MuSCLe: A Multi-Strategy Contrastive Learning Framework for Weakly Supervised Semantic Segmentation

Kunhao Yuan 1  Gerald Schaefer 1  Yu-Kun Lai 2  Yifan Wang 1  Xiyao Liu 3  Lin Guan 1  Hui Fang 1
Loughborough University 1  Cardiff University 2  Central South University 3

Abstract

Weakly supervised semantic segmentation (WSSS) has gained significant popularity since it relies only on weak labels such as image level annotations rather than 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 the current state-of-the-art on the widely used PASCAL VOC 2012 dataset.

1. Introduction

Semantic segmentation is a well-established computer vision task that aims to assign object labels to individual pixels of an image. Various applications, such as autonomous driving, precision agriculture and medical image analysis [14, 36, 37], have become possible thanks to recent developments in deep learning (DL)-based semantic segmentation. However, generalising DL models to wider applications is difficult since it requires high-quality pixel level annotations that are costly to obtain. To address this issue, weakly supervised semantic segmentation (WSSS) approaches use inexpensive weak labels, typically image level annotations, to achieve fine-grained semantic segmentation [47, 38, 23, 1, 43, 6].

Since the introduction of class activation maps (CAMs) [47] a lot of effort has focussed on improving DL-based WSSS. One type of approach is to introduce extra cues, such as points [4], scribbles [27, 42], or bounding boxes of objects [33, 11, 21], to yield stronger constraints to supervise the learning. Another group of methods utilises either global context correlations [43] or local pixel correlations [2, 1, 48] 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 [23, 2, 1, 7]. 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 [43] or sub-category clustering in feature space [6] to enhance object representations, the distinctiveness between contrastive object sample pairs is still under-explored.

Inspired by recent success of contrastive learning frameworks [32, 17, 8, 15], 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 Figure 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 contributions 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 the current state-of-the-art in WSSS on the PASCAL VOC 2012 dataset [13], and does so with a lighter network architecture (i.e., fewer tuneable parameters).

We will release the code for MuSCLe on GitHub after the blind review process.

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 [23, 1, 43]. 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 points [4], scribbles [27], or bounding boxes [21].

To achieve pixel level segmentation from image level annotations, [23] 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 saliency maps, some recent approaches utilise class activation maps as pseudo labels [47, 38]. However, early CAM methods only highlight the most distinctive parts of objects, leading to insufficient segmentation performance. More recent approaches explore adversarial erasing [44], adjacent affinity transformations [2, 1], self-supervised attention [43], sub-category mining [6], and boundary exploration [7] to enhance the quality of pseudo labels.

2.2. Contrastive learning

Contrastive learning [32, 3, 17] originates from self-supervised learning [12, 31, 35] 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, [22] uses image labels to improve the learned features by maximising the distances between paired samples belonging to different classes (i.e., positive samples) while minimising those of the same object pairs (i.e., negative samples).

Introduction of pixel level contrast allows to place constraints on feature maps for better generalisation [45]. [43] utilises pixel-level contrast from positive samples after geo-transformations to extract so-called equivariant features, while [29] employs 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.
2.3. Boundary enhancement

Exploitation of object boundaries is another promising option to enhance WSSS performance. In [23], a constraint-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. [2] proposed 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. The network designed in [1] predicts pixel displacements and boundary probabilities to directly obtain an affinity matrix for boundary enhancement, whereas in [7], 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, many 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 then 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 to provide supervision on an encoder-decoder-based network for semantic image segmentation. Inspired by the recent success of employing contrastive learning to improve feature representations, we design a multi-level contrastive learning framework to enforce multiple constraints to learn more reliable and distinctive feature representations. 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 Figure 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 \( x_i \in X \) and its label \( y_i = (y_i;1, y_i;2, \ldots, y_i;K) \), a \( K \)-dimensional multi-hot vector to represent the presence of \( K \) objects, in each training batch \( x_i \) and \( x_j \) form a positive pair if \( y_j = y_i \). Conversely, \( x_i \) and \( x_j \) form a negative pair if \( y_j \cap y_i = \emptyset \).

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

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

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 [17, 8], our \( \mathcal{L}_{\text{imc}} \) 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 [22] in the Supplemental Material.

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 [45], 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 Figure 2, we obtain 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 [5, 43].

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

\[
M' = SAM(M) = \text{softmax}(g_1(M)^T \times g_2(M)) \times g_3(M),
\]

where...
where $g_1(\cdot), g_2(\cdot),$ and $g_3(\cdot)$ denote individual linear projections and $M$ is the CAM response map. The alignment between CAM and SAM in our contrastive loss further improves the generalisation of the feature representations.

The loss term we employ is

$$L_{\text{pixc}} = -\frac{1}{HW} \sum_{k=1}^{HW} \cos(u_k, \text{sg}(v'_k)),$$  

(3)

where $u'_k, v_k$ are feature vectors from overlapping regions in $M'$ and $\tilde{M}$, respectively, $H$ and $W$ are the height and width of the regions and $\text{sg}(\cdot)$ denotes the stop gradient operator which avoids interference of cross-optimisation [15, 9]. This design also reduces the computational cost, thus improving the efficiency of our method.

### 3.2.3 Pairwise regional contrast

In addition to 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 Figure 3, to provide sufficient flexibility while keeping complexity low, we divide the CAM response map into static (e.g. $2 \times 2$) 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. We use the earth mover’s distance [10, 29] to match and compare the feature maps from paired patches.

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

$$M_{\text{bg}} = 1 - \max_{1 \leq c \leq C-1} (M_c),$$  

(4)

with

$$M_c = \frac{\exp(M_c)}{\sum_{1 \leq c \leq C-1} \exp(M_c)},$$  

(5)

where $M_c$ belongs to one of the $C - 1$ foreground activation maps and $M_{\text{bg}}$ represents the estimated background activation map. The concatenation of all foreground activation maps and the background activation map yields the background-included CAM ($M \in \mathbb{R}^{C \times H \times W}$). We refer the reader to [46] for details on marginal weight generation in the EMD computation.

Our pairwise regional contrastive loss term is defined as

$$L_{\text{prc}} = \arg\min_{(a,b)} \text{EMD}(p_a, \text{sg}(\tilde{p}_b)),$$  

(6)

where $p_a \in \{p\}_1^A \subset M'$ and $\tilde{p}_b \in \{\tilde{p}\}_1^B \subset \tilde{M}$ are feature vectors from paired patches generated from the original feature maps $M'$ and $\tilde{M}$, respectively. As for pixel level contrastive loss, we use the stop gradient operator $\text{sg}(\cdot)$ 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) [30], focal loss [28], and a pair loss [26] to address sample imbalance and over-confidence of negative sample issues, and combine them to form a hybrid classification loss (HCL) term

\[ L_{hcl}(y, \hat{y}) = L_{bce}(y, \hat{y}) + L_{focal}(y, \hat{y}) + L_{pair}(y, \hat{y}), \]

(7)

which improves WSSS performance compared to using individual terms.

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

\[ L_{MCL} = L_{hcl} + L_{imc} + L_{pixc} + L_{prec}. \]

(8)

3.3. BEACON decoder

When the decoder of a typical WSSS network is trained, 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 [2] or conditional random fields [24], 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 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.

![Figure 4](https://example.com/fig4.png)

Figure 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.

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 [18] 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 Figure 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 quantise them to 8 directions corresponding to the 8-neighborhood of a pixel. 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 (Figure 4(c)) and the dense map from the decoder to define a boundary contrastive loss term. As illustrated in Figure 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 \( \text{Sign} (\cdot) \) to identify if the similarity values calculated from the dense feature map \( \hat{y} \) coincide with those from the soft pseudo mask \( y \) by comparing their scores to a threshold \( \tau \).

![Algorithm 1](https://example.com/alg1.png)

**Algorithm 1**: BEACON algorithm. InOutDiv(·) is described in the Supplemental Material.

**Input**: orientation map \( M \); dense feature map \( \hat{y} \); soft pseudo mask \( y \); parameters \( \text{steps}, k \)

**Output**: BEACON loss \( L_{\text{beacon}} \)

// select in/out-ward point sets from \( M \):
\( \Phi, \Psi \leftarrow \text{InOutDiv}(M, \text{steps}); \)
randomly select \( k \) samples from \( \Phi, \Psi \) as \( I, O; \)
// anchor \( I, O \) back onto \( y \) and \( \hat{y} \) to yield inward set \( (I^d, I^m) \) and outward set \( (O^d, O^m) \) on dense feature map and pseudo mask:
\( I^m \leftarrow y^I; y^O;\)
\( I^d \leftarrow \hat{y}^I; \hat{y}^O;\)
// calculate similarity matrix \( S \) for the two sets:
\( S^d \leftarrow f(S \circ (I^d), O^d); \)
\( S^m \leftarrow f(I^m, O^m); \)
// obtain point-wise signs:
\( \text{sign}^I, \text{sign}^O \leftarrow \text{Sign}(S^m_{i,o}, S^d_{i,o}); \)
// calculate and return loss:
\( L_{\text{beacon}} \leftarrow \frac{1}{|I|} \sum_{o \in O} \log(\text{sign}^O \cdot \frac{1}{|I|} \sum_{i \in I} S^d_{i,o}) + \frac{1}{|I|} \sum_{i \in I} \log(\text{sign}^I \cdot \frac{1}{|O|} \sum_{o \in O} S^d_{i,o}); \)
The sign determines the direction of optimisation imposed on similarity as shown in Algorithm 2. 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 cases, point-wise signs are obtained and are used to calculate the boundary contrastive loss.

The loss function for this training stage is expressed as

\[
L_{seg} = L_{ce} + \lambda L_{beacon}
\]

where parameter \( \lambda \) allows for balancing between the global pixel-wise cross entropy loss \( L_{ce} \) and the near-boundary pixel representation enhancement.

### 4. Experimental Results

#### 4.1. Implementation details

Our experiments are conducted on the PASCAL VOC 2012 dataset [13] with 20 foreground classes and 1 background class. Following [16, 43], 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 pairwise cost matrix and use Sinkhorn iteration [10] for fast computation of the EMD.

Our MuSCL encoder comprises an EfficientNet [39] and a BiFPN decoder [40]. 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 [34]. The input of image level contrast and classification head is resized while keeping the original image aspect ratio and padded to \( 448 \times 448 \), while the pixel level contrast and pairwise regional contrast heads use random crops of size \( 224 \times 224 \) as inputs. CRF and affinity refinement [2] are executed after SAM output to generate pseudo masks.

#### 4.2. Improved CAM quality

We quantitatively evaluate the effectiveness of each component of our MuSCL approach in Table 1. From there, it is evident that each proposed module leads to a notable performance increase. Following common practice [2, 43, 6], 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 we can see, our MCL clearly outperforms the other methods, improving the CAM of AffinityNet [2] by 10.4% and the result of SEAM [43] by 3.0%. Although the improvement with affinity refinement is less pronounced compared to that for raw CAMs, we still obtain the highest mIoU of 64.6%. We conjecture that this is because our CAM trained by MCL is denser and more continuous, and thus the affinity transformation barely enhances local feature representation with adjacent context.

In addition, we have visualised the learned representations from SEAM and our MCL using t-SNE dimensionality reduction [41]. As can be seen from the obtained results in Figure 5, the classes are better separated in the MCL visualisation, while for SEAM we observe significant overlap between classes and higher variation within each class.
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 2.0%/1.3% boost with/without BEACON.

We perform a thorough ablation study, with results listed in Table 4, on our BEACON module to show the impact of different hyper-parameters and the effectiveness of BEACON. Since larger values of \( \lambda \) 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 \( \lambda \) 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 \( \tau \), a halfway division of the similarity scores (i.e., \( \tau = 0.5 \)) provides only a small improvement compared to a dynamic threshold, \( \mu_m \), 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 good performance/efficiency trade-off. Overall, the best results are obtained by combining \( \mu_m \) 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, tuneable parameters and supervision in Table 5. From there, we see that on the val set we achieve performance on-par with SC-CAM [6] with merely 40% of their learnable parameters, while outperforming all other methods. We also notice that even the smaller MuSCLe-b5 outperforms most other SOTA methods.

Importantly, we also observe MuSCLe to yield better performance on the test set. With an mIoU of 66.7% it is slightly superior to the previous best method of [7] and this

| backbone | # BiFPN | BEACON | mIoU |
|----------|---------|--------|------|
| EfficientNet-b3 | 1 | no | 60.7 |
| EfficientNet-b3 | 2 | no | 61.3 |
| EfficientNet-b3 | 3 | no | 63.2 |
| EfficientNet-b5 | 3 | no | 63.8 |
| EfficientNet-b7 | 3 | no | 64.5 |
| EfficientNet-b3 | 3 | yes | 64.1 |
| EfficientNet-b5 | 3 | yes | 65.2 |
| EfficientNet-b7 | 3 | yes | **66.1** |

Table 3. Ablation study for segmentation network architecture. mIoU reflects segmentation performance on val set.

| \( \lambda \) | steps | \( k \) | \( \tau \) | single scale mIoU | multi scale mIoU |
|-------------|-------|------|-------|----------------|----------------|
| 0           | n/a   | n/a  | n/a   | 60.3           | 60.5           |
| 0.05        | 7     | 128  | 0.5   | 60.5           | 60.4           |
| 0.05        | 7     | 64   | \( \mu_m \) | 60.1           | 60.6           |
| 0.05        | 7     | 128  | \( \mu_m \) | **61.7**       | 61.6           |
| 0.05        | 7     | 256  | \( \mu_m \) | 60.7           | 61.5           |
| 0.05        | 5     | 128  | \( \mu_m \) | 60.0           | 61.7           |
| 0.05        | 5     | 256  | \( \mu_m \) | 59.7           | 61.1           |
| 0.1         | 7     | 128  | \( \mu_m \) | 60.4           | 60.9           |

Table 4. BEACON ablation study. mIoU reflects the segmentation performance on val set. \( \mu_m \) denotes the mean of similarity matrix derived from the soft mask. In the result columns, the left and right values denote MuSCLe-b5 and MuSCLe-b7 results, respectively.

| method | parameters | label validation | test |
|--------|------------|------------------|------|
| AffinityNetCVPR2018 [2] | 268.7 | 1 | 61.7 | 63.7 |
| FickleNetCVPR2019 [25] | 306.4 | 1+S | 64.9 | 65.3 |
| OAACCV2019 [19] | 172.8 | 1+S | 65.2 | 66.4 |
| SEAMCVPR2020 [43] | 218.8 | 1 | 64.5 | 65.7 |
| SC-CAMCVPR2020 [6] | 204.5 | 1 | **66.1** | 65.9 |
| BESSECCV2020 [7] | 105.0 | 1 | 65.7 | 66.6 |
| LayerCAMTP2021 [20] | 218.4 | 1 | 63.0 | 64.5 |
| MuSCLeb5 | 44.9 | 1 | 65.2 | 66.7 |
| MuSCLeb7 | 81.7 | 1 | **66.1** | **67.3** |

Table 5. Comparison with SOTA WSSS methods in terms of mIoU on VOC2012 val and test set. I=image level label; I+S=image level label + saliency map. Parameters are counted on classification network and segmentation network together.
Table 6. Category performance comparison on PASCAL VOC2012 val set.

| method        | bike | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | milk | person | plant | sheep | sofa | train | tv  |
|---------------|------|------|------|------|------|--------|-----|-----|-----|-------|-----|-------|-----|-------|------|--------|------|-------|------|-------|-----|
| AffinityNet [2] | 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 [25] | 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 [43]      | 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 [6]     | 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 [7]        | 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 |
| MuSCLe-b7      | 88.4 | 75.5 | 31.8 | 76.3 | 50.3 | 71.7   | 85.8| 77.4 | 72.1 | 33.4  | 90.2 | 49.1  | 75.6| 82.8  | 73.6 | 70.4  | 49.0 | 85.9  | 47.7 | 65.3  | 36.1 |

with 57% fewer parameters, based on an EfficientNet-b5 backbone. Further deepening the model to an EfficientNet-b7 structure, we obtain a new best mIoU of 67.3% despite requiring fewer parameters than previous approaches and using only image level labels (in contrast to some other methods such as [25, 19] which rely on stronger supervision based on image labels in combination with saliency maps).

Looking at the class-wise performance on the val set in Table 6, MuSCLe gives the best result for 6 classes, more than any other method (SC-CAM is best for 4 categories). In particular, for the cow, sheep, and sofa 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 TVs in the VOC dataset often appear together with other objects such as benches and tables. Further insight and examples are discussed in the Supplemental Material.

### 4.5. Qualitative segmentation results

We show some representative qualitative segmentation results obtained from MuSCLe-b7 in Figure 6, from which we can observe that detailed object boundaries are properly recovered (columns 1-2). For multi-label scenarios (columns 3-4), our model correctly distinguishes the instances of each category, while multiple object instances at different scales and locations are also recognised (column 5), demonstrating the efficacy of the dynamic cropping and matching strategy.

Figure 7 gives some typical samples from the SBD dataset [16], 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.

### 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, with significantly fewer parameters, MuSCLe achieves SOTA performance on the PASCAL VOC2012 dataset, while ablation studies and visualisations further illustrate the efficacy and efficiency of our proposed approach. Notably, this is achieved on a single GPU, unlike most existing work in the area. In future work, we will investigate extending our model to multi-GPU settings to further boost the performance.
References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2209–2218, 2019. 1, 2, 3, 7

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 4981–4990, 2018. 1, 2, 3, 5, 6, 7, 8

[3] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In Neural Information Processing Systems, pages 15535–15545, 2019. 2

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What’s the point: Semantic segmentation with point supervision. In European Conference on Computer Vision, pages 549–565. Springer, 2016. 1, 2

[5] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Genet: Non-local networks meet squeeze-excitation networks and beyond. In International Conference on Computer Vision Workshops, pages 0–0, 2019. 3

[6] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In IEEE Conference on Computer Vision and Pattern Recognition, pages 8991–9000, 2020. 1, 2, 6, 7, 8

[7] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In European Conference on Computer Vision, pages 347–362. Springer, 2020. 1, 2, 3, 7, 8

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning, pages 1597–1607. PMLR, 2020. 1, 3

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In IEEE Conference on Computer Vision and Pattern Recognition, pages 15750–15758, 2021. 4

[10] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. Neural Information Processing Systems, 26:2292–2300, 2013. 4, 6

[11] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In IEEE International Conference on Computer Vision, pages 1635–1643, 2015. 1

[12] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In IEEE International Conference on Computer Vision, pages 1422–1430, 2015. 2

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 2, 6

[14] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Gläsner, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems, 22(3):1341–1360, 2021. 1

[15] Jean-Bastien Grill, Florian Strub, Florent Aléch, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. Neural Information Processing Systems, 33, 2020. 1, 4

[16] Bharath Hariharan, Pablo Arbeláez, Labonir Bourdev, Subhransu Maji, and Ittenda Malik. Semantic contours from inverse detectors. In IEEE International Conference on Computer Vision, pages 991–998. IEEE, 2011. 6, 8

[17] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In International Conference on Machine Learning, pages 4182–4192. PMLR, 2020. 1, 2, 3

[18] FG Irwin et al. An isotropic 3x3 image gradient operator. Presentation at Stanford AI Project, 2014(02), 1968. 5

[19] Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, and Hong-Kai Xiong. Integral object mining via online attention accumulation. In IEEE International Conference on Computer Vision, pages 2070–2079, 2019. 7, 8

[20] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. LayerCAM: Exploring hierarchical class activation maps for localization. IEEE Transactions on Image Processing, 30:5875–5888, 2021. 7

[21] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 876–885, 2017. 1, 2

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. arXiv preprint arXiv:2004.11362, 2020. 2, 3

[23] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In European Conference on Computer Vision, pages 695–711. Springer, 2016. 1, 2, 3

[24] John D. Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In International Conference on Machine Learning, pages 282–289, 2001. 5

[25] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Bernardo Avila Pires. Improving pairwise ranking for multi-label image classification. In IEEE Conference on Computer Vision and Pattern Recognition, pages 5267–5276, 2019. 7, 8

[26] Yuncheng Li, Yale Song, and Jiebo Luo. Improving pairwise ranking for multi-label image classification. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3617–3625, 2017. 5

[27] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribblesupervised convolutional networks for semantic segmentation. In IEEE Conference on Computer
