CLAWS: Clustering Assisted Weakly Supervised Learning with Normalcy Suppression for Anomalous Event Detection

Muhammad Zaigham Zaheer[1,2], Arif Mahmood[3], Marcella Astrid[1,2], and Seung-Ik Lee[1,2]

[1] University of Science and Technology, Daejeon, South Korea.
[2] Electronics and Telecommunication Research Institute, Daejeon, South Korea.
{mzz, marcella.astrid}@ust.ac.kr, the_silee@etri.re.kr
[3] Information Technology University, Ferozpur Road, Lahore, Pakistan
arif.mahmood@itu.edu.pk

Abstract. This is supplementary material for the paper, providing an extended overview on our proposed normalcy suppression mechanism for weakly supervised anomalous event detection. The document also provides visualizations of suppression output from our trained model, as well as the discussion on its behaviour. A comparison with conventional attention mechanism is also provided to highlight the main differences as well as the gains in performance by using the proposed suppression approach.

1 Normalcy Suppression Module (NSM)

The role of normalcy suppression module in our proposed CLAWS Net is two-fold:

- It learns to suppress the normal portions of an input batch.
- It holds back the backbone network from producing high anomaly scores for all input features in the presence of noisy labels.

The ideal behavior of NSM is to minimize the output values as much as possible if an input batch corresponds to a normal video.In the case that an input batch corresponds to an anomalous, the NSM minimizes its values along the normal segments. Another noticeable property that an NSM should posses is to learn abnormal behavior based on the temporal order of the events in an input batch.

In the following sections, we define several possible configurations of our suppression module and analyze the performances. This study verifies the effectiveness of our proposed architecture by presenting various comparisons of these configurations. Also, we present a detailed analysis highlighting significant differences between our approach and the existing attention mechanisms which have been widely used for various other problems.

Table 1: Area under the curve (AUC) comparison of various normalcy suppression (NS) configurations on UCF Crime dataset.

Normalcy suppression	**AUC %**
Element-wise NS (CLAWS Net)	83.03
Temporal NS	81.24
Features NS	77.95
CLAWS Net without any NSM	76.81

1.1 Which type of normalcy suppression is better?

Element-wise NS: Keeping in view the above mentioned properties of an ideal normalcy suppression module (NSM), in the CLAWS Net, we propose to utilize an NSM which calculates probabilities temporally in an element wise fashion (Fig. 1(a)). Our proposed technique, referred to as Element-wise NS in the rest of the supplementary material, provides more freedom to the NSMs in minimizing values if features belong to a normal portion of an input video, hence complementing the backbone network (BBN) to produce low anomaly scores.
Temporal NS: Another possible choice for normalcy suppression is to calculate suppression values temporally without element-wise application. It means one value is computed for each feature vector within a batch, as shown in Fig. 1(b).
Features NS: Furthermore, in order to provide a contrastive comparison of why learning temporal information is necessary for our proposed NSM, we also experiment with a normalcy suppression which doesn't consider temporal information. This setting, referred to as Features NS (shown in Fig. 1(c)), computes suppression values along each feature vector of the input batch. Detailed performance evaluation and discussion on each of these configurations is provided below:
Quantitative comparison: Table 1 summarizes the frame level AUC performance of these three configurations. It can be seen that the element-wise NS outperforms the other two counterparts with a noticeable margin. It is interesting to observe that the performance of temporal NS is relatively closer to the element-wise NS which is due to the reason that these two are quite similar in essence. Both learn to minimize the effects of normal features towards anomaly scoring however, the element-wise NS performs it with the additional freedom to operate at each dimension of the input feature vectors in a batch. Another important observation is that both of these suppression mechanisms utilizing temporal properties outperform the third scheme, features NS, significantly.
Qualitative comparison: Detailed visualizations of the output on an anomalous and a normal batch by the above mentioned suppression mechanisms are provided in Figs. 3 & 4. Among these, (a)-(c) are the output of NSM-1 and (d)-(f) are the output of NSM-2 (For NSM-1 and NSM-2, please see Fig. 1 of the main manuscript). It can be seen that temporal NS, similar to element-wise NS, learns to suppress the effects of normal features within an input batch. In contrast, features NS computes its values across one feature vector at a time,

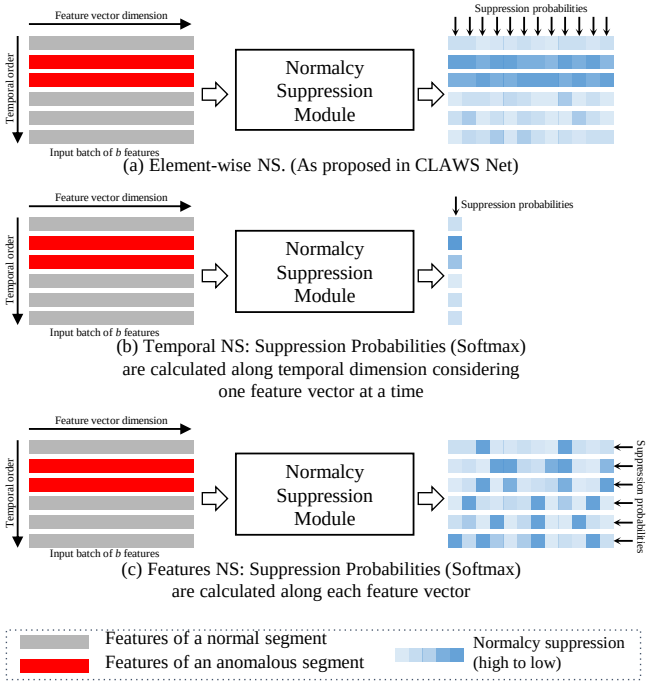


Fig. 1: Visualization of the three types of suppression mechanisms including element-wise NS (as proposed in the CLAWS Net), temporal NS and features NS.

hence it does not provide the desired effect of suppressing the features corresponding to normal events. Instead, it assists the backbone network to learn noisy labels for each individual input feature by essentially taking the form of a feature-dimension reduction mechanism. Therefore, AUC performance of the features NS is not much higher (1.14%) than the system without any suppression. This small difference can be attributed to the dimensionality reduction property which assists the backbone network while computing scores.

The element-wise NS, as proposed in the CLAWS Net, takes the advantage of both temporal and the feature level suppression. Due to its freedom to select various elements of features along the temporal order, such mechanism learns to minimize its values for the normal input as well as it also possesses the capability of learning to reduce the effect of non-contributing elements of a feature vector which helps the backbone network in producing better scores. Further visualization of the output from the NSM-1 and NSM-2 of our trained network on several normal and abnormal test batches are shown in Fig. 5.

Table 2: AUC comparison of our proposed multiplicative suppression mechanism with the residual suppression. The residual suppression does not produce the desired effect of minimizing anomaly scores for the normal regions, hence demonstrates low performance.

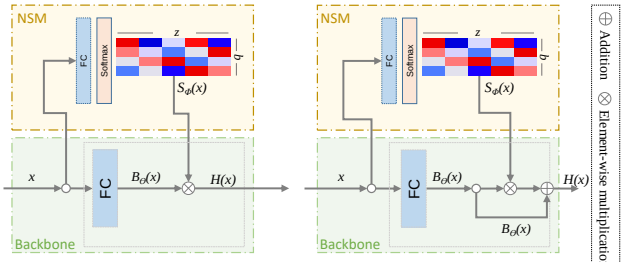
Normalcy suppression	**AUC %**
Multiplicative suppression (CLAWS Net)	83.03
Residual suppression	77.91
CLAWS Net without any NSM	76.81

1.2 Normalcy Suppression Vs. Attention?

Compared to attention [4, 1, 3, 6, 2, 5], attributed to the rare occurrence of anomalies, our proposed formulation approaches the problem in terms of suppressing certain features as opposed to highlighting [6, 2, 5]. In addition, we define the problem by relying on the special characteristics of the training labels in which we have *noise-free* annotations for normal videos and *noisy* annotations for anomalous videos. Therefore, given an input batch x, we calculate the suppressed results $H(x)$ by performing an element-wise multiplication $\otimes$ between NSM output $S_\phi(x)$ and backbone output $B_\theta(x)$ as:

$$H(x) = S_\phi(x) \otimes B_\theta(x), \tag{1}$$

where ϕ and θ represent the parameters of NSM and backbone, respectively (Fig. 2(a)). For the conventional attention, such multiplication have been reported to produce the undesirable effect of dissipating features inside a model [5]. It is because attention computes probabilities which, when multiplied directly with the features, can reduce the values significantly. Therefore, various attention mechanisms are based on residual connections in which attention-applied features are added back to the original features [4, 5]. In order to experiment with



(a) Multiplicative Suppression (CLAWS Net)

(b) Residual Suppression

Fig. 2: (a) Multiplicative suppression, as proposed in CLAWS Net. (b) Residual suppression devised based on the attention mechanism proposed in [5].

such *residual-attention-like* mechanism in our normalcy suppression modules, we design a formulation in which suppressed results $H(x)$ are calculated as:

$$H(x) = B_\theta(x) \oplus S_\phi(x) \otimes B_\theta(x), \tag{2}$$

where $\oplus$ is an addition operation. We refer to this scheme as residual suppression (Fig. 2(b)).

Table 2 shows a comparison of the results between residual suppression and our multiplicative approach, demonstrating importance of the latter. Compared with the model without any NSM, the residual suppression shows only a slight improvement of 1.1% whereas the proposed multiplicative suppression shows an improvement of 6.22%.

We believe that the superiority of multiplicative suppression is based on two factors: first, because we approach the problem as features suppression instead of highlighting, our proposed method exploits the property of multiplicative suppression to reduce the impact of features inside the network during a forward pass. It means, in case of a normal input batch (where noise-free labels are available), the suppression module learns to minimize its output, hence dissipating the features and helping the backbone network to produce low anomaly scores. On the other hand, in the case of an input with anomalous features having noisy labels, even if the backbone network tries to produce high anomaly scores on all these features, it gets limited by the NSM which cannot produce high values across a whole batch. Hence, to reduce the overall training loss, our multiplicative configuration forces the NSM towards learning to suppress only the normal features, consequently assisting the backbone network to produce high anomaly scores on anomalous portions of an input batch.

Second, given the forward pass in Equation 1, gradients of the multiplicative suppression with respect to the backbone parameters are given as:

$$\frac{\partial H(x)}{\partial \theta} = S_\phi(x) \frac{\partial B_\theta(x)}{\partial \theta} \tag{3}$$

Whereas, based on the forward pass in Equation 2, gradients of the residual suppression with respect to the backbone parameters can be computed as:

$$\frac{\partial H(x)}{\partial \theta} = \frac{\partial B_\theta(x)}{\partial \theta} + S_\phi(x) \frac{\partial B_\theta(x)}{\partial \theta} \tag{4}$$

In the case of multiplicative suppression (Equation 3), it can be seen that the $S_\phi(x)$, as a normalcy suppressor, prevents wrong gradients from flowing into the backbone when the network encounters noisy labels in anomalous videos (some segments are normal). In the residual suppression case, since the gradients $\frac{\partial B_\theta(x)}{\partial \theta}$ in Equation 4 are not suppressed by $S_\phi(x)$, the loss from noisy labels can flow into the backbone network, consequently degrading its performance. This partial suppression is particularly the reason why residual suppression only achieves a slightly better performance than the network without any NSM (Table 2). On the other hand, our proposed multiplicative approach, which suppresses the gradients to minimize the impact of noisy labels, achieves significant performance gains (See Figure 6).

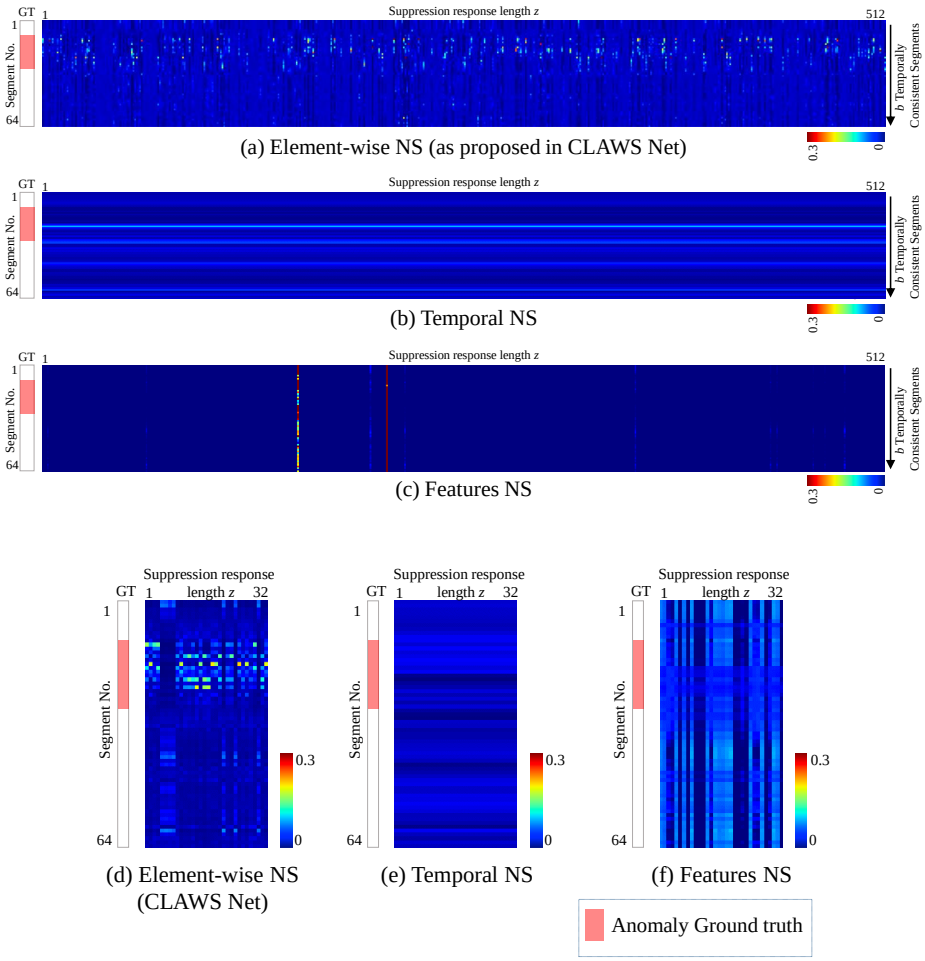


Fig. 3: Output (S_ϕ) comparison of various NSM-1 (a)-(c) and NSM-2 (d)-(f) configurations. z is the output dimension of FC layer corresponding to each NSM and b is the input batch size. Actual temporal NS output by both modules is of size $1 \times b$ however, it is repeated to create a $z\times$ b vector for better and consistent visualization.

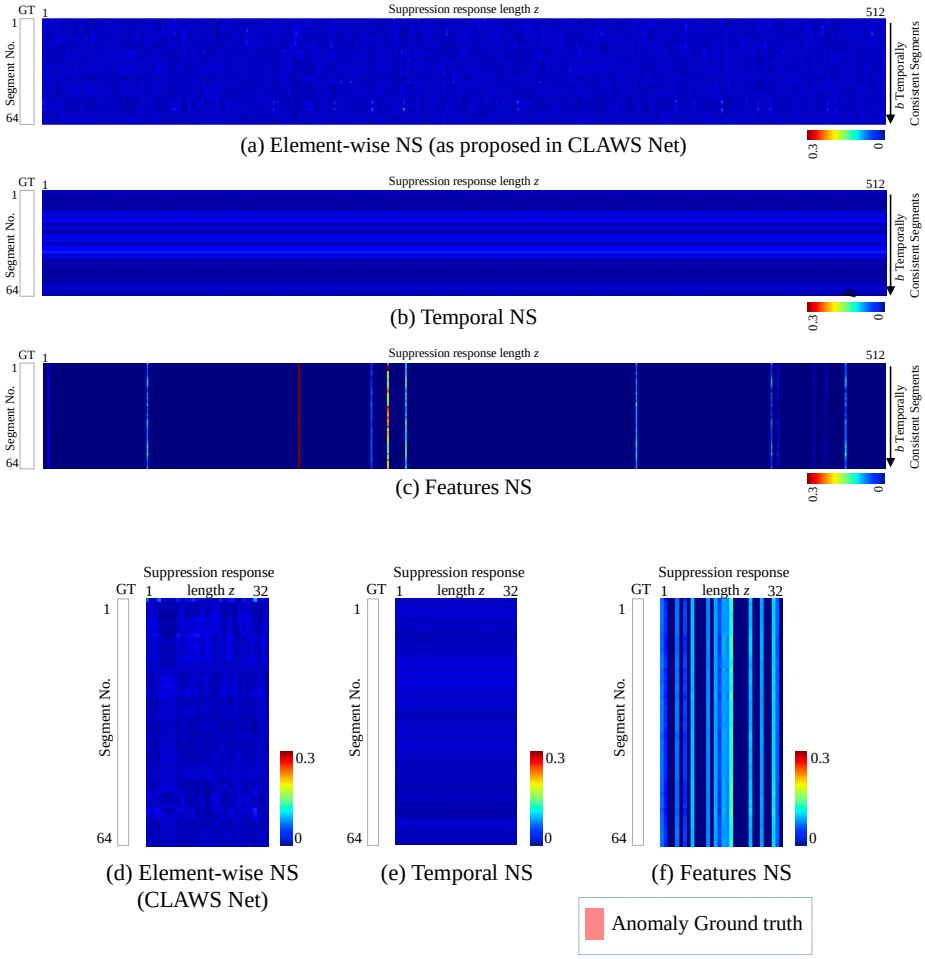


Fig. 4: Output (S_ϕ) comparison of various NSM-1 (a)-(c) and NSM-2 (d)-(f) configurations on a normal batch.

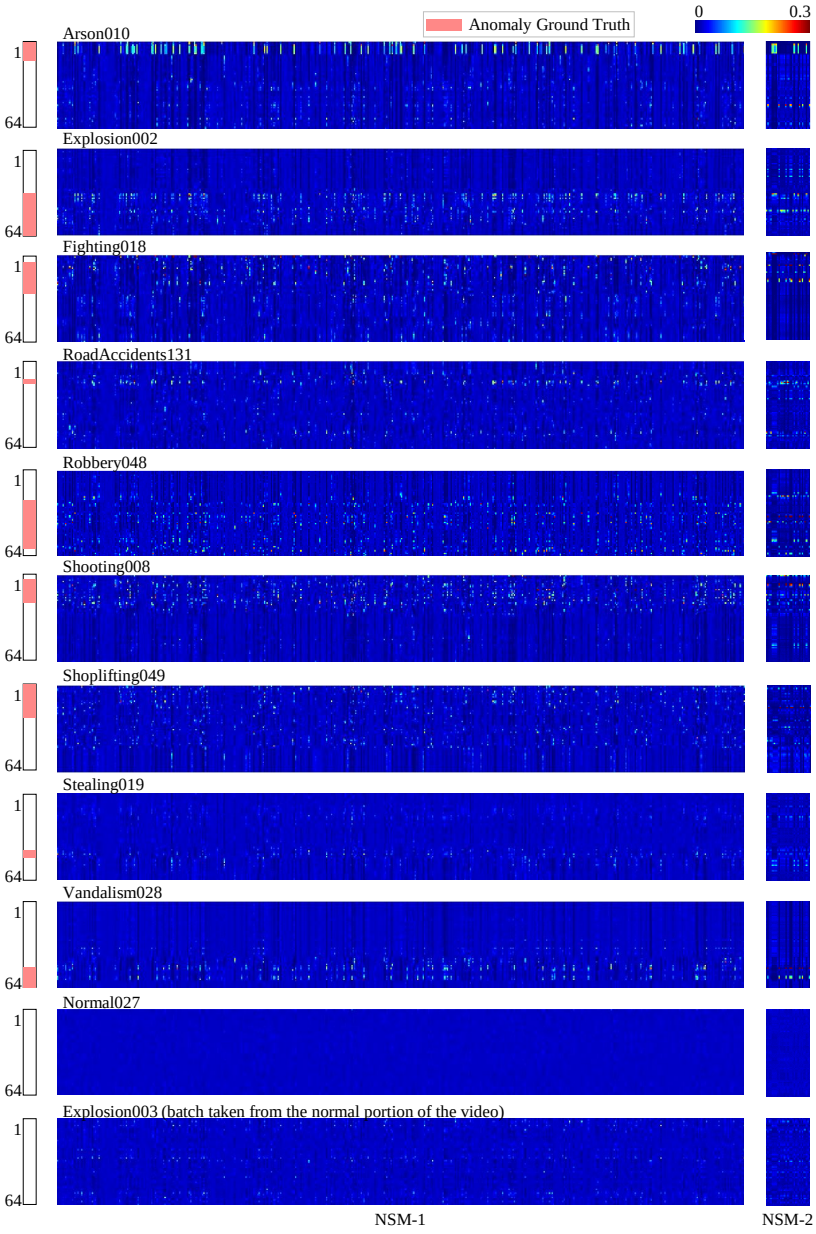


Fig. 5: Visualization of the softmax output S_ϕ of NSM-1 and NSM-2 under our proposed element-wise multiplicative configuration, on several batches taken from normal and anomalous test videos.

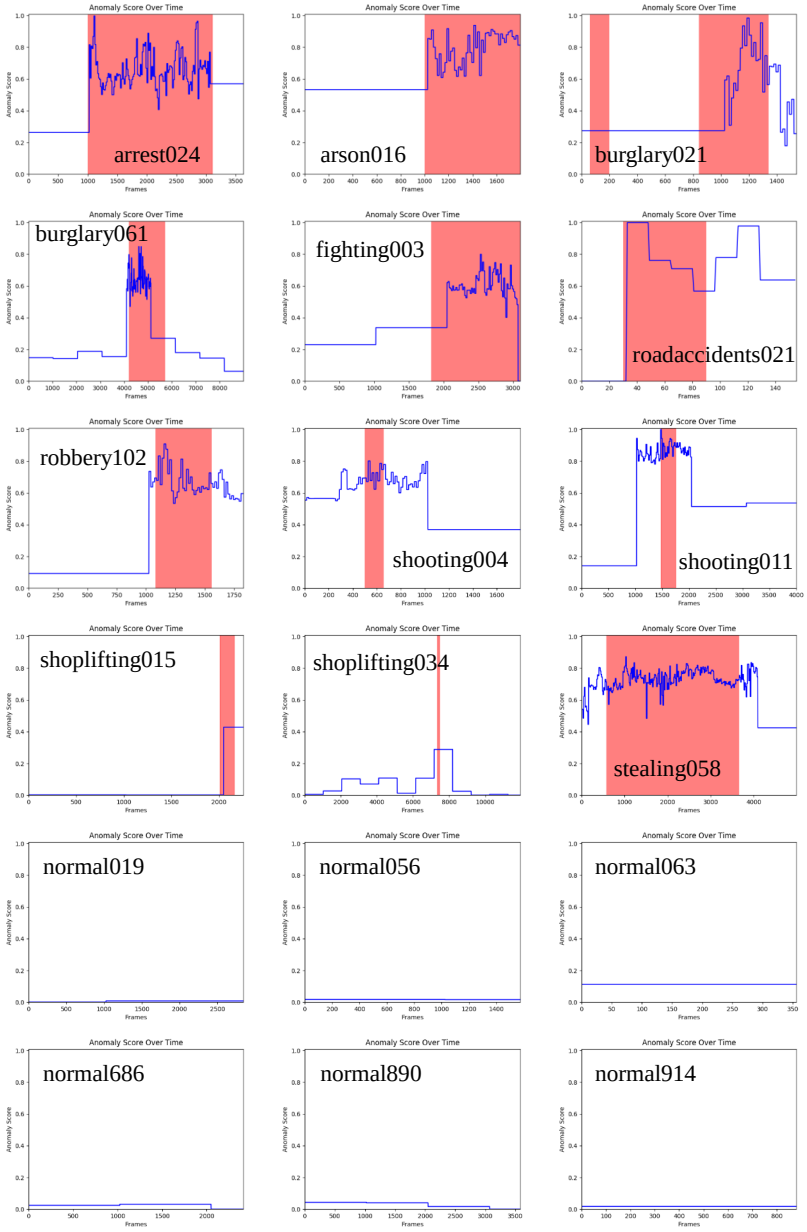


Fig. 6: Anomaly scores of the proposed CLAWS Net on different videos from UCF Crime Dataset. Note that in some videos, actual anomalous frames than the annotated ones as the annotation is only for the event itself. For Example, in shooting011 video, abnormal situation starts around 1000 and continues much later than the annotated window which only contains the shooting event.

References

1. Chen, X., Xu, C., Yang, X., Tao, D.: Attention-gan for object transfiguration in wild images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 164–180 (2018)
2. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
3. Shen, Y., Ni, B., Li, Z., Zhuang, N.: Egocentric activity prediction via event modulated attention. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 197–212 (2018)
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
5. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3156–3164 (2017)
6. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)