# Light Annotation Fine Segmentation: Histology Image Segmentation based on VGG Fusion with Global Normalisation CAM

No Author Given

No Institute Given

**Abstract.** Deep learning has been widely used to segment tumour regions in stained histopathology images. However, precise annotations are expensive and labour-consuming. To reduce the manual annotation workload, we propose a light annotation-based fine-level segmentation approach for histology images based on a VGG-based Fusion network with Global Normalisation CAM. The experts are only required to provide a rough segmentation annotation on the images, and then accurate fine-level segmentation boundaries can be produced using this method. To validate the proposed approach, three histopathology datasets with rough and fine quality segmentation annotation are built. The fine quality labels are used only as ground truth in evaluation. The VFGN-CAM method includes three main components: an annotation enhancement to boost boundary accuracy and model generalisability; a VGG Fusion module that integrates multi-scale tumour features; and a Global Normalisation CAM module that combines local and global gradient information of tumour regions. Our VGG fusion and Global Normalisation CAM outperform the existing methods with a Dice of 84.188%. The final improvement for our proposed methods against the original rough labels is around 22.8%. The codes are released at:xxx.

**Keywords:** Segmentation · Tumor · Annotation improvement.

## 1 INTRODUCTION

Cancer is the most deadly illness in the world due to it capability to generate distant metastases. Digital pathology scanners can provide whole slide image (WSI) with a very high resolution (e.g. 80000 × 150000). Stained WSIs are the gold standard for diagnosing cancer and predicting tumour reoccurrence and other potential deterioration. However, manual tumour segmentation is expensive and time consuming for pathologists. Therefore, the automatic segmentation method is essential for efficient and accurate tumour classification on WSIs.

Several challenges exist in labelling tumour regions. Compared with carefully hand-drawn boundaries that describe exactly the tissue structures, pathologists tend to mark tumour parts with rough smooth curves in practice, which will save substantial time for marking. The rough markings are informative but could be misleading in model training to some extent since these boundaries include inevitable. In addition, Intrinsic variance in the tumour, the variances between patients, and technical variances generated in slicing, staining, and scanning cause inaccurate manual tumour segmentation annotation.

To relieve the dependence on segmentation annotation, many methods have been proposed for weakly supervised segmentation (WSS) purposed including image-level [2], scribble-based [4], point-based [10] and iterative based methods [11]. Class activation mapping(CAM) [14] with global average pooling (GAP) is a simple yet effective technique for weakly-supervised segmentation. Wang et.al propose consistency regularization on predicted CAMs from various transformed images to provide self-supervision [12]. Durand et.al jointly aim at aligning image regions for gaining spatial invariance and learning strongly localized features[1]. Similar to CAM, adversarial erasing is an efficient way to represent objects partly according to the peak responses of classes [7, 3]. Recently, Multi-branch WSS methods are proposed to segment objects more preciously such as complex attention modules [5], cross-image mining [8] and siamese networks [12]. Most of the existing methods are designed by combining a series of modules including training classifiers, visualizing activation maps and re-training segmentation networks.

Inspired by the efficacy of WSS methods, to reduce the dependency on accurate tumour annotations and minimise pathologists' workload of marking on WSIs, we build two kinds of annotations including fine quality labels(F-label), and poor quality labels(P-label). The purpose of this work is to exploit a large amount of P-labels for training and use a few F-labels for testing.

In this work, we propose a VGG-based fusion network with global normalization CAM (VFGN-CAM). Our contributions are threefold. (1)We refine the P-labels based on k-means clustering and soft label. This annotation refinement process ensures the annotation accuracy of tumour boundaries and enhances the subsequent model generalizability. (2) A VGG-based fusion module (VF-Net) is proposed based on VGG16. Multi-scale features are fused together for patch-based tumour classification. (3) A global normalization CAM (GN-CAM) module is presented to integrate gradient information both in the global whole image and local patches, to acquire the position features in distinguishing the tumour and background.

## 2 METHODS

The overall framework is shown in fig. 1. The rough annotation is first processed in the annotation enhancement (AE) module which employs the k-means clustering algorithm to improve the annotation for network training with soft labels. Then we propose a VGG-based fusion classification network based on VGG16 to exploit multi-scale features for fine-grained patch-based classification. After network training, the information of the last convolution layer of the network is extracted and calculated by a GN-CAM which combines the normal CAM result and a global normalization CAM result by specific weights. At last, the output heat-map for each patch is embedded into the whole slide image and then

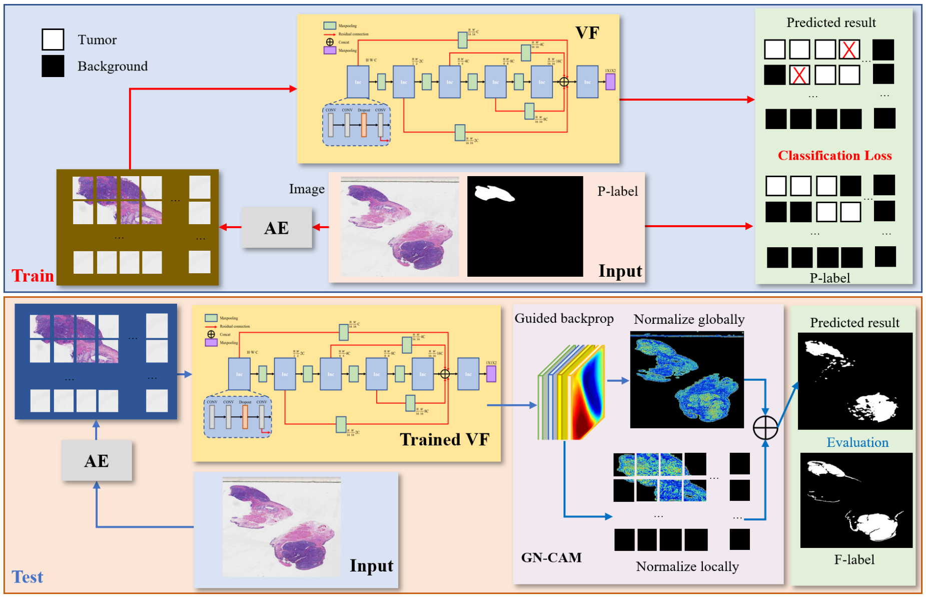

**Fig. 1.** The structure of VFGN-CAM. VF-Net are trained with the data pre-processed by annotation enhancement (AE), GN-CAM are used in test stage to acquire more accurate result.

goes through a convolutional CRFs-based noise eliminator (NE) to smooth the boundary of the generated annotation and eliminate the noise.

### 2.1 Annotation enhancement

To reduce the inaccuracy of the rough annotations, we propose an annotation enhancement (AE) module based on k-means clustering, and a soft label modifier to refine cancer annotations, especially in tumor marginal regions. The rough annotation $Y_0$ marked by experts delineates non-tumour regions from tumour regions. $K$-means clustering cluster together pixels with similar features together to create label $Y_1$ that is a refined version of the original tumour boundaries in $Y_0$. The intersection point set $\hat{Y} = Y_1 \cap Y_0$ is considered as the refined ground truth. In addition, to ensure a highly efficient model training and boost the model generalizability, patch-based soft labels [13] are generated by a sliding window with the size of (512,512) as shown in algorithm 1.

As several errors still exist in tumour boundaries and especially in isolated tiny tumour regions, the errors will be propagated if we directly train models according to pixel-wise refined annotations $Y$. In this case, we design a patch-based classification model with GN-CAM supervised by soft labels, to reduce the error effects around tumour boundaries.

**Algorithm 1** Generate soft label on the refined whole slide annotation $\hat{Y}$

```
repeat
    Assuming the centre of the sliding window is $(p, q)$, the proportion of tumour areas $f_{p,q}$ is calculated on the adjusted annotation $Y$, where $\mathbb{I}$ is a binary function to discriminate whether one region in the sliding window belongs to tumor.
        $$f_{p,q} = \sum_{i=p-256}^{p+256} \sum_{j=p-256}^{p+256} \frac{\mathbb{I}(tumor, i, j)}{512^2}$$
    Patches $X$ extracted by the sliding window are stored and marked with soft label $Y$
        $$Y = \begin{cases} 1-\sigma, & f_{p,q} > \theta \\ \sigma, & others \end{cases}$$
until moving the sliding window across all refined annotation boundaries.
```

### 2.2 VF Classification Network

Convolution-based design is capable of inferring accurate local features (texture, boundary and greyscale) with few features. VGG is a universal backbone for image feature extraction, which has been widely applied to classification, detection and segmentation tasks for medical images. [6]

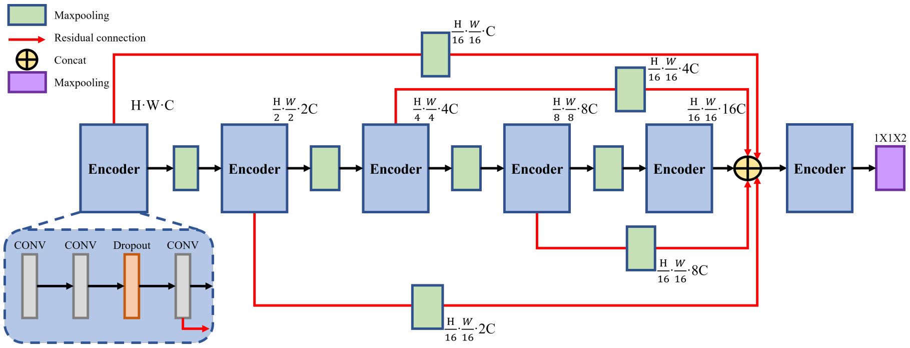
Maxpooling
Residual connection
Concat
Maxpooling
H·W·C
$\frac{H}{2}\cdot\frac{W}{2}\cdot 2C$
$\frac{H}{4}\cdot\frac{W}{4}\cdot 4C$
$\frac{H}{8}\cdot\frac{W}{8}\cdot 8C$
$\frac{H}{16}\cdot\frac{W}{16}\cdot 16C$
$\frac{H}{16}\cdot\frac{W}{16}\cdot C$
$\frac{H}{16}\cdot\frac{W}{16}\cdot 4C$
$\frac{H}{16}\cdot\frac{W}{16}\cdot 8C$
$\frac{H}{16}\cdot\frac{W}{16}\cdot 2C$
Encoder
Encoder
Encoder
Encoder
Encoder
Encoder
1X1X2
CONV
CONV
Dropout
CONV

**Fig. 2.** The network structure of VF.

Convolution-based design is capable of inferring accurate local features (texture, boundary and greyscale) with few features. VGG is a universal backbone for image feature extraction, which has been widely applied to classification, detection and segmentation tasks for medical images. [6]Inspired by the effectiveness and lightweight of VGG 16, we apply VGG 16 as our base model for tumour classification, increasing a series of residual connections among convolutions and design a multi-scale fusion module to ensure accurate classification of tiny tumours. Fig. 2 illustrates the detailed framework of our VF method. Each block contains three convolution layers with residual connections to ensure the stability of network back-propagation. One dropout layer is inserted after the second convolution layer to increase the network generalization. In addition, a multi-scale feature fusion module is presented to fuse feature maps generated by all Max-pooling layers. All feature maps are resampled to the same size as the feature map from the final Max-pooling. These features are concatenated together and pass through a convolution layer.

### 2.3 Global Normalised Class Activation Mapping

Global Normalised Class Activation Mapping (GN-CAM) is a new way of combining feature maps using the gradient signal. Inspired by assigning an importance factor to each neuron by the gradient of G-CAM, this paper proposes a global normalized CAM that extracts the guided gradient features $R^l$ flowing out the last convolution layer. The $l^{th}$ and $(l+1)^{th}$ layers are the last two layers of the VF. Denote the $i^{th}$ feature map of the $(l+1)^{th}$ layer as $f_i^{l+1}$, the $i^{th}$ gradient map of the $l+1$ layer as $R_i^{l+1}$ and the output map is $f^{\text{out}}$ . The gradient map of $l+1$ layer is calculated by

$$f_i^{l+1} = \text{relu}\left(f_i^l\right) = \max\left(f_i^l, 0\right), \tag{1}$$

$$R_i^{l+1} = \frac{\partial f^{\text{out}}}{\partial f_i^{l+1}}. \tag{2}$$

The guided gradient map of the $l$ layer $R^l$ is calculated by:

$$R_i^l = \left(f_i^l > 0\right) \cdot \left(R_i^{l+1} > 0\right) \cdot R_i^{l+1}. \tag{3}$$

All guided gradient maps $R$ from the same WSI are stored in a queue $Q_1$. Then we normalise each map in $Q_1$ with the global mean and standard deviation. These processed maps $R^{'}$ are stored in a new queue $Q_2$. Also, assuming $(w, h)$ is the spatial position of a gradient map $R$, every pixel $R_{i,w,h}^l$, $w \in W, h \in H$ is normalized locally and recalculated by:

$$\mu_i^l = \frac{\sum_{w=1}^{W} \sum_{h=1}^{H} R_{i,w,h}^l}{WH} \tag{4}$$

$$s_i^l = \sqrt{\frac{\sum_{w=1}^{W} \sum_{h=1}^{H} \left(R_{i,w,h}^l - \mu_i^l\right)^2}{(WH)^2}} \tag{5}$$

$$R_i^{l^{''}} = \frac{R_i^l - \mu_i^l}{s_i^l} \tag{6}$$

The locally normalised maps $R^{''}$ from the same WSI are stored in a queue $Q_3$. Two normalised gradient maps $R_i^{'}$ and $R_i^{''}$ from $Q_2$ and $Q_3$ are added together by order. The final segmentation results $M$ is calculated by

$$M_i = \frac{R_i^{'} + R_i^{''}}{2}. \tag{7}$$

### 2.4 Noise Eliminator

After model training and the CAM process, the generated masks are more accurate, but some noise remains. This is because the tumour regions are calculated in the region of 512 × 512 pixels, so the predicted boundary is very sharp. Also, the isolated tumour cells and fine details in boundaries are often not considered in human manual labelling. Thus, to resemble manual segmentation, a post-processing step using convolutional CRFs[9] is developed to ensure the segmentation boundary is medically relevant. The output of convolutional CRFs has more smooth boundaries and less noise, especially inside the tumour region.

## 3 Experiment and Result

### 3.1 Data introduction and training details

We train and evaluate our framework on three tumour datasets including basal cell cancer (BCC), squamous papilloma (SP), and seborrheic keratosis cancer (SKC) datasets. All three datasets are skin cancer data. The common challenge of a skin cancer dataset is that the boundary of the tumour region is difficult to identify. So the rough annotations on this kind of dataset will further influence the performance of the segmentation network. In the training process, to reduce the requirement for memory and accelerate the training process, we cut all the whole slide tumour images into patches the size of (512,512). The Adam optimizer is used with a learning rate of 0.0001 and a step learning scheduler with step size=60 and $\gamma = 0.95$. The loss function is cross entropy. It takes around 10 hours to train our model and test the results on a NVIDIA RTX 3080 GPU.

### 3.2 Evaluation and results

There are five widely used measurement parameters used in this evaluation: sensitivity, specificity, accuracy, IOU and dice coefficient. IOU and dice coefficient are widely used to comprehensively evaluate the segmentation performance of the target network. First, we discuss the performance of annotation enhancement and our proposed VF network. Tab. 1 demonstrates the evaluation about the annotation enhancement (AE) and network. All network results with annotation enhancement have a better performance against the same network without annotation enhancement. It is reasonable to believe that our proposed AE has

**Table 1.** Network results with or without annotation enhancement.

| | Sensitivity(%) | Specificity(%) | Accuracy(%) | IOU(%) | Dice(%) |
|---|---|---|---|---|---|
| VGG | 70.480 | 97.735 | 93.997 | 61.077 | 75.358 |
| VGG-AE | 74.592 | **97.972** | 94.819 | 64.602 | 78.032 |
| VF | 75.267 | 97.869 | 94.915 | 64.522 | 77.766 |
| **VF-AE (ours)** | **80.947** | 97.703 | **95.813** | **68.538** | **81.090** |

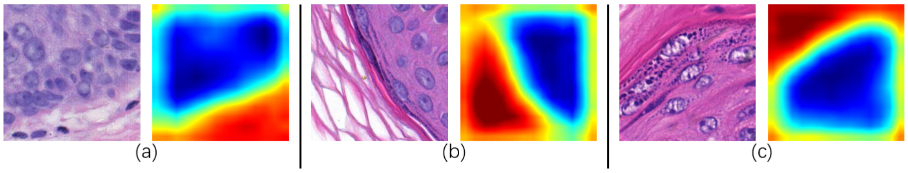

**Fig. 3.** Output patches of GN-CAM for three dataset: (a) basal cell cancer (BCC); (b) squamous papilloma (SP); (c) seborrheic keratosis cancer (SKC).

a non-negligible effect in a weakly trained segmentation task, especially in the poor quality annotation situation. Also, our proposed VF network has an average of 3 percent improvement against the VGG network. Which proves the VFGN-CAM structure is more suitable for this work.

**Table 2.** Results of different CAM based on annotation enhancement.

| | | Sensitivity(%) | Specificity(%) | Accuracy(%) | IOU(%) | Dice(%) |
|---|---|---|---|---|---|---|
| VGG | CAM | 69.386 | **98.212** | 94.010 | 61.574 | 75.697 |
| | **GN-CAM** | 79.799 | 97.732 | 95.628 | 67.629 | 80.367 |
| VF | CAM | 78.890 | 97.790 | 95.597 | 67.569 | 80.316 |
| | **GN-CAM** | **83.003** | 97.615 | **96.029** | **69.507** | **81.865** |

CAM aims to extract the information in the convolution layer to explain the results of network training. In this work, information is extracted from the last convolution layer of the trained network. We propose a global normalization CAM in tumour segmentation task, the result of this CAM of patches is shown in fig. 3. We explore the differences between two kinds of CAM, the segmentation result are shown in tab. 2. Using annotation enhancement or not, our proposed GN-CAM achieve better performance in all parameter against the normal CAM.

**Table 3.** Results of noise eliminator under annotation enhancement and GN-CAM.

| | Sensitivity(%) | Specificity(%) | Accuracy(%) | IOU(%) | Dice(%) |
|---|---|---|---|---|---|
| VGG | 79.799 | 97.732 | 95.628 | 67.629 | 80.367 |
| VGG-NE | 81.961 | **98.106** | 96.187 | 70.837 | 82.626 |
| VF | 83.003 | 97.615 | 96.029 | 69.507 | 81.865 |
| **VF-NE (ours)** | **85.461** | 98.000 | **96.600** | **72.963** | **84.188** |

Two ablation studies are presented on the tab. v and the tab.4. As shown in tab. 3, average segmentation improvements on three datasets are around 3 % with the proposed noise eliminator, regardless of whether we use VGG or VF. Tab. 4 shows the comparison of the P-label and the predicted segmentation evaluated against the F-label. All the machine learning results are generated under the annotation enhancement and noise eliminator by GN-CAM. The generated mask presents a great improvement against the P-label. Among all the

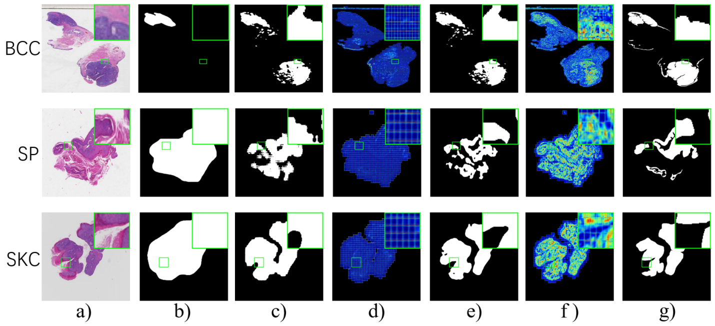

**Fig. 4.** Some examples of CAM output heat map and tumor segmentation results: (a) original image; (b) Poor quality label (P-label); (c) annotation for VGG; (d) heat map for VGG; (e) annotation for VF; (f) heat map for VF; (g) Fine quality label (F-label).

results, our proposed VF and GN-CAM with annotation enhancement and noise eliminator achieve the best result. Fig. 4 shows the result of CAM heat map output and final annotations after noise eliminator. Compare to poor quality labels (P-label), our method generates more accurate and detailed boundaries. The proposed VF methods lead to an improvement of 22.846% in Dice coefficient against the P-label, which proves the success of the VFGN-CAM framework. Also, the heat map generated by GN-CAM shows a significant visual correlation to the tumour area, meaning that the output segmentation can be used for medical assessment tasks which have roughly annotated training sets.

**Table 4.** Segmentation results compared to original P-label by F-label as ground truth.

| | | Sensitivity(%) | Specificity(%) | Accuracy(%) | IOU(%) | Dice(%) |
|---|---|---|---|---|---|---|
| | P-label | 57.601 | 97.898 | 87.972 | 45.048 | 61.342 |
| CAM | VGG | 70.821 | **98.907** | 94.473 | 65.037 | 78.600 |
| | VF | 80.569 | 98.329 | 96.162 | 71.001 | 82.759 |
| **GN-CAM** | VGG | 81.961 | 98.106 | 96.187 | 70.837 | 82.626 |
| | **VF** | **85.461** | 98.000 | **96.600** | **72.963** | **84.188** |

## 4 Conclusion

In this paper, we explore a new patch-based tumour segmentation method supervised by rough annotations called VFGN-CAM. More specifically, an annotation enhancement is presented to progressively refine the annotations, which ensures accuracy in tumour boundary shape. A VF net is used to classify the patches. We also propose a GN-CAM to integrate global and local gradient information of tumour regions. Experiments on three tumour datasets show the effectiveness and superiority of our model. In future, more weakly supervised work will be proposed based on our P-labels and our method.

## References

1. Durand, T., Mordan, T., Thome, N., Cord, M.: Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 642–651 (2017)
2. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7014–7023 (2018)
3. Kweon, H., Yoon, S.H., Kim, H., Park, D., Yoon, K.J.: Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6994–7003 (2021)
4. Luo, X., Hu, M., Liao, W., Zhai, S., Song, T., Wang, G., Zhang, S.: Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision. arXiv preprint arXiv:2203.02106 (2022)
5. Qin, J., Wu, J., Xiao, X., Li, L., Wang, X.: Activation modulation and recalibration scheme for weakly supervised semantic segmentation. arXiv preprint arXiv:2112.08996 (2021)
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale visual recognition. arXiv preprint arXiv:1409.1556 (2014)
7. Stammes, E., Runia, T.F., Hofmann, M., Ghafoorian, M.: Find it if you can: end-to-end adversarial erasing for weakly-supervised semantic segmentation. In: Thirteenth International Conference on Digital Image Processing (ICDIP 2021). vol. 11878, p. 1187825. International Society for Optics and Photonics (2021)
8. Sun, G., Wang, W., Dai, J., Van Gool, L.: Mining cross-image semantics for weakly supervised semantic segmentation. In: European conference on computer vision. pp. 347–365. Springer (2020)
9. Teichmann, M.T., Cipolla, R.: Convolutional crfs for semantic segmentation. arXiv preprint arXiv:1805.04777 (2018)
10. Tian, K., Zhang, J., Shen, H., Yan, K., Dong, P., Yao, J., Che, S., Luo, P., Han, X.: Weakly-supervised nucleus segmentation based on point annotations: A coarse-to-fine self-stimulated learning strategy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 299–308. Springer (2020)
11. Wang, X., You, S., Li, X., Ma, H.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1354–1362 (2018)
12. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12275–12284 (2020)
13. Yuan, L., Tay, F.E., Li, G., Wang, T., Feng, J.: Revisiting knowledge distillation via label smoothing regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3903–3911 (2020)
14. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)