IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020.DOI

Extracting Structured Supervision from Captions for Weakly Supervised Semantic Segmentation

DANIEL R. VILAR[1], AND CLAUDIO A. PEREZ[1], (Senior Member, IEEE)
[1]Department of Electrical Engineering and Advanced Mining Technology Center, Universidad de Chile, Santiago 8370451, Chile

Corresponding author: Claudio A. Perez (e-mail: clperez@ing.uchile.cl).

This work was supported by ANID (Agencia Nacional de Investigación y Desarrollo) under Grants FONDECYT 1191610, and FONDEF ID16I20290, by the Department of Electrical Engineering, and Advanced Mining Technology Center (CONICYT Project AFB180004), Universidad de Chile.

ABSTRACT Weakly supervised semantic segmentation (WSSS) methods have received significant attention in recent years, since they can dramatically reduce the annotation costs of fully supervised alternatives. While most previous studies focused on leveraging classification labels, we explore instead the use of image captions, which can be obtained easily from the web and contain richer visual information. Existing methods for this task assigned text snippets to relevant semantic labels by simply matching class names, and then employed a model trained to localize arbitrary text in images to generate pseudo-ground truth segmentation masks. Instead, we propose a dedicated caption processing module to extract structured supervision from captions, consisting of improved relevant object labels, their visual attributes, and additional background categories, all of which are useful for improving segmentation quality. This module uses syntactic structures learned from text data, and semantic relations retrieved from a knowledge database, without requiring additional annotations on the specific image domain, and consequently can be extended immediately to new object categories. We then present a novel localization network, which is trained to localize only these structured labels. This strategy simplifies model design, while focusing training signals on relevant visual information. Finally, we describe a method for leveraging all types of localization maps to obtain high-quality segmentation masks, which are used to train a supervised model. On the challenging MS-COCO dataset, our method moves the state-of-the-art forward significantly for WSSS with image-level supervision by a margin of 7.6% absolute (26.7% relative) mean Intersection-over-Union, achieving 54.5% precision and 50.9% recall.

INDEX TERMS Image captions, semantic segmentation, weakly supervised.

I. INTRODUCTION

THE goal of semantic segmentation is to classify each pixel in an image with a unique label, chosen from a fixed set of object categories. This task constitutes a fundamental step in a vast number of modern computer vision applications, including autonomous driving [1], medical imaging [2]–[5], scene understanding [6], and remote sensing [7]. In recent years, remarkable success has been achieved in semantic segmentation thanks to developments in the use of Convolutional Neural Networks (CNNs) [8], [9], which have revolutionized the state-of-the-art in this and various other image processing tasks, such as object detection [10], classification [11], [12], and object recognition [13], [14]. However, large volumes of images are usually required to train these models effectively, which in the case of semantic segmentation under a fully-supervised setting means annotating each image with pixel-accurate masks. The prohibitively high cost of manually generating this type of supervision for large datasets severely limits the expansion of modern semantic segmentation models to more diverse and complex application domains.

These limitations have motivated various recent publications that report attempts to train segmentation models using weaker, less expensive supervision, including bounding boxes [15], [16], scribbles [17], points [18], or only image-level supervision [19]–[27]. The latter is particularly challenging, as no localization information is available, but it

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

also presents the most dramatic reduction in annotation costs, and consequently has received significant attention from the research community. Most modern approaches that tackle weakly supervised semantic segmentation (WSSS) with image-level supervision employ a two-stage procedure [21]–[27], leveraging Class Activation Maps (CAMs) [28] produced by a convolutional classifier to generate pseudo-ground truth segmentation masks, and then using those masks to train a supervised segmentation model. Development of these methods has been largely motivated and driven by the existence of several large-scale image datasets annotated with classification labels, such as ImageNet [29], MS-COCO [30], and PASCAL VOC [31]. However, this is not the case for most image domains or object categories, and the cost of annotating datasets of larger scales remains very high, even for this weaker form of supervision.

In this paper, we investigate the use of natural language captions as an alternative form of image-level supervision, which has the advantage of being easily obtainable in large quantities from the web for a wide range of image types [32]. Captions also contain additional information about context and complementary visual attributes, which can be used to improve localization cues. Despite these advantages, only a small number of research initiatives have focused on this task, and existing weakly supervised methods based on classification labels cannot be applied directly to work with natural language supervision.

A possible way to generate localization cues similar to CAMs utilizing image captions was proposed in [33]. Inspired by recent publications on the task of visual grounding [34]–[37], their model [33] projects images and text segments into a shared multi-modal embedding space, which preserves semantic relations. The authors [33] then identified references to relevant object categories in the training captions by simply detecting class names, and localized the corresponding text snippets into the paired images by matching representations in the embedding space. The resulting activation maps were then applied analogously to CAMs to train a supervised segmentation model.

However, by training their model [33] to localize arbitrary text directly, non-visual and irrelevant information is introduced into the supervision, which hinders training. The model also has to learn that multiple grammatical structures can be equivalent, and that multiple different concepts could refer to the same object category. For example, the MS-COCO category "person" could be mentioned with a synonym, e.g., "individual"; a hyponym, e.g., "man"; or a holonym, e.g., "baseball team". In addition to the extra burden that this imposes on the training of the localization model, there is no direct way of retrieving this knowledge to guide the generation of segmentation masks. Thus, a separate method is required to associate text snippets to specific object categories, or to use additional information during this stage.

In this paper, we present a method to address the previously mentioned problems, which is summarized in Fig. 1. Instead of training a model to localize text directly in images,

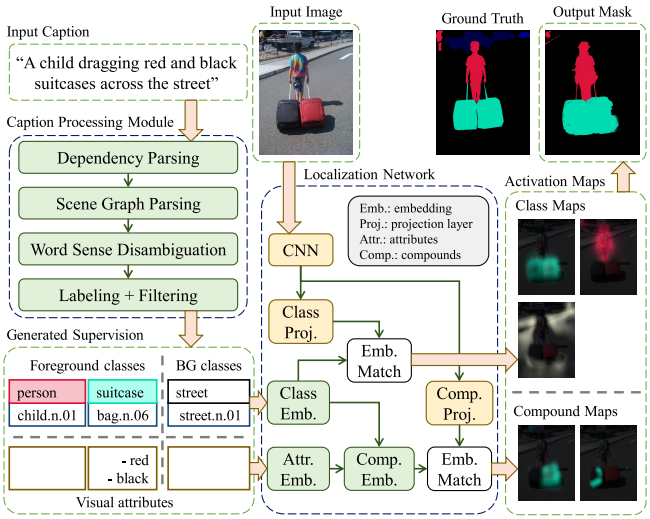


FIGURE 1. Overview of the proposed approach. A caption processing module combines syntactic structures and knowledge-based semantic relations to extract information from image captions that can be used for semantic segmentation. The extracted supervision consists of foreground and background object category labels, along with their respective visual attributes. This supervision is used to train a localization network that learns separate semantic embeddings for both single classes and compound class-attribute pairs. All activation maps generated by the trained localization network can then be leveraged to generate pseudo-ground truth segmentation masks.

we developed a caption processing module to first extract a structured supervision from natural language descriptions. This module uses the syntactic structures of the captions to identify mentions of objects and their attributes, and takes advantage of a general knowledge database to explicitly assign words to useful semantic labels.

The resulting supervision is used to train a novel localization network, that naturally extends the classifiers normally used for generating CAMs. This model can then be used to produce activation maps for relevant as well as contextual object categories, in addition to categories enriched with specific visual attributes. Also, by simplifying the supervision of the localization network, the training procedure can focus on relevant, *localizable* information, improving the quality of the generated maps. Finally, we propose a method that takes advantage of all types of activation maps generated by our model to obtain accurate segmentation masks on the training set, substantially improving performance of the downstream segmentation model.

In summary, the proposed approach presents the following three main novel contributions: (1) A caption processing module that extracts useful visual information from image captions, including accurate labels for foreground as well as background categories, and their respective visual attributes. (2) A localization network that can be trained using the extracted supervision to obtain activation maps for classes and compound class-attribute pairs. (3) A methodology to generate high-quality segmentation masks, taking advantage of all activation maps produced by the proposed localization network. We performed extensive experiments to validate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

all components of the proposed approach, and show that our method outperforms the state-of-the-art for WSSS with image-level supervision by a margin of 7.6% absolute mIoU (26.7% relative) on the challenging MS-COCO dataset.

The rest of the paper is organized as follows: In Section II we present our literature review for WSSS and other related tasks. Our proposed method is described in detail in Section III, whereas the experimental setup used to evaluate our approach is detailed in Section IV. The results from our experiments and the corresponding discussion are presented in Section V. Finally, a summary of the main conclusions derived from the work is presented in Section VI.

II. RELATED WORK

In this section, we review previous studies from different research areas, which relate to various aspects of our proposed approach. For better understanding, and to highlight some of the main contributions of our work, Table 1 summarizes some significant characteristics of existing models to compare with our method. This includes the ability to use captions as supervision, the requirement of additional supervision at either training or test time, and the capacity to generate segmentation masks for novel images. For studies that also proposed methods for extracting useful semantic labels from captions, we indicate the general strategy used in each case, and the types of semantic labels that can be extracted with each method. For WSSS methods, we also indicate which of these types of semantic labels can be used to guide mask generation. The rest of this section presents these comparisons in detail.

A. WSSS USING IMAGE CAPTIONS

In our literature review, we found only one previous study [33] that tackled WSSS using image captions. Their approach followed previous studies in WSSS with classification labels by adopting a two-stage procedure, in which a model trained with weak supervision was first used to generate segmentation masks on the training set, and then these masks served as supervision for training a segmentation model. To implement the first stage, the authors followed previous studies on visual grounding by training a model to learn a visual-semantic embedding (VSE) space, optimized to encode images and arbitrary text jointly. The text is encoded using pre-trained word2vec [47], and the model was trained using binary cross-entropy loss, with negative snippets from the dataset sampled randomly. Text segments from input captions were assigned to relevant object classes via exact matches of the class names, and were then projected along with images into the VSE to generate activation maps, which were combined to obtain the final segmentation masks.

The method proposed in [33] constitutes a promising approach for applying captions to WSSS, but also presents significant drawbacks. The simple exact match heuristic used to assign semantic labels fails to detect synonyms and subtypes of classes, and can result in false positives for classes with polysemic names. Our caption processing module is designed to handle these more complex cases, and, additionally, can detect complementary visual information. On the other hand, by training their model to localize arbitrary text, non-visual information that hinders training is introduced into the supervision. Furthermore, false negatives can appear when sampling short caption segments randomly, since these phrases can often describe regions of unpaired images correctly. By contrast, our localization network encodes only classes and compound class-attribute pairs, which focuses training on relevant visual information, while utilizing only the labels produced by our caption processing module prevents label multiplicity and false negatives during training.

B. WSSS USING CLASSIFICATION LABELS

Our work is also closely related to previous studies in WSSS using classification labels. Most modern approaches to this task use confident regions retrieved from Class Activation Maps (CAMs) [28] produced by a convolutional classifier to initialize segmentation masks, which are then used to train a supervised segmentation model. As CAMs tend to highlight only the most discriminative regions for each class, most recent efforts have focused on refining the artificial masks [22], [25], [27], [39], [48], or regularizing the segmentation model to improve learning from incomplete or imperfect supervision [21], [23], [49], [50].

In [21], a combination of three loss functions was proposed to train the segmentation model to simultaneously classify labeled pixels correctly, expand foreground regions, and constrain to low-level image boundaries. This approach was later expanded in [23] by utilizing the predictions of the segmentation model to label confident pixels in the initially sparse training masks iteratively. In [22], an iterative training procedure for the classifier was proposed, in which the most confident pixels for each class were progressively erased from the training images, thus forcing the model to expand to less discriminative regions. Similar strategies have also been used in more recent studies [39], [48]. In [39], the training and erasing procedures were performed simultaneously in an end-to-end fashion, whereas in [48] a Self-Erasing Network was introduced to prevent CAMs from expanding to background regions in later erasing iterations. A different approach was explored in [24], which consisted of appending a pyramid of atrous convolutions with different dilation rates to the classifier to propagate activations from discriminative regions to neighboring pixels. A generalization of this approach was presented in [26], in which a stochastic dropout layer was used to achieve a similar diffusion effect. In [51], dilated convolutions were combined with an attention mechanism on features from different scales to improve the resulting CAMs. In [25], an additional model was trained using CAMs to generate pixel affinity matrices, which were then used to refine the CAMs using random walk optimization. A similar approach was presented in [27], in which an iterative training procedure was employed to learn to generate segmentation masks and pixel affinity matrices jointly.

In this work, we focus on a different but complemen-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

TABLE 1. Comparison of the proposed approach with previous state-of-the-art studies in related research areas.

Paper(s)	Image-level supervision	Additional supervision	Caption to semantic labels				Mask generation labels			Test-time annotations	Output
			Method	FG	Attr.	BG	FG	Attr.	BG		
WSSS with labels											
[19], [21], [25], [27], [38]	labels	-	·	·	·	·	✓	✗	✗	-	SM
[22]–[24], [26], [39]–[43]	labels	saliency	·	·	·	·	✓	✗	✗	-	SM
Visual grounding											
[35]–[37]	captions	-	·	·	·	·	·	·	·	captions	HM
WSOD with captions											
Cap2Det [44]	captions + labels	-	Classifier	✓	✗	✗	·	·	·	-	BB
Scene Attributes [45]	captions	-	EM + SG	✓	✓	✗	·	·	·	-	BB
ZSSS with captions											
Cap2Seg [46]	captions	full masks	·	·	·	·	·	·	·	captions	SM
WSSS with captions											
TAM-Net [33]	captions	-	EM	✓	✗	✗	✓	✓*	✗	-	SM
Ours	captions	-	CPM (ours)	✓	✓	✓	✓	✓	✓	-	SM

EM: exact match **FG**: foreground class labels ✓: supported **SM**: segmentation masks * Implicit in text segments
SG: scene graphs **Attr.**: visual attributes ✗: not supported **HM**: attention heatmaps
CPM: caption processing module **BG**: background class labels ·: not applicable **BB**: bounding boxes

tary problem: generating accurate segmentation masks using natural language captions instead of classification labels, which is not possible using these previous methods. Our method also introduces an effective way to take advantage of complementary labels for background categories and visual attributes for mask generation, which cannot be leveraged with existing systems. Some of these previous studies also used class-agnostic saliency maps to generate background cues [22]–[24], [26], [39], while others used saliency maps to guide and refine foreground regions [40]–[43]. However, this introduces dependencies on off-the-shelf models that require stronger supervision. Our method, by contrast, uses only captions, and does not require any extra supervision.

C. VISUAL GROUNDING

Our work is also related to the task of visual grounding [34]–[37], and particularly to those studies that model this problem as finding a relevance heatmap in an image for a given query phrase [35]–[37]. Similar to these studies, we also train a model to learn a VSE space, and then use this model to generate localization heatmaps guided by semantic concepts. However, unlike our approach, these models do not define a way to generate segmentation masks on novel images, and require input captions at test time to produce attention heatmaps. We also restrict our VSE space to encode only object categories and compound class-attribute pairs, which allows us to dispense with the recurrent networks and contrastive losses used by these previous methods.

D. WEAKLY SUPERVISED OBJECT DETECTION USING IMAGE CAPTIONS

Recent studies have also explored the related task of weakly supervised object detection (WSOD) using only image captions as supervision [44], [45]. Similar in spirit to our proposed caption processing module, these studies have explored methods for extracting useful visual information explicitly from captions as a structured set of labels. In [44], a label inference module was proposed for detecting relevant object categories mentioned in captions, which was implemented as a classifier neural network with captions as inputs. However, that model requires a dataset of paired captions and ground truth classification labels for training, which is expensive to generate, and cannot be used to extract visual attributes or to discover background categories. In [45], an approach similar to ours was used for detecting objects and attributes in captions using a scene graph parser. However, their model relies on the exact match heuristic to assign semantic labels to objects, and, as such, cannot handle more complex cases. Additionally, neither of the previously mentioned studies presents a way for utilizing the extracted information for segmentation, relying instead on bounding boxes generated by off-the-shelf models to train a detection model in a multiple instance learning (MIL) framework.

E. ZERO-SHOT SEMANTIC SEGMENTATION USING IMAGE CAPTIONS

In [46], captions were used as supervision to train a novel three-branch network for semantic segmentation under a zero-shot setting, which aims to segment semantic categories not seen during training. However, this method required full supervision for most semantic categories during training, as well as input captions at both the training and test times. We, by contrast, focus on the weakly supervised setting, where no pixel-level supervision is available, and our proposed model does not require any type of annotations to generate segmentation masks at test time.

F. SCENE GRAPHS

Scene graphs represent the contents of an image in terms of a set of objects, their attributes, and their relations. These

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

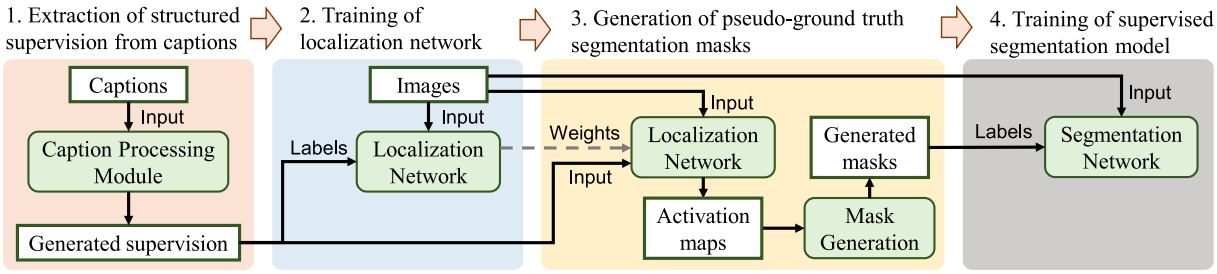


FIGURE 2. Overview of the full training procedure of the proposed approach. First, a novel caption processing module is used to extract structured visual supervision from natural language descriptions, which is then used to train a proposed localization network. Once trained, the localization network generates different types of activation maps, guided by the same structured supervision, which are then combined to obtain artificial segmentation masks on the training set. Finally, these masks are used to train a supervised segmentation model.

representations have been used for multiple computer vision applications, including image retrieval [52], image captioning [53], and image generation [54]. In our work, we employ the rule-based scene graph parser from [55] to extract object candidates and their corresponding attributes from captions, and make use of this information to train a localization network for WSSS.

G. WORDNET

WordNet [56] is a large lexical database for the English language, structured as a graph in which nodes correspond to *synsets* (sets of words sharing the same meaning), and edges encode different types of conceptual-semantic relations, from which our model leverages the following:

- Hyponymy: X is a hyponym of Y (equivalently, Y is a hypernym of X) if X is a type of Y (e.g., "dog" is a hyponym of "mammal").
- Meronymy: X is a meronym of Y (equivalently, Y is a holonym of X) if X is a part or member of Y (e.g., "branch" is a meronym of "tree", and "tree" is a meronym of "forest").

WordNet has also been applied to several computer vision tasks, including object detection [57], image retrieval [58], and zero-shot segmentation [59], [60]. In our work, we use WordNet to assign object candidates to relevant semantic labels for WSSS, and to easily construct a similar set of labels for background categories discovered from the dataset.

III. PROPOSED METHOD

Fig. 2 shows an overview of the training procedure followed in the proposed approach. First, our caption processing module extracts a structured representation of the relevant visual information contained in image captions (Section III-A). Then, this supervision is used to train the proposed localization network (Section III-B). Once trained, this model is used to generate pseudo-ground truth segmentation masks on the training set (Section III-C). Finally, the generated masks can be used to train any supervised segmentation model. The details of the supervised model used in our experiments are presented in Section IV-C4.

A. CAPTION PROCESSING MODULE

Our aim in developing the proposed caption processing module is to consolidate mentions of different object categories, and their attributes, at a semantic level that is useful for WSSS, while suppressing non-visual or ambiguous information present in the image captions. For this purpose, a syntactic parsing stage is applied to identify mentions of objects and their attributes using a scene graph representation [52], [55]. This is followed by a semantic parsing stage which takes advantage of WordNet [56] to assign semantic labels to detected objects. Fig. 3 shows an overview of the entire process. The syntactic and semantic parsing stages are described in Sections III-A1 and III-A2, respectively. The resulting supervision is then formalized in Section III-A3.

1) Syntactic Parsing

We implement the syntactic parsing stage of our module using a modified version of a public Python implementation[1] of the rule-based scene graph parser proposed in [55]. This parser is composed of the following two processing steps:

Dependency Parsing. The first step corresponds to a standard language processing pipeline, including tokenization, part-of-speech tagging, and dependency tree parsing of the input caption [61], as illustrated in Fig. 3.

Scene Graph Parsing. The second step uses syntactic patterns in the dependency tree to identify objects, their attributes, and their relations explicitly. In particular, we follow [55] by identifying all noun phrases as potential object candidates, and apply rules for parsing their corresponding attributes, but skip node replication based on quantifiers, and pronoun resolution. Relations between objects detected by the parser are ignored. Instead of simply selecting the head of each noun phrase as the class of the corresponding object, as done in [55], we identify an "extended head" of the phrase as the longest substring that both (1) contains the head noun of the phrase, and (2) is defined in WordNet. This extended head is used later to infer the corresponding object category.

It should be noted that, contrary to the strategy of simply identifying positive categories by matching class names [33], our method allows filtering out cases where class names

[1] https://github.com/vacancy/SceneGraphParser

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

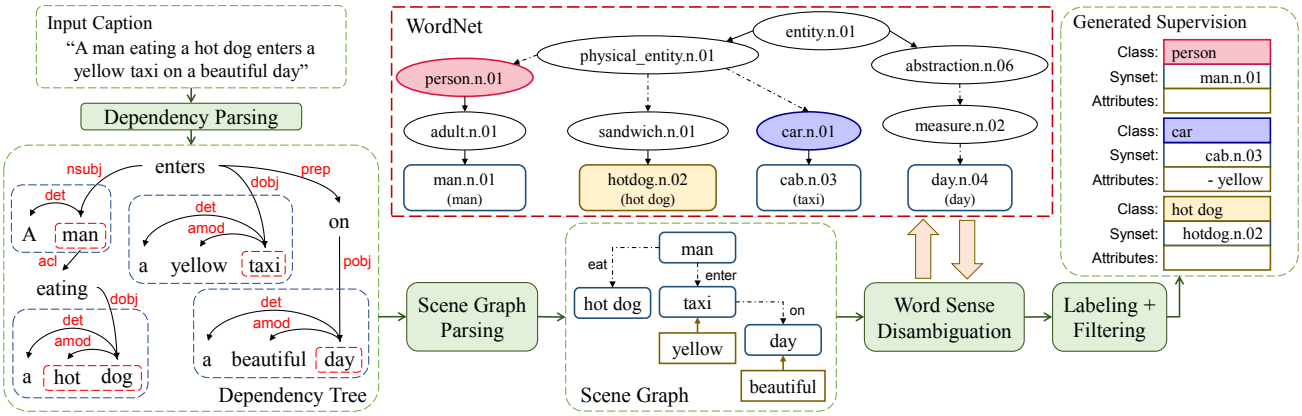


FIGURE 3. Overview of the proposed caption processing module. First, a dependency tree is generated from the input caption, and a rule-based scene graph parser is used to identify object candidates and their respective visual attributes. The extended heads of the objects' noun phrases are then mapped into the WordNet graph using a word sense disambiguation model. Finally, object candidates belonging to useful semantic categories are identified using the WordNet graph, and the rest are discarded.

are used with different parts of speech (e.g., "orange" as an adjective, or "train" as a verb, instead of using them as nouns). Additionally, by considering only the extended heads, we can handle cases in which compound nouns change the meaning of the head (e.g., "microwave oven" refers to the class "microwave" and not "oven", whereas "hot dog" does not refer to the class "dog").

2) Semantic Parsing

Word Sense Disambiguation. For the purpose of identifying the corresponding WordNet synset [56] for each object candidate, we employ the state-of-the-art Word Sense Disambiguation (WSD) model from [62]. Concretely, we utilize the pre-trained model based on BERT [63], but modify it slightly to handle compound nouns by restricting sense prediction for each noun phrase to the set of possible synsets associated with its extended head. We then choose the argmax over this set, using the scores predicted for the leftmost token of the extended head, following the strategy of the original authors for handling words split in multiple tokens by BERT.

This step is crucial both for preventing false positives (e.g., depending on the context, "mouse" could refer to a rodent instead of an electronic device, as is intended for the class "mouse"), and for identifying synonyms (e.g., class "couch" could be denoted equivalently as "sofa").

Labeling and Filtering. Finally, we leverage the WordNet graph to assign each object proposal to a semantic label, selected from a pre-defined set $\mathcal{C}$. Each label in $\mathcal{C}$ is associated with a specific synset, such that any candidate with a synonym, hyponym, or holonym of a labeled synset can be assigned to the same class. For example, "penguin" can be identified as a subtype of the class "bird", and "people" as a collection of objects in the class "person". Hyponymy conflicts are resolved by choosing the labeled hypernym with the shortest path distance in the graph, whereas holonymy relations are only considered if all meronyms of the labeled synset belong to the same class. Objects that do not match any label in $\mathcal{C}$ are discarded.

In practice, WSSS tasks define only a set of relevant or "foreground" categories, $\mathcal{C}_{\mathrm{fg}}$. However, other categories mentioned in the captions provide valuable contextual information, and can be used to improve the segmentation masks. Since all pixels that do not correspond to classes in $\mathcal{C}_{\mathrm{fg}}$ should be assigned to a special "background" category by the final model, we call these "background" classes, and give them a separate set of labels $\mathcal{C}_{\mathrm{bg}}$. It is important to emphasize that the definition of background is arbitrary and depends on the application. Thus, $\mathcal{C}$ can be obtained as the union of both sets, $\mathcal{C} = \mathcal{C}_{\mathrm{fg}} \cup \mathcal{C}_{\mathrm{bg}}$. The set $\mathcal{C}_{\mathrm{bg}}$ could be constructed manually by selecting useful background categories similarly to $\mathcal{C}_{\mathrm{fg}}$. We also propose a simple assisted procedure to facilitate this process, which takes advantage of the WordNet graph to filter out non-visual synsets easily, and to assign useful semantic labels iteratively. The details of this procedure, as well as the final list of 72 background classes used in our experiments, are detailed in Appendix A-A.

In addition to categories, visual attributes are also mapped to a set of pre-defined labels, $\mathcal{A}$, and unlabeled attributes are discarded. Since our goal is to use attributes to guide localization, we restrict attributes to a set of 40 words describing low-level visual features, such as colors, materials (e.g., "wooden", "plastic", "metal"), and textures (e.g., "striped", "dotted", "furry"). This type of low-level information has been extensively used in image processing applications such as segmentation [64], retrieval [65], and classification [66]. The full list of selected attributes is included in Appendix A-B.

3) Generated Supervision

After processing all captions using the proposed procedure, the resulting dataset $\mathcal{D}$ can be defined as a set of tuples of the form:

$$\mathcal{D} = \{(\mathbf{I}_n, \mathcal{C}_n^{\mathrm{fg}}, \mathcal{C}_n^{\mathrm{bg}}, \{\mathcal{A}_{n,c}\}_{c \in \mathcal{C}_n})\}_{n=1}^{N}, \tag{1}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

where $\mathbf{I}_n$ is the n-th image of the dataset; N is the total number of images; $\mathcal{C}_n^{\text{fg}} \subseteq \mathcal{C}_{\text{fg}}$ and $\mathcal{C}_n^{\text{bg}} \subseteq \mathcal{C}_{\text{bg}}$ are the sets of positive foreground and background category labels for the n-th image, respectively; $\mathcal{C}_n = \mathcal{C}_n^{\text{fg}} \cup \mathcal{C}_n^{\text{bg}}$ is the set of all positive classes for the n-th image; and $\mathcal{A}_{n,c} \subseteq \mathcal{A}$ is the set of positive attributes for a given class $c \in \mathcal{C}_n$. We alternatively refer to the set of all positive class-attribute pairs for a given image as $\mathcal{P}_n$, i.e.,

$$\mathcal{P}_n = \{(c, a) \mid c \in \mathcal{C}_n, a \in \mathcal{A}_{n,c}\}. \tag{2}$$

B. LOCALIZATION NETWORK

1) Architecture

Following previous studies [33], [35]–[37], the proposed localization network learns to map images and semantic concepts into a shared multi-modal embedding space. To do this, we first assign each object category c to a semantic embedding $\boldsymbol{e}_c^{\text{cls}} \in \mathbb{R}^{d_{\text{cls}}}$, and each attribute a to an embedding $\boldsymbol{e}_a^{\text{attr}} \in \mathbb{R}^{d_{\text{attr}}}$. Compound category-attribute pairs $(c, a) \in C \times A$ are encoded using compound embeddings $\boldsymbol{e}_{c,a}^{\text{comp}} \in \mathbb{R}^{d_{\text{comp}}}$, computed as:

$$\boldsymbol{e}_{c,a}^{\text{comp}} = \boldsymbol{W}_{\text{comp}} \left[\boldsymbol{e}_c^{\text{cls}}; \boldsymbol{e}_a^{\text{attr}}\right] + \boldsymbol{b}_{\text{comp}}, \tag{3}$$

where $\boldsymbol{W}_{\text{comp}} \in \mathbb{R}^{d_{\text{comp}} \times (d_{\text{cls}} + d_{\text{attr}})}$ and $\boldsymbol{b}_{\text{comp}} \in \mathbb{R}^{d_{\text{comp}}}$ are, respectively, the weights and biases of an affine transformation, and $[\cdot\,;\cdot]$ denotes vector concatenation.

The input image $\mathbf{I}_n$ is encoded into a feature map $\mathbf{F} \in \mathbb{R}^{d_{\text{enc}} \times h \times w}$, with d_{enc} channels and spatial dimensions (h, w), using a fully-convolutional neural network $\phi_{\text{CNN}}(\cdot\,; \boldsymbol{\theta}_{\text{CNN}})$ parameterized by $\boldsymbol{\theta}_{\text{CNN}}$. We evaluate several standard convolutional architectures for implementing ϕ_{CNN}, as detailed in Section IV-C2.

We then employ two independent 1×1 convolutional layers to project $\mathbf{F}$ into separate embedding spaces for single classes and compound class-attribute pairs, to avoid competition between the two types of representations. This results in a class feature map $\mathbf{F}_{\text{cls}} = f_{\text{cls}}(\mathbf{F}; \boldsymbol{\theta}_{\text{cls}}) \in \mathbb{R}^{d_{\text{cls}} \times h \times w}$, and a compound feature map $\mathbf{F}_{\text{comp}} = f_{\text{comp}}(\mathbf{F}; \boldsymbol{\theta}_{\text{comp}}) \in \mathbb{R}^{d_{\text{comp}} \times h \times w}$. A global pooling operation is then applied separately over $\mathbf{F}_{\text{cls}}$ and $\mathbf{F}_{\text{comp}}$ to obtain the final visual embeddings $\boldsymbol{v}_{\text{cls}} \in \mathbb{R}^{d_{\text{cls}}}$ and $\boldsymbol{v}_{\text{comp}} \in \mathbb{R}^{d_{\text{comp}}}$, respectively. Following [36], we employ a WELDON pooling layer to leverage both positive and negative evidence [67].

Given a pair of corresponding visual and semantic embeddings, we can then compute their similarity score using a simple inner product:

$$s_c^{\text{cls}} = \langle \boldsymbol{v}_{\text{cls}}, \boldsymbol{e}_c^{\text{cls}} \rangle, \tag{4}$$

$$s_{c,a}^{\text{comp}} = \langle \boldsymbol{v}_{\text{comp}}, \boldsymbol{e}_{c,a}^{\text{comp}} \rangle. \tag{5}$$

It is worth noting that this model is reduced to the usual classification network used for generating CAMs when only foreground object categories are used, except for an extra projection layer, since the class embeddings are mathematically equivalent to a fully-connected layer without bias.

2) Training

The training procedure of the proposed localization network is supervised by the following loss function:

$$\ell = \lambda_{\text{fg}} \ell_{\text{fg}} + \lambda_{\text{bg}} \ell_{\text{bg}} + \lambda_{\text{comp}} \ell_{\text{comp}}, \tag{6}$$

where ℓ_{fg} and ℓ_{bg} supervise single class predictions for foreground and background classes, respectively; ℓ_{comp} supervises compound pair predictions; and λ_{fg}, λ_{bg}, and λ_{comp} are scalar parameters that balance the contribution of each component.

For ℓ_{fg}, we adopt the following weighted variation of the usual Binary Cross-Entropy (BCE) loss:

$$\ell_{\text{fg}} = -\frac{1}{|\mathcal{C}_{\text{fg}}|} \left[\sum_{c \in \mathcal{C}_n^{\text{fg}}} w_c^+ \log(\sigma(s_c^{\text{cls}})) + \sum_{c' \in \mathcal{C}_{\text{fg}} \setminus \mathcal{C}_n^{\text{fg}}} w_{c'}^- \log(1 - \sigma(s_{c'}^{\text{cls}})) \right] \tag{7}$$

where $\sigma(\cdot)$ represents the sigmoid function, such that $\sigma(x) = 1/(1 + e^{-x})$. The scalar weights w_c^+ and w_c^- are inversely proportional to the frequency of positive and negative examples for class c, respectively. These weights are normalized to ensure that all classes contribute the same accumulated weight across the dataset as in the unweighted case, where $w_c^+ = w_c^- = 1$. Similar approaches have been used previously for classification and detection [68], [69], and we find that this simple strategy leads to better results in practice.

ℓ_{bg} is computed in the same way as ℓ_{fg}, but using background categories, i.e., by replacing $\mathcal{C}_{\text{fg}}$ by $\mathcal{C}_{\text{bg}}$ and $\mathcal{C}_n^{\text{fg}}$ by $\mathcal{C}_n^{\text{bg}}$ in (7). Similarly, ℓ_{comp} is computed as:

$$\ell_{\text{comp}} = -\frac{1}{|\mathcal{P}_n| + |\tilde{\mathcal{P}}_n^-|} \left[\sum_{(c,a) \in \mathcal{P}_n} \tilde{w}_a^+ \log(\sigma(s_{c,a}^{\text{comp}})) + \sum_{(c',a') \in \tilde{\mathcal{P}}_n^-} \tilde{w}_{a'}^- \log(1 - \sigma(s_{c',a'}^{\text{comp}})) \right]. \tag{8}$$

In this case, we consider only a subset $\tilde{\mathcal{P}}_n^-$ of negative compound pairs, constructed as follows: for each positive compound $(c, a) \in \mathcal{P}_n$, we select all pairs with negative attributes given by $\{(c, \tilde{a}) \mid \tilde{a} \in \mathcal{A} \setminus \mathcal{A}_{n,c}\}$, and randomly sample 10 pairs with negative classes from $\{(\tilde{c}, a) \mid \tilde{c} \in \mathcal{C} \setminus \mathcal{C}_n\}$. This strategy lowers the penalty for erroneous class prediction in compound pairs, enabling the model to rely on low-level attributes to highlight less discriminative, but relevant regions. The weights $\tilde{w}_a^+$ and $\tilde{w}_a^-$ are computed analogously to w_c^+ and w_c^-, respectively, but considering the total number of non-empty $\mathcal{A}_{n,c}$ that contain attribute a across the dataset, instead of the number of images labeled with a given class. In this case, class frequency is ignored.

C. MASK GENERATION

In this section we describe how to generate segmentation masks using the outputs of the trained localization network.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

Similar to CAMs [28], we obtain localization heatmaps $\boldsymbol{H}_c^{\text{cls}} \in \mathbb{R}^{h \times w}$ for each positive class $c \in \mathcal{C}_n$ by computing a weighted sum over the channels of the feature map $\mathbf{F}_{\text{cls}}$. In our case, the weights are given by the elements of the corresponding class embedding $\boldsymbol{e}_c^{\text{cls}}$, instead of a fully-connected classifier. Analogously, we can compute compound heatmaps $\boldsymbol{H}_{c,a}^{\text{comp}}$ by projecting embeddings $\boldsymbol{e}_{c,a}^{\text{comp}}$ over the feature map $\mathbf{F}_{\text{comp}}$ in the same way. We then combine all maps associated with a given class c via pixel-wise sum, such that:

$$\tilde{\boldsymbol{M}}_c(i,j) = \boldsymbol{H}_c^{\text{cls}}(i,j) + \sum_{a \in \mathcal{A}_{n,c}} \boldsymbol{H}_{c,a}^{\text{comp}}(i,j), \qquad (9)$$

where i and j indicate spatial coordinates. The resulting maps are thresholded at 0, and normalized by dividing its maximum activation to obtain the final class maps $\boldsymbol{M}_c \in \mathbb{R}^{h \times w}$:

$$\boldsymbol{M}_c(i,j) = \frac{\text{ReLU}(\tilde{\boldsymbol{M}}_c(i,j))}{\max_{k,l} \text{ReLU}(\tilde{\boldsymbol{M}}_c(k,l))}. \qquad (10)$$

For the purpose of computing the activation map for the background class, we combine two different types of cues. First, we follow previous work [25], [33] by computing a background map $\boldsymbol{M}^-_{\langle \mathbf{bg} \rangle}$ using negative information given by the activation maps of foreground categories:

$$\boldsymbol{M}^-_{\langle \mathbf{bg} \rangle}(i,j) = \left\{ 1 - \max_{c \in \mathcal{C}_n^{\text{fg}}} \boldsymbol{M}_c(i,j) \right\}^{\alpha}, \qquad (11)$$

such that pixels with low scores for all foreground categories are assigned higher background probabilities. The parameter $\alpha \geqslant 1$ controls the magnitude of the activations.

Additionally, we compute a second activation map $\boldsymbol{M}^+_{\langle \mathbf{bg} \rangle}$ using positive information given by the predicted maps for background categories. For this, we first compute $\tilde{\boldsymbol{M}}^+_{\langle \mathbf{bg} \rangle}$ as the pixel-wise sum over the unnormalized background maps:

$$\tilde{\boldsymbol{M}}^+_{\langle \mathbf{bg} \rangle}(i,j) = \sum_{c \in \mathcal{C}_n^{\text{bg}}} \tilde{\boldsymbol{M}}_c(i,j). \qquad (12)$$

$\tilde{\boldsymbol{M}}^+_{\langle \mathbf{bg} \rangle}$ is then normalized as in (10) to obtain $\boldsymbol{M}^+_{\langle \mathbf{bg} \rangle}$, which is combined with $\boldsymbol{M}^-_{\langle \mathbf{bg} \rangle}$ via pixel-wise maximum to obtain the final background map $\boldsymbol{M}_{\langle \mathbf{bg} \rangle}$, such that:

$$\boldsymbol{M}_{\langle \mathbf{bg} \rangle}(i,j) = \max\{\boldsymbol{M}^-_{\langle \mathbf{bg} \rangle}(i,j), \gamma \cdot \boldsymbol{M}^+_{\langle \mathbf{bg} \rangle}(i,j)\}. \qquad (13)$$

The scalar parameter $\gamma < 1.0$ is added to prevent background classes from dominating over foreground objects.

The localization maps are then upscaled to the resolution of the input image $\mathbf{I}_n$, and refined using dense Conditional Random Fields (dCRF) [70]. Finally, the segmentation mask $\boldsymbol{S}$ is obtained by selecting the argmax for each pixel:

$$\boldsymbol{S}(i,j) = \operatorname*{argmax}_{c \in \mathcal{C}_n^{\text{fg}} \cup \{\langle \mathbf{bg} \rangle\}} \boldsymbol{M}_c(i,j). \qquad (14)$$

The complete procedure is summarized in Algorithm 1.

Algorithm 1 Segmentation Mask Generation

Input: Class heatmaps $\{\boldsymbol{H}_c^{\text{cls}} \mid c \in \mathcal{C}_n\}$
Compound heatmaps $\{\boldsymbol{H}_{c,a}^{\text{comp}} \mid (c,a) \in \mathcal{P}_n\}$
Output: Segmentation mask $\boldsymbol{S}$

for each $c \in \mathcal{C}_n$ **do**
  Compute $\tilde{\boldsymbol{M}}_c$ from $\boldsymbol{H}_c^{\text{cls}}$,$\{\boldsymbol{H}_{c,a}^{\text{comp}} | a \in \mathcal{A}_{n,c}\}$ (Eq.:9)
end for
for each $c \in \mathcal{C}_n^{\text{fg}}$ **do**
  Normalize $\tilde{\boldsymbol{M}}_c$ to obtain $\boldsymbol{M}_c$ (Eq.: 10)
end for
Compute $\boldsymbol{M}^-_{\langle \mathbf{bg} \rangle}$ from $\{\boldsymbol{M}_c \mid c \in \mathcal{C}_{\text{fg}}\}$ (Eq.: 11)
Compute $\tilde{\boldsymbol{M}}^+_{\langle \mathbf{bg} \rangle}$ from $\{\tilde{\boldsymbol{M}}_c \mid c \in \mathcal{C}_{\text{bg}}\}$ (Eq.: 12)
Normalize $\tilde{\boldsymbol{M}}^+_{\langle \mathbf{bg} \rangle}$ to obtain $\boldsymbol{M}^+_{\langle \mathbf{bg} \rangle}$ (Eq.: 10)
Compute $\boldsymbol{M}_{\langle \mathbf{bg} \rangle}$ from $\boldsymbol{M}^-_{\langle \mathbf{bg} \rangle}$ and $\boldsymbol{M}^+_{\langle \mathbf{bg} \rangle}$ (Eq.: 13)
Upscale all score maps $\{\boldsymbol{M}_c \mid c \in \mathcal{C}_n^{\text{fg}} \cup \{\langle \mathbf{bg} \rangle\}\}$
Refine score maps $\{\boldsymbol{M}_c \mid c \in \mathcal{C}_n^{\text{fg}} \cup \{\langle \mathbf{bg} \rangle\}\}$ using dCRF
Compute $\boldsymbol{S}$ from $\{\boldsymbol{M}_c \mid c \in \mathcal{C}_n^{\text{fg}} \cup \{\langle \mathbf{bg} \rangle\}\}$ (Eq.: 14)

IV. EXPERIMENTS

A. DATASETS

MS-COCO. Most of the reported experiments were performed using the MS-COCO dataset [30], which contains 123k images, each annotated with 5 different textual captions, as well as instance-level segmentation masks for 80 object categories. Following previous studies [23], [27], [33], [43], we used the train2014 split with 83k images for training, and reserved the val2014 split with 41k images exclusively for testing. To adjust the hyperparameters of our model, we followed the same validation strategy as in [33], which consisted in evaluating the quality of the generated masks on the training set. However, we restricted this evaluation to use only a subset of 4k images from train2014 (less than 5% of the total), to reduce the dependency on the ground truth masks of the training set. To prevent confusion with the val2014 set used for evaluation, we refer to this set exclusively as $trainval4$k.

Unless stated otherwise, all reported experiments with MS-COCO used the 2014 splits, as described above. However, to facilitate the comparison with [71] and future studies, we also report results using the 2017 splits. These 2017 splits assign 118k images for training, leaving 5k for testing.

PASCAL VOC. With the purpose of evaluating the performance of our approach further, we also performed additional experiments using the PASCAL VOC 2012 dataset [31]. This dataset is annotated with segmentation masks for 20 object categories, that correspond to a subset of the classes from MS-COCO. Since this dataset is not annotated with captions, we use PASCAL VOC only for evaluation, employing its validation set with 1449 images.

YouTube-Objects. The YouTube-Objects [72] dataset is composed of videos collected from YouTube, which were retrieved by querying 10 object classes from PASCAL VOC. Since YouTube-Objects is not annotated with captions, we use this dataset only for evaluation, as in the case of PAS-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

CAL VOC. For this evaluation we use the subset of frames manually annotated with pixel-level masks provided by [73].

B. EVALUATION

As in previous studies [23], [27], [33], [43], performance is reported in terms of pixel-wise mean Intersection-over-Union (mIoU), averaged across all categories.

C. IMPLEMENTATION DETAILS

1) Caption Processing Module

All MS-COCO object categories were mapped to their corresponding WordNet synsets, with ambiguity resolved manually for class names with more than one possible meaning. In cases where a single category is assigned in practice to objects of more than one type, all relevant synsets were annotated (e.g., objects tagged as class "tv" include instances of both "television_receiver.n.01" and "computer_monitor.n.01"). In the case of categories without a corresponding synset in WordNet ("stop sign" and "potted plant"), only exact matches were used for detection. We also extended WordNet by mapping a few common neologisms to their closest synsets, such as "smartphone" to "cellular_telephone.n.01", and "macbook" to "laptop.n.01".

2) Localization Network

We used the slightly modified ResNet-50 [74] architecture from [75] as the backbone for the localization network for most of our experiments, since it presents a good trade-off between performance and model size. For a more direct comparison with previous studies, we also report results using the VGG-16 [76] and ResNet-38 [77] architectures, adapted for WSSS with dilated convolutions as described in [25]. All reported results used the ResNet-50 backbone, unless stated otherwise, and all backbones were initialized from weights pre-trained on ImageNet [29].

The model hyperparameters and training details used in our experiments are summarized in Table 2. We multiplied the learning rate by a factor of 10 for all parameters in $\boldsymbol{\theta}_{\mathrm{cls}}$, $\boldsymbol{\theta}_{\mathrm{comp}}$, $\boldsymbol{W}_{\mathrm{comp}}$, $\boldsymbol{b}_{\mathrm{comp}}$, and the semantic embeddings, which were all trained from scratch. The first two convolutional blocks of all backbones were not modified during training.

Following standard practice for data augmentation [22], [24], [26], [33], a random transformation was applied to each image before being passed as input for the model during training. In each case, a rectangular crop was generated by choosing a random scale between 60% and 100% of the area of the original image, with a random aspect ratio between $3/4$ and $4/3$, and a random position within the image. The resulting cropped region was then resized to a square of size 321×321 pixels. Horizontal flipping was also applied with a probability of 50%. For testing and validation, only the original images were used as inputs for the model The sets of images used for training and testing remained disjoint, as explained in Section IV-A.

TABLE 2. Localization network hyperparameters and training details.

Parameter	Symbol	Value
Class embedding size	d_{cls}	512
Attribute embedding size	d_{attr}	512
Compound embedding size	d_{comp}	512
Number of training epochs	-	15
Minibatch size	-	18
Weight decay	-	0.0005
Optimizer	-	SGD
Initial learning rate	-	0.01
Learning rate policy	-	poly
Polynomial scheduling momentum	-	0.9
Foreground classes loss weight	λ_{fg}	1.0
Background classes loss weight	λ_{bg}	0.1
Compounds loss weight	λ_{comp}	0.1
Random crop scale range	-	$[60\%, 100\%]$
Random crop aspect ratio range	-	$[3/4, 4/3]$
Random flip probability	-	50%
Training input size	-	321×321

3) Mask Generation

The parameter α from (11) was set independently for each backbone, following the strategy described in Section IV-A, resulting in a value of 8 for ResNet-50, of 4 for VGG-16, and of 5 for ResNet-38. The value of γ for computing the background map was set to 0.7 in all experiments. The CRF parameters were all set to their default values [70].

4) Segmentation Network

We use the pseudo-ground truth masks generated by our model to train the VGG-16-based DeepLab-ASPP model presented in [9] with single-scale input, re-implemented in PyTorch. We trained this model for 10 epochs using minibatches of size 18 and the balanced seed loss from [23], but with all pixels labeled. The base learning rate was increased from 0.001 to 0.01, and we added 2k steps of linear warm-up [78] at the start of training to prevent instability early on. We find that this setting improves model performance significantly, as is shown in Section V-F3. All other hyperparameters were the same as in the original implementation [9].

V. RESULTS

In this section we present and discuss our experimental results. We begin by reporting our main results in Section V-A, in which we present the comparison of our full method with the current state-of-the-art for WSSS. The following subsections then report complementary experiments performed to assess the impact of each of the main novel components of the proposed approach. In Section V-B we evaluate the quality of the labels for foreground objects extracted by our model. Then, in Section V-C we evaluate the impact of each supervision component produced by our caption processing module in terms of segmentation quality on MS-COCO. This analysis is expanded in Section V-D, in which we evaluate the generalization ability of our model across various datasets under a transfer learning setting. In Section V-E we report experiments in which we replaced our localization network with a visual grounding model, to evaluate the impact of our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

approach on model training and mask generation separately. In Section V-F, we report additional ablation studies to evaluate the impact of several design choices, and parameter settings. Finally, in Section V-G, some qualitative results for our approach are presented.

A. COMPARISON WITH THE STATE-OF-THE-ART

1) MS-COCO 2014

Table 3 summarizes our main results, together with previous state-of-the-art results for WSSS on MS-COCO using the 2014 splits for training and evaluation.

WSSS with image captions. The results most directly comparable to those of our proposed approach are those reported for TAM-Net [33], since this is the only previously published method for addressing WSSS using only image captions. We observe that our VGG-16-based model outperforms their TAM-Net based on the same architecture by a margin of 5.3% mIoU, even without the use of Deep Seeded Region Growing (DSRG) [23], which is a technique used to expand confident seed regions iteratively through an expensive retraining scheme. Even more notably, the version of our model with a ResNet-38 backbone improves the state-of-the-art for this task substantially, by a margin of 7.6% mIoU, from 0.285 to 0.361 mIoU. Our model based on the much smaller ResNet-50 also reaches a very similar mIoU, offering a good balance between performance and model size. These results demonstrate the effectiveness of our approach for leveraging image captions for WSSS.

WSSS with other types of supervision. For a more complete evaluation of our method, we also present state-of-the-art results for models trained using other types of weak supervision. In particular, we observe that even our VGG-16-based model outperforms all previous methods trained using only ground truth classification labels, showing the effectiveness of our approach, and of image captions as supervision for this task. Our method also presents comparable results to those reported in [43] for SGAN, despite the fact that this model leverages much stronger supervision extensively in the form of class-agnostic saliency maps, generated by a model trained with pixel-level supervision.

2) MS-COCO 2017

To complement our analysis, and facilitate comparison with [71], as well as for future work, we also present results using the 2017 splits of MS-COCO, which are summarized in Table 4. We observe that the results for VGG-16 and ResNet-50 remain similar to those obtained with the 2014 splits, whereas the results for ResNet-38 improve further to 0.370 mIoU. This indicates that the larger ResNet-38 architecture benefits the most from the increased number of training images. Additionally, we observe that our approach outperforms [71] even using the smaller ResNet-50, again showing the effectiveness of the proposed method.

TABLE 3. Comparison of weakly supervised semantic segmentation methods on the MS-COCO 2014 validation set in terms of mIoU.

Method	Backbone	Image-level supervision	Additional supervision	mIoU
Journal articles				
BFBP [19]	VGG-16	labels	-	0.204
WAILS [38]	VGG-16	labels	-	0.225
IAL [27]	VGG-16	labels	-	0.277
SGAN [43]	VGG-16	labels	saliency	0.336
Conference articles				
SEC [21]	VGG-16	labels	-	0.224
DSRG [23]	VGG-16	labels	saliency[a]	0.260
TAM-Net [33]	VGG-16	captions	-	0.216
TAM-Net [33] + DSRG [23]	VGG-16	captions	saliency[a]	0.269
TAM-Net [33]	ResNet-38	captions	-	0.285
TAM-Net [33] + DSRG [23]	ResNet-38	captions	saliency[a]	0.277
Ours	VGG-16	captions	-	0.322
Ours	ResNet-50	captions	-	0.357
Ours	ResNet-38	captions	-	0.361

[a]Background cues only.

TABLE 4. Comparison of weakly supervised semantic segmentation methods on the MS-COCO 2017 validation set in terms of mIoU.

Method	Backbone	Image-level supervision	Additional supervision	mIoU
Journal articles				
MGCFF [71]	ResNet-101	labels	-	0.281
Ours	VGG-16	captions	-	0.324
Ours	ResNet-50	captions	-	0.356
Ours	ResNet-38	captions	-	0.370

TABLE 5. Evaluation of foreground class labels retrieved from captions of the train2014 set of MS-COCO using the proposed caption processing module, compared with the exact match baseline.

Method	mIoU	Precision	Image recall	Pixel recall
Baseline	0.521	0.904	0.560	0.762
Proposed	0.579	0.911	0.616	0.826

B. EVALUATION OF GENERATED SEMANTIC LABELS

We begin our ablation studies by evaluating the quality of the labels for foreground categories retrieved from captions by our model. We use the ground truth classification labels from MS-COCO as reference, and present results in terms of mIoU, precision, and (image) recall. We also report results using pixel recall, which takes the sizes of the objects into account by considering the number of pixels annotated with a given label across ground truth masks, instead of the number of images. As shown in Table 5, the proposed caption processing module improves performance for all metrics with respect to the exact match baseline [33], especially in terms of mIoU and recall. The large margins between image and pixel recall also indicate that captions are significantly more likely to mention larger objects, as is expected.

For a more detailed comparison of both methods, we also present the differences in IoU for each class in Fig. 4. It

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

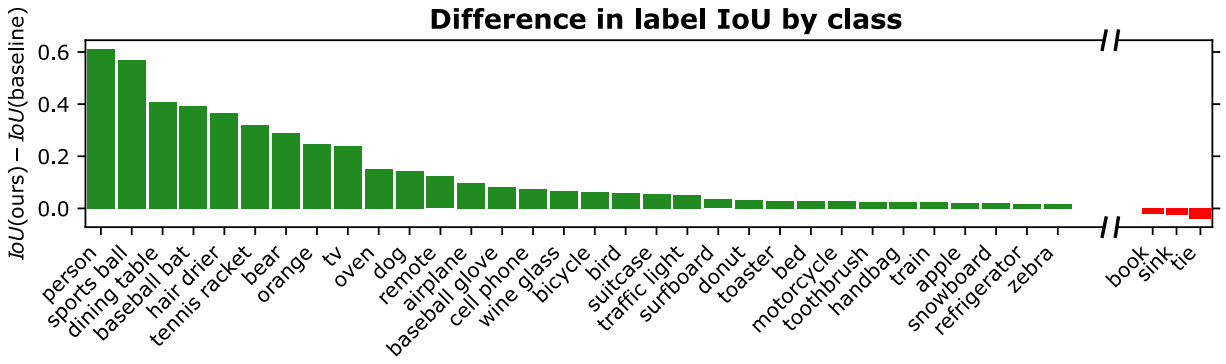


FIGURE 4. Difference in label detection IoU by class when using the proposed caption processing module instead of the exact match strategy of previous methods. For clarity, only the 35 classes with largest absolute difference are shown.

TABLE 6. Effect of different supervision components over model performance on MS-COCO.

Foreground classes	Attributes	Background classes	Generated supervision (train set)			Segmentation model (val set)		
			Precision	Recall	mIoU	Precision	Recall	mIoU
Baseline	-	-	0.454	0.558	0.324	0.440	0.536	0.308
Proposed	-	-	0.492	0.589	0.359	0.507	0.513	0.338
Proposed	✓	-	0.491	0.611	0.368	0.505	0.524	0.344
Proposed	-	✓	0.547	0.526	0.359	0.554	0.491	0.352
Proposed	✓	✓	0.549	0.543	0.369	0.560	0.497	0.357
GT	-	-	0.429	0.674	0.349	0.448	0.536	0.330

can be observed that the proposed method improves IoU for almost all MS-COCO classes. It is especially useful for increasing recall for categories with multiple common synonyms and hyponyms, such as "sports ball" or "person", and for improving precision for classes with names that are often used to form compounds that do not belong to the same class, such as with "bear" in "teddy bear", "oven" in "toaster oven", or "dog" in "hot dog". The syntactic parsing step can also improve precision for classes with names that can be used as attributes, such as "orange" as a color, or "apple" as the brand. Only a small fraction of the classes show slight drops in IoU, corresponding mostly to names with multiple meanings that the syntactic parsing model can sometimes assign to erroneous part-of-speech tags, such as "tie", "sink", or "book".

C. ANALYSIS OF SUPERVISION COMPONENTS

We experimented by training our localization network using different supervision components produced by our caption processing module, namely foreground categories, background categories, and visual attributes. Table 6 summarizes the results, both for the generated supervision on the training set, as well as for the final segmentation model on the validation set. As a baseline, we consider only labels for foreground classes, retrieved from captions by detecting mentions of the class name, or its plural form, as in [33]. We also present results using the ground truth classification labels.

1) Foreground Categories

As shown in Table 6, using the improved labels for foreground object categories produced by our method results in a substantial improvement in the quality of the final segmentation model compared to the exact match strategy. This corresponds to a 9.7% relative improvement, from 0.308 to 0.338 mIoU on the validation set. The classes with the largest increase in segmentation IoU, such as "person" ($+54.4\%$ IoU), "dining table" ($+25.6\%$ IoU), and "orange" ($+21.0\%$ IoU), also usually correspond to those with the largest gain in label IoU. These results highlight the effectiveness of our approach for extracting relevant object categories from captions and improving segmentation performance.

2) Attributes

By introducing attributes, the quality of the generated masks improved to 0.368 mIoU, boosting performance on the validation set to 0.344 mIoU. Fig. 5 illustrates the effect of this supervision. Qualitatively, we observe that compound activation maps can help to extend single class maps to less discriminative parts and instances of the same category guided by low-level visual features, improving the coverage of the resulting masks. For example, in the second row of Fig. 5 we observe that the activation map for the class "cat" highlights only the most discriminative parts of the head, resulting in an incomplete mask. However, by including the map for the compound concept ("cat", "furry"), which highlights mostly the fur of the animal, the resulting mask is expanded to cover a larger portion of the object. This is especially useful for categories with a large number and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

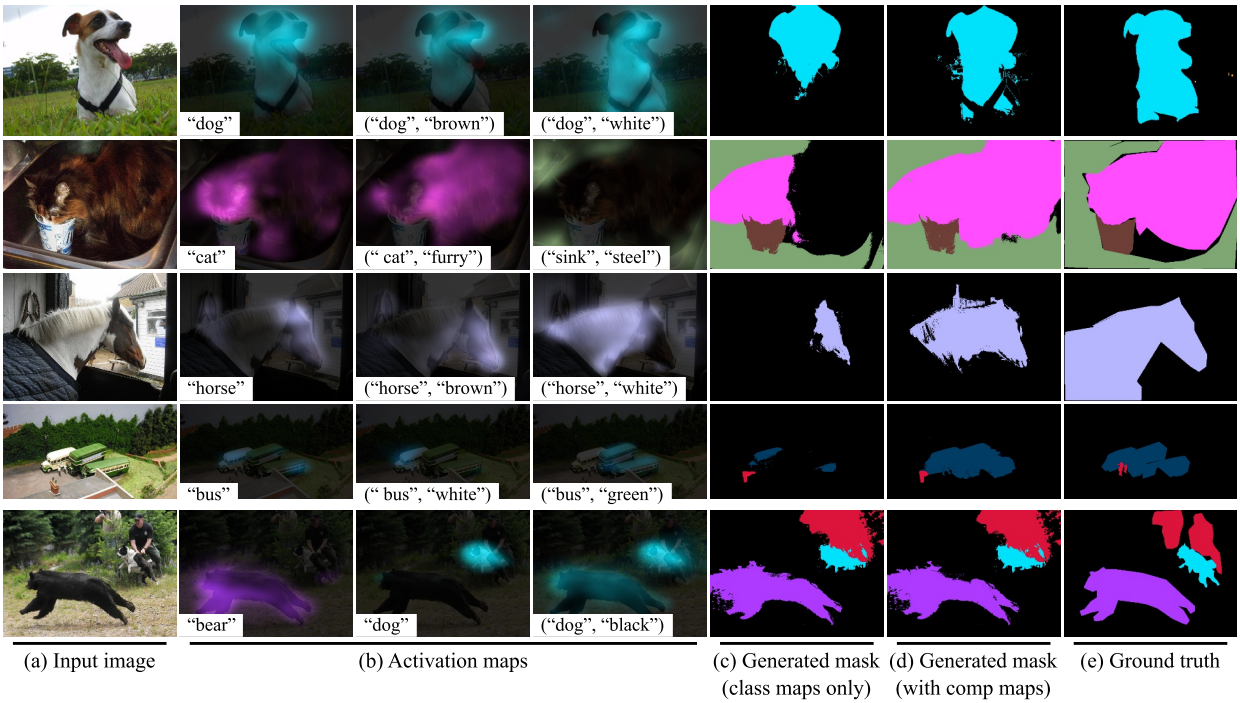


FIGURE 5. Visualization of the effect of compound activation maps over the generated segmentation masks.

variety of associated attributes, that also correspond to large and complex objects such as "cow" (+4.2% IoU on the validation set), "train" (+3.1% IoU), and "horse" (+2.5% IoU), or small, harder to localize objects such as "frisbee" (+4.7% IoU) and "donut" (+3.3% IoU). We note that the model can sometimes incorrectly highlight objects of similar categories guided by their attributes, as shown in the last row of Fig. 5 with the map for ("dog", "black"), which also partially highlights the black bear in the foreground. This is a consequence of primarily penalizing attribute predictions during training of the compound embeddings. In these cases, we rely on single class activation maps and dCRF to correct mislabeled regions. In practice, we observe that compound maps increase recall and mIoU of the generated masks, without hurting precision significantly.

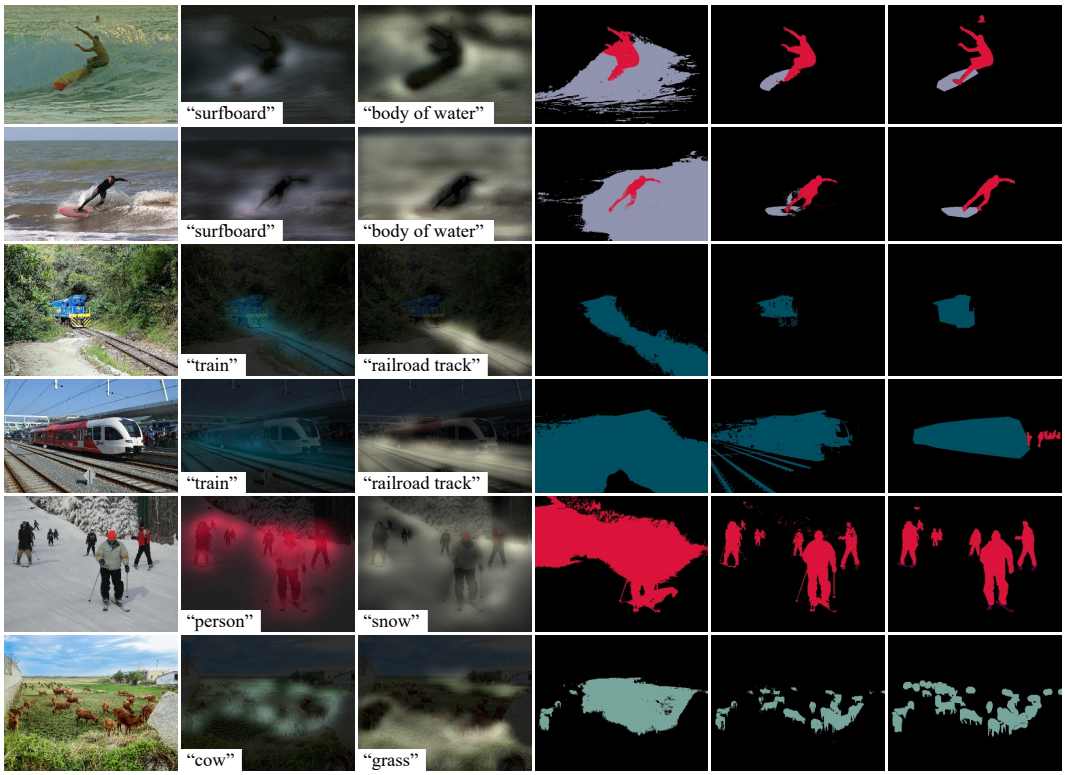
3) Background Categories

We observe that attention maps for complementary categories help to separate foreground objects from their surroundings, especially in cases with many instances of the same class cluttered together. It is also particularly useful for refining foreground classes that present high co-occurrence with specific background objects, such as "train" with "train track", or "surfboard" with "water". In these cases, CAMs for the foreground category will tend to highlight the background objects as well, as these provide useful information for classification, despite being undesirable for segmentation. By including the background class explicitly, the model can take advantage of examples in which the objects appear separately to learn to differentiate both classes. This results in overlapping but distinct activation maps that can be combined to better localize the foreground category. Several examples of this are shown in Fig. 6. Despite the fact that including background categories mostly improved the precision of the generated masks, while roughly preserving mIoU on the training set, the results on the validation set improved significantly, from 0.344 to 0.357 mIoU. This result could indicate that removing consistently erroneous regions from the supervision has a greater impact over the learned model than removing some correct foreground regions randomly.

4) Comparison with Ground Truth Labels

We also experimented by training our model using ground truth classification labels, obtaining 0.330 mIoU on the validation set, as shown in Table 6. This result is better by a margin of 2.2% mIoU than that obtained by using the exact match baseline, but is surprisingly worse than training with class labels retrieved by our caption processing module, despite their relatively low recall, as is discussed in Section V-B. This could be because classes that are missing from image captions tend to be harder to identify, such as when there are small or partially occluded objects. If the model fails to localize an object, its label acts essentially as a false positive, adding erroneous foreground regions that hurt segmentation quality. These results indicate that the proposed approach is robust to incomplete captions, and can, in fact, benefit from the information about object saliency that is encoded implicitly in this type of supervision. By leveraging visual attributes and background categories we can further improve results to 0.357 mIoU, showing the potential of the richer visual and contextual information present in image captions as supervision for this task.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access



(a) Input image (b) Foreground activation map (c) Background activation map (d) Generated mask (without BG maps) (e) Generated mask (with BG maps) (f) Ground truth

FIGURE 6. Visualization of the effect of background class activation maps over the generated segmentation masks.

TABLE 7. Transfer learning experiment results, in terms of segmentation performance on the validation set of PASCAL VOC 2012, and YouTube-Objects datasets. All models were trained on the MS-COCO train2014 set using all 81 MS-COCO semantic categories.

Foreground classes	Attr.	Background classes	Test mIoU	
			PASCAL VOC	YouTube-Objects
Baseline	-	-	0.448	0.528
Proposed	-	-	0.512	0.538
Proposed	✓	-	0.524	0.550
Proposed	-	✓	0.533	0.584
Proposed	✓	✓	0.548	0.597
GT	-	-	0.503	0.517

D. TRANSFER LEARNING EXPERIMENTS

For evaluating the performance of our approach further, we performed additional experiments using the PASCAL VOC, and YouTube-Objects datasets. Since these datasets are not annotated with image captions, which are required for training our model, we performed these experiments under a transfer learning setting. More precisely, we trained our model using the train2014 set of MS-COCO as reported in the previous sections, and then evaluated its performance on the validation set of PASCAL VOC, and YouTube-Objects. For adapting our model to the new datasets, we simply discarded all the weights in the final segmentation layer that did not correspond to any of the semantic categories in the target domain, without performing any additional re-training stages. We did not find any previous work that had reported results under this setting to use as comparisons, so we trained several different baselines based on our proposed approach to serve as references, as shown in Table 7.

1) PASCAL VOC Dataset

We observe that all the models achieve significantly better results on PASCAL VOC than on MS-COCO, as is expected, since PASCAL VOC contains fewer classes, and objects are usually larger, and easier to identify than those in MS-COCO. We obtained similar performance improvements for the various components of the proposed approach, further validating the conclusions derived from our experiments on MS-COCO. In particular, we observed that using our caption processing module to detect relevant object categories, instead of the exact match baseline used in previous studies, improved performance substantially, in this case from 0.448 mIoU to 0.512 mIoU. Using the relevant categories detected by our model also led to better results than using the ground truth classification labels, which in this case yielded 0.503 mIoU. Our results were further improved by leveraging both attributes and background categories, with our full model achieving 0.548 mIoU on the validation set of PASCAL VOC in this transfer learning setting.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

TABLE 8. Results on MS-COCO validation set for experiments combining the visual grounding model from [36] with our proposed caption processing module.

Foreground classes	Full phrases	Background classes	Precision	Recall	mIoU
Baseline	-	-	0.436	0.539	0.307
Proposed	-	-	0.460	0.536	0.321
Proposed	✓	-	0.484	0.529	0.334
Proposed	✓	✓	0.510	0.511	0.342

2) YouTube-Objects Dataset

In the case of the YouTube-Objects dataset, using the foreground labels detected by our model also improved performance compared to the case using the exact match baseline, but by a smaller margin of 1.0% mIoU. This is because the 10 semantic categories defined in this dataset have unambiguous names that usually work well with the simple exact match strategy. Taking advantage of complementary information extracted by our model improves performance substantially, in this case resulting in a relative increase in mIoU of 11.0%. These results are also 15.5% higher than those obtained using the ground truth labels, further underscoring the usefulness of image captions as supervision for WSSS, and of our approach for taking advantage of this information.

Finally, these transfer learning results show that the proposed approach also improves the generalization ability of the final model to different datasets.

E. EFFECT OF EACH MODULE

For the purpose of understanding the effect of each proposed module better, we performed additional experiments employing the pre-trained model from Engilberge *et al.* [36], which constitutes the state-of-the-art in heatmap-based weakly supervised visual grounding. We first experimented by grounding class names using the method described in the original paper, and adapted the resulting heatmaps to WSSS following [33] by subtracting the average of the min and max values for each heatmap, setting background scores to 0, and selecting the argmax for each pixel. Applying this method to categories retrieved from captions with the exact match heuristic yielded an mIoU of 0.234 on the $trainval4\mathrm{k}$ set. We also experimented using the pipeline described in Section III-C for grounding, normalizing, and generating negative background cues from the embeddings produced by this model. The parameter α from (11) was set to 4 following the same strategy used for the localization network. This method improved mIoU to 0.272. Refining the maps using dCRF further improved mIoU on the training set to 0.309. We therefore used this setting as baseline for our experiments, which are summarized in Table 8.

By using the foreground class labels produced by our caption processing module, performance on the validation set improved from 0.307 to 0.321 mIoU compared to the baseline. This can be improved further to 0.334 mIoU by also grounding the full noun phrases associated with each label, and combining the resulting maps following the strategy proposed for handling compound maps, as in (9). By additionally leveraging maps for background categories as described in Section III-C, results were improved to 0.342 mIoU. In total, taking full advantage of the information extracted by our caption processing module resulted in a relative improvement of 11.4% mIoU on the validation set, showing the effectiveness of the extracted supervision, and of the proposed combination strategies.

Furthermore, we observe that this final result is still significantly lower than the 0.357 mIoU obtained with our localization network. This is despite the fact that the model from Engilberge *et al.* [36] uses a much more powerful ResNet-152 architecture, and is significantly more expensive to train due to the use of recurrent networks and a hard-negative mining contrastive loss. These results highlight the importance of filtering caption information for learning to generate high-quality activation maps for WSSS.

TABLE 9. Effect of different attribute types over semantic segmentation performance on MS-COCO.

Attribute type:	None	Low-level only	All adjectives	All verbs	All comp. nouns
mIoU (train):	0.359	0.368	0.355	0.356	0.357
mIoU (val):	0.338	0.344	0.337	0.336	0.337

TABLE 10. Effect of balanced binary cross-entropy loss over segmentation performance on the MS-COCO validation set.

Loss function	Precision	Recall	mIoU
Standard BCE	0.586	0.466	0.351
Balanced BCE	0.560	0.497	0.357

TABLE 11. Effect of learning rate and warm-up phase during training of the segmentation model, in terms of mIoU over the MS-COCO validation set.

Learning rate	Warm-up	Precision	Recall	mIoU
0.001	-	0.526	0.470	0.326
0.001	2k steps	0.525	0.468	0.328
0.01	2k steps	0.560	0.497	0.357

F. ADDITIONAL ABLATION STUDIES

1) Effect of Attribute Types

Most of our experiments were performed using a subset of visual attributes describing low-level visual features, such as colors, materials, and textures. Table 9 shows segmentation results on the training and validation sets when employing other types of attributes detected by our caption processing module, such as all the adjectives, verbs, and compound nouns. We observe that all other unfiltered attribute types actually hurt performance slightly compared to when only class labels are used. This can be explained by the fact that these attributes describe mostly non-local information, such as quantities or actions, that the model fails to localize accurately, thus degrading the quality of the generated masks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

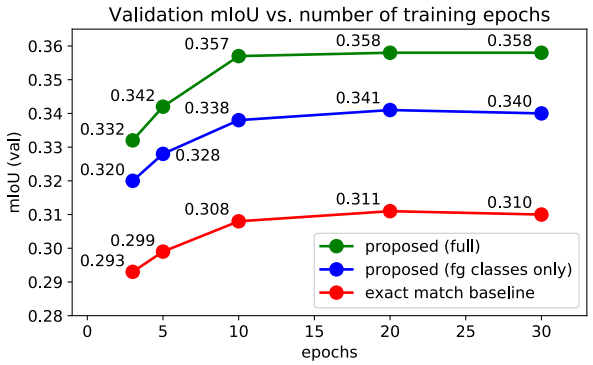


FIGURE 7. Effect of the number of training epochs on the performance of the final segmentation model.

2) Effect of Weighted BCE Loss

With the purpose of assessing the effect of the proposed weighted binary cross-entropy loss, we retrained our model based on ResNet-50 using regular BCE loss (equivalent to setting $w_c^+ = w_c^- = 1$ in (7) and $\tilde{w}_a^+ = \tilde{w}_a^- = 1$ in (8)). The results are presented in Table 10, showing a drop of 0.6% mIoU on the validation set when employing standard BCE loss, mostly affecting a small subset of less frequent classes, showing the effectiveness of this simple strategy to handle class imbalance. Qualitatively, the weighted BCE also allows the model to learn to localize compound pairs with infrequent attributes, although this has little impact over the performance of the final model.

3) Effect of the Learning Rate on the Segmentation Model

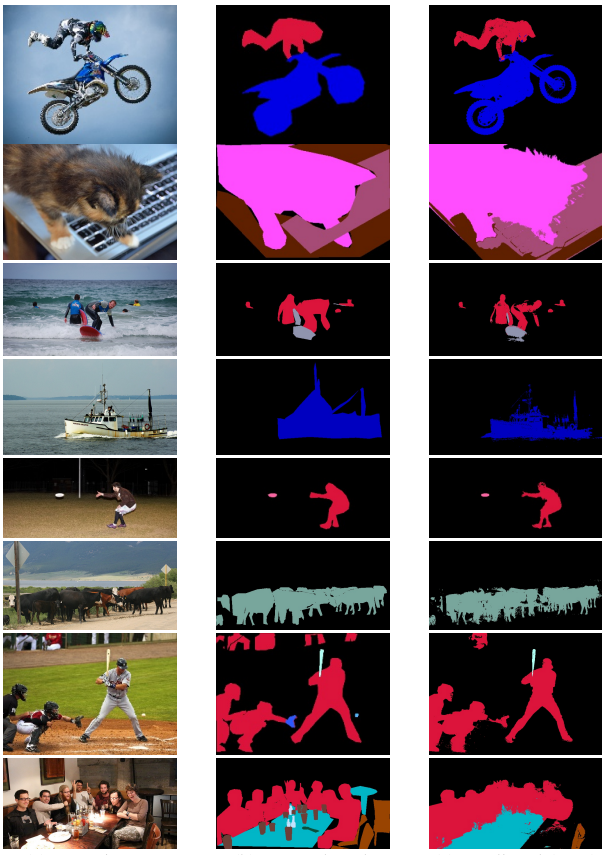
Table 11 shows the results over the validation set for different settings of the learning rate during training of the segmentation model. We observe that by increasing the learning rate used in [9] from 0.001 to 0.01 and adding 2k steps of linear warm-up at the beginning of training, we were able to improve the performance of the final model substantially. Adding the warm-up phase by itself had little impact over performance, but prevented the model from diverging early on when increasing the learning rate. These results show that this is a simple and inexpensive way of boosting performance on the validation set. It is worth noting that, even without this modification, our ResNet-50 model still outperforms the previous state-of-the-art by a margin of 4.3% mIoU.

4) Effect of the Number of Training Iterations on the Segmentation Model

Fig. 7 shows the effect of the number of training epochs on the performance of the final segmentation model on the validation set of MS-COCO. We observe that increasing the total number of iterations generally improves performance, but stabilizes after the 10 epochs used in our experiments.

G. QUALITATIVE RESULTS

Fig. 8 shows some qualitative results on the validation set of MS-COCO, obtained using the segmentation model trained with the pseudo-ground truth masks generated by our localization network with the ResNet-38 backbone. We observe that the resulting model is capable of localizing a wide range of object categories accurately across different scales. The last two rows show typical failure cases, in which the model fails to localize small or less prominent objects that appear in cluttered scenes, which are frequently omitted from the training captions, and consequently from the training masks. This includes the baseball glove and ball in the penultimate row, and the cups and bottles in the last row.



(a) Input image (b) Ground truth (c) Predicted (ours)

FIGURE 8. Examples of segmentation results on the MS-COCO 2014 validation set obtained using the proposed approach.

VI. CONCLUSIONS AND FUTURE WORK

A. CONCLUSIONS

In this paper, we presented a comprehensive methodology to tackle weakly supervised semantic segmentation using only image captions. The key component of our approach is a caption processing module, that leverages syntactic structures and knowledge-based semantic relations to extract visual information from captions, without requiring any additional annotations. We presented a novel localization network that can be trained using the extracted supervision to generate activation maps both for single classes, and for class-attribute compound pairs. Finally, we described a method to leverage all types of maps generated by this network to obtain high-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3076074, IEEE Access

TABLE 12. List of all 72 WordNet synsets used for defining background categories in the reported experiments, that extend the 80 foreground object categories defined in MS-COCO.

Index	Synset	Index	Synset	Index	Synset	Index	Synset	Index	Synset
81	street.n.01	96	tray.n.01	111	shower.n.01	126	painting.n.01	141	egg.n.02
82	plate.n.04	97	snow.n.01	112	lamp.n.02	127	engine.n.01	142	rice.n.01
83	tree.n.01	98	meat.n.01	113	cart.n.01	128	candle.n.01	143	goat.n.01
84	grass.n.01	99	ramp.n.01	114	tarmacadam.n.02	129	trail.n.02	144	hay.n.01
85	body_of_water.n.01	100	rock.n.01	115	curb.n.01	130	branch.n.02	145	pool.n.01
86	sky.n.01	101	shelf.n.01	116	rug.n.01	131	rack.n.05	146	ceiling.n.01
87	sidewalk.n.01	102	runway.n.04	117	basket.n.01	132	tile.n.01	147	roadway.n.01
88	floor.n.01	103	cabinet.n.01	118	home_plate.n.01	133	graffito.n.01	148	leash.n.01
89	pole.n.01	104	cheese.n.01	119	streetlight.n.01	134	curtain.n.01	149	roof.n.01
90	mirror.n.01	105	bathtub.n.01	120	step.n.04	135	potato.n.01	150	pepper.n.04
91	flower.n.01	106	pan.n.01	121	tomato.n.01	136	crossing.n.05	151	chest_of_drawers.n.01
92	track.n.09	107	shrub.n.01	122	bun.n.01	137	onion.n.01	152	soup.n.01
93	box.n.01	108	platform.n.01	123	post.n.04	138	highway.n.01		
94	railroad_track.n.01	109	countertop.n.01	124	pot.n.01	139	carriage.n.02		
95	light.n.02	110	pen.n.01	125	doorway.n.01	140	platter.n.01		

TABLE 13. List of all 40 visual attributes used in the reported experiments.

white	pink	tile/tiled	furry	leafy
black	grey/gray	wood/wooden	cloudy	glass
red	purple	grass/grassy	paved	brick
blue	gold/golden	rusty/rusted	snowy	dirt
green	beige	ceramic	muddy	stone
brown	silver	rock/rocky	fluffy	steel
yellow	checkered	metal/metallic	plastic	cement
orange	stripe/striped	concrete	sandy	leather

quality segmentation masks, that are effective to train a supervised model.

We presented several experimental results that show the advantages of our approach. Using the proposed caption processing module to detect foreground categories, the improvement relative to the exact match heuristic used in previous methods was 9.7% mIoU on the validation set of the MS-COCO database. We also showed that by including complementary visual information in the form of attributes and background categories, results were further improved by 5.6% relative mIoU. Additional experiments showed that, by restricting its supervision to relevant visual information, our simple localization network can outperform a much more complex visual grounding model trained to localize arbitrary text, by 4.4% relative mIoU. Finally, we showed that our best model advances the state-of-the-art for WSSS with image-level supervision on MS-COCO significantly, by a margin of 7.6% absolute (26.7% relative) mIoU. The proposed approach constitutes a powerful framework for utilizing image captions as supervision to train segmentation models. It could increase the effective number of training examples available for this task dramatically, by enabling the use of images paired with this type of annotation which are freely available on the web.

B. FUTURE WORK

The proposed approach could be improved in the future by addressing the problem of correcting mislabeled images that arises from incomplete or incorrect captions, and that currently results in inaccurate pseudo-ground truth masks. Thus, an interesting line of research would be to model the training of the localization network as a problem of classification with noisy labels [79], [80], using the predictions of the same model to iteratively refine the labels initialized from captions.

Another promising improvement could be extending the proposed approach to the more complex problem of weakly supervised instance segmentation. Whereas existing approaches for this task are based on splitting class masks obtained from CAMs into instance-level segments [75], models based on captions could take advantage of attribute and hyponym information to obtain activation maps closer to the instance-level, facilitating the mask generation procedure. Additionally, the presence of explicit quantifiers could be exploited to introduce information about the number of instances for each class to be segmented in each training image.

.

APPENDIX A DETAILS OF MS-COCO ATTRIBUTES AND BACKGROUND CATEGORIES

A. BACKGROUND CATEGORIES

We propose a simple iterative procedure to initialize the set of background categories $\mathcal{C}_{\text{bg}}$ using the initially unlabeled synsets present in the dataset, i.e., those not assigned to any foreground class in $\mathcal{C}_{\text{fg}}$ as described in Section III-A2. Given that these synsets initially contain non-visual information, as well as excessive granularity, further filtering is required to construct $\mathcal{C}_{\text{bg}}$, which is performed as follows:

1) We discard all hypernyms of labeled synsets in $\mathcal{C}_{\text{fg}}$, since these describe categories of objects that are too diverse visually (e.g., “furniture”, “animal”, etc.), as well as all of their meronyms, given that these refer to parts of labeled objects (e.g., “wing”, “wheel”, etc.).
2) We also discard hyponyms of synsets for “abstraction” (describing mostly non-visual information), “part” (since these refer to incomplete objects), as well as “room” and “location” (which usually describe whole scenes).

3) We select the remaining unlabeled synset with the highest frequency in the dataset, and assign it to a new label, extending $\mathcal{C}_{\text{bg}}$. We then assign all of its hyponyms and holonyms to the same label, and discard all of its hypernyms and meronyms, as explained in step 1).
4) Step 3) is repeated until no synset in the dataset exceeds a certain frequency threshold, T, which is set at $T = 200$ for MS-COCO.

This procedure prioritizes labeling at semantic levels which are most frequently used by annotators to describe objects, while preserving consistency by exploiting the WordNet graph. The resulting set can then be further refined according to the application. For example, in the case of MS-COCO we also discard all hyponyms of "clothing.n.01", as these are labeled together in practice with the person or animal wearing them. To make replication easier, we include the final list of 72 background categories used in our experiments in Table 12.

B. ATTRIBUTES

Additionally, Table 13 includes all the selected attribute types describing low-level visual information, which are used in most of our experiments.

REFERENCES

[1] W. Wang, Y. Fu, Z. Pan, X. Li, and Y. Zhuang, "Real-Time Driving Scene Semantic Segmentation," *IEEE Access*, vol. 8, pp. 36 776–36 788, 2020.
[2] W. Wang, Y. Wang, Y. Wu, T. Lin, S. Li, and B. Chen, "Quantification of full left ventricular metrics via deep regression learning with contour-guidance," *IEEE Access*, vol. 7, pp. 47 918–47 928, 2019.
[3] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *J. Digit. Imag.*, vol. 32, no. 4, pp. 582–596, aug 2019.
[4] P. Yin, R. Yuan, Y. Cheng, and Q. Wu, "Deep Guidance Network for Biomedical Image Segmentation," *IEEE Access*, vol. 8, pp. 116 106–116 116, 2020.
[5] W. Shen, W. Xu, H. Zhang, Z. Sun, J. Ma, X. Ma, S. Zhou, S. Guo, and Y. Wang, "Automatic segmentation of the femur and tibia bones from X-ray images based on pure dilated residual U-Net," *Inverse Problems Imag.*
[6] J.-Y. Sun, S.-W. Jung, and S.-J. Ko, "Lightweight Prediction and Boundary Attention-Based Semantic Segmentation for Road Scene Understanding," *IEEE Access*, vol. 8, pp. 108 449–108 460, 2020.
[7] B. Zhang, S. Li, X. Jia, L. Gao, and M. Peng, "Adaptive Markov random field approach for classification of hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 5, pp. 973–977, sep 2011.
[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jun 2015, pp. 3431–3440.
[9] 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. Pattern Anal. Mach. Intell*, vol. 40, no. 4, pp. 834–848, apr 2018.
[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, jun 2017.
[11] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–15, aug 2020.
[12] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–13, aug 2020.
[13] D. P. Benalcazar, J. E. Zambrano, D. Bastias, C. A. Perez, and K. W. Bowyer, "A 3D iris scanner from a single image using convolutional neural networks," *IEEE Access*, vol. 8, pp. 98 584–98 599, 2020.
[14] C. A. Perez, P. A. Estevez, F. J. Galdames, D. A. Schulz, J. P. Perez, D. Bastias, and D. R. Vilar, "Trademark image retrieval using a combination of deep convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, jul 2018, pp. 1–7.
[15] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple Does It: Weakly supervised instance and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jul 2017, pp. 1665–1674.
[16] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, dec 2015, pp. 1635–1643.
[17] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jun 2016, pp. 3159–3167.
[18] A. L. Bearman, O. Russakovsky, V. Ferrari, and F.-F. Li, "What's the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 549–565.
[19] F. S. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, J. M. Alvarez, and S. Gould, "Incorporating network built-in priors in weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1382–1396, jun 2018.
[20] Y. Feng, L. Wang, and M. Zhang, "Weakly-supervised learning of a deep convolutional neural networks for semantic segmentation," *IEEE Access*, vol. 7, pp. 91 009–91 018, 2019.
[21] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 695–711.
[22] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jul 2017, pp. 6488–6496.
[23] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7014–7023.
[24] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7268–7277.
[25] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4981–4990.
[26] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "FickleNet: Weakly and semi-supervised semantic image segmentation using stochastic inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 5267–5276.
[27] X. Wang, S. Liu, H. Ma, and M.-H. Yang, "Weakly-supervised semantic segmentation by iterative affinity learning," *Int. J. Comput. Vis.*, 2020.
[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jun 2016, pp. 2921–2929.
[29] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jun 2009, pp. 248–255.
[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.
[31] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, jun 2010.
[32] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2018, pp. 2556–2565.
[33] J. Sawatzky, D. Banerjee, and J. Gall, "Harvesting information from captions for weakly supervised semantic segmentation," in *Proc. ICCV Workshops*, oct 2019, pp. 4481–4490.
[34] F. Zhao, J. Li, J. Zhao, and J. Feng, "Weakly supervised phrase localization with multi-scale anchored transformer network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 5696–5705.

[35] F. Xiao, L. Sigal, and Y. J. Lee, "Weakly-supervised visual grounding of phrases with linguistic structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jul 2017, pp. 5253–5262.

[36] M. Engilberge, L. Chevallier, P. Pérez, and M. Cord, "Finding beans in burgers: Deep semantic-visual embedding with localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3984–3993.

[37] H. Akbari, S. Karaman, S. Bhargava, B. Chen, C. Vondrick, and S.-F. Chang, "Multi-level multimodal common semantic space for image-phrase grounding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 12 476–12 486.

[38] H. Zhou, K. Song, X. Zhang, W. Gui, and Q. Qian, "WAILS: Watershed algorithm with image-level supervision for weakly supervised semantic segmentation," *IEEE Access*, vol. 7, pp. 42 745–42 756, 2019.

[39] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: guided attention inference network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 9215–9223.

[40] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, nov 2017.

[41] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jul 2017, pp. 5038–5047.

[42] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1354–1362.

[43] Q. Yao and X. Gong, "Saliency guided self-attention network for weakly and semi-supervised semantic segmentation," *IEEE Access*, vol. 8, pp. 14 413–14 423, 2020.

[44] K. Ye, M. Zhang, A. Kovashka, W. Li, D. Qin, and J. Berent, "Cap2Det: Learning to amplify weak caption supervision for object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, oct 2019, pp. 9685–9694.

[45] A. Jerbi, R. Herzig, J. Berant, G. Chechik, and A. Globerson, "Learning object detection from captions via textual scene attributes," sep 2020. [Online]. Available: http://arxiv.org/abs/2009.14558

[46] G. Tian, S. Wang, J. Feng, L. Zhou, and Y. Mu, "Cap2Seg: Inferring semantic and spatial context from captions for zero-shot image segmentation," in *Proc. ACM Int. Conf. Multimedia*, oct 2020, pp. 4125–4134.

[47] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 3111–3119.

[48] Q. Hou, P. Jiang, Y. Wei, and M.-M. Cheng, "Self-erasing network for integral object attention," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, 2018, pp. 549–559.

[49] M. Tang, F. Perazzi, A. Djelouah, I. B. Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised CNN segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 524–540.

[50] M. Tong, W. Li, X. Ren, X. Yu, and W. Lin, "Weakly-Supervised Semantic Segmentation With Regional Location Cutting and Dynamic Credible Regions Correction," *IEEE Access*, vol. 8, pp. 204 378–204 388, 2020.

[51] L. Xu, M. Bennamoun, F. Boussaid, and F. Sohel, "Scale-aware feature network for weakly supervised semantic segmentation," *IEEE Access*, vol. 8, pp. 75 957–75 967, 2020.

[52] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jun 2015, pp. 3668–3678.

[53] N. Xu, A.-A. Liu, J. Liu, W. Nie, and Y. Su, "Scene graph captioner: Image captioning based on structural visual representation," *J. Vis. Commun. Image Represent.*, vol. 58, pp. 477–485, 2019.

[54] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1219–1228.

[55] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in *Proc. Workshop Vis. Lang.*, 2015, pp. 70–80.

[56] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, nov 1995.

[57] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2017-Janua, dec 2017, pp. 6517–6525.

[58] B. Barz and J. Denzler, "Hierarchy-based image embeddings for semantic image retrieval," in *Proc. Winter Conf. Appl. Comput. Vis. (WACV)*, jan 2019, pp. 638–647.

[59] H. Zhao, X. Puig, B. Zhou, S. Fidler, and A. Torralba, "Open vocabulary scene parsing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, oct 2017, pp. 2021–2029.

[60] D. Golub, A. El-Kishky, and R. Martin-Martin, "Leveraging pretrained image classifiers for language-based segmentation," in *Proc. Winter Conf. Appl. Comput. Vis. (WACV)*, mar 2020, pp. 1999–2008.

[61] M.-C. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and C. D. Manning, "Universal Stanford dependencies: A cross-linguistic typology," in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, 2014, pp. 4585–4592.

[62] L. Vial, B. Lecouteux, and D. Schwab, "Sense vocabulary compression through the semantic knowledge of WordNet for neural word sense disambiguation," in *Proc. Global Wordnet Conf.*, 2019.

[63] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[64] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "ExFuse: Enhancing feature fusion for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, sep 2018, pp. 273–288.

[65] P. Liu, J.-M. Guo, C.-Y. Wu, and D. Cai, "Fusion of deep learning and compressed domain features for content-based image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5706–5717, dec 2017.

[66] J. E. Tapia and C. A. Perez, "Clusters of features using complementary information applied to gender classification from face images," *IEEE Access*, vol. 7, pp. 79 374–79 387, 2019.

[67] T. Durand, N. Thome, and M. Cord, "WELDON: Weakly supervised learning of deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jun 2016, pp. 4743–4752.

[68] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2017-Janua, jul 2017, pp. 2642–2651.

[69] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jun 2016, pp. 5375–5384.

[70] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with gaussian edge potentials," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, 2011, pp. 109–117.

[71] F. Meng, K. Luo, H. Li, Q. Wu, and X. Xu, "Weakly supervised semantic segmentation by a class-level multiple group cosegmentation and foreground fusion strategy," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4823–4836, dec 2020.

[72] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jun 2012, pp. 3282–3289.

[73] S. D. Jain and K. Grauman, "Supervoxel-consistent foreground propagation in video," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 656–671.

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[75] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2209–2218.

[76] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, sep 2014.

[77] Z. Wu, C. Shen, and A. van den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, 2019.

[78] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 558–567.

[79] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 2309–2318.

[80] P. Chen, B. Liao, G. Chen, and S. Zhang, "Understanding and utilizing deep neural networks trained with noisy labels," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 1062–1070.

DANIEL R. VILAR was born in Caracas, Venezuela, in 1995. He received the B.S. degree in electrical engineering from Universidad de Chile, Santiago, Chile, in 2017, where he is currently pursuing the M.S. degree in electrical engineering. He is currently a Research Assistant at the Image Processing Laboratory, Department of Electrical Engineering, Universidad de Chile. His main research interests include deep learning, image segmentation, image retrieval, and visual-semantic embeddings.

CLAUDIO A. PEREZ received the B.S. degree and the P.E. title in Electrical Engineering and the M.S. degree in Biomedical Engineering, all from Universidad de Chile in 1980 and 1985, respectively. He was a Fulbright student at the Ohio State University where he obtained a Presidential Fellow in 1990 and received the Ph.D. degree in 1991. He was a visiting scholar at UC, Berkeley in 2002 through the Alumni Initiatives Award Program from Fulbright Foundation. He is a Professor at the Department of Electrical Engineering, Universidad de Chile. He was the Department Chairman from 2003 to 2006 and Director of the Office of Academic and Research Affairs at the School of Engineering, Universidad de Chile from 2014 to 2018. His research interests include biometrics, image processing applications, convolutional neural networks and pattern recognition. He is a Senior Member of the IEEE, Systems, Man and Cybernetics and IEEE-CIS societies.