Pattern Recognition 137 (2023) 109298

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

A multi-strategy contrastive learning framework for weakly supervised semantic segmentation

Kunhao Yuan[a], Gerald Schaefer[a], Yu-Kun Lai[b], Yifan Wang[a], Xiyao Liu[c], Lin Guan[a], Hui Fang[a,*]

[a] Loughborough University, UK
[b] Cardiff University, UK
[c] Central South University, China

ARTICLE INFO

Article history:
Received 16 June 2022
Revised 15 December 2022
Accepted 31 December 2022
Available online 2 January 2023

Keywords:
Weakly supervised learning
Representation learning
Contrastive learning
Semantic segmentation

ABSTRACT

Weakly supervised semantic segmentation (WSSS) has gained significant popularity as it relies only on weak labels such as image level annotations rather than the pixel level annotations required by supervised semantic segmentation (SSS) methods. Despite drastically reduced annotation costs, typical feature representations learned from WSSS are only representative of some salient parts of objects and less reliable compared to SSS due to the weak guidance during training. In this paper, we propose a novel Multi-Strategy Contrastive Learning (MuSCLe) framework to obtain enhanced feature representations and improve WSSS performance by exploiting similarity and dissimilarity of contrastive sample pairs at image, region, pixel and object boundary levels. Extensive experiments demonstrate the effectiveness of our method and show that MuSCLe outperforms current state-of-the-art methods on the widely used PASCAL VOC 2012 dataset.

1. Introduction

Deep learning (DL)-based semantic segmentation is a well-established computer vision task that has been widely used in various pattern recognition applications, *e.g.*, autonomous driving [1], medical imaging [2] and satellite image analysis [3]. However, generalising DL models to wider applications is difficult since they require high-quality pixel-level annotations that are costly to obtain. One way to address this issue focuses on how to distill and transfer knowledge from either pre-trained networks [4–6] or intrinsic data connections [7], while another is to develop weakly supervised methods that use inexpensive weak labels to achieve fine-grained tasks. In particular, weakly supervised semantic segmentation (WSSS) approaches use coarse labels, typically image-level annotations, to achieve pixel-wise semantic segmentation [8–10].

Since the introduction of class activation maps (CAMs) [8], a lot of effort has focused on improving DL-based WSSS. One type of approach is to introduce extra cues, such as saliency maps [11,12], scribbles [13], or bounding boxes of objects [14], to yield stronger constraints to supervise the learning. Another group of methods utilise either global context correlations [10] or local pixel correlations [9,15] to enhance image level WSSS.

Despite continuously improved performance of image level WSSS methods, most approaches focus on maximising inter-class variations of feature representations belonging to different classes [9,11,16]. Consequently, their segmentation results only identify the most salient parts of objects since these are sufficient to optimise their defined loss functions. Although some recent work explores pixel correlations [10] or sub-category clustering in feature space [17] to enhance object representations, the distinctiveness between contrastive object sample pairs is still under-explored.

Inspired by recent success of contrastive learning frameworks [18–20], in this paper, we propose a multi-level contrastive learning strategy to further enhance both the feature representation and mapping function of image-level WSSS by embedding contrastive learning metrics at image, region, pixel and object boundary levels. As illustrated in Fig. 1, we use image level contrast between different objects as well as region and pixel level contrast extracted from overlapping regions of the same objects to improve the object feature representation. We further propose a boundary-based contrast extracted from the pseudo labels to enhance our decoder for better segmentation. Overall, the contribu-

* Corresponding author.
E-mail address: H.Fang@lboro.ac.uk (H. Fang).

https://doi.org/10.1016/j.patcog.2022.109298

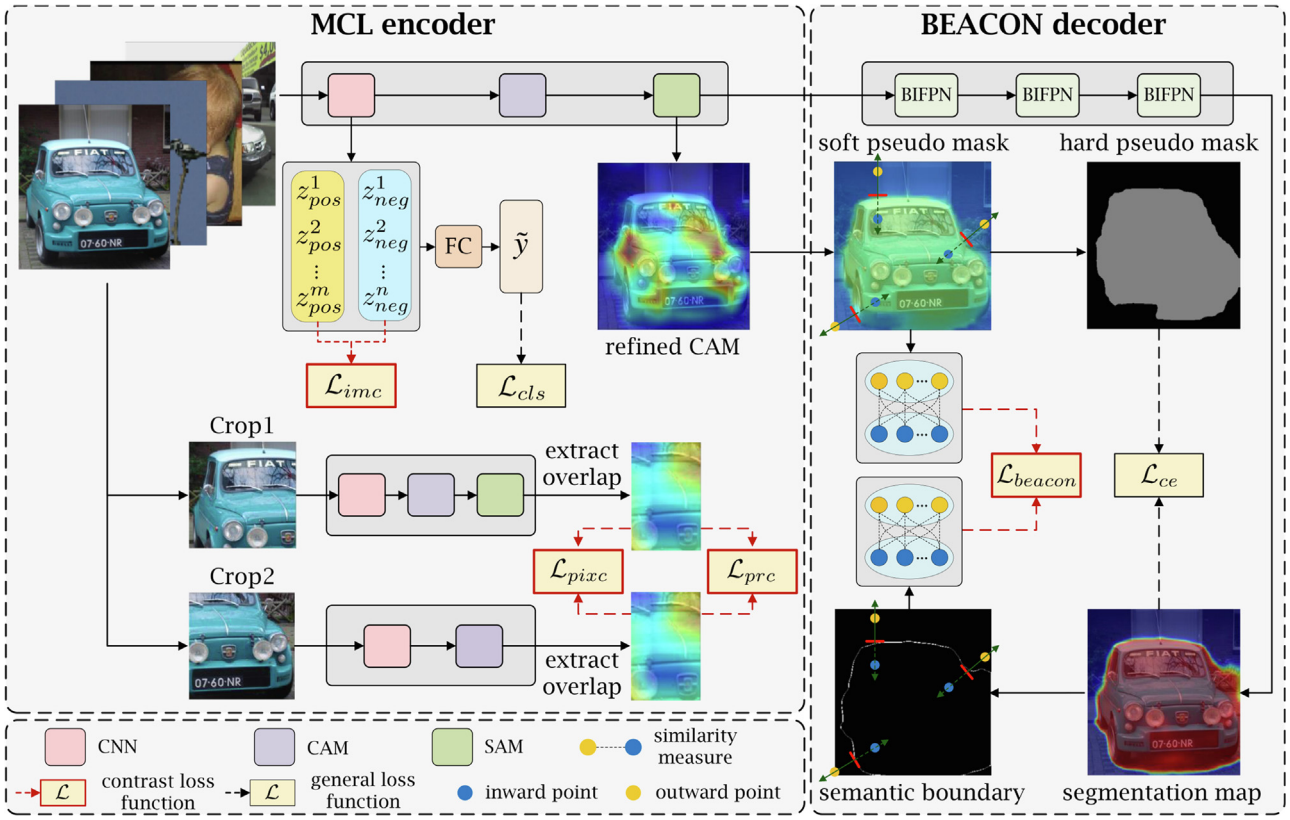

Fig. 1. Our proposed MuSCLe framework, composed of an MCL encoder and a BEACON decoder, exploits different levels of contrast information to enhance both the feature representation extracted from the encoder and the mapping function of the decoder for better WSSS performance. CAM=class activation map; SAM=spatial attention module; BiFPN=bi-directional feature pyramid network.

tions of our Multi-Strategy Contrastive Learning (MuSCLe) framework for WSSS are:

- We propose a multi-contrast learning (MCL) encoder to improve both the generalisation and distinctiveness of object feature extraction. In addition to the classification loss term used in a typical WSSS method, we explore image level contrast (IMC), pixel level contrast (PIXC), and pairwise regional contrast (PRC) based on overlapping regions of paired randomly cropped object patches to enhance the encoder representation.
- We design a novel boundary-based contrastive learning method, Boundary Enhancement viA Contrastive Orientation Navigation (BEACON), to enhance the decoder by learning features across boundaries extracted from pseudo masks, which are derived from our improved activation maps.
- Extensive experiments show that MuSCLe outperforms current state-of-the-art methods in WSSS on the PASCAL VOC 2012 dataset [21].

Our code is made available at https://github.com/SCoulY/MuSCLe.

2. Related work

2.1. Weakly supervised semantic segmentation

WSSS refers to segmenting semantic objects in images at the pixel level when only weak labels are available for training [10,11,15]. Our focus in this paper is to perform WSSS with only the weakest supervision cue in the form of image labels rather than additional cues such as saliency maps [11,12], scribbles [13], or bounding boxes [14].

To achieve pixel level semantic segmentation from image level annotations, [11] is one of the pioneering works to use saliency maps generated by a classification network as seeds (*i.e.*, pseudo labels) to guide the training of a segmentation network. To enhance initial seeds, some recent approaches utilise class activation maps as pseudo labels [8]. However, early CAM methods only highlight the most distinctive parts of objects, leading to insufficient segmentation performance. To address this, follow-up works introduce gradient-refined CAMs [22] and class-aware cross-entropy loss [23] to improve CAM performance. More recent approaches explore adversarial erasing [24], adjacent affinity transformations [9,15], self-supervised attention [10], sub-category mining [17], boundary exploration [16] and adversarial climbing [25] to enhance the quality of pseudo labels. Most recently, [12] and [26] perform online updates on the pseudo labels while training through incremental checkpoints and decomposition of classification and segmentation branches.

2.2. Contrastive learning

Contrastive learning [18] originates from self-supervised learning [27] and aims to learn generalised feature representations of an image from positive and negative sample pairs. To further explore the pairwise contrast information of image level annotations, [28] use image labels to improve the learned features by maximising the distances between paired samples belonging to different classes (*i.e.*, negative samples) while minimising the distances between same object pairs (*i.e.*, positive samples).

Introduction of pixel level contrast allows to place constraints on feature maps for better generalisation [29] and maintenance of fine details [30]. Wang et al. [10] utilise pixel-level contrast from positive samples after geo-transformations to extract so-

called equivariant features, while [31] employ two feature maps from different Siamese heads as two sets of marginal probability distributions and the earth mover's distance (EMD) to minimise the distance between paired patches in the two sets.

Ke et al. [32] integrates four types of semantic relationships into a uniform pixel-wise contrastive learning paradigm. However, their low-level similarity relies on extra supervision from a pre-trained edge detector and segmentation network to generate coarse segments, rendering the method inflexible and not image-label supervised only. In contrast, both our multi-strategy contrastive learning framework and CAM refinement module are trained strictly with image-level labels. Additionally, we impose contrastive constraints not only at pixel-level, but also at image-, regional- and boundary-levels.

2.3. Boundary enhancement

Exploitation of object boundaries is another promising option to enhance WSSS performance. In [11], a constrain-to-boundary loss is introduced to align a conditional random field (CRF) with the output of the trained network to support more detailed object segmentation[9]. propose an affinity network to generate consistent outputs on pixels that share similar semantics by constructing an affinity matrix to enhance object segmentation results, especially at boundaries. Ye et al. [33] leverages a set of guided matting filters to make edges sharper at body contours while fuzzier in hair regions. The network designed in [15] predicts pixel displacements and boundary probabilities to directly obtain an affinity matrix for boundary enhancement, whereas in [16], boundary annotations are extracted from an attention-pooling CAM and used to train a boundary exploration net (BENet) to identify object boundaries. These boundary maps form constraints to propagate pixels between salient semantic regions and their corresponding boundaries. Despite its effectiveness, various heuristic parameters need to be set at the training stage to enable BENet to distinguish real object boundaries from low-level edges.

3. Approach

In the following, after clarifying the motivation of our work, we present our novel MuSCLe framework in detail, introducing its multi-contrast learning (MCL) encoder and its Boundary Enhancement viA Contrastive Orientation Navigation (BEACON) decoder.

3.1. Motivation

State-of-the-art WSSS methods use enhanced CAMs to generate pseudo labels in order to provide supervision on an encoder-decoder-based network for semantic image segmentation. Inspired by the recent success [19,34] of employing contrastive learning to improve feature representations as well as to promote dense downstream tasks' performance (*i.e.*, detection and segmentation), we design a multi-level contrastive learning approach to enforce constraints to learn more reliable feature representations and distinctive semantics at both encoder and decoder for better segmentation. In particular, paired samples for contrastive learning are extracted at image level, pixel level, regional level and boundary level in order to ensure consistency of features in the same object classes while maximising distances between different object categories. This simple yet effective strategy facilitates the generation of high-quality pseudo labels as well as improves the encoder-decoder network for better segmentation performance.

3.2. MCL encoder

The encoder in a WSSS network not only extracts salient feature representations to be used in its decoder but generates pseudo masks to provide additional cues for fine-grained segmentation. As illustrated in Fig. 1, we propose contrastive learning loss terms to build our multi-contrast learning encoder, which can generate generalised feature representations and high-quality pseudo masks.

3.2.1. Image level contrast

Given a query sample x_i in the batched images X from dataset Ω and its label $y_i = (y_{i;1}, y_{i;2}, \ldots, y_{i;K})$ as K-dimensional multi-hot vector representing the presence of K object categories with present categories indicated by ones and absent categories by zeros, in each training batch. We propose a novel way to process contrast pairs in each batch which significantly increases the efficiency of contrastive learning compared to Siamese networks. To measure similarity of both positive and negative pairs, we extract image embeddings by average pooling of the feature maps from the last convolutional layer of the CNN feature extractor and using the dot product to calculate scores.

The image-level contrastive learning loss term we employ is calculated as

$$\mathcal{L}_{imc} = -\log\left(\frac{\sum_{Z^+}\exp(z_i\cdot\tilde{z}_i)}{\sum_{Z^+}\exp(z_i\cdot\tilde{z}_i)+\sum_{Z^-}\exp(z_i\cdot z_j)}\right), \tag{1}$$

where z_i is the vector embedding of the query sample, $\tilde{z}_i \in Z^+$ represents each embedding of positive samples, and $z_j \in Z^-$ represents each embedding of negative samples. In contrast to [18,19], our $\mathcal{L}_{imc}$ does not rely on augmented views to generate positive samples which significantly reduces memory consumption. Additionally, to alleviate single positive pair bias and to enforce batch-wise attention on positive pairs, we compute the integral of $\exp(z_i\cdot\tilde{z}_i)$ before taking the logarithm. We show, empirically and theoretically, that this leads to more effective training compared to the loss term from Khosla et al. [28] in Appendix A.

3.2.2. Pixel level contrast

Given two random regions cropped from an image, pixel level contrastive learning aims to maximise the feature similarity of pixels in their overlapping region even though their representations are not exactly the same due to their distinct contexts within the receptive field. As highlighted in [29], pixel level contrast imposes pixel-wise feature consistency to enhance the feature representations for its dense-prediction downstream task (*i.e.*, image segmentation in this paper).

As illustrated in Fig. 2, we obtain the pixel level contrastive loss by calculating the similarity between paired pixel-wise features in the overlapping region from two types of feature maps, the original CAM and a spatial attention module (SAM) map [10,35].

The SAM utilises a global self-attention mechanism, capable of exploiting long-range contexts, to enhance the CAM, and is obtained as

$$M' = SAM(M) = \mathrm{softmax}(g_1(M)^T \times g_2(M)) \times g_3(M), \tag{2}$$

where $g_1(\cdot)$, $g_2(\cdot)$, and $g_3(\cdot)$ denote individual linear projections, and M is the CAM response map.

To alleviate inconsistencies and learn generalised features of the overlapped area from crops of CAM and SAM [20,34], we employ an alignment using the pixel-wise contrastive loss

$$\mathcal{L}_{pixc} = -\frac{1}{HW}\sum_{k=1}^{HW}\cos(u'_k, \mathrm{sg}(\tilde{v}_k)), \tag{3}$$

where u'_k and $\tilde{v}_k$ are feature vectors from overlapping regions in M' and $\tilde{M}$ extracted from SAM and CAM, respectively, H and W are the height and width of the regions, and $\mathrm{sg}(\cdot)$ denotes the stop gradient operator which avoids interference of cross-optimisation [20,34]. This design also reduces the computational cost, thus improving the efficiency of our method.

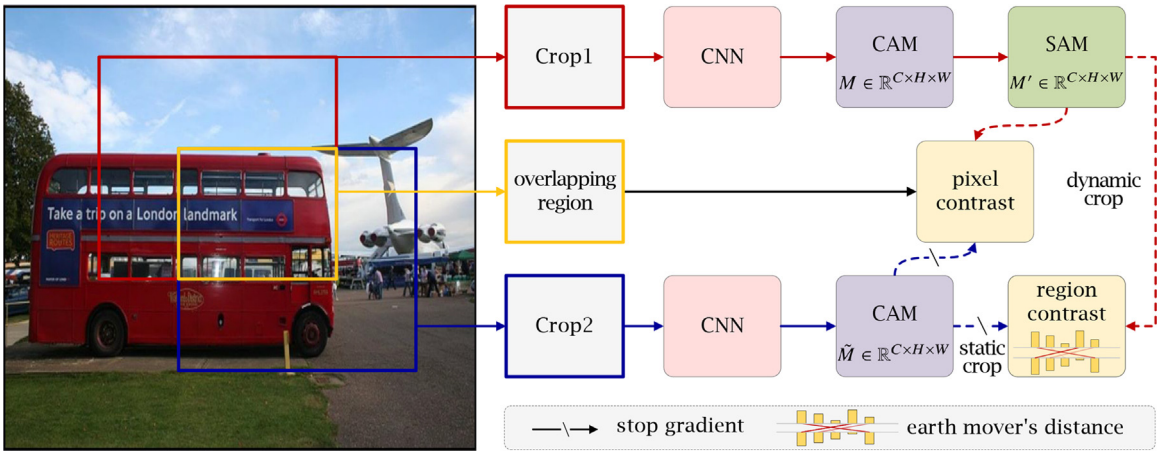

Fig. 2. Illustration of pixel level contrast and pairwise regional contrast. Blocks of the same colour share the same weights.

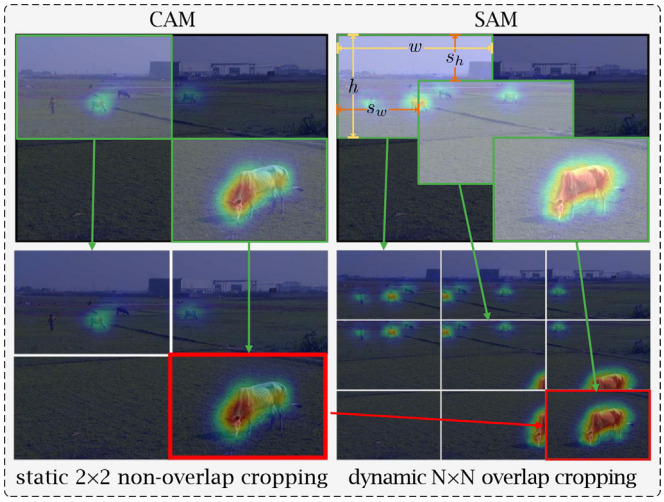

Fig. 3. Illustration of dynamic cropping and matching.

3.2.3. Pairwise regional contrast

In addition to the image level and pixel level contrastive loss terms, we propose a novel dynamic pairwise regional contrastive loss term to further enhance the scale-invariant characteristics of the extracted features. As illustrated in Fig. 3, to provide sufficient flexibility while keeping complexity low, we divide the CAM response map into static (*e.g.*, 2×2) non-overlapping patches. To impose feature consistency from paired objects of different scales, we introduce four parameters, patch width w, patch height h, and horizontal and vertical sliding strides s_w and s_h, to randomly sample regions at different scales from the SAM feature map. Although metrics like KL-divergence and cosine similarity are popular for aligning two well-defined probability distributions, the earth mover's distance (EMD) [36] is more suitable to align the crops of CAM and SAM from potentially different spatial locations and, more importantly, of differing sizes.

To obtain a more reliable EMD measure, we aim to avoid the bias introduced from background response maps, and estimate the background activation map as

$$M_{bg} = 1 - \max_{1\le c\le C-1}(M_c), \tag{4}$$

with

$$M_c \leftarrow \frac{\exp(M_c)}{\sum_{1\le c\le C-1}\exp(M_c)}, \tag{5}$$

where M_c belongs to one of the $C-1$ foreground activation maps and M_{bg} represents the estimated background activation map. The concatenation of all foreground activation maps and the background activation map yields the background-included CAM ($M \in \mathbb{R}^{C\times H\times W}$).

Our pairwise regional contrast can be seen as a transportation problem if we consider a cropped static patch from CAM as a shipper and a dynamic matched patch from SAM as a receiver. The goal is to find an optimal path that minimises the global transportation cost (*i.e.*, the discrepancy or dissimilarity in our case). The optimal match is solved by the earth mover's distance (EMD) and once the best match is found, we again minimise the dissimilarity between patches that have the lowest transportation cost as

$$\mathcal{L}_{prc} = \underset{(a,b)}{\operatorname{argmin}}\,\mathrm{EMD}(p_a, \mathrm{sg}(\tilde{p}_b)), \tag{6}$$

where $p_a \in \{p\}_1^{\mathrm{A}} \subset M'$ and $\tilde{p}_b \in \{\tilde{p}\}_1^{\mathrm{B}} \subset \tilde{M}$ are feature vectors from paired patches generated from the original feature maps M' and $\tilde{M}$, respectively. As in pixel level contrastive loss, we use the stop gradient operator $\mathrm{sg}(\cdot)$ to avoid cross-optimisation.

3.2.4. Overall loss function

In addition to the loss terms introduced above, we use the well-established multi-label multi-class classification loss (*i.e.*, binary cross entropy loss), focal loss [37], and a pair loss [38] to address sample imbalance and over-confidence of negative sample issues, and combine them to form a hybrid classification loss (HCL) term

$$\mathcal{L}_{hcl}(y,\hat{y}) = \mathcal{L}_{bce}(y,\hat{y}) + \mathcal{L}_{focal}(y,\hat{y}) + \mathcal{L}_{pair}(y,\hat{y}), \tag{7}$$

which improves WSSS performance compared to using individual terms.

The overall loss function of our MCL encoder is then defined as

$$\mathcal{L}_{MCL} = \mathcal{L}_{hcl} + \mathcal{L}_{imc} + \mathcal{L}_{pixc} + \mathcal{L}_{prc}. \tag{8}$$

3.3. BEACON decoder

During training of a typical WSSS network, pseudo labels generated from the encoder output are used to supervise the learning process. Consequently, these pseudo labels are key to the final segmentation performance. Although some recent work improves the quality of pseudo labels by introducing an extra processing stage, such as an AffinityNet [9] or conditional random fields [39], to enhance implicit boundary smoothness, the resulting hard masks lead to supervision bias during training. To alleviate this and achieve more consistent segmentation results across

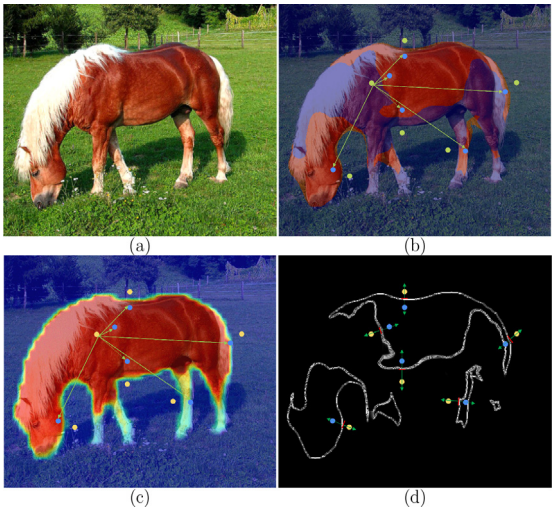

Fig. 4. Illustration of inward/outward point set division. (a) original image; (b) segmentation map during training; (c) pseudo mask; (d) in/out-ward division based on boundary map.

object boundaries, we propose a novel boundary contrastive loss term, named Boundary Enhancement viA Contrastive Orientation Navigation (BEACON), to further improve our segmentation network.

The detailed algorithm for BEACON is given in Algorithm 1. We first form two boundary candidate point sets, an inward point set and an outward point set. To do so, we apply the Sobel operator on the segmentation map, and identify object boundary points as those points that exhibit the top 20% largest gradient magnitudes. This allows to build reliable sets to select paired samples for contrastive learning. Note that the obtained boundary map (see Fig. 4(d) for an example) has a strong semantic meaning and is different from applying the Sobel operator directly on the input image. We obtain the gradient directions of the boundary points and uniformly divide the continuous gradient orientation from 0 to 2π into 8 equally sized bins to fit an 8-neighbourhood of a pixel (see Algorithm 2). Based on a step parameter, we then calculate a displacement from a boundary point along the gradient direction as well as the opposite direction to generate candidate points for the two sets.

Having obtained the inward and outward boundary point sets, we use the soft pseudo masks generated from our MCL encoder (Fig. 4(c)) and the dense map from the decoder to define a boundary contrastive loss term. As illustrated in Fig. 4(b), the segmentation map is far from perfect at the early training stages. Thus, we calculate the point-wise one-to-all similarity between the two sets on both the dense feature map and the pseudo mask to enhance the object boundary feature consistency. In particular, we define a sign function Sign$(\cdot)$ to identify if the similarity values calculated from the dense feature map $\tilde{y}$ coincide with those from the soft pseudo mask y by comparing their scores to a threshold τ.

The sign determines the direction of optimisation imposed on similarity as shown in Algorithm 3. Intuitively, if $S^d < \tau$, the query in-out pair is recognised as dissimilar and thus a positive edge (P) is assigned. Furthermore, if $S^m < \tau$ is also satisfied, a true positive (TP) case is identified, yielding a similarity suppression (positive sign) to make them more dissimilar. Integrating TP, FP, FN, and TN

Algorithm 2: In-/out-ward division function InOutDiv$(\cdot)$.

Input : orientation map M; step size s
Output: inward point set Φ; outward point set: Ψ
$\Phi \leftarrow \varnothing$, $\Psi \leftarrow \varnothing$;
$P \leftarrow$ points in M with gradient magnitude in top 20%;
for p *in* P **do**
  // retrieve gradient orientation of p:;
  $\vec{o} \leftarrow M_p$;
  // quantise $\vec{o}$ into 8-directional vector:;
  $\vec{q} \leftarrow$ `Quantise`$(\vec{o})$;
  // get inward and outward points:;
  obtain inward point ϕ by stepping from p along $-\vec{q}$ with step size s;
  obtain outward point ψ by stepping from p along $\vec{q}$ with step size s;
  // add new points to the sets:;
  append ϕ to Φ;
  append ψ to Ψ;
end

Algorithm 1: BEACON algorithm.

Input : orientation map M; dense feature map $\tilde{y}$; soft pseudo mask y; parameters *steps*, k
Output: BEACON loss $\mathcal{L}_{beacon}$
// select in/out-ward point sets from M:;
$\Phi, \Psi \leftarrow$ `InOutDiv`$(M,\ steps)$;
randomly select k samples from Φ, Ψ as I, O;
// anchor I, O back onto y and $\tilde{y}$ to yield inward set (I^d, I^m) and outward set (O^d, O^m) on dense feature map and pseudo mask:;
$I^m,\ O^m \leftarrow y^I, y^O$;
$I^d,\ O^d \leftarrow \tilde{y}^I,\ \tilde{y}^O$;
// calculate similarity matrix S for the two sets:;
$S^d \leftarrow f($`sg`$(I^d), O^d)$;
$S^m \leftarrow f(I^m, O^m)$;
// obtain point-wise signs:;
$sign^I, sign^O \leftarrow$ `Sign`$(S^m_{i,o},\ S^d_{i,o})$;
// calculate and return loss:;
$\mathcal{L}_{beacon} \leftarrow$
$\frac{1}{|O|}\sum_{o\in O}\log(sign^O \cdot \frac{1}{|I|}\sum_{i\in I} S^d_{i,o}) + \frac{1}{|I|}\sum_{i\in I}\log(sign^I \cdot \frac{1}{|O|}\sum_{o\in O} S^d_{i,o})$

Algorithm 3: Sign$(\cdot)$ function.

Input : mask similarity matrix $S^m_{i,o}$; feature similarity matrix $S^d_{i,o}$
Output: point-wise signs $sign^I, sign^O$
$S^m_I \leftarrow \frac{1}{|O|}\sum_{o\in O} S^m_{i,o}$;
$S^d_I \leftarrow \frac{1}{|O|}\sum_{o\in O} S^d_{i,o}$;
$FP^I \leftarrow$ `AND`$(\mathbb{I}(S^m_I > \tau), \mathbb{I}(S^d_I < \tau))$;
$FN^I \leftarrow$ `AND`$(\mathbb{I}(S^m_I < \tau), \mathbb{I}(S^d_I > \tau))$;
$TP^I \leftarrow$ `AND`$(\mathbb{I}(S^m_I < \tau), \mathbb{I}(S^d_I < \tau))$;
$TN^I \leftarrow$ `AND`$(\mathbb{I}(S^m_I > \tau), \mathbb{I}(S^d_I > \tau))$;
convert FP^I, FN^I, TP^I and TN^I into binary values {-1, 1};
// assign negative to actual condition negative cases:;
$TN^I \leftarrow -TN^I$;
$FP^I \leftarrow -FP^I$;
// compute signs for inward set:;
$sign^I \leftarrow FN^I \cup TP^I \cup TN^I \cup FP^I$;
// compute signs for outward set:;
$sign^O \leftarrow FN^O \cup TP^O \cup TN^O \cup FP^O$;

cases, point-wise signs are obtained and are used to calculate the boundary contrastive loss.

The loss function for this training stage is defined as

$$\mathcal{L}_{seg} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{beacon}, \tag{9}$$

where parameter λ allows for balancing between the global pixel-wise cross entropy loss $\mathcal{L}_{ce}$ and the near-boundary pixel representation enhancement.

4. Experimental results

4.1. Implementation details

Our experiments are conducted on the PASCAL VOC 2012 dataset [21] with 20 foreground classes and 1 background class. Following [10,40], we build an augmented training set with 10,582 images. During classification training, only the 20 foreground class logits are taken into consideration, while the background class activation map is estimated for pairwise regional contrast. We use cosine similarity to compute the transportation cost matrix and use Sinkhorn iteration [41] for fast computation of the EMD. We set equal weights for the HCL, IMC, PIXC, and PRC loss terms and enable them one after another at fixed epoch intervals to avoid potential interference.

Our MuSCLe implementation comprises an EfficientNet encoder and a BiFPN decoder [42]. To efficiently scale up the model, we use batch sizes of 16, 8 and 6 for EfficientNet-b3, EfficientNet-b5, and EfficientNet-b7, respectively, with the same decoder which has 3 BiFPN layers. Experiments are conducted on an RTX 3090 GPU using PyTorch. Unless otherwise noted, all presented results are averaged over 3 random runs.

The input of image level contrast and classification head is resized while keeping the original image aspect ratio and padded to 448×448, while the pixel level contrast and pairwise regional contrast heads use random crops of size 224×224 as inputs. CRF and random walk refinement [9,15] are executed after SAM output to generate pseudo masks.

4.2. Improved CAM quality

We quantitatively evaluate the effectiveness of each component of our MuSCLe approach in Table 1. From there, it is evident that each proposed module leads to a notable performance increase. Following common practice [9,10,17], test time augmentation (TTA) with multi-scale inference gives a further improvement of 2.5%-3%. Compared to ordinary CAM methods, our multi-contrast learning encoder improves CAM quality by a large margin (+6.8%).

Table 2 compares the pseudo label quality of our method with other state-of-the-art (SOTA) approaches. As is evident, our MCL clearly outperforms the other methods, improving the CAM of AffinityNet [9] by 10.4% and the result of SEAM [10] by 3.0%. Although the improvements with affinity/IRN refinement are less pronounced compared to those for raw CAMs, we obtain the highest mIoU of 64.9% and 66.8% (with Wide-ResNet38 and EfficientNet backbones, respectively). We conjecture that because our CAM trained by MCL is denser and more continuous, the affinity transformation barely enhances local feature representation with adjacent context. Note that we opt for EfficientNet as our backbone mainly for its processing efficiency and light-weightedness.

In addition, we generate visualisations of the learned representations from SEAM and our MCL using t-SNE dimensionality reduction. As can be seen from the results shown in Fig. 5, the classes are better separated in the MCL visualisation, while for SEAM we can observe significant overlap between classes and higher variation within each class.

Table 1
Ablation study for MCL encoder. mIoU (%) reflects pseudo CAM quality on *train* set. HCL=hybrid classification loss; IMC=image level contrast; PIXC=pixel level contrast; PRC=pairwise regional contrast.

HCL	IMC	PIXC	PRC	single scale mIoU	multi-scale mIoU
				48.5	51.6
✓				53.3	55.7
✓	✓			54.3	57.2
✓	✓	✓		54.8	57.6
✓	✓	✓	✓	**55.3**	**58.4**

Table 2
Pseudo label quality on *train* set in terms of mIoU. ∗=random walk with affinity refinement; †=random walk with IRN refinement.

method	backbone	CAM	CAM∗	CAM†
AffinityNet	ResNet50	48.0	58.1	–
IRN	ResNet50	48.3	59.3	66.5
MCL	ResNet50	52.6	60.9	63.1
SC-CAM	Wide-ResNet38	50.9	63.4	–
SEAM	Wide-ResNet38	55.4	63.6	–
MCL	Wide-ResNet38	58.3	**64.9**	66.6
MCL	EfficientNet	**58.4**	64.6	**66.8**

Table 3
Ablation study for segmentation network architecture. mIoU reflects segmentation performance on *val* set with IRN refined pseudo label.

backbone	# BiFPN	BEACON	mIoU
EfficientNet-b3	1	no	63.2
EfficientNet-b3	2	no	64.1
EfficientNet-b3	3	no	64.9
EfficientNet-b5	3	no	65.5
EfficientNet-b7	3	no	65.8
EfficientNet-b3	3	yes	65.2
EfficientNet-b5	3	yes	65.9
EfficientNet-b7	3	yes	**66.6**

4.3. Semantic segmentation training

To investigate the impact of the decoder architecture on segmentation training using synthesised pseudo labels from the encoder, we test MuSCLe with different backbones, different numbers of BiFPN layers, and with and without BEACON. From Table 3, we notice that BEACON leads to a consistent performance increase, while densifying BiFPN layers also gives notable improvement. In addition, scaling up the encoder backbone from b3 to b7 gives a 1.4%/0.9% boost with/without BEACON.

We perform a thorough ablation study, with results reported in Table 4, on our BEACON module to show the impact of different hyper-parameters and the effectiveness of BEACON. Since larger values of λ in Eq. (9) put more focus on near-boundary pixel enhancement and boundary map generation relies on accurate pixel-wise segmentation, as expected, too extreme λ values do not lead to an improvement. For the step size walking towards the gradient orientation, we observe an optimal value of 7 with fewer steps not supporting sufficiently distinctive in/out-ward feature representation and more steps exceeding tiny object boundaries when selecting inward points along the inverse gradient orientation. Turning to the similarity threshold τ, a halfway division of the similarity scores (*i.e.*, $\tau = 0.5$) provides only a small improvement compared to a dynamic threshold, μ_m, which is obtained as the mean of the similarity matrix derived from the soft mask. Selecting $k = 128$ candidates of in-/outward pairs results in a

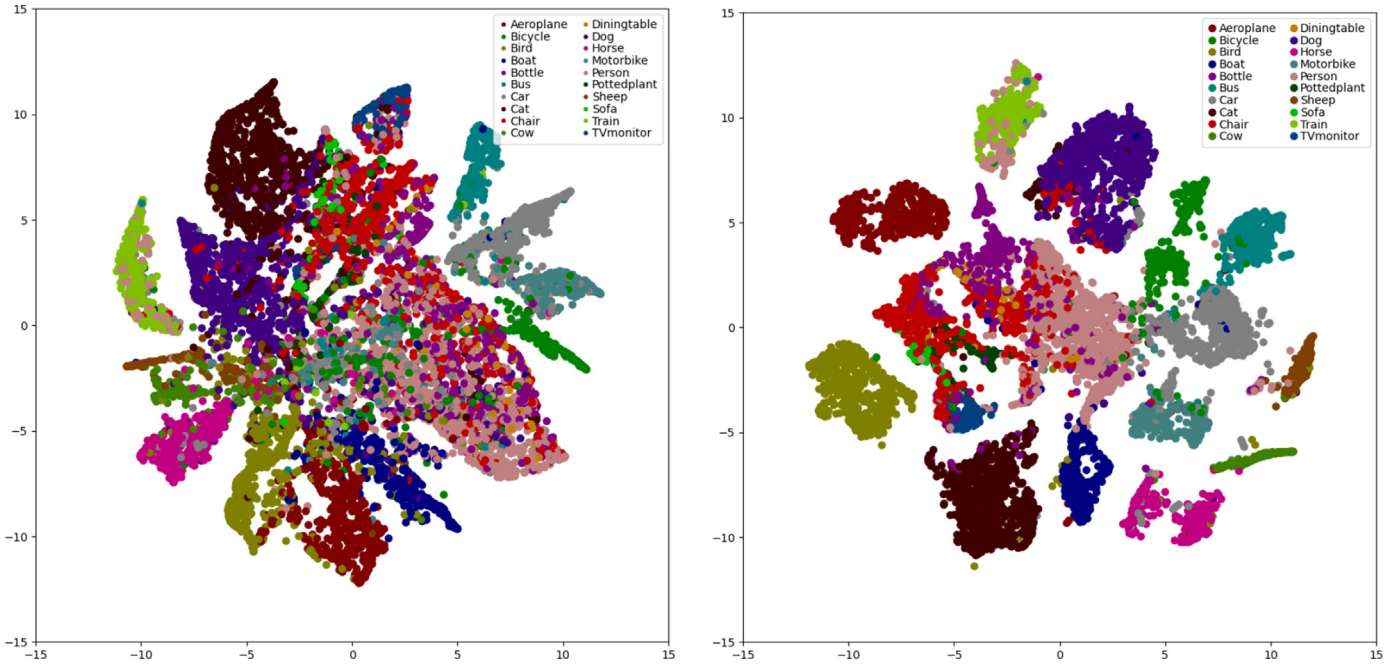

Fig. 5. t-SNE visualisations of the learned representations of encoder for SEAM (left) and our MCL (right).

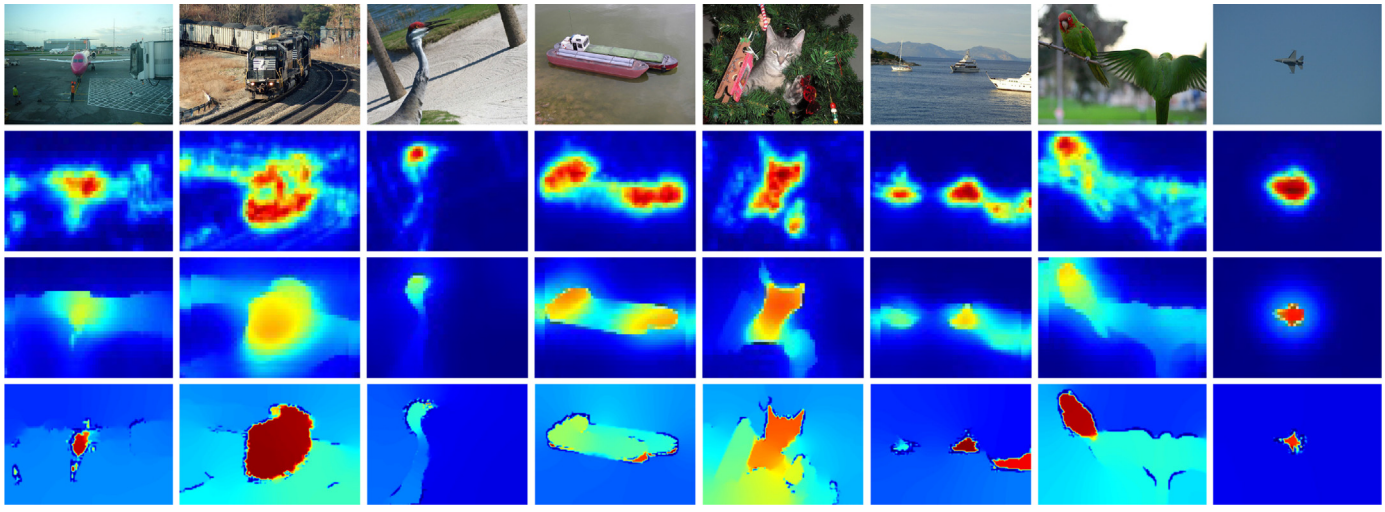

Fig. 6. Visualisations of SAMs before and after affinity/IRN refinement. From top to bottom: original image; SAM output; affinity refined SAM output; IRN refined SAM output.

good performance/efficiency trade-off. Overall, the best results are obtained by combining μ_m with $\lambda = 0.05$, 7 steps, and $k = 128$.

4.4. Comparison with SOTA

We compare MuSCLe with current SOTA methods in terms of performance and supervision in Table 5. From there, we see that on the *val* set, MuSCLe-b5 achieves performance on-par with LIID [47], surpassing most other methods. Further deepening our model to MuSCLe-b7, our model outperforms all compared methods on the *test* set using only image level labels (in contrast to some other methods such as [43–45] which rely on stronger supervision based on image labels in combination with saliency maps), while only inferior to the previous best AdvCAM [25] on the *val* set.

Looking at the class-wise performance on the *val* set in Table 6, MuSCLe gives the best result for 9 classes, more than any other method (AdvCAM is best for 7 categories). In particular, for the *cow, dog, horse*, and *sheep* classes, the performance is vastly superior to other approaches. On the other hand, worse performance is obtained on the *tv* category. This is because we enforce contextual feature enhancement in our proposed method while TV monitors in the VOC dataset often appear together with other uncategorised objects such as benches and keyboards.

4.5. Qualitative results

We qualitatively show the performance of affinity and IRN refinement [9,15] and compare with the global spatial attention from Cao et al. [35]. From Fig. 6, we can see that object boundaries after affinity refinement enhance the SAM maps. Furthermore, IRN refinement generates sharper edges which is more beneficial than affinity refinement when used as prior in the BEACON decoder. In our experiments, we perform affinity/IRN refinement on SAM output with 4/8 iterations to yield the pseudo semantic segmentation label.

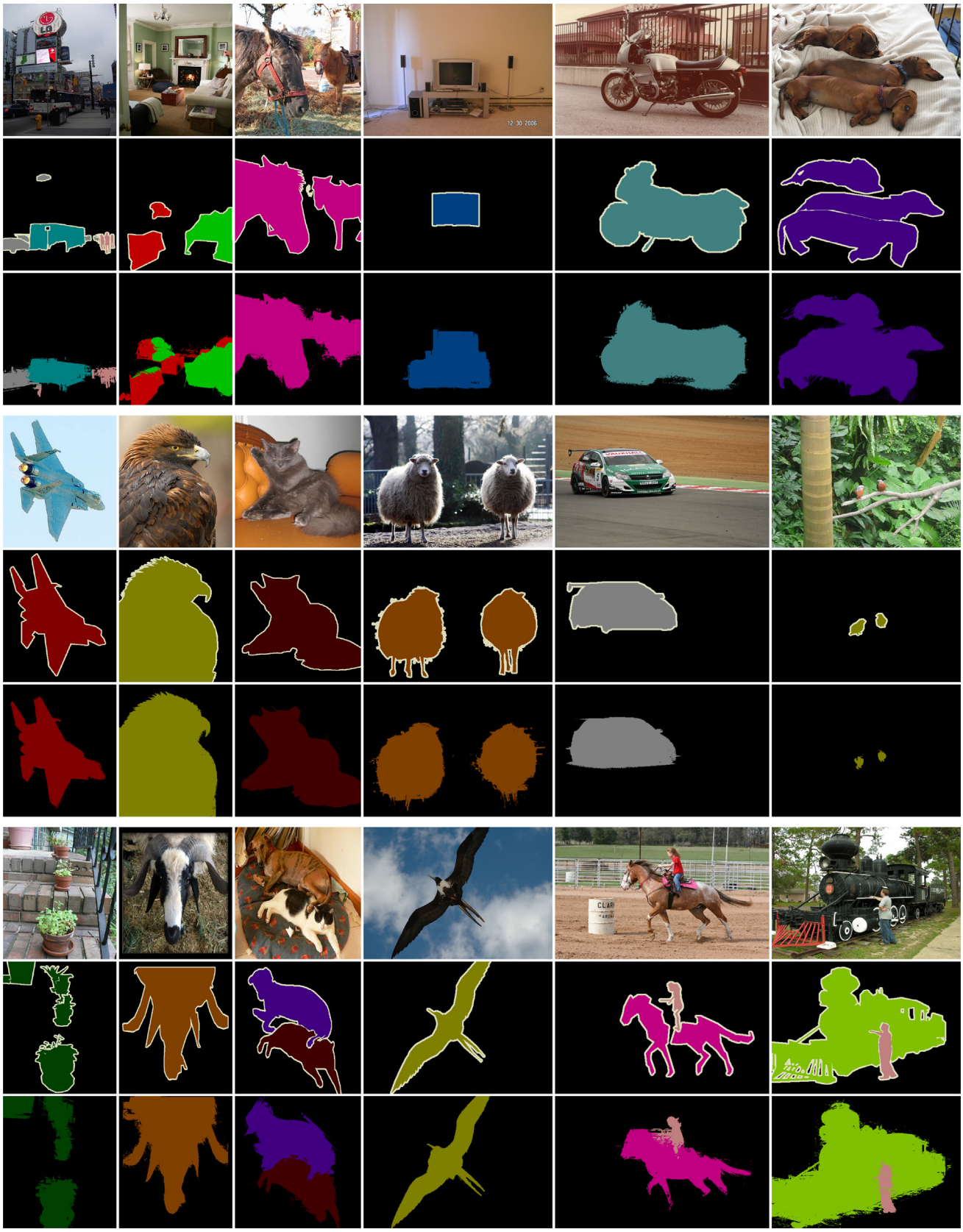

Fig. 7. Example segmentation results on PASCAL VOC2012 *val* set. From top to bottom: original image; ground truth; segmentation result; original image; ground truth; segmentation result; original image; ground truth; segmentation result.

Table 4
BEACON ablation study of segmentation performance on *val* set. $*$=segmentation with affinity refined pseudo labels; †=segmentation with IRN refined pseudo labels. μ_m denotes the mean of similarity matrix derived from the soft mask. For each result, the left and right values denote MuSCLe-b5 and MuSCLe-b7 performance, respectively.

λ	steps	k	τ	mIoU*		mIoU†	
0	n/a	n/a	n/a	63.8	64.1	65.5	65.8
0.05	7	128	0.5	64.1	63.8	65.5	66.2
0.05	7	64	μ_m	63.8	63.9	65.6	66.1
0.05	7	128	μ_m	65.2	66.1	**65.8**	**66.6**
0.05	7	256	μ_m	64.4	64.6	65.5	65.9
0.05	5	128	μ_m	64.5	65.1	65.6	66.1
0.05	9	128	μ_m	63.9	64.2	65.5	66.3
0.1	7	128	μ_m	64.7	64.3	65.5	66.0

Table 5
Comparison with SOTA WSSS methods on VOC2012 *val* and *test* sets. I=image level label; I+S=image level label + saliency map. For RRM, the one-stage result in the paper is given to allow for a fair comparison.

method	label	val mIoU	test mIoU
FickleNet_CVPR19[43]	I+S	64.9	65.3
OAA_CVPR19 [44]	I+S	65.2	66.4
OAA++$^+_{\text{TPAMI21}}$ [45]	I+S	66.1	67.2
AffinityNet_CVPR18 [9]	I	61.7	63.7
SEAM_CVPR20 [10]	I	64.5	65.7
SC-CAM_CVPR20 [17]	I	66.1	65.9
BES_ECCV20 [16]	I	65.7	66.6
LSISU_PR21 [12]	I	61.2	62.5
LayerCAM_TIP21 [46]	I	63.0	64.5
AdvCAM_CVPR21 [25]	I	**68.1**	68.0
RRM_PR22 [26]	I	65.4	65.3
LIID_TPAMI22 [47]	I	66.5	67.5
MuSCLe-b5	I	65.9	67.4
MuSCLe-b7	I	66.6	**68.8**

We show some representative qualitative segmentation results obtained from MuSCLe-b7 in Fig. 7, from which we can observe that detailed object boundaries are properly recovered. For multi-label scenarios, our model correctly distinguishes the instances of each category, while multiple object instances at different scales and locations are also correctly recognised, demonstrating the efficacy of the dynamic cropping and matching strategy.

Figure 8 shows some typical examples from the SBD dataset [40], illustrating the impact BEACON has on the obtained semantic boundaries. It is apparent that the semantic boundaries detected with BEACON are more complete and noise-robust compared to those without BEACON.

In Fig. 9, we show some failure cases of TVs, which, as noted above, is where MuSCLe performs relatively inferior compared to other methods. We find that the segmented masks also cover some other objects, such as keyboards and TV stands, which have a high co-occurrence rate with TVs in the dataset. This could be addressed by either adding extra supervision cues or by having more training data that only contains single objects (*i.e.* without commonly co-occurred objects).

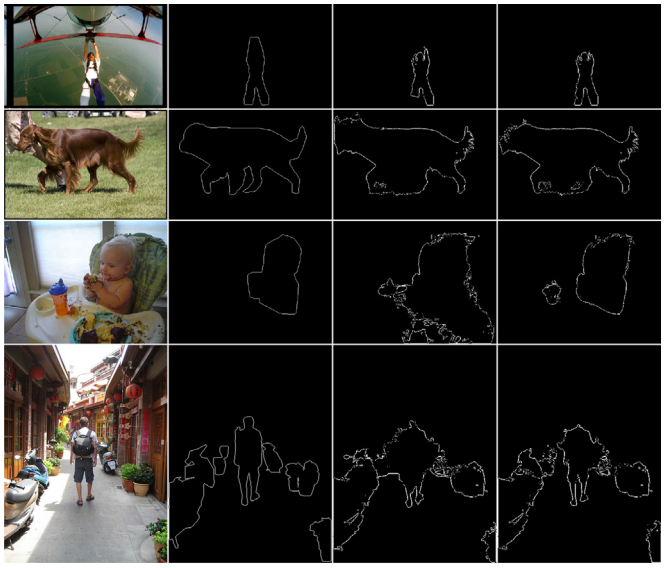

Fig. 8. Example boundary results on SBD *trainval* set. From left to right: original image; class-agnostic semantic boundary label; semantic boundary without BEACON; semantic boundary with BEACON.

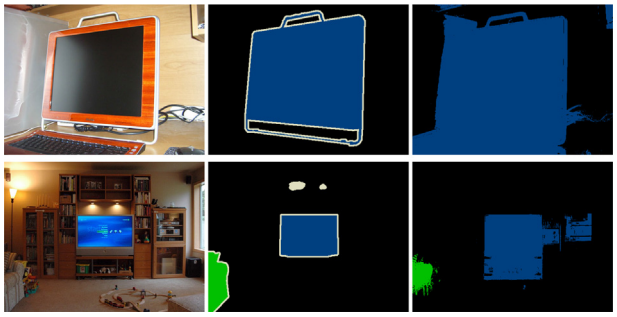

Fig. 9. Example segmentation results of TVs. From left to right: original image; ground truth; segmentation result.

5. Conclusions

In this paper, we exploit only image-level annotation to accomplish weakly supervised semantic segmentation. For this, we have proposed a novel MuSCLe framework which comprises an MCL encoder and a BEACON decoder. The former is designed to improve the initial CAM response via contrastive learning at different levels, while the latter aims to explicitly enhance feature representations around object boundaries through a contrastive scheme. Extensive experiments have demonstrated that, MuSCLe achieves SOTA performance on the PASCAL VOC2012 dataset, while ablation studies and visualisations further illustrate the efficacy of our proposed approach. Notably, this is achieved on a single GPU, unlike most existing work in the area, demonstrating the efficiency of our method. In future work, we will investigate online pseudo-labelling and refinement for WSSS to further boost the training efficiency.

Table 6
Category performance comparison on PASCAL VOC2012 *val* set.

method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv
AffinityNet	88.2	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2	80.4	62.0	70.4	73.7	42.5	70.7	42.6	68.1	51.6
FickleNet	89.5	76.6	32.6	74.6	51.5	71.7	83.4	74.4	83.6	24.1	73.4	47.4	78.2	74.0	68.8	73.2	47.8	79.9	37.0	57.3	**64.6**
SEAM	88.8	68.5	33.3	85.7	40.4	67.3	78.9	76.3	81.9	29.1	75.5	48.1	79.9	73.8	71.4	**75.2**	48.9	79.8	40.9	58.2	53.0
SC-CAM	88.8	51.6	30.3	82.9	53.0	**75.8**	88.6	74.8	86.6	**32.4**	79.9	**53.8**	82.3	78.5	70.4	71.2	40.2	78.3	42.9	66.8	58.8
BES	88.9	74.1	29.8	81.3	53.3	69.9	**89.4**	**79.8**	84.2	27.9	76.9	46.6	78.8	75.9	72.2	70.4	50.8	79.4	39.9	65.3	44.8
AdvCAM	**90.0**	**79.8**	**34.1**	82.6	**63.3**	70.5	**89.4**	76.0	87.3	31.4	81.3	33.1	82.5	80.8	**74.0**	72.9	50.3	82.3	42.2	**74.1**	52.9
MuSCLe-b7	87.7	71.3	31.1	**86.7**	51.8	68.5	84.6	79.5	**88.1**	22.3	**83.6**	51.8	**86.1**	**83.0**	**74.0**	64.6	**51.1**	**84.8**	**44.8**	63.3	40.6

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my code in the paper.

Appendix A. Gradient analysis of image level contrast

We provide both a formal proof and empirical evaluation to show how our proposed image level contrastive loss term $\mathcal{L}_{imc}$ outperforms the classical supervised contrastive loss term $\mathcal{L}_{sup}$ from Khosla et al. [28] in limited batchsize settings.

We compare our proposed image level contrastive loss

$$\mathcal{L}_{imc} = -\log \frac{\sum_{Z^+} \exp(z_i \cdot \tilde{z}_i)}{\sum_{Z^+} \exp(z_i \cdot \tilde{z}_i) + \sum_{Z^-} \exp(z_i \cdot z_j)}, \tag{A.1}$$

where $\tilde{z}_i \in Z^+, z_j \in Z^-$, to the original supervised contrastive learning

$$\mathcal{L}_{sup} = -\sum_{Z^+} \log \frac{\exp(z_i \cdot \tilde{z}_i)}{\exp(z_i \cdot \tilde{z}_i) + \sum_{Z^-} \exp(z_i \cdot z_j)}, \tag{A.2}$$

used in [28].

The gradient of our proposed term w.r.t. z_i is

$$\begin{aligned}\frac{\partial \mathcal{L}_{imc}}{\partial z_i} &= \frac{\partial}{\partial z_i}\{\log[\sum_{Z^+} \exp(z_i \cdot \tilde{z}_i) + \sum_{Z^-} \exp(z_i \cdot z_j)]\} \\ &\quad - \frac{\partial}{\partial z_i}[\log \sum_{Z^+} \exp(z_i \cdot \tilde{z}_i)] \\ &= \frac{\sum_{Z^+} \tilde{z}_i \exp(z_i \cdot \tilde{z}_i) + \sum_{Z^-} z_j \exp(z_i \cdot z_j)}{\sum_{Z^+} \exp(z_i \cdot \tilde{z}_i) + \sum_{Z^-} \exp(z_i \cdot z_j)} - \frac{\sum_{Z^+} \tilde{z}_i \exp(z_i \cdot \tilde{z}_i)}{\sum_{Z^+} \exp(z_i \cdot \tilde{z}_i)},\end{aligned} \tag{A.3}$$

where $Z^+ = \{\tilde{z}_i\}_1^N$ and $Z^- = \{z_j\}_1^M$ denote the positive and negative sets, respectively. The normalised dot product requires $\|z_i\|_2$, $\|\tilde{z}_i\|_2$, $\|z_j\|_2$, $\|z_i \cdot \tilde{z}_i\|_2$ and $\|z_i \cdot z_j\|_2 \in [0, 1]$ so that $\angle z_i \tilde{z}_i, \angle z_i z_j \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

When $Z^+ = \varnothing$, *i.e.* $N = 0$, then

$$0 \le \left\| \frac{\partial \mathcal{L}_{imc}}{\partial z_i} \right\|_2 = \left\| \frac{\sum_{Z^-} z_j \exp(z_i \cdot z_j)}{\sum_{Z^-} \exp(z_i \cdot z_j)} \right\|_2 \le 1, \tag{A.4}$$

and when $Z^+ \neq \varnothing$, *i.e.* $N \neq 0$, then

$$\left\| \frac{\partial \mathcal{L}_{imc}}{\partial z_i} \right\|_2 \le \frac{\|z_j - \tilde{z}_i\|_2 \sum_{Z^-} \exp(z_i \cdot z_j)}{\sum_{Z^+} \exp(z_i \cdot \tilde{z}_i) + \sum_{Z^-} \exp(z_i \cdot z_j)}. \tag{A.5}$$

When $\|\tilde{z}_i\|_2 = \|z_j\|_2 = 1, \tilde{z}_i \perp z_i$ and $z_j \parallel z_i$, Eq. (A.5) becomes an equality and the gradient magnitude reaches its maximum such that

$$0 \le \left\| \frac{\partial \mathcal{L}_{imc}}{\partial z_i} \right\|_2 \le \begin{cases} 1, & \text{if } N = 0, \\ \frac{\sqrt{2}Me}{N+Me}, & \text{if } N \neq 0. \end{cases} \tag{A.6}$$

We follow the same procedure to calculate the derivative of the original supervised contrastive loss as

$$\begin{aligned}\frac{\partial \mathcal{L}_{sup}}{\partial z_i} &= \frac{\partial}{\partial z_i} \sum_{Z^+} \log[\exp(z_i \cdot \tilde{z}_i) + \sum_{Z^-} \exp(z_i \cdot z_j)] \\ &\quad - \frac{\partial}{\partial z_i} \sum_{Z^+} \log[\exp(z_i \cdot \tilde{z}_i)] \\ &= \sum_{Z^+} \frac{\tilde{z}_i \exp(z_i \cdot \tilde{z}_i) + \sum_{Z^-} z_j \exp(z_i \cdot z_j)}{\exp(z_i \cdot \tilde{z}_i) + \sum_{Z^-} \exp(z_i \cdot z_j)} - \sum_{Z^+} \tilde{z}_i.\end{aligned} \tag{A.7}$$

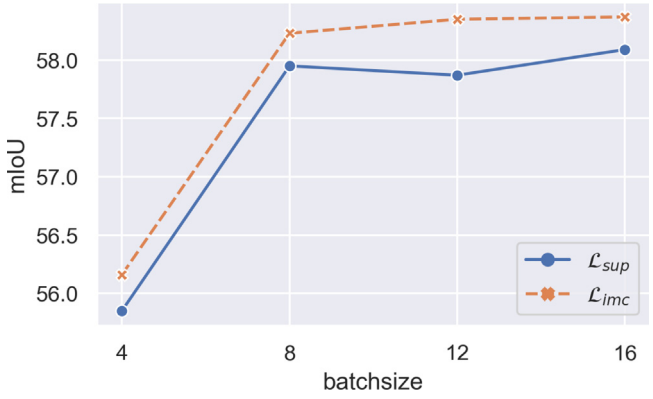

Fig. A.10. CAM performance of $\mathcal{L}_{sup}$ and $\mathcal{L}_{imc}$ under different batchsizes.

When $Z^+ = \varnothing$, *i.e.* $N = 0$, then

$$\frac{\partial \mathcal{L}_{sup}}{\partial z_i} = \mathbf{0}, \tag{A.8}$$

and when $Z^+ \neq \varnothing$, *i.e.* $N \neq 0$, then

$$\left\| \frac{\partial \mathcal{L}_{sup}}{\partial z_i} \right\|_2 \le \sum_{Z^+} \frac{\|z_j - \tilde{z}_i\|_2 \sum_{Z^-} \exp(z_i \cdot z_j)}{\exp(z_i \cdot \tilde{z}_i) + \sum_{Z^-} \exp(z_i \cdot z_j)}. \tag{A.9}$$

When $\|\tilde{z}_i\|_2 = \|z_j\|_2 = 1, \tilde{z}_i \perp z_i$ and $z_j \parallel z_i$, Eq. (A.9) becomes an equality and the gradient magnitude reaches its maximum such that

$$0 \le \left\| \frac{\partial \mathcal{L}_{sup}}{\partial z_i} \right\|_2 \le \begin{cases} 0, & \text{if } N = 0, \\ \frac{\sqrt{2}NMe}{1+Me}, & \text{if } N \neq 0. \end{cases} \tag{A.10}$$

It is proved that the gradient vanishing point of Eq. (A.7) is that there are **no** positive samples within a batch (*i.e.* $N = 0$), and thus the maximum value of their gradient magnitude equals the minimum of 0 in Eq. (A.10). In contrast, our loss $\mathcal{L}_{imc}$ is capable of minimising the negative sample pair similarity regardless of whether there exist positive samples. Simply put, the gradient is still valid and within the range of 0 to 1.

As one of the advantages of our proposed method is removing the Siamese architecture for a more efficient contrastive learning process with no augmented samples added in our training batches, the design of our image level contrastive loss term thus provides better robustness during training with relatively smaller batch sizes compared to those SOTA contrastive learning methods. This is further witnessed when running experiments with four tested batchsizes as illustrated in Fig. A.10.

References

[1] J. Wu, Z. Wen, S. Zhao, K. Huang, Video semantic segmentation via feature propagation with holistic attention, Pattern Recognit. 104 (2020) 107268.
[2] O. Ronneberger, P. Fischer, 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.
[3] K. Yuan, X. Zhuang, G. Schaefer, J. Feng, L. Guan, H. Fang, Deep-learning-based multispectral satellite image segmentation for water body detection, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14 (2021) 7422–7434.
[4] Y. Yang, J. Qiu, M. Song, D. Tao, X. Wang, Distilling knowledge from graph convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 7074–7083.
[5] Y. Jing, Y. Yang, X. Wang, M. Song, D. Tao, Amalgamating knowledge from heterogeneous graph neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 15709–15718.
[6] X. Yang, Z. Daquan, S. Liu, J. Ye, X. Wang, Deep model reassembly, Neural Inform. Process. Syst. (2022).
[7] Y. Yang, Z. Feng, M. Song, X. Wang, Factorizable graph convolutional networks, Neural Inform. Process. Syst. 33 (2020) 20286–20296.
[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[9] J. Ahn, S. Kwak, Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4981–4990.
[10] Y. Wang, J. Zhang, M. Kan, S. Shan, X. Chen, Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 12275–12284.
[11] A. Kolesnikov, C.H. Lampert, Seed, expand and constrain: three principles for weakly-supervised image segmentation, in: European Conference on Computer Vision, Springer, 2016, pp. 695–711.
[12] W. Luo, M. Yang, W. Zheng, Weakly-supervised semantic segmentation with saliency and incremental supervision updating, Pattern Recognit. 115 (2021) 107858.
[13] W. Lu, D. Gong, K. Fu, X. Sun, W. Diao, L. Liu, Boundarymix: generating pseudo-training images for improving segmentation with scribble annotations, Pattern Recognit. 117 (2021) 107924.
[14] A. Khoreva, R. Benenson, J. Hosang, M. Hein, B. Schiele, Simple does it: weakly supervised instance and semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 876–885.
[15] J. Ahn, S. Cho, S. Kwak, Weakly supervised learning of instance segmentation with inter-pixel relations, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2209–2218.
[16] L. Chen, W. Wu, C. Fu, X. Han, Y. Zhang, Weakly supervised semantic segmentation with boundary exploration, in: European Conference on Computer Vision, Springer, 2020, pp. 347–362.
[17] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, M.-H. Yang, Weakly-supervised semantic segmentation via sub-category exploration, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 8991–9000.
[18] O. Henaff, Data-efficient image recognition with contrastive predictive coding, in: International Conference on Machine Learning, PMLR, 2020, pp. 4182–4192.
[19] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
[20] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., Bootstrap your own latent-a new approach to self-supervised learning, Neural Inform. Process. Syst. 33 (2020).
[21] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012.
[22] W. Hui, C. Tan, G. Gu, Y. Zhao, Gradient-based refined class activation map for weakly supervised object localization, Pattern Recognit. 128 (2022) 108664.
[23] Y. Wang, J. Zhang, M. Kan, S. Shan, Learning pseudo labels for semi-and-weakly supervised semantic segmentation, Pattern Recognit. (2022) 108925.
[24] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, S. Yan, Object region mining with adversarial erasing: a simple classification to semantic segmentation approach, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1568–1576.
[25] J. Lee, E. Kim, S. Yoon, Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 4071–4080.
[26] B. Zhang, J. Xiao, Y. Wei, K. Huang, S. Luo, Y. Zhao, End-to-end weakly supervised semantic segmentation with reliable region mining, Pattern Recognit. 128 (2022) 108663.
[27] M. Noroozi, P. Favaro, Unsupervised learning of visual representations by solving jigsaw puzzles, in: European Conference on Computer Vision, Springer, 2016, pp. 69–84.
[28] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, Neural Inform. Process. Syst. 33 (2020) 18661–18673.
[29] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, H. Hu, Propagate yourself: exploring pixel-level consistency for unsupervised visual representation learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 16684–16693.
[30] Y. Jing, Y. Liu, Y. Yang, Z. Feng, Y. Yu, D. Tao, M. Song, Stroke controllable fast style transfer with adaptive receptive fields, in: European Conference on Computer Vision, 2018, pp. 238–254.
[31] S. Liu, Z. Li, J. Sun, Self-EMD: self-supervised object detection without imagenet, arXiv preprint arXiv:2011.13677(2020).
[32] T.-W. Ke, J.-J. Hwang, S.X. Yu, Universal weakly supervised segmentation by pixel-to-segment contrastive learning, in: International Conference on Learning Representations, 2021.
[33] J. Ye, Y. Jing, X. Wang, K. Ou, D. Tao, M. Song, Edge-sensitive human cutout with hierarchical granularity and loopy matting guidance, IEEE Trans. Image Process. 29 (2019) 1177–1191.
[34] X. Chen, K. He, Exploring simple siamese representation learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 15750–15758.
[35] Y. Cao, J. Xu, S. Lin, F. Wei, H. Hu, GCNet: non-local networks meet squeeze-excitation networks and beyond, in: International Conference on Computer Vision Workshops, 2019. pp. 0–0
[36] Y. Rubner, C. Tomasi, L.J. Guibas, A metric for distributions with applications to image databases, in: IEEE International Conference on Computer Vision, IEEE, 1998, pp. 59–66.
[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[38] Y. Li, Y. Song, J. Luo, Improving pairwise ranking for multi-label image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3617–3625.
[39] J.D. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: International Conference on Machine Learning, 2001, pp. 282–289.
[40] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in: IEEE International Conference on Computer Vision, IEEE, 2011, pp. 991–998.
[41] M. Cuturi, Sinkhorn distances: lightspeed computation of optimal transport, Neural Inform. Process. Syst. 26 (2013) 2292–2300.
[42] M. Tan, R. Pang, Q.V. Le, EfficientDet: scalable and efficient object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10781–10790.
[43] J. Lee, E. Kim, S. Lee, J. Lee, S. Yoon, FickleNet: weakly and semi-supervised semantic image segmentation using stochastic inference, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5267–5276.
[44] P.-T. Jiang, Q. Hou, Y. Cao, M.-M. Cheng, Y. Wei, H.-K. Xiong, Integral object mining via online attention accumulation, in: IEEE International Conference on Computer Vision, 2019, pp. 2070–2079.
[45] P.-T. Jiang, L.-H. Han, Q. Hou, M.-M. Cheng, Y. Wei, Online attention accumulation for weakly supervised semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2021).
[46] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, Y. Wei, LayerCAM: exploring hierarchical class activation maps for localization, IEEE Trans. Image Process. 30 (2021) 5875–5888.
[47] Y. Liu, Y.-H. Wu, P. Wen, Y. Shi, Y. Qiu, M.-M. Cheng, Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 44 (3) (2022) 1415–1428.

Kunhao Yuan received his BSc and MSc degree from Northeastern University, China, 2017 and Loughborough University, 2018 respectively. From later 2018 to first quarter of 2020, he was an algorithmic engineer in UnionBigData, Chengdu, China. He is now a PhD student at Loughborough University. His research interests include computer vision, self-supervised learning and representation learning.

Gerald Schaefer gained his PhD in Computer Vision from the University of East Anglia. He worked at the Colour & Imaging Institute, University of Derby, in the School of Information Systems, University of East Anglia, in the School of Computing and Informatics at Nottingham Trent University, and in the School of Engineering and Applied Science at Aston University before joining the Department of Computer Science at Loughborough University. His research interests are mainly in the areas of computer vision, colour image analysis, medical imaging, and computational intelligence. He has published extensively in these areas with a total publication count of about 500, has been invited as keynote or tutorial speaker to numerous conferences, is the organiser of various international workshops and special sessions at conferences, and the editor of several books, conference proceedings and special journal issues.

Yu-Kun Lai is a Professor at School of Computer Science and Informatics, Cardiff University, UK. He received his bachelor's and PhD degrees in Computer Science from Tsinghua University, in 2003 and 2008 respectively. His research interests include computer graphics, computer vision, geometric modeling and image processing. He is on the editorial boards of Computer Graphics Forum and The Visual Computer. For more information, visit https://users.cs.cf.ac.uk/Yukun.Lai.

Yifan Wang was born in Hunan Province, in 1994. He received the BS degree in Information Management and Information System from Beijing Technology and Business University in 2016 and received the MS degree in Computer Science from Central South University in 2019. He is currently a PhD candidate in Loughborough University. His research interests include information security, Deep Learning and Computer Vision.

Xiyao Liu was born in Hunan Province, in 1987. He received the BS degree and the PhD degree from School of Electrical Engineering and Computer Sciences, Department of Microelectronics, Peking University in 2008 and 2015. He is now an associate professor in School of Computer Science and Engineering, Central South University, Changsha, China. His research interests include information security, multimedia technology, computational vision, and deep learning.

Lin Guan is a Senior Lecturer in the Department of Computer Science at Loughborough University. Her research interests focus on performance modelling/evaluation of heterogeneous computer networks and systems (or system of systems); QoX-QoS (Quality of Service, Quality of Resilience, Quality of Experience) analysis, provisioning and enhancements; caching, edge/fog Computing; vehicular ad-hoc networks (VANET); software defined networks (SDN); cloud computing and security, mobile computing, wireless and wireless sensor networks; multimedia systems and Model Based System Engineering (MBSE)with QoS attributes. She has published over 100 journal and conference papers and she has been serving as guest editor for several international journals, such as those published by Elsevier and Springer. She is currently on Editorial Board of Elsevier Journal of Systems and Software (ranked #2 for SE venues in Google Scholar) and Editor in Elsevier Simulation Modelling Practice and Theory. During her PhD, she was awarded the British Federation of Women Graduates Foundation Main Grant in 2004. She then held two EPSRC/industry CASE awards, one EPSRC/BAE EngD projects and two industrial subcontracts on feasibility study and consultancy. She received a prestigious award as Royal Society Industry Fellow and EPSRC KTA project. One of her current projects works on EPSRC/Rolls Royce funded 4 years project Model Based System Engineering (MBSE) with QoS Attributes.

Hui Fang received the BS degree from the University of Science and Technology, Beijing, China, in 2000 and the PhD degree from the University of Bradford, U.K., in 2006. He is currently with the Computer Science Department at Loughborough University. Before, he has carried out research at several world-leading universities, such as University of Oxford and Swansea University. His research interests include computer vision, image/video processing, pattern recognition, machine learning, data mining, scientific visualisation, visual analytics, and artificial intelligence. During his career, he has published more than 100 journal and conference papers.