the 512 feature maps obtained after the convolutional layer that follows the Pyramid Pooling Module, which contains both local and global representations.

B. Motion network

In contrast to the appearance network, the motion network obtains a model $\mathbb{M}$ again making use of PSPNet, but in this case from the optical flow $\mathbb{O}$. In particular, we convert the optical flow vectors [27] into a 3-channel colour-coded optical flow image $\mathbb{O}$ [15] and train PSPNet to produce a foreground segmentation from the colour-coded optical flow $\mathbb{O}$. In order to get the model $\mathbb{M}$, we select the 512 feature maps at the same layer as done in the appearance network.

C. Unified representation

The intuition behind using two networks with independent inputs (appearance and motion) is to benefit from the complementary of appearance and motion exhibited by moving objects [28]. Therefore, we exploit such complementarity by combining both feature maps following a multiplicative fusion (see the Combination stage in Figure 1) similarly to [29], where the authors multiply different sources of CNNs (appearance and motion) to enable the amplification or suppression of feature activations based on their agreement. In our approach, this fusion consists of a 1×1 convolution applied to both the appearance features $\mathbb{A}$ and the motion features $\mathbb{M}$ followed by an element-wise multiplication of both sets of feature maps

(which have the same dimensionality) to produce the unified encoding $\mathbb{R}$. Applying the 1×1 convolutions helps to control the dimensionality while learning how to combine the feature maps before their multiplicative combination.

D. Foreground and background encoding

Unlike many VOS literature which jointly process all features [15], we introduce an attention mechanism that splits the feature maps into foreground and background to better guide the learning process and obtain a better VOS model (see FG-BG encoding stage in Figure 1). In particular, we use an foreground mask $\mathcal{S}^e$ (obtained from a state-of-the-art algorithm) which is downsampled to match the size of $\mathbb{R}$. Then, we split $\mathbb{R}$ into foreground $\mathbb{F} = \mathbb{R} \times \mathbb{S}^e$ and background $\mathbb{B} = \mathbb{R} \times (1 - \mathbb{S}^e)$ representations according to the guidance provided by the independent foreground $\mathcal{S}^e$. On the one hand, multiplying the feature maps $\mathbb{R}$ by the independent foreground $\mathcal{S}^e$ helps focusing on important spatial areas of the features by zeroing responses associated to background areas in $\mathbb{F}$, thus implementing an attention mechanism. On the other hand, maintaining information from the background regions through $\mathbb{B}$ helps in the segmentation process, which may be useful as it assures a background representation is maintained, especially when errors in $\mathcal{S}^e$ lead to the suppression of important foreground responses. Subsequently, $\mathbb{F}$ and $\mathbb{B}$ are individually processed by four convolutional layers to separately model foreground and background representations before concatenating them as presented in the *FG-BG encoding* stage in Figure 1. Finally, this set of concatenated foreground and background representations contain a high-level joint encoding of appearance and motion that is fed to four additional convolutional layers for decoding in order to compute the final prediction (see Decoding stage in Figure 1), i.e. the foreground segmentation mask $\mathcal{S}$.

E. Architecture details

Our architecture is the fully-convolutional network shown in Figure 1. After the appearance and motion networks (both using the PSPNet architecture), there are two convolutional layers that use 512 1×1 kernels to process the feature maps $\mathbb{A}$ and $\mathbb{M}$ (both have 1/8 the original image resolution). Then, the multiplication of $\mathcal{S}^e$ is preceded by a downsampling performed through average pooling with stride 8. Regarding the processing of $\mathbb{F}$ and $\mathbb{B}$: before concatenating both we use four convolutional layers, three with depth 256 followed by batch normalization and ReLU and the final one with depth 128 to reduce the number of feature maps. As it is well known that reducing the feature map resolution may result in coarse predictions [30], we use dilated convolutions (1, 2, 4 and 8 depths) to aggregate context [31] while preserving the already reduced resolution (1/8). Finally, we apply four convolutional layers with dilated convolutions (128, 64, 64 and 1 depths). We use 3×3 kernels throughout (unless otherwise stated).

IV. Experimental work

A. Datasets and metrics

We evaluate the proposed approach on the DAVIS 2016 dataset [11] which covers many challenges of unconstrained VOS. We use all available test videos (20) whose length ranges from 40 to 100 frames with 854×480 resolution. A unique foreground object is annotated in each frame. We use three standard performance measures [11] to account for region similarity (Jaccard index $\mathcal{J}$, i.e. intersection-over-union ratio between the segmented foreground mask and the ground-truth mask), for contour accuracy (F-score $\mathcal{F}$ between contour pixels of the segmented and the ground-truth masks) and for temporal stability $\mathcal{T}$ of foreground masks (by associating temporal smooth and precise transformations to good foreground segmentations). These measures are computed frame-by-frame and then averaged to compute a sequence-level score, which are also averaged to get dataset-level performance. Note that we do not consider DAVIS 2017 dataset as it is semi-supervised whereas our approach is unsupervised.

B. Implementation details

The training procedure consists of three stages. First, we train the appearance network to perform semantic segmentation on the PASCAL VOC 2012 dataset [32] using a total of 10582 training images. Second, we train the motion network to perform foreground segmentation based on optical flow data [27] using the annotations provided by [15] for 84929 frames of ImageNet-Video dataset [33]. Third, we freeze both the appearance and the motion networks and train the rest of the network on 22 of the 30 training video sequences of DAVIS 2016 dataset [11] (we use 8 sequences for validation). For this last step, masks of an independent algorithm are needed, so we use available foreground masks in DAVIS 2016[1] (the 13 algorithms evaluated in [11] and the algorithm proposed in [34]). We train the appearance and the motion networks for 30k and 90k iterations, respectively, using the "poly" learning rate policy described in [2]. However, for the third step we train the network for 20 epochs reducing the learning rate by ten each five epochs (starting with the value 0.1) and we select the best model using the performance in the validation set. We use batch size 8, data augmentation (random Gaussian blur, sized crops, rotations and horizontal flips), cross-entropy loss and *Kaiming* weight initialization [35] in all training steps.

C. Evaluation

We compare our approach against 6 recent and top-performing VOS alternatives (both unsupervised and semi-supervised): ARP [7], CTN [14], FSEG [15], MSK [17], OFL [36] and VPN [37]. To provide a fair comparison, we also use the segmentation masks of these alternatives for our proposal (indicated by *). Table I presents the average performance for the 20 test sequences of DAVIS 2016 in terms of Jaccard index $\mathcal{J}$, contour F-score $\mathcal{F}$ and temporal stability $\mathcal{T}$. By comparing the independent foreground masks and the proposed approach

[1]https://davischallenge.org/davis2016/soa_compare.html

TABLE I
COMPARATIVE RESULTS FOR DAVIS 2016. (*) INDICATES THE PROPOSED
APPROACH USING THE CORRESPONDING FOREGROUND MASK. ↑ (↓)
MEANS HIGHER (LOWER) IS BETTER. BOLD DENOTES BETTER
PERFORMANCE FOR OUR APPROACH.

| Algorithms	$\mathcal{J}$ (↑)		$\mathcal{F}$ (↑)		$\mathcal{T}$ (↓)	
	mean	std.	mean	std.	mean	std.
ARP	.7609	.1125	.7051	.1111	.5363	.3446
ARP*	**.8069**	**.0631**	**.8121**	**.0799**	**.3930**	**.2675**
CTN	.7304	.1048	.6886	.1147	.3682	.2176
CTN*	**.7933**	**.0627**	**.7996**	**.0775**	.3969	2762
FSEG	.7068	.0804	.6524	.1036	.4456	.2809
FSEG*	**.7834**	**.0673**	**.7950**	**.0773**	**.3938**	**.2566**
MSK	.7955	.0817	.7519	.0929	.3226	.1978
MSK*	**.8012**	**.0718**	**.8097**	**.0821**	.3805	.2660
OFL	.6738	.1231	.6279	.1330	.3608	.2123
OFL*	**.7844**	**.0764**	**.7926**	**.0933**	.3898	.2805
VPN	.6984	.0902	.6513	.1002	.4923	.2630
VPN*	**.7765**	**.0709**	**.7920**	**.0792**	**.3918**	**.2551**

Fig. 2. Performance in DAVIS 2016 dataset of the non-guided representations learnt (NG1 and NG2) compared to the improvements achieved when guiding through independent foreground masks.

(*), we outperform all of them in terms of $\mathcal{J}$ and $\mathcal{F}$. Regarding the temporal stability $\mathcal{T}$, we improve for 3 of the 6 algorithms. The reason $\mathcal{T}$ is not improved for CTN, MSK, and OFL is that these are implicitly focused on transferring the same segmentation frame to frame, thus resulting in an inherent temporal stability that is difficult to improve upon from and unsupervised perspective. The improvements achieved are due to both the potential of our learned representations and the guiding mechanism introduced through the independent masks. As it is not possible to separate both facts from the results in Table I, we perform an additional experiment to highlight the potential of the learned representations. In particular, we have repeated the third step of the training process (see Subsection IV-B) without making use of independent foreground masks. We explore two non-guided alternatives, the first one implies removing the FG-BG encoding (see Figure 1), thus passing the representation $\mathbb{R}$ directly to the decoder (NG1 alternative). The second alternative consists in using a dummy foreground mask where all pixels contain foreground (i.e. $\mathbb{F}$ and $\mathbb{B}$ contain the same information), thus avoiding the use of independent foreground masks to guide the segmentation while using a network (NG2 alternative) with the same number of parameters as the one proposed. Figure 2 presents the performance in terms of $\mathcal{J}$ and $\mathcal{F}$ for this two non-guided alternatives (NG1 and NG2), demonstrating that non-guided representations perform worse than the guided ones, i.e. those learnt with separate foreground-background representation.

Furthermore, directly comparing NG1 and NG2 performance (around $\mathcal{J} = 0.76$) with the selected alternatives in Table I shows that we obtain competitive results (only MSK with $\mathcal{J} = 0.7955$ is better than NG1 and NG2).

Figure 3 gives some examples of foreground segmentations. The first column presents an example where the proposed mask (green) is capable of solving the mistakes of the non-guided representations (red) due to the guiding introduced by an independent algorithm (blue). The second and third columns show examples where the proposed approach (green) outperforms the independent algorithm (blue) due to the guid-
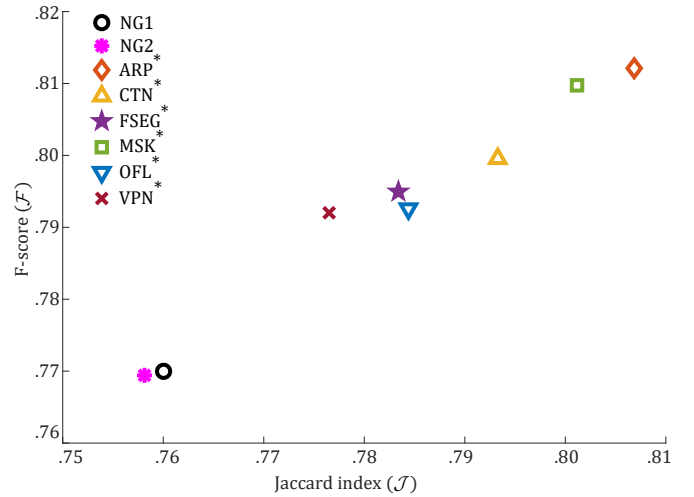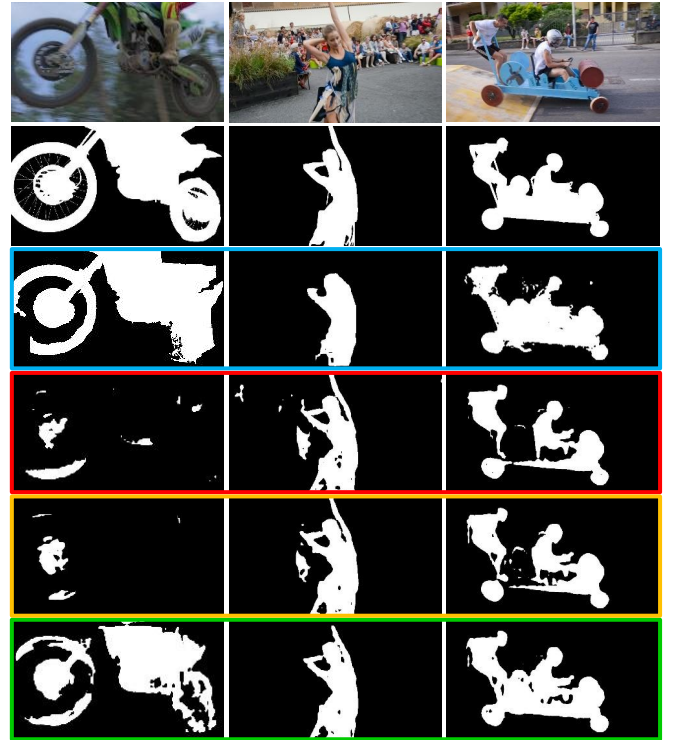

Fig. 3. Foreground segmentation examples. From top to bottom: image to segment, ground-truth and independent (blue), non-guided 1 (red) and 2 (orange) and proposed guided (green) foreground segmentation masks. From left to right, examples for *motocross-jump* (ARP), *dance-twirl* (FSEG) and *soapbox* (VPN) video sequences. Note that the masks with no independent guiding (red and orange) do not depend on a particular algorithm.

ing scheme over representations that, without guiding (red), were lacking of object parts or introducing false positives.

V. CONCLUSIONS

This paper proposes a VOS approach which takes advantage of independent (i.e. state-of-the-art) algorithm results to guide

the segmentation process. This strategy enables learning an enhanced colour and motion representation for VOS due to specific attention on foreground and background classes. The experimental work validates the utility of our approach to improve foreground segmentation performance. Future work will explore the use of feedback strategies to induce the foreground-background separation from the produced result and not from independent algorithms.

REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[2] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239.

[3] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient Object Detection: A Benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[4] S. Wu, M. Nakao, and T. Matsuda, "SuperCut: Superpixel Based Foreground Extraction With Loose Bounding Boxes in One Cutting," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1803–1807, 2017.

[5] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the Objectness of Image Windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012.

[6] S. Jain, B. Xiong, and K. Grauman, "Pixel Objectness," *CoRR*, vol. abs/1701.05349, 2017.

[7] Y. J. Koh and C. S. Kim, "Primary Object Segmentation in Videos Based on Region Augmentation and Reduction," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7417–7425.

[8] Y. J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1995–2002.

[9] D. Zhang, L. Yang, D. Meng, D. Xu, and J. Han, "SPFTN: A Self-Paced Fine-Tuning Network for Segmenting Objects in Weakly Labelled Videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5340–5348.

[10] A. Faktor and M. Irani, "Video Segmentation by Non-Local Consensus voting," in *Proceedings of British Machine Vision Conference (BMVC)*, 2014.

[11] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] K. Maninis, S. Caelles, J. Pont-Tuset, and L. V. Gool, "Deep Extreme Cut: From Extreme Points to Object Segmentation," *CoRR*, vol. abs/1711.09081, 2017.

[13] S. Chen, Q. Zhou, and H. Ding, "Learning Boundary and Appearance for Video Object Cutout," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 101–104, 2014.

[14] W. D. Jang and C. S. Kim, "Online Video Object Segmentation via Convolutional Trident Network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7474–7483.

[15] S. D. Jain, B. Xiong, and K. Grauman, "FusionSeg: Learning to Combine Motion and Appearance for Fully Automatic Segmentation of Generic Objects in Videos," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2126.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[17] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning Video Object Segmentation from Static Images," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3491–3500.

[18] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[20] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollïċœr, "Focal Loss for Dense Object Detection," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[21] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of Transferability for a Generic ConvNet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1790–1802, 2016.

[22] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the Devil in the Details: Delving Deep into Convolutional Nets," in *Proceedings of British Machine Vision Conference (BMVC)*, 2014.

[23] Q. Huang, C. Xia, C. Wu, S. Li, Y. Wang, Y. Song, and C. Kuo, "Semantic segmentation with reverse attention," in *Proceedings of British Machine Vision Conference (BMVC)*, 2017.

[24] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 350–359.

[25] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, "Learning attentions: residual attentional Siamese Network for high performance online visual tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[26] E. Mohedano, K. McGuinness, X. Giró-i Nieto, and N. E. O'Connor, "Saliency Weighted Convolutional Features for Instance Search," in *International Conference on Content-Based Multimedia Indexing (CBMI)*, 2018, pp. 1–6.

[27] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[28] P. Tokmakov, K. Alahari, and C. Schmid, "Learning Video Object Segmentation with Visual Memory," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4491–4500.

[29] E. Park, X. Han, T. L. Berg, and A. C. Berg, "Combining multiple sources of knowledge in deep CNNs for action recognition," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–8.

[30] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[31] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," *CoRR*, vol. abs/1511.07122, 2015.

[32] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[33] O. D. J. Russakovsky, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[34] N. Miċœrki, F. Perazzi, O. Wang, and A. Sorkine-Hornung, "Bilateral Space Video Segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 743–751.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[36] Y. H. Tsai, M. H. Yang, and M. J. Black, "Video Segmentation via Object Flow," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3899–3908.

[37] V. Jampani, R. Gadde, and P. V. Gehler, "Video Propagation Networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3154–3164.