Improving the Efficiency of CMOS Image Sensors through In-Sensor Selective Attention

Tianyi Zhang[1], Kishore Kasichainula[1], Dong-Woo Jee[2], Injune Yeo[1], Yaoxin Zhuo[3], Baoxin Li[3], Jae-sun Seo[1], Yu Cao[1]
[1]School of Electrical, Computer and Energy Enginnering, Arizona State University, Tempe, AZ, USA
[2]Department of Electrical and Computer Engineering, Ajou University, Suwon, South Korea
[3]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA
E-mail: {tzhan177, ycao}@asu.edu

2023 IEEE International Symposium on Circuits and Systems (ISCAS) | 978-1-6654-5109-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCAS46773.2023.10181835

***Abstract*—Inspired by the selective attention mechanism in human vision, we propose to introduce a saliency-based processing step in the CMOS image sensor, to continuously select pixels corresponding to salient objects and feedback such information to the sensor, instead of blindly passing all pixels to the sensor output. To minimize the overhead of saliency detection in this feedback loop, we propose two techniques: (1) saliency detection with low-precision, down-sampled grayscale images, and (2) Optimization of the loss function and model structure. Finally, we pad the minimum number of pixels around the selected pixels to maintain the accuracy of object detection (OD). Our method is experimented with two types of OD algorithms on three representative datasets. At the similar OD accuracy with the full image, our proposed selective feedback method successfully achieves 70.5% reduction in the volume of output pixels for BDD100K, which translates to 4.3× and 3.4× reduction in power consumption and latency, respectively.**

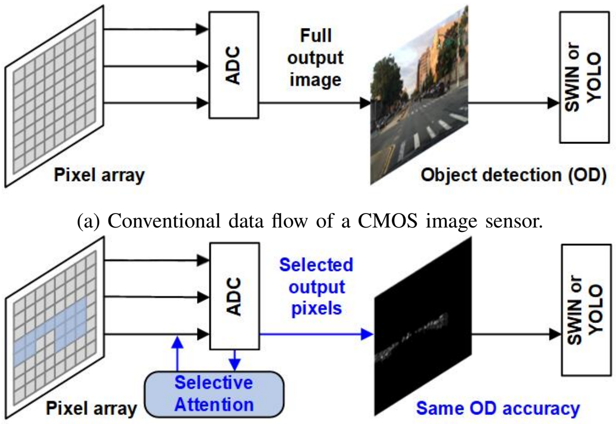
***Index Terms*—selective attention, saliency detection, image sensor, object detection, power consumption, latency**

I. Introduction

The complexity and resolution of CMOS image sensors are ever increasing, leading to larger data volume and higher cost [1]. As shown in Fig. 1(a), current design separates the pixel generation at the frontend and data processing off the sensor, with a sequential path through the analog-to-digital converter (ADC). With larger image size, such a design results in longer output latency, lower throughput, and higher power consumption, especially on the ADC.

In comparison, the human visual system employs the selective attention mechanism to solve the fundamental conflict between image size and throughput: it quickly scans the image, localizes the region of interest depending on the context (i.e., saliency), and selectively outputs the salient pixels only [2], [3]. Through selective attention, human vision dynamically adjusts the region of interest, and minimizes the output data volume without degrading the quality of further image analysis (e.g., object detection from the selected pixels).

Inspired by this mechanism, we propose to introduce a new feedback control on the image sensor, selective attention (Fig. 1(b)), which will continuously detect the salient region, feedback the selection to reduce the number of output pixels, and in turn, enhance both the throughput and energy efficiency. To meet the high-throughput demand by the sensor, we systematically reduce the complexity and output data volume through selective attention achieved by the following steps:



(a) Conventional data flow of a CMOS image sensor.

(b) Proposed method to reduce the output data volume.

Fig. 1: Comparison of the feedforward flow in conventional sensors and our proposed method of the feedback selection.

1) *Input scaling*: Instead of the full RGB image, we aggressively scale down its size and precision for selection.
2) *Algorithm optimization*: Based on the U-Net model, we significantly reduce its size and further adjust its loss function to emphasize the coverage of object pixels.
3) *Output padding*: This guarantees the accuracy and robustness of object detection (OD) with selected pixels.

The proposed method is experimented on three datasets, MSRA10K, COCO2017 and BDD100K, demonstrating significant reduction in sensor output pixels while maintaining similar OD accuracy for all datasets.

II. Selective attention

Selective attention shares a similar goal as saliency detection. Yet it emphasizes more on the preservation of true object pixels, rather than the separation between the object and background pixels. Therefore, conventional saliency detection algorithms, such as semantic segmentation or object detection with the full image, are overly expensive for selective attention. In this section, we aggressively *scale down the input image and the selection model, under the constraint of the OD accuracy post the selection*. Fig. 2 presents the processing flow, along with the pseudo-codes in Algorithm 1.

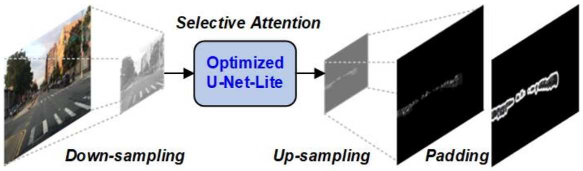


Fig. 2: The processing flow of the selective attention module.

A. Scaling of Input Image

Since the target of selective attention is less strict than the segmentation between the object and the background, it is possible to reduce the precision and size of the input image for efficient selection, with the tolerance of false object pixels.

Algorithm 1 Selective Attention

```
function $Down-sample(x, r)$
    $R_g, G_g, B_g = x[..., 0], x[..., 1], x[..., 2]$
    $R_l, G_l, B_l = gammmaExpansion(R_g, G_g, B_g)$
    $x_{HG} = 0.2126R_l + 0.7152G_l + 0.0722B_l$
    max down-sample $x_{HG}$ to $x_{LG}$ with ratio $r$
    return $x_{LG}$
end function
Randomly initialize U-Net-Lite model parameters $\theta$
for each epoch t = 1, 2, ... do
    $I_{LG} = Downsample(I_{HC}, r)$
    $p = f_{\theta_t}(I_{LG})$
    $\mathcal{L}(\theta) = -ylog(p) - \beta(1-y)log(1-p)$
    $\theta_{t+1} = \theta_t - \alpha\nabla_\theta\mathcal{L}(\theta)$
end for
for each image $I^1, I^2, ..., I^K$ do
    $M_L^k = f_\theta(I^k)$                   ▷ Generate the saliency map
    $M_H^k = upsample(M_L^k)$ ▷ Upsample the saliency map
    $M_P^k = padding(M_H^k)$            ▷ Pad the saliency map
    $I_S^k = I_{HC}^k \times M_P^k$                  ▷ Select salient pixels
end for
Randomly initialize object detection model parameters $\theta'$
for each epoch t = 1, 2, ... do
    Update model $\theta'$ with $I_S$
end for
Output the prediction of object classes and bounding boxes
with trained model, and evaluate mAP
```

Grayscale Conversion: In the human visual system, the chromatic information is sent to the parvocellular lateral geniculate nucleus after an image is projected onto the retina, while the achromatic portion is sent to the superior colliculus for selective attention [3]. Based on this observation, we first convert the RGB image (i.e., I_{HC}) to the grayscale (i.e., I_{HG}) to reduce the workload and precision. Similar to the transformation in [2], [3], we use the following conversion to preserve the original RGB information:

$$C_l = \begin{cases} \frac{C_g}{12.92} & C_g \leq 0.04045 \\ (\frac{C_g+0.055}{1.055})^{2.4} & C_g > 0.04045 \end{cases} \tag{1}$$

$$I_{HG} = 0.2126R_l + 0.7152G_l + 0.0722B_l \tag{2}$$

where C_l denotes the linear RGB color values after gamma expansion; R_l, G_l, B_l are red, green, and blue channel linear values, respectively; C_g represents the gamma-compressed I_{HC} values, and I_{HG} is the calculated grayscale value.

Down-sampling: For selective attention, high-resolution details, such as the texture and the exact shape, may not be required to localize the salient pixels [2]. Therefore, we propose to further transform the grayscale image (i.e., I_{HG}) to a lower-resolution grayscale image (i.e., I_{LG}). Different from that in [2], [3], we use max down-sampling to select the maximum value as the down-sampled value. It has a lower cost in implementation than other interpolation-based methods.

B. Algorithm Optimization

The scaling-down of the input image enables the simplification of the neural network model for selective attention. At this step, we start from one of the commonly used segmentation models, U-Net [4], simplify its model structure under the constraint of comparable Intersection-over-Union (IoU), and further optimize its loss function to bias toward lower false negative selection (i.e., increase the coverage of salient pixels).

Model simplification: The original U-Net consists of a symmetrical encoder and decoder [4]. We use VGG13 as the backbone of the encoder [5], and reduce the number of channels and layers to build up our own U-Net-Lite. With the down-sampled grayscale image as the input, we first reduce the number of input channels of the first layer from three to one. Then the maximum number of channels across all layers is limited at 32. Finally, we experiment with fewer number of convolutional layers until the degradation of IoU.

Loss function: The result of selective attention can be evaluated by four categories: True Positive (TP) for correct selection of object pixels, False Positive (FP) for false selection of object pixels, False Negative (FN) for incorrect prediction of background pixels, and True Negative (TN) for correct prediction of background pixels. The quality of training depends on the balance of these four metrics in the loss function. To secure the selection of object pixels, we deliberately reduce FN by using the weighted cross entropy for training:

$$\mathcal{L} = -ylog(p) - \beta(1-y)log(1-p) \tag{3}$$

where $\mathcal{L}$ is loss, y is the ground truth label, p is the prediction and β is a parameter to adjust the importance of FN and TP, e.g., a smaller β emphasizes more on FN and TP.

C. Padding of Selected Output

The simplification of the input image and the selection algorithms inevitably cause information loss (e.g., loss of object pixels), even though we carefully monitor the model quality. Such loss results in the degradation of object detection using the selected RGB pixels, as compared to OD with the full RGB image. To compensate for that, we add a number of pixels from the background to the object, along the border of the selected region (i.e., padding), since the incorrect selections happen mostly around the boundary between the salient object and the background.

In summary, these three steps aim to achieve efficient and robust selection, with the minimum computation cost and overhead in output pixels. The optimization of parameters in each step are tuned on three datasets. The quality is monitored by IoU in salient object detection, the accuracy of object detection, and other metrics.

III. Experimental Results

Three popular datasets are used in this study: MSRA10K for salient object detection, COCO2017 [6] for image segmentation, and BDD100K [7] for autonomous driving. The optimization process is evaluated by multiple metrics, including mean absolute error (MAE), true positive rate (TPR), F-measure (e.g., $F_{0.3}$), maxF, and IoU that is defined as:

$$IoU = \frac{TP}{TP + FN + FP} \quad (4)$$

Besides these metrics for saliency detection, we further evaluate the quality of object detection after padding, using mean Average Precision (mAP) as the metric. The goal is to minimize the output pixel volume after selective attention; meanwhile, we should also achieve the minimal degradation of mAP in object detection, with selected pixels plus the padding.

A. Image Scaling for Selective Attention

Table I evaluates selection attention with down-sampled, grayscale images. The original U-Net is utilized to assess the quality with various metrics. As compared to the original RGB image, we observe that the dimension of the image can be reduced by 4×4 in BDD100K (i.e., 16× reduction in the number of pixels), without any degradation of selection attention. 3.125×3.125 and 5×5 down-sampling ratios work for the MSRA10K and COCO2017 datasets, respectively. In addition, our study is based on the grayscale image, which further reduces the data volume by 3× from the original RGB scale. This result matches the study of the human visual system in [2], with a higher accuracy by our machine learning model (i.e., U-Net). We adopt the down-sampled, grayscale images for further experiments on model optimization.

B. Model Optimization

With the reduction of input image size, we are able to simplify the U-Net structure into a lighter version, which we call "U-Net-Lite". We reduce the number of convolution layers to five in the decoder, and holistically decrease the number of channels in each layer. Table II presents the results on BDD100K. Based on the tradeoff between model simplicity and the IoU performance, we select the structure for U-Net-Lite in our study.

TABLE II: Simplification of model structure on BDD100K.

Selective Attention Model	Model Structure*	IoU
U-Net (Original)	(64, 128, 256, 512, 512)	0.796
Model 1	(64, 128, 128, 256, 256)	0.794
Model 2	(16, 32, 32, 64, 64)	0.739
U-Net-Lite	**(16, 32, 32, 32, 32)**	**0.742**
Model 3	(16, 16, 32, 32, 32)	0.710
Model 4	(16, 16, 16, 16, 32)	0.684

*Model structure denotes the number of channels in the encoder. While the original U-Net has two convolution layers in each decoder level, our models only have one layer.

Moreover, we adjust the weighted cross-entropy in the loss function of U-Net-Lite (Section II-B) to maximize the coverage of true object pixels (i.e., higher TP and lower FN). While a lower β in Eq. (3) improves the coverage, it also leads to more selected pixels as the overhead. We adopt $\beta = 0.001$ for BDD100K, and $\beta = 0.1$ for the other two datasets, which achieves a marginal loss in segmentation.

C. Accuracy of Object Detection

Finally, we evaluate the accuracy of object detection with the selected pixels only, with a black background (Fig. 1). Two OD algorithms are used, the vision transformer, Swin Transformer (SWIN) [8], and the convolution-based YOLO [9]. To compensate for the loss of fine object features through the selection process, we uniformly pad the selected region with an extra number of pixels. Fig. 3 presents the OD results under various amounts of padding. Based on the tradeoff between pixel overhead and mAP, we select 20 extra pixels for padding. With the padding to guarantee the OD accuracy, we only need to select 69.4%, 72.9%, 29.5% of the entire image area, for MSRA10K, COCO2017, and BDD100K, respectively, which promises higher processing efficiency.

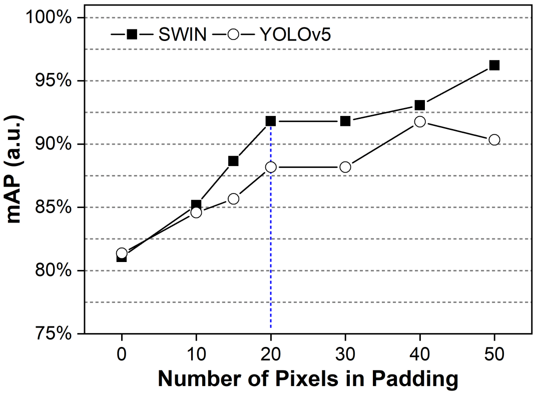


Fig. 3: Padding improves the mAP in object detection, for both SWIN and YOLO models on BDD100K. The mAP values are normalized to that from the full RGB image, for each OD algorithm.

TABLE I: Evaluation of down-sampled, grayscale images in selective attention. U-Net is used in this experiment.

Dataset	**MSRA10K**				**COCO 2017**			**BDD100K**		
Image	**Image size**	**$F_{0.3}$**	**MAE**	**maxF**	**Image size**	**IoU**	**TPR**	**Image size**	**IoU**	**TPR**
Original (RGB)*	400 × 400	0.865	0.064	0.876	640 × 640	0.625	0.732	1280 × 720	0.752	0.794
Down-sampled (Grayscale)	128 × 128	0.893	0.050	0.904	128 × 128	0.645	0.786	320 × 180	0.807	0.872
Down-sampling Ratio	3.125 × 3.125				5 × 5			4 × 4		

* For COCO 2017 and BDD100K, grayscale images are used as the original one for testing because of the big dataset size.

IV. Hardware Efficiency

The selective attention module effectively reduces the data volume from the sensor, which in turn reduces the sequential workload of the ADC to generate the output data. To implement the selective attention module on the same chip with the pixel array, we plan to adaptively control the ADC precision (i.e., low precision, down-sampled output for selection, and full precision for regular output) and pipeline the operation between selection and image output. In this section, we evaluate the improvement of computation and data-path efficiency to justify the feasibility for future implementation.

A. Computation Cost of Selective Attention

The computation cost of selection attention should be minimized, such that the addition of it in the sensor data-path does not restrict the throughput, as proposed in Fig. 1.

We develop two main techniques to address the efficiency issue: scaling down the input image and simplification of the network, as explained in previous sections. Table III summarizes the results of three datasets. Even after the padding, selective attention is able to reduce the number of pixels to 29.5-72.9% of the original image. The reduction will be more significant when the object ratio is smaller. Together with 3× reduction by the grayscale and 420× reduction in model size, the computation cost is reduced to 0.035-0.089%.

TABLE III: Improvement of the efficiency in selective attention.

Pixels (a.u.)*	**MSRA10K**	**COCO 2017**	**BDD100K**
Object pixels	22.2%	29.8%	10.7%
Selection	30.4%	59.7%	21.3%
Post padding	69.4%	72.9%	29.5%
Model size (a.u.)*	0.24%	0.24%	0.24%
Computation cost (a.u.)*	0.089%	0.035%	0.054%

* Pixels are normalized to the full image; model size is normalized to that of U-Net; computation cost is based on the number of operations, normalized to that of U-Net with the full RGB image.

B. Improvement of Sensor Performance

The reduction of output data volume proportionally reduces the cost of the sequential ADC operations. To quantitatively evaluate such benefit, we extract a scalable power consumption (P) and latency (T) model from design data of a CMOS image sensor [1], [10], as the functions of the number of pixels, pixel generation, ADC operation, and digital data readout :

$$P = N_R \cdot N_C \cdot (P_{PIX} + P_{ADC} + P_{RD}) \quad (5)$$

$$T = N_R \cdot (T_{PIX} + T_{ADC} + T_{RD}) \quad (6)$$

where N_R is the number of pixel rows, N_C is the number of pixel columns, and P_{PIX}, P_{ADC}, P_{RD}, T_{PIX}, T_{ADC}, T_{RD} denote the power consumption and latency for pixel generation, ADC and digital readout, respectively.

Based on this model, we extrapolate the latency and power saving with the selective attention module. Fig. 4 presents the result on BDD100K. With a larger image size, both power consumption and latency dramatically increase due to sequential ADC and data buffering. With a lower volume of output pixels, we expect to save the power consumption by 4.3× and latency by 3.4×. These results confirm the potential of selection attention for high-throughput imaging.

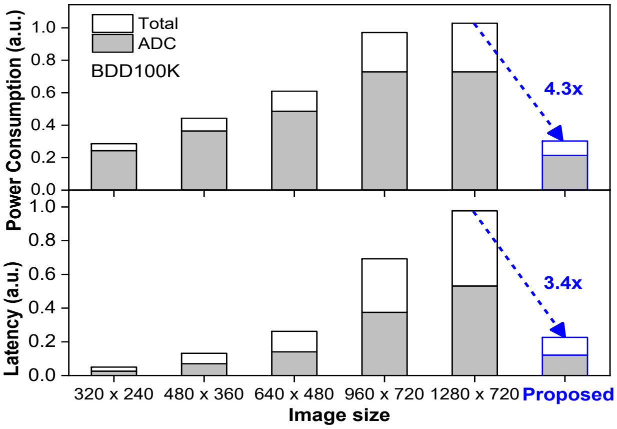


Fig. 4: The incorporation of selective attention effectively reduces the power consumption and latency in image processing.

V. Summary

By implementing the selective attention mechanism, our proposed work considerably reduces the data volume and improves the processing efficiency. With image down-sampling, model optimization, and reasonable padding, our method requires only 69.4%, 72.9% and 29.5% of the full image on OD task, for MSRA10K, COCO2017, and BDD100K, respectively. As a result, the method saves 4.3× power consumption and 3.4× latency of a CMOS image sensor on BDD100K.

References

[1] J. Choi, S. Park, J. Cho, and E. Yoon, "An energy/illumination-adaptive cmos image sensor with reconfigurable modes of operations," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 6, pp. 1438–1450, 2015.

[2] S. Yohanandan, A. Song, A. G. Dyer, A. Faragasso, S. Roy, and D. Tao, "Fast efficient object detection using selective attention," *arXiv e-prints*, pp. arXiv–1811, 2018.

[3] S. Yohanandan, A. Song, A. G. Dyer, and D. Tao, "Saliency preservation in low-resolution grayscale images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 235–251.

[4] 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.

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

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[7] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[10] J. Choi, J. Shin, D. Kang, and D.-S. Park, "Always-on cmos image sensor for mobile and wearable devices," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 1, pp. 130–140, 2015.