BenderNet and RingerNet: Highly efficient line segmentation deep neural network architectures for ice rink localization

Pascale Walters
Mehrnaz Fani
David Clausi
Alexander Wong
University of Waterloo
Email: pascale.walters@uwaterloo.ca

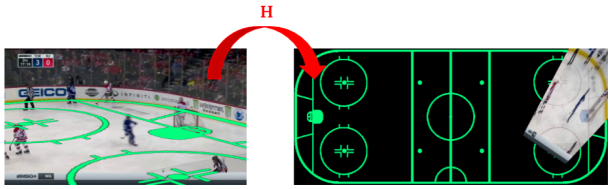


Fig. 1: The absolute positions of ice hockey players on the ice surface is determined with a homography transform H. Segmentation of the lines on the ice, shown on the left, is often an intermediate step for homography estimation.

Abstract

A critical step for computer vision-driven hockey ice rink localization from broadcast video is the automatic segmentation of lines on the rink. While the leveraging of segmentation methods for sports field localization has been previously explored, the design of deep neural networks for segmenting ice rink lines has not been well studied. Furthermore, the exploration of efficient architecture designs is very important given the operational requirements of real-time sports analytics. Motivated by this, BenderNet and RingerNet, two highly efficient deep neural network architectures, have been designed specifically for ice rink line segmentation. RingerNet consists of two highly compact generative adversarial networks, designed to segment ice rink and rink lines in a sequential manner. BenderNet is a lightweight convolutional neural network architecture that comprises dilated depthwise separable convolutions to minimize architectural and computational complexity. Experiments on a dataset of annotated NHL broadcast video demonstrate high accuracy while maintaining high model efficiency, thus making the proposed methods well-suited for real-time ice hockey rink localization.

1 Introduction

Ice hockey is a popular sport, with several professional leagues operating, such as the NHL (National Hockey League) and CHL (Canadian Hockey League). Video analytics of hockey games can be used to provide teams with an advantage over their competitors, whereby they can gather more data about game events. These data can be used to influence coaching strategies and management decisions. In addition, the data can increase fan engagement as sports consumption becomes more digital [1]. The sports analytics market is rapidly growing and is anticipated to reach revenues of $4.5 billion by 2024 [2].

Data generated in real-time can be used by the coaches and players to adapt their play to their opponents. It also allows for live data to be generated for sports betting.

Data generated in real-time can be used by the coaches and players to adapt their play to their opponents. It also allows for live data to be generated for sports betting.

With many recent developments in the field of computer vision, automatic generation of sports analysis data from video has become possible. Existing computer vision solutions analyze video feed from several cameras placed in calibrated locations throughout the arena [3]. While this technique can be effective, it requires specialized hardware to be deployed at all arenas where games are played. In situations where this may not be possible, analytics derived from broadcast footage is an appropriate substitute.

Automatic analysis of a hockey game from broadcast video incorporates several tasks, including 2D-to-3D projection, player tracking, pose estimation and event detection. Once analysis is performed on the source video, higher level information (e.g., optimal strategy) can be determined.

Sports field localization is required to determine the absolute positions of the players and the puck on the ice, regardless of the broadcast camera's position. There have been methods developed to perform this analysis [4–7], and several of these methods require segmentation of the lines on the playing field as an intermediate step (Fig. 1). The resulting edge maps are used for further processing, such as for vanishing point estimation [5] or dictionary lookup [4, 7].

Despite the relative successes of these methods, there seems to be little focus into the selection of the semantic segmentation methods and justification for their use, but more into the downstream analysis [4, 5, 7]. This work deals solely with the line segmentation problem from hockey broadcast video, and proposes two efficient deep networks to solve it.

This work details two methods for approaching real-time performance for line segmentation from hockey broadcast video. BenderNet and RingerNet are small networks that achieve high accuracy on our annotated dataset from NHL games.

The contributions of this work are two lightweight semantic segmentation networks that effectively detect the lines on the playing surface from hockey broadcast video. BenderNet is two conditional generative adversarial networks (GANs) and RingerNet is segmentation network that uses dilated depthwise separable convolutions.

BenderNet achieves a mean intersection over union (mIOU) score of 31.12 with 2.8 million parameters and RingerNet achieves an mIOU score of 55.69 with 0.78 million parameters on the test split of the labelled dataset used in this work [8]. This opens the door for further research into small networks for line segmentation as an intermediate step for homography estimation.

2 Related work

Works related to small semantic segmentation networks and sports field localization are reviewed here.

2.1 Line segmentation for sports field localization

In the literature, there have been several papers that attempt to solve the problem of sports field localization. Of these published methods, there are some that require segmenting the lines and outline of the ice surface as an intermediate step [4, 5, 7]. To our knowledge, there have not been any published methods that intensively explore the line segmentation component.

The sports field localization methods in the literature have some shortcomings, however, including include methods that only report performance on frames from soccer broadcast video [4, 7] and a lack of availability of the source code [7].

Line segmentation from broadcast soccer games differs from the same task with hockey for several reasons. First, the players are much smaller in relation to the field markings in soccer games than they are for hockey. The soccer playing field is much larger than the ice surface in a hockey rink and the broadcast camera tends to be further from the soccer field. This means that the broadcast camera for soccer games captures a larger area of the field. The typical position of the broadcast camera in professional hockey rinks means that the boards on the near side of the rink tend to be captured, which occludes the players and rink markings that are on the near side of the ice. The game of hockey is more dynamic and faster than that of soccer, which means that the broadcast camera for hockey pans and zooms faster and more often. Finally, soccer games tend to be played outside, which means that they face variable lighting conditions depending on the configuration of the stadium and the time of day that the game is played. This may lead to some parts of the field in the shade and some players throwing a shadow. The variability can be within a stadium, depending on the time of day, of between stadiums, with different configuration. Hockey games may have a similar problem, as rinks across the professional leagues

Fig. 2: Three frames from the hockey line segmentation dataset. Occlusions from the near side boards can be seen in the left and centre frames and from overlayed broadcast graphics can be seen in the right frame. Pixels belonging to the line class are in blue.

tend to have variable lighting conditions, which would remain consistent over the course of a day. Example hockey broadcast frames can be seen in Fig. 2. For these reasons, methods for line segmentation that work well for soccer may not translate to an effective solution for hockey games.

Furthermore, in order to reduce inference time to approach real-time performance, the network to perform line segmentation should have few parameters [9]. VGGNet16 used by Homayounfar *et al.* [5] has 135 million parameters [10].

A method is needed that can achieve real-time or near real-time performance and also works well with a small training dataset. Due to the time-consuming nature of collecting accurate line segmentation data, our dataset contains approximately 1500 images.

2.2 Semantic segmentation

Semantic segmentation is a dense prediction task in which each pixel in an input image is assigned to a class [8, 11]. Long *et al.* in 2015 were the first to demonstrate that fully convolutional networks, inspired by CNNs for other visual tasks, could be trained end-to-end with supervised pretraining and be used to obtain state-of-the-art results [8].

Recent semantic segmentation methods that achieve state-of-the-art segmentation accuracy use spatial pyramid pooling [11, 12]. These CNNs that achieve the highest accuracy typically require billions of FLOPs, hundreds of layers and thousands of channels [13]. There are several techniques that can make these large CNNs much smaller while maintaining acceptable accuracy, which include network compression, low-bit representation, and lightweight CNNs [14].

3 Methodology

3.1 Dataset

Annotated hockey broadcast video frames were collected from ten NHL games in various arenas. Two or three clips of approximately 30 seconds were retrieved from one period from each game. Frames were then sampled at a rate of 1.5 fps.

Annotations were performed by collecting ground truth homography transforms by annotating matching points on frames from the broadcast video and a scaled model of an NHL-sized hockey rink. 1550 good quality annotated frames are used in the dataset. Three sample frames from the dataset are shown in Fig. 2.

Ground truth segmentations are extracted based on the ground truth homography [5]. The rink template was warped according to the ground truth homography and used as the segmentation mask.

3.2 BenderNet

The lines are segmented from the ice in a two step process [4]. First, the playing surface is segmented from the boards and spectators, then the lines on the playing surface are segmented from the masked frame. Both steps use simultaneously trained conditional adversarial networks [15]. Performing segmentation in two steps prevents any confusion from line-like structures on the boards or in the crowd [4].

BenderNet's architecture is based on Isola *et al.*'s pix2pix conditional adversarial network, for use in image translation tasks [15]. Line segmentation in this context can also be thought of as an image translation task, where the model of the rink is warped so that it overlays the lines in the frame. For both GANs, the generator and discriminators are U-Net shaped. The architecture of BenderNet is shown in Fig. 3.

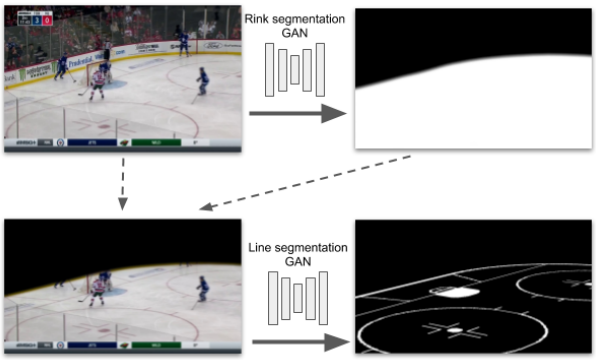


Fig. 3: BenderNet architecture. The output of the rink segmentation GAN is used as a mask on the original frame. This combined image is used as input to a line segmentation GAN that isolates the lines.

3.3 RingerNet

A network to achieve real-time performance for line segmentation from hockey broadcast video was developed by extending an ESPNetv2 segmentation backbone [14]. The use of dilated depthwise convolutional blocks allows for a reduction in the size of the network without a large reduction in accuracy.

In a standard convolution, the number of parameters that must be learned is $n^2 c\hat{c}$, where c is the number of input channels, $n \times n$ is the size of the effective receptive field, and $\hat{c}$ is the number of output channels.

In depth-wise dilated separable convolutions, the convolution operation is factored into two steps: 1) depth-wise dilated convolutions and 2) point-wise convolution. The first convolutions are performed on each input channel with a dilation rate of r, which gives a receptive field of $n_r \times n_r$, where $n_r = (n-1)(r+1)$. In the second convolution step, a linear combination of the channels is learned. This reduces the number of parameters that must be learned to $n^2 c + c\hat{c}$.

RingerNet has an architecture that comprises alternating strided EESP (extremely efficient spatial pyramid of depthwise dilated separable convolutions) and EESP modules, as described by [14]. The architecture of the segmentation network is shown in Fig. 4.

4 Experimental Setup

Semantic segmentation of lines in hockey broadcast video with RingerNet and BenderNet were evaluated on the annotated hockey dataset. The results are reported in Table 1. Train and validation splits were obtained by randomly splitting the dataset into 4 and 3 games, respectively, with approximately the same number of frames in each split. This ensures that there are games present in both splits that occur in different arenas, thereby allowing for validation of the method in varying lighting conditions and with different broadcasters.

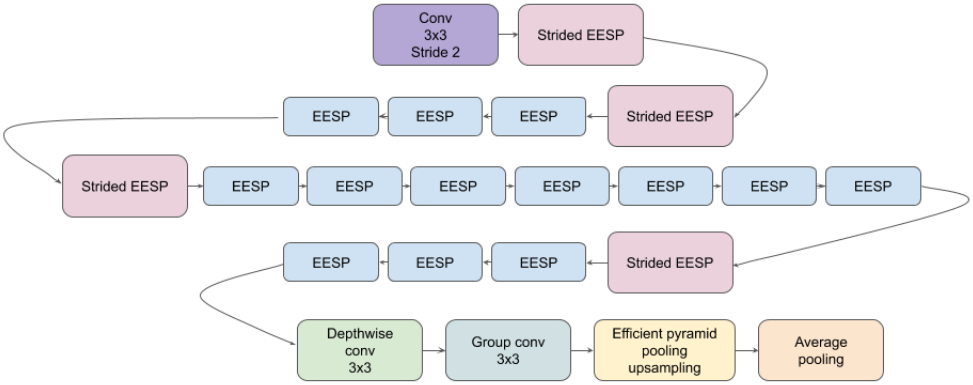


Fig. 4: Network architecture of RingerNet. The EESP and Strided EESP modules extremely efficient spatial pyramid of depthwise dilated separable convolutions.

4.1 BenderNet

The input to each GAN is two 256×256 images. For the rink segmentation GAN, the inputs are the original frame and a mask of the playing surface. For the line detection GAN, the inputs are the frame with everything other than the playing surface masked out and the line segmentation mask.

The two GANs have a U-Net backbone [16] and are trained for 100 epochs with a binary cross entropy loss. Training is also done with Adam optimizer [17] with an initial learning rate of 0.0002 and momentum of 0.5.

Experiments: Performance of BenderNet is evaluated on two tasks: 1) line segmentation with the 2GAN architecture, and 2) rink segmentation with the first segmentation network.

4.2 RingerNet

RingerNet performs line segmentation in one step. It is based on ESPNetv2 and its use of dilated depthwise separable convolutions.

Training is performed with cross entropy loss and stochastic gradient descent with momentum and weight decay as the optimizer. Momentum is set to 0.9 and weight decay is 4×10^{-5}. A hybrid learning rate scheduler, as described by Mehta *et al.* [14], varies the learning rate during training. Initial learning rate is 0.009 and has 61 epochs of a cyclic learning rate policy before switching to linear.

The network has an additional scale parameter, which refers to the a scaling factor for the number of channels used throughout the model. In this segmentation model, the scale parameter was 2. All lines are assigned to the same class and the rink and stands are assigned to the background. Both classes are weighted evenly.

Training is performed in two steps. First the network is trained on 256×256 images, then 384×384 images [14].

Experiments: The performance of RingerNet is evaluated on two tasks: 1) line segmentation on the frames directly extracted from the broadcast video, and 2) line segmentation on frames with the ice surface pre-segmented.

In the second experiment, the effects of preprocessing the video frames on the performance of the segmentation network were observed. This was inspired by BenderNet, in which the playing surface is segmented before performing detection on the lines. This line segmentation is performed with the ground truth annotations of the rink surface, where the spectators and boards are masked out before the frame is fed into the network.

5 Results and Discussion

Results for all experiments are reported in Table 1.

BenderNet performs segmentation of the lines in two stages. In the first step, the playing surface is segmented from the surrounding area and achieves an mIOU of 98.96. Performance is significantly reduced in the second step, where the lines are segmented from the masked frame and an mIOU of 31.12 is obtained. This shows that ice segmentation is an easier problem than line segmentation, likely due to occlusions and shadows from the players and the fact that the lines occupy a small area on the ice surface.

RingerNet trained and tested on the broadcast video frames was able to obtain an mIOU of 55.69, which is higher than the segmentation achieved with BenderNet. These are promising, as a much smaller network can be used to obtain acceptable results for further processing to estimate homography. Sample output of RingerNet can be seen in Fig. 5d.

When analyzing the RingerNet results, a prior segmentation of the ice surface (i.e., masking out of the spectators and boards with ground truth rink mask) increases the performance of the line segmentation to an mIOU of 60.08. This shows that further accuracy gains can be obtained by preprocessing the frames to have a prior knowledge of the rink surface. An interesting next step would be to combine the highly accurate rink segmentation component of BenderNet and the efficient line segmentation method of RingerNet.

The reported mIOU of RingerNet is higher than that of BenderNet. The frames inferred with BenderNet tend to have more continuous lines, but it is easily confused by players and other markings on the ice. BenderNet has more discontinuous lines, but can discriminate better from distracting elements of the frames. Since the frames come from broadcast video, there may be additional graphics included, such as the game clock, advertisements, and scores from other games. RingerNet does a better job of avoiding these regions, even without having to initially segment the playing surface, as with BenderNet.

In Fig. 5c, the red arrows show spurious detections in the BenderNet output. While the lines are more continuous in this figure, as compared to the RingerNet output in Fig. 5d, these spurious detections may be more deleterious in the downstream processing methods [4, 5, 7]. For example, feature extraction, as performed in [4, 7] can include these regions. These spurious regions likely lead to the lower mIOU score for this method. To achieve better performance with BenderNet, a further post-processing step may be required to remove these methods before extracting features.

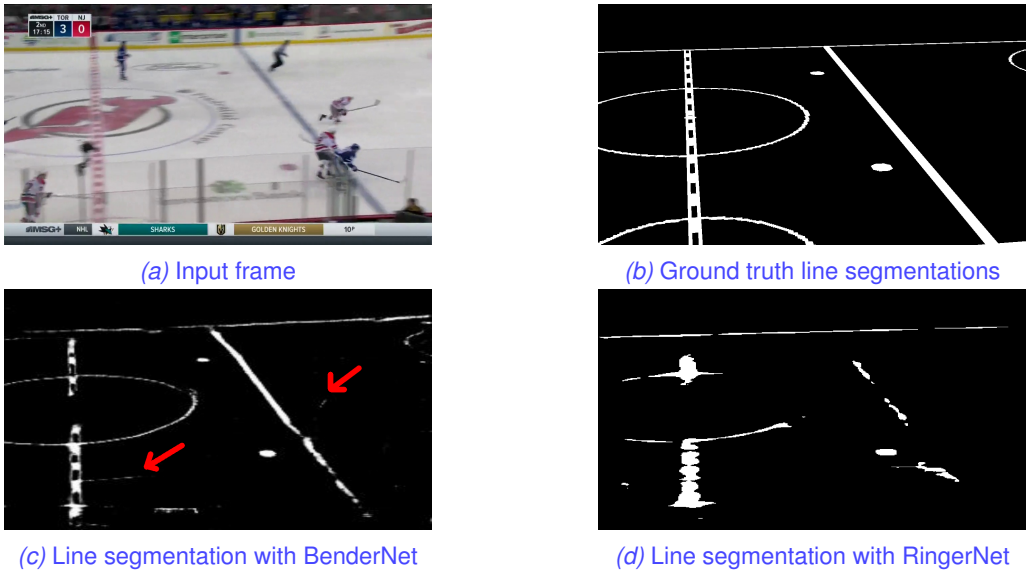
These results are encouraging to perform further experiments to investigate small networks for line segmentation. Further research could be done to observe the effects considering temporal aspects, since the input is a video.

6 Conclusion

BenderNet and RingerNet are two lightweight deep segmentation networks that achieve state-of-the-art results on the rink and line segmentation task for hockey broadcast video. In addition, a review of methods for performing sports field localization was performed, and there are several methods that require line segmentation as an intermediate step. The fundamental differences between hockey and other broadcast sports, especially soccer, means that a different, task-specific approach needs to be taken. The two efficient methods proposed in this paper can be used in sports field localization pipelines to achieve low-latency inference.

References

[1] V. Viswanathan, “Why ai is the next frontier in sports fan engagement and revenue,” Aug 2019. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2019/08/16/why-ai-is-the-next-frontier-in-sports-fan-engagement-and-revenue



(a) Input frame

(b) Ground truth line segmentations

(c) Line segmentation with BenderNet

(d) Line segmentation with RingerNet

Fig. 5: Example results for line segmentation with RingerNet and BenderNet. Inference is performed on a single frame from a broadcast video of an NHL hockey game. The red arrows in the BenderNet output show spurious detections.

Table 1: Network sizes and performances of three segmentation methods on the NHL broadcast video dataset. The results in **bold** are the best results in network size and segmentation performance for the rink seg. + line seg. task.

Method	Task	Parameters [M]	mIOU
Homayounfar *et al.* [5]	line seg.	135 [10]	-
BenderNet	rink seg. + line seg.	2.8	31.12
BenderNet	rink seg.	2.8	98.96
RingerNet	rink seg. + line seg.	**0.78**	**55.69**
RingerNet	line seg.	0.78	60.08

[2] "Global $4.5 billion sports analytics market forecasts up to 2024," Dec 2018. [Online]. Available: https://www.businesswire.com/news/home/20181205005823/en/Global-4.5-Billion-Sports-Analytics-Market-Forecasts

[3] J. Lemire, "Nhl currently testing sportlogiq as optical tracking partner," Apr 2019. [Online]. Available: https://www.sporttechie.com/nhl-testing-sportlogiq-optical-tracking-partner-data/

[4] J. Chen and J. J. Little, "Sports camera calibration via synthetic data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[5] N. Homayounfar, S. Fidler, and R. Urtasun, "Sports field localization via deep structured models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5212–5220.

[6] W. Jiang, J. C. G. Higuera, B. Angles, W. Sun, M. Javan, and K. M. Yi, "Optimizing through learned errors for accurate sports field registration," 2019.

[7] R. A. Sharma, B. Bhat, V. Gandhi, and C. Jawahar, "Automated top view registration of broadcast football videos," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 305–313.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[9] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," *arXiv preprint arXiv:1611.06440*, 2016.

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

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[13] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.

[14] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9190–9200.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.