1. Cover Form

Insert the title here

■ Authors: Jungsun Im, Subin An, Soochahn Lee

■ Corresponding Author: Soochahn Lee

■ Affiliation: Department of Electronics Engineering, Kookmin University

■ Please check your area of study below.

⃞ Broadcast content – voice/audio ⃞ Broadcast content - video/image ⃞ Content management and application

⃞ Broadcasting system ⃞ Broadcasting system - modulation and transmission

⃞ Broadcasting communication convergence ⃞ Broadcasting industry technique

■ Paper classification

⃞ Experimental paper: A paper that implements the content of the paper's proposal and presents its experimental results.

⃞ Production paper: A paper that introduces the results and effects of producing systems, devices, performances, etc. by synthesizing various existing technologies.

⃞ Standard paper: A paper introducing the standard's architecture, use cases, requirements, standard specifications, etc.

⃞ Survey paper: A paper that summarizes and classifies the research contents of existing research or technologies.

⃞ Policy paper: A paper on policy research and proposals for industrial development and innovation in the field of broadcasting media technology.

■ Key words (maximum of five key words in English)

weakly supervised semantic segmentation, affinity enhancement, prototype exploration, self-supervised learning, image-specifc

■ Address including zip code : 02707

■ Phone number: cell phone number: 01024395865

■ E-mail address: ijs\_1016@kookmin.ac.kr

■ Please indicate the type of paper below.

⃞ Special Paper (Please specify the title.)

⃞ Special Paper(Broadcast and Media Technology(영문특집호))

⃞ Regular Paper

⃞ Letter Paper (short, four-page paper; it can be submitted on-line as a J2 Letter Paper.)

2. Requirements for Submitting a Paper

■ All authors of a paper must register for membership and pay the membership fee.

Membership site:

<http://acoms.atit.co.kr:8090/ksbembc/index.jsp?publisher_cd=ksbembc&cid=&cid_year=2013&cid_seq=A&lang=kor>

■ Account number for payment: Citibank Korea, 124-50340-241 (The Korea Institute of Broadcast and Media Engineers)

3. Submission of Paper

■All authors must comply with the instructions regarding how to submit a paper.

- The instructions are provided in our Online Submission System.

- All pages should be in one column and numbered.

- Figures and tables should be included in the body of the paper; their titles and contents should be written in English.

- To comply with the Korean Federation of Science and Technology Societies, the references should be written in English according to the instructions on how to submit a paper. The titles of Korean journals should be translated into English.

- The summary should be approximately 500 Korean characters in order to register it with Scopus. It should not be longer or shorter than the designated length.

■ Before uploading the paper, please check the spelling and word spacing.

- The file to be submitted should be in earlier versions of Hangul 2010 or in Word 2010.

**Author Checklist**

**Title of the paper: AFFINITY ENHANCED IMAGE-SPECIFIC PROTOTYPES FOR**

**WEAKLY SUPERVISED SEMANTIC SEGMENTATION**

**Before you submit your manuscript, please review the following items and mark applicable items with ✔.**

1. General details of preparation

☑ The manuscript includes the original copy of the manuscript, a checklist, an agreement for copyright transfer, and an author introduction.

☑ The manuscript is prepared in Korean or English, and English terms are used when necessary (provided that paper titles, author names, institutions to which authors belong, abstracts, and explanations of figures and tables in English accompany those in Korean).

☑ The manuscript begins with a cover page followed by a title page, an abstract in Korean, an abstract in English, the body (including necessary figures and tables), an appendix, and references in order of precedence.

☑ The body begins with an introduction and ends with conclusions.

☑ The manuscript is prepared using a PC with a Hangeul word processor in a single column, and the length of the manuscript is less than 12 pages (A4 size) including figures and tables (around 1,000 letters per page, double spaced).

☑ In manuscripts in English, only the first letter in each sentence is capitalized (except for proper nouns), and the remaining letters are written in lowercase.

2. Cover page

☑ On the cover page of the manuscript, the article title (including the title in English) and detailed area of the study are provided. Moreover, the submitter’s address, postal code, phone number, and e-mail address must be written.

☑ In the case of names in English, given names are written first followed by family names.

3. Title page

☑ The first page (the page after the cover page) is a title page, and the names of the authors or the organization to which the authors belong are not written; only the title (including the title in English) is written.

4. Abstract and keywords

☑ Both an abstract in Korean (less than 500 letters) and an abstract in English (less than 200 words) are included in the paper.

☑ Five or fewer keywords are written below the abstract in English.

5. Body

☑ Page numbers are included in the paper. Numbers for chapters are written in Roman numerals (Ⅰ, Ⅱ, Ⅲ, Ⅳ, …), and numbers for sections are written in Arabic numerals (1, 2, 3, …).

☑ Figures and tables are presented on the pages in which they are mentioned if possible; however, if they must be presented on a separate page, they are presented at the beginning of the next page. The titles of figures are written below the figures, and the titles of tables are written above the tables in both Korean and English.

☑ With regard to figures or photos, a sufficiently high resolution is maintained so that they can be identified.

6. References

☑ References are written in the order in which they are cited in the body in the manner shown below for both those in Korean and in English, and journal and book names are italicized.

A. Journal: author name, article title, journal name, volume, issue, page, month, year (year followed by month in the case of references in Korean)

B. Book: author name, book name, publishing company name, location, page, year

☑ All references with English descriptions (not the whole paper, just title, author names, etc.) are written in English, and references without English descriptions are written in both Korean and English by translating the Korean into English.

☑ The numbers of references cited in the body shall be written in superscript and in square brackets after cited sentences[1].

7. Ethics

☑ The author(s) of this article has (have) checked the Korean Institute of Broadcast and Engineers’ Internal Regulations for Paper Ethics and will observe the ethics regulations.

약지도 의미론적 영상 분할을 위한 개별 이미지 별 유사성 강화 프로토타입 기법

AFFINITY ENHANCED IMAGE-SPECIFIC PROTOTYPES FOR

WEAKLY SUPERVISED SEMANTIC SEGMENTATION

요 약.

많은 약지도 학습 의미론적 영상 분할 방법은 이미지 분류를 위해 훈련된 네트워크에서 추출한 픽셀 수준 피쳐를 활용한다. 클래스 활성화 맵 생성, 픽셀 유사성을 특징 유사성으로 정의하고, 특징 군집을 기반으로 클래스 별 특징 프로토타입을 생성하는 데 사용 된다. 본 논문은 이전 연구들을 향상시키기 위해 친화도 기반 세분화를 이미지 별 클래스 별 프로토타입 생성에 통합하는 방법을 제안하며, 이로 인해 대표적인 강도가 크게 향상 된다. 이러한 프로토타입은 개선된 의사 레이블을 만들어내며, 궁극적으로 세그멘테이션을 개선한다. 실험 결과는 기준선 방법과 비교하여 상당한 개선이 있으며, 최근 최첨단 방법과 유사한 수준의 결과를 나타낸다. 논문 코드는 <https://github.com/IJS1016/AE_SIPE>에서 제공한다.

Abstract

Many weakly supervised semantic segmentation methods rely on the pixel-level features extracted from networks trained for image classification. These features can be used to create class activation maps for semantic scores, define pixel affinity as feature similarities, and construct per-class feature prototypes based on feature clustering. This paper proposes a method that enhances previous works by incorporating affinity-based refinement into the generation of image-specific per-class prototypes, resulting in significantly improved representative strength. These prototypes then lead to improved pseudo-labels, ultimately improving segmentations. Experimental results show significant improvements compared to baseline methods, and are on par with recent state-of-the-art methods. The code is available at <https://github.com/IJS1016/AE_SIPE>

Keywords

weakly supervised semantic segmentation, affinity enhancement, prototype exploration, self-supervised learning, image-specifc

I. Introduction

The goal of weakly supervised semantic segmentation (WSSS) is to learn how to generate pixel-level labels from limited supervision, usually in the form of image-level class labels [2]. The introduction of the Class Activation Map (CAM) [3] was a significant advancement towards achieving this goal, as it provides a means of generating pixel-level per-class scores based on image classification. However, it has been observed that meaningful CAM scores are often only assigned to a selective number of the most discriminative pixels, leading to limitations in directly using CAM as a segmentation solution.

Nonetheless, CAM proves to be a highly efficient technique for utilizing image-level annotations to make pixel-level predictions. It has frequently served as a base upon which multiple methods have been proposed to enhance and optimize the acquisition of pixel-level class probabilities.

One approach is to erase [5] or suppress [6] the more discriminative regions, further mine discriminative pixels. Another approach is to assign the limited discriminative regions as seeds and expand them into full segmentation labels using conventional region growing algorithms [7, 8], based on the similarities of local pixel values. Further methods extended this approach by incorporating pixel adaptive refinement [4], random walks on semantic features [9], or multitask inference of displacement and class boundary [9, 10].

Many recent methods are based on self-supervised learning. A contrastive learning framework, with positive image pairs defined by pairing an image with its linear transform and negative pairs of different images, were applied in [11, 12].

Another approach uses network features to create a per-class feature prototype-based alternative score map, providing supervision to guide the network towards generating consistent features with pixel affinities and image-level class labels [1].

Combining these methods with others has shown benefits, as seen in recent works [11, 12, 1]. The improved CAM-like score maps generated by these methods are used to enhance pixel affinities and generate pseudo-labels [13, 10], which are used to train a fully supervised semantic segmentation network [14].

In this paper, we propose a method to incorporate pixel-adaptive mask refinement (PAMR) [4] so that pixel affinity is maximized when generating score maps within the self-supervised image-specific prototype exploration (SIPE) method [1]. Experimental results demonstrate that our proposed method provides substantial improvements over the baseline method SIPE. We also propose additional modifications that further improve quantitative results.

II. PROPOSED METHOD

1. Framework

A visual summary of the proposed method is presented in Fig. 2. The baseline method, SIPE [1], comprises 1) the encoder module which generates pixel-level features, 2) the image classification module, which provides image-level supervision and generates the CAM, 3) the structure analysis module, which generates semantic structure seeds, and 4) the prototype module, which generates image-specific per-class prototype features and pixel-level per-class scores, denoted as image-specific CAM (IS-CAM). We note that we are using our own terminology, which we believe provides a more intuitive understanding of the framework.

In the proposed method, we incorporate the affinity enhancement (AE) module to the framework. In the AE module, a refined IS-CAM is generated, which is then used to generate refined region seeds, which are used to refine the prototypes and generate an improved IS-CAM. To aid the description of the AE module in 2.2, we provide a brief summary of the modules of SIPE [1] as follows:

**Encoder** comprises a backbone CNN, pre-trained on image classification. The feature tensor generated from this encoder E for the input image I is denoted as F = E(I), and each feature vector at grid coordinate (i, j) is denoted as fij.

**Classification** comprises a layer to compute the CAM, and the global average pooling layer to connect the CAM with the image-level supervision through the classification loss.

**Structure analysis** relates the spatial distribution of features to the CAM to create seeds for prototypes. The pixel-wise spatial structure of grid coordinate (i, j) is first defined as Sij = ReLU(Sc(fij, F)), with the cosine similarity function Sc(·) being broadcast for the elements of F. This is then compared to CAM to determine the semantic structure seed label, SSij = arg maxk IoU(Sij, Mk) denoting intersection-over-union with the CAM Mk for the kth class.

**Prototypes** pk are generated as pk = 1 |Sk | ∑ (i,j)∈Sk fij, where Sk = {(i, j)|SSij == k} denotes the set of coordinates with SSij == k. That is, pk is the mean of the features with seed label k. The IS-CAM ̃Mk is defined as ̃Mkij = ReLU(Sc(fij, pk)).

**Training loss** comprises the classification loss, defined as the cross-entropy between ground truth and the inferred image-level labels, and the general-specific consistency (GSC) loss, defined as the pixel-level L1 distance between the initial CAM Mk and refined IS-CAM ̃Mk, for all classes k.

1. Affinity Enhanced Image-specific CAM

Within the AE module, PAMR [4], which is essentially bilateral filtering [15] on the semantic labels, is applied to the refined IS-CAM ̃Mk. PAMR is defined as follows:

PAMR( ̃Mk ij ) = ∑ (p,q)∈N αij,pq ̃Mk pq , (1)

where the affinity kernel αpq is a function based on the differences in image pixel values αij,pq = 1 / W exp ( − ||Iij −Ipq ||2 / σ2 ), with a normalization term W ensuring that ∑ (p,q)∈N αpq = 1. N denotes the local neighborhood of (i, j), which is defined as a combination of multiple 3×3 windows with varying dilation rates.

The further refined CAM ˆMk ij is obtained by iteratively applying PAMR n times, as ˆMk ij = PAMRn( ̃Mk ij ). Using ˆMk ij , we redefine seed labels as ˆSSij = arg maxk ˆMk ij to compute AE prototypes as ˆpk = 1 / | ˆS∥ | ∑ (i,j)∈ ˆSk fij , where ˆSk = {(i, j)| ˆSSij == k}. The final affinity-enhanced IS-CAM, which we term AE-IS-CAM, is computed as ˇMk ij = ReLU(Sc(fij , ˆpk)). Examples that highlight the improvements from Sk to ˆSk and from ̃Mk to ˇMk are depicted in Fig. 1.

1. Additional Modifications

We also apply further minor modifications regarding the details of encoded features, normalization of refined (AE) IS-CAM, and rescaling of background scores. We observed these modifications result in small improvements in the quantitative evaluations.

**Structure Analysis with Hierarchical Features**: In SIPE [1], only features from the last layer (semantic features) are used in structure analysis, while the concatenation of projected features generated from all internal layers (hierarchical features) are used in prototype and IS-CAM generation. However, we use hierarchical features for structure analysis as well as prototype, IS-CAM, and AE-IS-CAM generation.

**IS-CAM Normalization:** As the cosine similarities between features and prototypes may not range from the full range of [0, 1], we apply min-max normalization on the AE-IS-CAM.

**Rescaling of Background Scores**: We observed background scores to be generally higher than the foreground class, as background regions may be more diverse in appearance. We thus rescale the background class activations by a factor of 0.8.

III. EXPERIMENTS

1. Experimental Settings

Implementation: The experiments were conducted on two Titan RTX GPUs, using an implementation based on the source code provided by the authors of [1], built on the PyTorch framework. The encoder module utilized a pre-trained ResNet-101 [16] as the backbone network. Training employed standard SGD optimization with a momentum of 0.9 and weight decay of 1e-4. The learning rate was set to 1e-2 for the pre-trained layers and 1e-1 for the layers in feature concatenation and the final classification layer. The PAMR process underwent 10 iterations, and a set of {1, 2, 4, 8, 12, 24} dilation rates defined N.

**Multi-stage Pipeline**: The complete segmentation pipeline consisted of three stages: 1) the proposed method for initial pseudo label construction, 2) the IRN [10] for refining the initial pseudo labels, 3) the DeepLabV3 [17] trained using the refined pseudo labels.

**Dataset:** The PASCAL VOC 2012 segmentation dataset, widely recognized as the standard benchmark for WSSS, was used. This dataset comprises 21 classes, including the background, with 1,464, 1,449, and 1,456 images in the train, validation, and test sets, respectively. To enhance training, the augmented train set containing 10,582 images [18] was used. Performance was evaluated using mean Intersection over Union (mIoU), and the mIoU score on the VOC test set was computed through the official evaluation server.

2. Comparative Evaluation

Quantitative evaluation results are summarized in Table 1. The incorporation of the AE module alongside the introduced modifications leads to improvements of 2.2% and 1.4% points over the baseline [1] on the validation and test sets, respectively.

Qualitative comparisons against the baseline are depicted in Figures 1 and 3. These examples illustrate instances where the proposed method yields enhanced segmentations by more effectively distinguishing both the background and the semantic classes of foreground objects.

3. Ablative Study

In this section, we delve into the specific effects of each of the proposed components: AE, structure analysis with hierarchical features (HF), IS-CAM normalization, and background rescaling, as part of ablative analysis, presented in Table 2. Our observations highlight that the primary improvements primarily stem from the AE module, with marginal enhancements arising from the supplementary modifications.

We also provide results from various combinations of PAMR, prototype generation, and IS-CAM generation, which constitute the submodules of the AE module, in Table 3. Notably, we found that iterations of PAMR or prototype and IS-CAM generation did not consistently yield improvements. The optimal results were achieved through the proposed AE module.

Ⅳ. DISCUSSION

In the comparative evaluation in Table 1, highlights that the RecurSeed method [23] attains the highest performance.

This method employs PAMR [4] for refining pseudo-labels, coupled with a *self-correlation map generation*(SCG) module [24]. Coincidentally, this SCG process, initially proposed for weakly supervised object localization, bears resemblance to the structure analysis module in SIPE [1].

Upon Further comparison revealed that while prototypes are used to generate pseudo-semantic segmentation labels in the proposed AE-SIPE, RecurSeed employs a decoder to infer these pseudo-labels. Additionally, while iterations improve results in RecurSeed, they do not consistently do so in the proposed method.

We believe that the prototype approach offers simplicity, while the decoder approach may offer greater capacity. There exists a relative scarcity of works that explicitly address the decoder structure within the self-supervised framework for WSSS, warranting further research. Additionally, we aim to identify refinement processes amenable to iteration for enhanced performance improvements.

\* The text within a figure should be written in English.

Figure 1. Enter the figure title (no period is needed).

Table 1. Enter the table title (no period is needed).

|  |  |  |  |
| --- | --- | --- | --- |
| Content in English | Content in English | Content in English | Content in English |
| Content in English |  |  |  |
| Content in English |  |  |  |
| Content in English |  |  |  |

References

[1] Q. Chen, L. Yang, J. Lai, and X. Xie, "Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation," in CVPR, 2022, pp. 4278–4288.

[2] S. Hong, S. Kwak, and B. Han, "Weakly supervised learning with deep convolutional neural networks for semantic segmentation: Understanding semantic layout of images with minimum human supervision," IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 39–49, 2017.

[3] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in CVPR, 2016, pp. 2921–2929.

[4] N. Araslanov and S. Roth, "Single-stage semantic segmentation from image labels," in CVPR, 2020, pp. 4252–4261.

[5] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in CVPR, 2017, pp. 6488–6496.

[6] B. Kim, S. Han, and J. Kim, "Discriminative region suppression for weakly-supervised semantic segmentation," AAAI, vol. 35, no. 2, pp. 1754–1761, May 2021.

[7] A. Kolesnikov and C.H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in ECCV, 2016, pp. 695–711.

[8] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in CVPR, 2018, pp. 7014–7023.

[9] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in CVPR, 2018, pp. 4981–4990.

[10] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in CVPR, June 2019.

[11] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in CVPR, 2020, pp. 12272–12281.

[12] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Weakly supervised semantic segmentation by pixel-to-prototype contrast," in CVPR, 2022, pp. 4310–4319.

[13] P. Kr¨ahenb¨uhl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in NeurIPS, 2011, vol. 24.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE Trans. PAMI, vol. 40, no. 4, pp. 834–848, 2018.

[15] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in ICCV, 1998, pp. 839–846.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[18] B. Hariharan, P. Arbel`aez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in ICCV, 2011, pp. 991–998.

[19] J. Lee, E. Kim, and S. Yoon, "Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4070–4078.

[20] H. Kweon, S.-H. Yoon, H. Kim, D. Park, and K.-J. Yoon, "Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation," in ICCV, 2021, pp. 6974–6983.

[21] F. Zhang, C. Gu, C. Zhang, and Y. Dai, "Complementary patch for weakly supervised semantic segmentation," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 7222–7231.

[22] M. Lee, D. Kim, and H. Shim, "Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds," in CVPR, 2022, pp. 4320–4329.

[23] S. Jo, I.-J. Yu, and K. Kim, "Recurseed and edgepredictmix: Single-stage learning is sufficient for weakly supervised semantic segmentation," 2022.

[24] X. Pan, Y. Gao, Z. Lin, F. Tang, W. Dong, H. Yuan, F. Huang, and C. Xu, "Unveiling the potential of structure preserving for weak