Weakly-supervised Video Scene Co-parsing

Guangyu Zhong[12⋆], Yi-Hsuan Tsai[1⋆], Ming-Hsuan Yang[1]

[1]UC Merced, [2]Dalian University of Technology
{gzhong,ytsai2,mhyang}@ucmerced.edu

Abstract. In this paper, we propose a scene co-parsing framework to assign pixel-wise semantic labels in weakly-labeled videos, i.e., only video-level category labels are given. To exploit rich semantic information, we first collect all videos that share the same video-level labels and segment them into supervoxels. We then select representative supervoxels for each category via a supervoxel ranking process. This ranking problem is formulated with a submodular objective function and a scene-object classifier is incorporated to distinguish scenes and objects. To assign each supervoxel a semantic label, we match each supervoxel to these selected representatives in the feature domain. Each supervoxel is then associated with a series of category potentials and assigned to a semantic label with the maximum one. The proposed co-parsing framework extends scene parsing from single images to videos and exploits mutual information among a video collection. Experimental results on the Wild-8 and SUNY-24 datasets show that the proposed algorithm performs favorably against the state-of-the-art approaches.

1 Introduction

Scene parsing, the task to assign labels for every pixel in images or videos [1,2,3], has attracted much attention in recent years. Many applications such as 3D layout estimation [4] and auto-driving [5] benefit from the results of scene parsing. However, existing scene parsing methods typically require large scale pixel-level annotated training images with fixed semantic categories and fully supervised [6] or retrieval-based [3,7] methods. Due to the restriction of category numbers and labor-intensive pixel-level annotations, it is not easy to directly apply these methods for videos with complex and dynamic scenes.

To relax the dependence on pixel-level annotations, several weakly-supervised methods [8,9,10] for video object segmentation have been proposed, in which only video-level semantic labels are given for each video. In these methods, segment-based classifiers are learned to distinguish objects from the background by extracting information with relevant and irrelevant frames from videos [8,9]. Object detectors are also used to help locate potential objects with specific semantics [10]. However, segment-based classifiers are suspecting to ambiguous training instances, and object detectors are not effective for parsing both scenes and objects in videos.

⋆ Both authors contribute equally to this work.

To address the above-mentioned issues, we extend scene parsing from single images to videos and propose a co-parsing framework to assign semantic labels in weakly-labeled videos. The proposed weakly-supervised method relaxes the constraints of large-scale annotations with fixed category numbers, while the co-parsing framework exploits information among a video collection to alleviate the ambiguity of individual segments. Considering the temporal consistency, we first segment each video into supervoxels [11,12]. Here, each supervoxel belongs to one or more parts of objects or scenes, which are quite different in terms of the contents (e.g., usually skies are smooth and objects are textured). Hence we develop a scene-object classifier to understand the contents of each supervoxel. Compared to previous methods using segment-based classifiers [8,13,14], our approach aims to learn a generalized scene-object classifier without the need to know specific semantic categories.

Since using the information within only one video is limited, we develop a co-parsing method to include videos that share the same labels. For each semantic category, we first collect all the supervoxels in videos with such label. We then select representative supervoxels through a submodular optimization problem guided by the scene-object classifier. These representative supervoxels selected in each category are further utilized in a matching process, in which we assign each supervoxel a potential to be a specific category by considering the similarities between the supervoxel and representative ones. Finally, a category is assigned to each supervoxel according to the maximum potential obtained from the matching process.

We demonstrate the effectiveness of the proposed weakly-supervised video co-parsing algorithm on the Wild-8 [13] and SUNY-24 [15] benchmark datasets. We first show the effectiveness of the proposed scene-object classifier incorporated in the submodular function for scene labeling. In addition, we extend the scene-object classifier to a detailed scene classifier with multiple categories, and show that the performance can be further improved. Overall, our experimental results show that the proposed algorithm performs favorably against the state-of-the-art methods in terms of visually quality and accuracy.

The main contributions of the proposed algorithm are summarized as follows. First, we propose a scene co-parsing framework for weakly-labeled videos, in which relations of supervoxels between different videos are exploited. Second, we formulate a submodular objective function to select representative supervoxels in semantics guided by a scene-object classifier from a collection of videos. Third, we propose an effective matching process to re-rank supervoxels in semantics and obtain final semantic labels in videos.

2 Related Work

Video Object Segmentation/Co-segmentation. In general, video object segmentation methods aim to detect and extract one or more dominant objects from a number of categories in image sequences [16,17,18,19,20,21,22,23]. These approaches achieve state-of-the-art performance by tracking segments [17], ex-

ploiting motion cues [18,24], propagating labels [19] or using spatial-temporal graph-based models [22]. However, these methods are not exploited to extract common objects from a video collection. Several video object co-segmentation methods have been proposed to separate common foreground objects from the background [25,26,27,28,29]. By analyzing the coherent motion and similar appearance in different videos, foreground objects from a collection of videos can be identified. However, these approaches usually assume that common objects appear in all the input videos, which is rarely true in real-world scenarios. In addition, these methods have limited capability in separating scenes in videos due to less motion information and large appearance variations of scenes. In contrast, the proposed method is able to parse scenes and objects with large appearance changes.

Object Segmentation in Weakly-labeled Videos. Weakly-supervised methods have recently been proposed for multi-class video segmentation [8,10,13,14]. Given the videos with video-level category labels, several learning-based approaches [8,14] use a large set of training samples to learn segment-based classifiers to distinguish foreground objects. To minimize the effect of ambiguous instances in the learning-based methods, pre-trained object detectors are incorporated to help locate objects [10]. However, object detectors can only locate instances from a number of known categories, rather than extract scenes in videos. Liu et al. [13] extend previous methods to a more challenging multi-class setting including objects and scenes. Based on the assumption that common objects in multiple videos should be similar in appearance, this method transfers video-level labels to each supervoxel via a nearest neighbor-based scheme. In contrast, the proposed algorithm introduces a more general and robust scene-object classifier without the need of training for specific categories, and considers the relations between different videos to facilitate the co-parsing task.

Image/Video Scene Parsing. Numerous approaches have been proposed for scene parsing [1,3,7,30,31,32,33]. These methods address the problem in single images via dense scene alignment [3], superpixel retrieval [7], neural networks [1,30] or context information [32]. However, directly applying single image parsing approaches to each video frame does not exploit temporal information and performs poorly in complex and dynamic scenes. Liu et al. [2] construct a conditional random field model to extract spatial-temporal information for video scene parsing, in which dense connections on supervoxel level and sparse object-level potentials are used for labeling. However, this method performs on single videos and requires manually pixel-wise labeled exemplars for initialization and propagation. In contrast, the proposed algorithm focuses on the co-parsing task from a video collection and requires only video-level labels. Recently, Chen et al. [34] propose a co-labeling task to parse scenes in multiple images, but without considering temporal connections. In addition, this method requires pixel-wise training samples and has a limitation that the training and test images should contain similar scenes. In contrast, the proposed algorithm exploits temporal consistency via supervoxels in a weakly-supervised fashion from a video collection.

3 Proposed Algorithm

3.1 Overview

Given a collection of videos with video-level labels, we aim to assign a semantic label to each pixel in image sequences. We formulate the labeling problem as a co-parsing task by simultaneously considering a video collection. To this end, our approach consists of two stages: (1) semantic supervoxel ranking via a submodular function: In this stage, we aim to discover representative semantic supervoxels for each category. We first segment weakly-labeled videos into supervoxels that maintain spatial-temporal consistency. For each semantic category, we construct a graph to connect the supervoxels collected from videos labeled with such category. To model the relations between supervoxels, we formulate it as a submodular optimization problem guided by a scene-object classifier based on the appearances and semantic information. The most representative supervoxels for each category are then extracted by solving this proposed submodular function; (2) scene co-parsing via region-based matching: In this stage, we aim to assign each supervoxel a semantic label by computing its category potentials. For each supervoxel, we compute similarities between it and representative ones for each category as its corresponding category potential. Each supervoxel is then assigned to a semantic label according to its maximum potential. Fig. 1 shows the main steps of the proposed algorithm.

3.2 Supervoxel Ranking via Submodular Function

Weakly-supervised video segmentation methods [2,13] usually transfer semantic labels to nearby regions spatially [13] or temporally [2] based on appearance features. However, globally searching the neighbors from all videos is likely to introduce redundant information and cause ambiguity when the videos share multiple semantic labels. In contrast, we start from each semantic label and aim to select representatives for each category. For each semantic category, we collect all the videos that share the same label and segment them into supervoxels. We then construct a graph where supervoxels are considered as nodes. We formulate a submodular optimization problem to select nodes that can represent each semantics.

Graph Construction. Given a collection of weakly-labeled videos V, we denote the full semantic label set as $\mathcal{L} = \{1, 2, \ldots, L\}$. For each category $l \in \mathcal{L}$, we collect videos containing l and segment them into supervoxels, which are denoted as $\mathcal{O}$. We construct a graph $G = (\mathcal{V}, \mathcal{E})$, in which each element $v \in \mathcal{V}$ is a supervoxel from $\mathcal{O}$ and the edge $e \in \mathcal{E}$ represents the pairwise relation between two supervoxels. To exploit the supervoxels that best represent a target category, we aim to select a subset $\mathcal{A}$ from $\mathcal{O}$.

Submodular Function. We model the supervoxel selection task as a facility location problem which can be solved by submodularity [35,36]. We design the

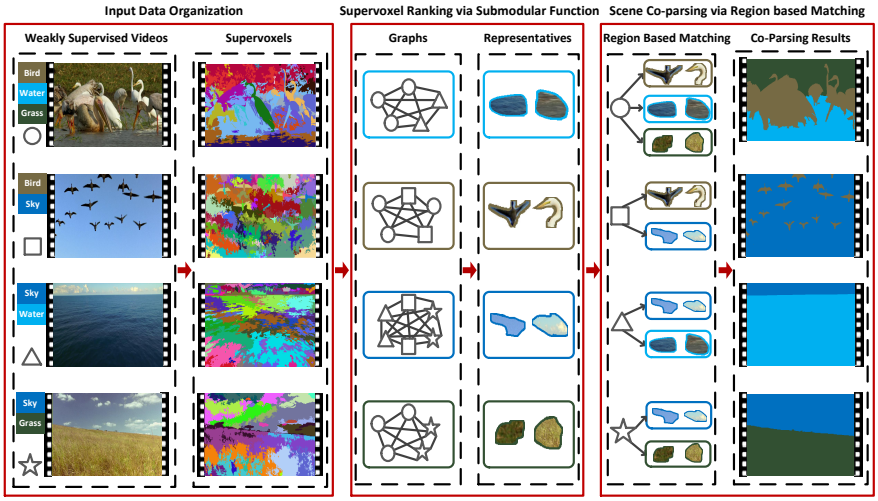

Fig. 1. Overview of the proposed algorithm. Given a collection of weakly-labeled videos, we aim to assign a semantic label to each pixel in image sequences. First, we segment each video into supervoxels. The supervoxels are illustrated by different patterns (e.g., circle represents all the supervoxels in the first video and rectangle represents the ones in the second video). We collect the supervoxels in videos that share the same semantic category and construct a graph. Each category is associated with a unique color (e.g., dark blue represents sky and green represents grass). We then formulate a submodular optimization problem to discover representative supervoxels for each category. Next, we match each supervoxel to the corresponding representatives and compute their similarities as the category potentials. Finally, a category is assigned to each supervoxel according to the maximum potential calculated during the matching process.

submodular objective function to find representative supervoxels that meet two criteria: (1) sharing high mutual similarities; (2) maintaining high probability to match the target category. To this end, we formulate the objective function with two terms, i.e., a facility-location term to show similarities among all the elements [23,37] and a semantic sensitive term to represent the potential of each element belongs to the target category. The formulation of facility-location (FL) term is defined by:

$$\mathcal{F}(\mathcal{A}) = \frac{1}{N_{\mathcal{V}}} \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{V}} w_{ij} - \sum_{i \in \mathcal{A}} \phi_i, \tag{1}$$

where ω_{ij} is the pairwise relation between a potential facility v_i and an element v_j. The cost of opening a facility is defined as ϕ_i and fixed to a constant σ (i.e., 1 in this work).

We represent the supervoxel v_i by a hierarchical convolutional neural network (CNN) feature vector f_i. For each supervoxel, we first extract CNN features in

each frame. The CNN features are computed by combining the first three convolutional layers [6] (i.e., 448-dimensional features). We then apply an average pooling method on all the frames and generate a feature vector for each superpixel. As f_i is extracted from hierarchical CNN layers which represent both visually fine-gained details and semantic information, features that share similar appearance and semantics should have higher mutual similarities.

To meet the first criteria, we define the pairwise relations ω_{ij} as the similarity between facilities and elements in the feature domain. This strategy encourages to select the node that well presents or is similar to its group elements so that the selected facilities in $\mathcal{A}$ are representative. We define the weight ω_{ij} of each edge e_{ij} in (1) as:

$$\omega_{ij} = S(v_i, v_j), \tag{2}$$

where $S(v_i, v_j)$ is the inner product of the features f_i and f_j, i.e., $\langle f_i, f_j \rangle$. The second term ϕ_i with a constant in (1) is to penalize excessive facilities. With the growth of $\mathcal{A}$, the cost of opening facilities becomes higher and thus it avoids selecting all the nodes.

However, videos within the same category usually contain various objects and scenes, and hence introduce large appearance variations and ambiguities between different categories. Therefore, it is not sufficient to use only facility-location term to discriminate different semantic information. To this end, we propose a unary term to represent the category sensitivity of each supervoxel. We define the proposed semantic sensitive (SS) term by:

$$\mathcal{U}(A) = \sum_{i \in \mathcal{A}} \psi_i, \tag{3}$$

where ψ_i denotes the potential of supervoxel v_i belonging to the target category. In the proposed algorithm, we apply classifiers to estimate these potentials. Considering large variations of semantic labels, learning classifiers for each category as [37] is time-consuming and labor-intensive. Hence we learn a generalized scene-object classifier based on the fully convolutional network (FCN) [6]. For each supervoxel with unknown category, we predict its category probabilities from the FCN output layer, i.e., scene and object. Then ψ_i for each supervoxel is computed in a way similar to the feature generation step, where the probability is first extracted from each frame and then averaged through all the frames.

Optimization for Supervoxel Ranking. To ensure that the selected facility set $\mathcal{A}$ shares more similarities with group elements and maintains high semantic sensitivities, we formulate the proposed submodular objective function of both facility-location term $\mathcal{F}(\mathcal{A})$ of (1) and semantic sensitive term $\mathcal{U}(\mathcal{A})$ of (3) by:

$$\begin{aligned} \max_{\mathcal{A}} \mathcal{C}(\mathcal{A}) = \max_{\mathcal{A}} \mathcal{F}(\mathcal{A}) + \lambda \mathcal{U}(\mathcal{A}), \\ \text{s.t. } \mathcal{A} \subseteq \mathcal{O} \subseteq \mathcal{V}, \mathcal{N}_{\mathcal{A}} \leq \mathcal{N}, \\ \mathcal{J}(\mathcal{A}^i) \geq 0, \end{aligned} \tag{4}$$

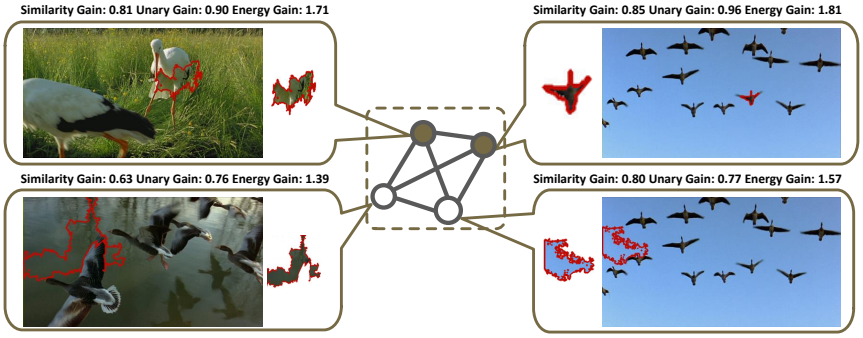

Fig. 2. Illustration of the proposed submodular function for selecting representatives. For the bird category, we show four supervoxels collected from three videos. The top two supervoxels are selected as the representatives as birds (denoted as circles with solid brown color). For each supervoxel, we show energy gain, similarity gain (FL term) and unary gain (SS term). As the video containing birds are usually accompanied with large sky or water regions, the supervoxels (e.g., the bottom two) in the scenes are likely to be similar to other regions and have high similarity gains. Owning to our scene-object classifier, the supervoxels containing the scenes (e.g., sky and water) are associated with lower unary gain in the object category (e.g., bird), and hence provide lower energy gains.

where $\mathcal{N}_{\mathcal{A}}$ is the number of selected facilities in $\mathcal{A}$, and $\mathcal{N}$ is the maximum number of $\mathcal{A}$. We set the energy gain $\mathcal{J}(\mathcal{A}^i)$ at the i-th iteration during the optimization as: $\mathcal{C}(\mathcal{A}^i) - \mathcal{C}(\mathcal{A}^{i-1})$. In addition, λ is the parameter to balance the contribution of two terms.

As the proposed objective function in (4) is the non-negative linear combination of two submodular terms, we can maximize $\mathcal{C}(\mathcal{A})$ via a greedy algorithm similar to [23]. We first initialize the facility set $\mathcal{A}$ as an empty set $\emptyset$. Then the element $a \in \mathcal{V} \setminus \mathcal{A}$ which leads to the maximum energy gain is added into $\mathcal{A}$. We iteratively select other elements and this absorbing process stops when either one of the following conditions is satisfied: (1) the maximum facility number is reached, i.e., $\mathcal{N}_{\mathcal{A}} > \mathcal{N}$; (2) the cost of opening facilities is larger than the gain from elements, i.e., $\mathcal{J}(\mathcal{A}^i) < 0$. In addition, due to the submodularity of the objective function, the optimization process can be sped up by an evaluation form as proposed in [38]. The process of selecting representatives for each category is presented in Algorithm 1 and the effectiveness of our submodular function is shown in Fig. 2.

3.3 Scene Co-parsing via Region-based Matching

Next, we aim to assign each supervoxel a semantic label by considering its relations to representatives in each category. Previous approaches generally formulate the labeling task as markov random field (MRF) [13] or conditional random

Algorithm 1 Representatives Selection for Each Category

Input: $G = (\mathcal{V}, \mathcal{E}), \mathcal{N}, \lambda$
Initialization: $\mathcal{A}^0 \leftarrow \emptyset$, $\mathcal{O}^0 \leftarrow \mathcal{V}$, $i \leftarrow 1$
loop
  $a^* = \arg\max_{\{\mathcal{A}^i \in \mathcal{V}\}} \mathcal{J}(\mathcal{A}^i)$, where $\mathcal{A}^i = \mathcal{A}^{i-1} \cup a$
  if $\mathcal{N}_{\mathcal{A}} > \mathcal{N}$ or $\mathcal{J}(\mathcal{A}^i) < 0$ **then**
    break
  end if
  $\mathcal{A}^i \leftarrow \mathcal{A}^{i-1} \cup a^*$, $\mathcal{O}^i \leftarrow \mathcal{O}^{i-1} \setminus a^*$
  $i = i + 1$
end loop
Output: $\mathcal{A} \leftarrow \mathcal{A}^i$, $\mathcal{O} \leftarrow \mathcal{O}^i$

field (CRF) [2] models that require additional optimization process to estimate category posteriors. In contrast, we propose an efficient way and predict the potential of each supervoxel by matching them to category representatives. We then assign each supervoxel a semantic label according to the maximum potential computed during the matching process.

Region-based Matching. The energy gain in the proposed submodular function can be utilized to estimate how likely each supervoxel belonging to which category as proposed in [39]. However, generating energy gains for all the elements in each category is ineffective. In addition, different submodular functions for various categories may differ in the graph size, element appearance and semantic distribution, which causes incomparable results when comparing energy gains between different categories. In this work, we propose to compute the category potentials for each supervoxel by a matching process, which is efficient and can reduce confusions between categories (see Fig. 3).

For the input video collection V, we denote all the supervoxels as $B = \{b_1, \ldots, b_M\}$, where M is the number of supervoxels. The corresponding video-level labels of B are denoted as $Y = \{y_1, \ldots, y_M\}$, where y_i is a semantic label set according to the video that b_i belongs to. For instance, y_i is identical to y_j if they belong to the same video. For each supervoxel b_i, we aim to compute its category potentials with respect to all the semantic labels $\mathcal{L}$. Hence we generate a M by L matrix P, where each element p_i^l denotes the potential of a supervoxel b_i belonging to a category l. To ensure that potentials between different categories are comparable, we compute P in the feature domain. For each category l, the p_i^l is computed as the average similarity between the feature f_i of supervoxel b_i and feature f_j^l of category representatives a_j^l in $\mathcal{A}_l$, where $\mathcal{A}_l$ is the corresponding facility set:

$$p_i^l = \begin{cases} \dfrac{1}{\mathcal{N}_l} \displaystyle\sum_{a_j^l \in \mathcal{A}_l} \langle f_i, f_j^l \rangle & l \in y_i, \\ -\infty & l \notin y_i, \end{cases} \tag{5}$$

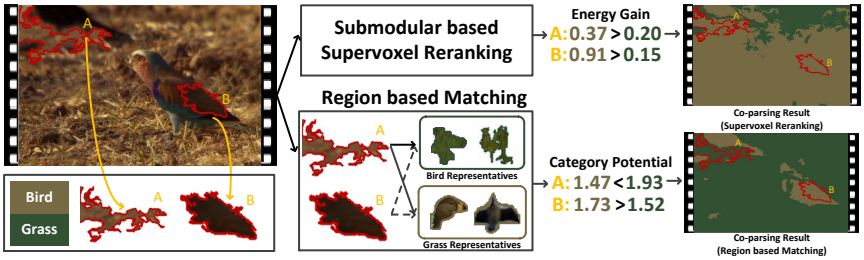

Fig. 3. Illustration of the proposed region-based matching process. Given a video with weakly-supervised labels (e.g., bird and grass), we first extract its supervoxels (e.g., A and B with red boundaries). We then compare them with representative supervoxels for each category and compute category potentials. By comparing these potentials, we can further assign a label to each supervoxel corresponding to the category with the maximum one (bottom row). In contrast, using the energy gain for comparing scores and assigning labels may produce wrong results, due to incomparable gains in different categories (upper row). Note that categories of the potential scores are associated with different colors.

where $\mathcal{N}_l$ denotes the number of representatives in $\mathcal{A}_l$. Note that if the video does not contain the target category l, p_i^l is assigned as the value $-\infty$, meaning that supervoxels in this video do not share any similarities with category l.

Scene Label Assignment. After matching supervoxels to representatives in each category, we assign a semantic label for each supervoxel b_i based on its category potential vector $P_i = [p_i^1, \ldots, p_i^L]$ as:

$$c_i = \underset{l \in \mathcal{L}}{\operatorname{argmax}}\ P_i(l). \tag{6}$$

Hence all the pixels within a supervoxel have the same assigned label c_i. The proposed region-based matching strategy to assign labels is illustrated in Fig. 3. Note that if there are existing multiple semantic labels with the same potential, we use their submodular energy gains for further comparisons. In addition, all the video-level labels in each video should be at least assigned to one of the supervoxels. For those labels that are not assigned to any supervoxel in the video, we ensure that at least top K (i.e., 15) supervoxels with high category potentials are re-assigned to that label.

4 Experimental Results

We evaluate the proposed scene co-parsing algorithm on the Wild-8 [13] and SUNY-24 [15] datasets with comparisons to the state-of-the-art methods, including CRANE [8], MIN [40], SVM [14], MIL [41] and WILD [13]. We use the same metrics in [13] for evaluation including average accuracy of each class,

Table 1. Video scene co-parsing results on the Wild-8 dataset. We measure the average accuracy of each class, average per category accuracy (aveAcc) and mean average precision (mAP). The highest score is marked in bold and the second highest score is marked with underlines.

Category	MIL	SVM	MIN	CRANE	WILD	Ours (similarity)	Ours (generalized)	Ours (detailed)
bird	31.5	42.5	48.1	47.8	53.0	35.8	41.5	**66.2**
water	79.3	74.5	75.2	76.5	77.3	60.5	72.5	**82.6**
sky	85.4	86.9	87.2	89.5	93.8	78.3	96.0	**98.2**
tree	41.1	45.5	36.7	42.8	50.1	93.0	93.0	**95.5**
grass	78.3	74.0	74.1	73.7	76.5	**81.0**	72.4	68.3
lion	2.1	16.6	15.4	19.3	21.3	47.1	**95.8**	91.5
sand	55.2	42.1	43.3	43.2	60.1	35.0	44.7	**75.3**
elephant	5.5	12.3	13.2	16.8	28.1	51.8	**87.1**	73.4
aveAcc	47.3	49.3	49.2	51.2	57.5	60.3	75.4	**81.4**
mAP	41.8	41.0	42.1	43.9	52.4	59.6	68.6	**78.6**

Table 2. Comparisons of using energy gains and region-based matching process for assigning final semantic labels on the Wild-8 dataset. We show two sets of results using different classifiers and measure the average per category accuracy (aveAcc) and mean average precision (mAP). The results consistently show the effectiveness of the proposed matching strategy. The highest scores are marked in bold.

Indicator	Energy gain (generalized)	Matching (generalized)	Energy gain (detailed)	Matching (detailed)
aveAcc	70.2	**75.4**	78.0	**81.4**
mAP	64.0	**68.6**	75.3	**78.6**

average per category accuracy (aveAcc), and mean average precision (mAP). More experimental results can be found in the supplementary material and the MATLAB codes will be made available to the public.

4.1 Experimental Settings

We use the streamGBH algorithm [12] at the fifteenth level with the default parameters to generate supervoxels in videos, For optimizing the submodular function in (4), the maximum number $\mathcal{N}$ of each category is set to 10, and the parameter λ is set to 1. For matching representatives in (5), we use the top-5 ranked supervoxels as $\mathcal{A}_l$.

To learn the scene-object classifier, we finetune a fully convolutional network (FCN) [6], which is a state-of-the-art algorithm for semantic segmentation. We follow the same setting used in [42] for collecting training images in the LMSun dataset [7], but merging all the categories into scene and object categories. The parameters are fixed in all the experiments.

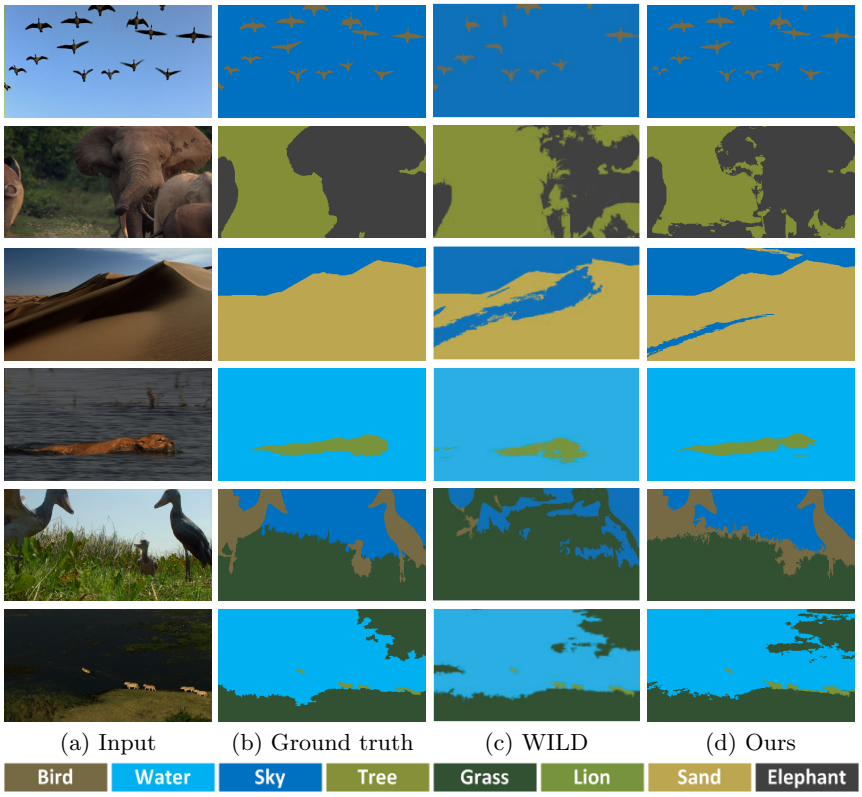

Fig. 4. Sample results of the proposed method (with the generalized scene-object classifier) and the WILD [13] method on the Wild-8 dataset. The results show that the proposed algorithm generates more complete and accurate results than WILD, especially on objects (lion and elephant). Each color indicates a semantic label and the legend is shown on the bottom.

4.2 Wild-8 Dataset

The Wild-8 dataset consists of 100 weakly-labeled videos, and 33 of them are with pixel-level annotations. It contains 8 categories including scenes (sky, tree, grass, sand and water) and objects (bird, lion and elephant). Each video is associated with multiple video-level labels and contains 30 frames with 640 × 480 resolution.

We first evaluate the contribution of the scene-object classifier in the submodular function. Table 1 shows that the proposed algorithm with the generalized scene-object classifier (second column from the right) significantly improves the performance of the state-of-the-art methods (e.g., more than 15% gain in terms of aveAcc and mAP). In addition, the scene-object classifier improves results

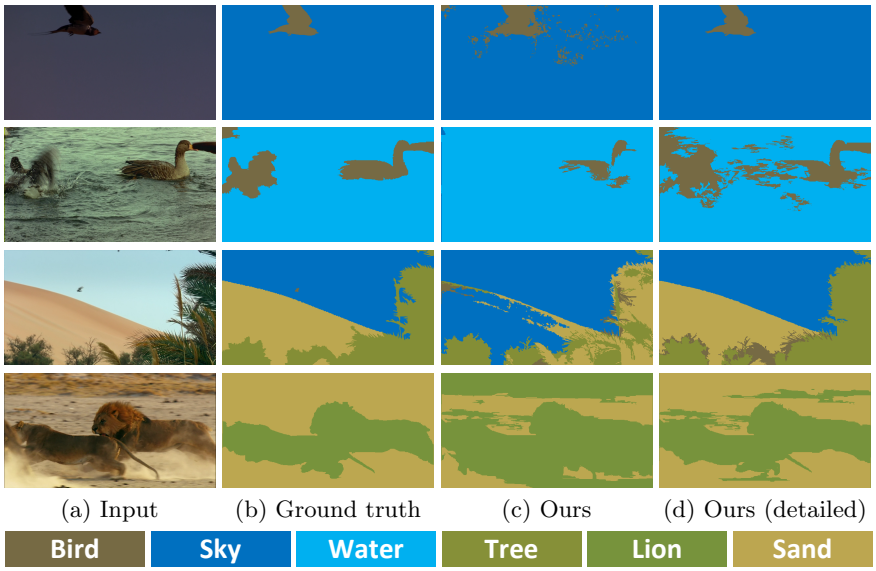

Fig. 5. Sample results of the proposed method with different classifiers on the Wild-8 dataset. The results in (c) and (d) are generated by the proposed method with the generalized and detailed scene-object classifier. The results show that the detailed scene classifier further improves the performance on both object (e.g, bird) and scene (e.g., sand) categories. Each color indicates a semantic label and the legend is shown on the bottom.

in most categories compared with only using the similarity term (third column from the right).

Overall, the proposed algorithm performs favorably on both object (e.g., lion and elephant) and scene (e.g., sky and tree) categories. The MIL method [40] achieves high accuracy in categories of water and grass since it uses the max-margin strategy, which contributes more to categories with larger regions while usually ignoring small objects (e.g., lion and elephant). The WILD scheme [13] performs well in the bird and sand categories. However, when objects and scenes have similar appearances, the smoothness assumption used in [13] may introduce ambiguity and thus it leads to low accuracy (e.g., lion, elephant and tree). In contrast, the proposed algorithm not only considers similarities between supervoxels but also utilizes a scene-object classifier, which is able to guide the submodular function for separating scenes and objects. Fig. 4 shows some results generated by the WILD method [13] and proposed algorithm with the generalized scene-object classifier. As the codes and results of the WILD method [13] are not available, we use the reported results from the original paper for illustration. The parsing results of the proposed algorithm are more complete and accurate, especially on object regions (lion and elephant). The results also demonstrate

the usefulness of the proposed scene-object classifier as it helps analyze video contents and discriminate objects from various scenes.

Next, we extend the scene-object classifier to a detailed scene classifier with multiple categories. To achieve this, we carry out another finetuned FCN by using the same training set as mentioned before, but keeping different scene categories (without merging to a single scene category). Here, we still keep a single object category due to the fact that objects usually vary a lot in different scenes while there are common scene categories appearing in different videos.

Table 1 (rightmost column) shows that the proposed algorithm with the detailed scene classifier achieves highest accuracy in terms of aveAcc (i.e., 81.4%) and mAP (i.e., 78.6%) when compared to all the other methods. As the detailed scene classifier provides more discriminative information for each scene category, the proposed algorithm further improves results in most categories, especially in bird and sand classes, which are significantly improved from 41.5% to 66.2% and from 44.7% to 75.3%, respectively. Fig. 5 shows sample results of the proposed algorithm guided by the proposed two classifiers, i.e., the generalized scene-object classifier and the detailed one.

Furthermore, we validate the effectiveness of the proposed region-based matching process by comparing it with directly using the energy gain obtained from the submodular optimization. When iteratively selecting supervoxels into the facility set in the submodular function for a target category, each supervoxel is associated with an energy gain to represent the potential to be the target category (see Fig. 3 for an illustration). Table 2 shows that compared with the method using energy gain, the matching process consistently improves the results with both generalized and detailed scene-object classifiers.

4.3 SUNY-24 Dataset

We also evaluate the proposed method on the challenging SUNY-24 dataset [15], which contains 24 categories and 8 videos. Each video is taken in one scene with motion of camera or objects, and contains 70 to 88 frames with pixel-level annotations. Each single video usually contains multiple categories with small semantic regions and complex scenes (see Fig. 6). The mutual information among the video collection is insufficient and ambitious. These factors make this dataset even challenging. We compare our method with the generalized and detailed scene-object classifiers to the CRANE [8] and WILD [13] methods. Fig. 6 shows the challenge in the SUNY-24 dataset and demonstrates the effectiveness of the proposed method with the generalized and detailed scene-object classifiers. Table 3 shows that the proposed method performs favorably on this challenging dataset. The results are improved with more than 7% gains in terms of aveAcc using the proposed generalized model, and more than 20% gains using the detailed scene classifier.

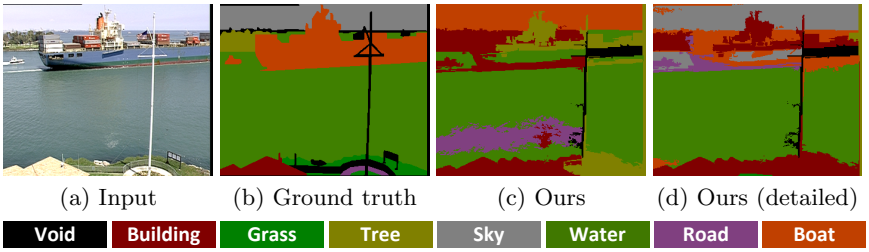

Fig. 6. Sample results of the proposed method with different classifiers on the SUNY-24 dataset. The results in (c) and (d) are generated by the proposed method with the generalized and detailed scene-object classifiers. The results show the challenges in the SUNY-24 dataset, i.e., some semantics in the video are small (e.g., building, grass and road) and the scenes are various and complex (e.g., the sample video contains 6 different scene labels). The proposed method with the generalized scene-object classifier successfully labels some of the semantics (e.g., void and water), and the performance is further improved by the detailed one (e.g., boat, water and sky). Each color indicates a semantic label and the legend is shown on the bottom.

Table 3. Video scene co-parsing in terms of the average per category accuracy (aveAcc) on SUNY-24.

Indicator	CRANE	WILD	Ours (generalized)	Ours (detailed)
aveAcc	13.8	14.1	21.6	35.42

5 Concluding Remarks

In this paper, we propose a scene co-parsing framework to assign semantic pixel-wise labels in weakly-labeled videos. We first extract representative supervoxels for different categories by a supervoxel ranking scheme. To relax the constraints of large-scale annotations with fixed category numbers, we incorporate the generalized scene-object classifier into the submodular objective function which provides guidance in distinguishing the contents of objects and scenes. By iteratively optimizing the submodular function, we select representatives that share mutual similarities and maintain higher probabilities to match the target category. To exploit semantic information among the video collection, we predict category potentials and assign semantic labels for each supervoxel by matching it to the selected representatives. Experimental results on the Wild-8 and SUNY-24 datasets show that the proposed algorithm performs favorably against the state-of-the-art approaches in terms of visual quality and accuracy.

Acknowledgments. This work is supported in part by the NSF CAREER grant #1149783, NSF IIS grant #1152576, and gifts from Adobe and Nvidia. G. Zhong is sponsored by China Scholarship Council and NSFC grant #61572099.

References

1. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2013) 1915–1929 1, 3
2. Liu, B., He, X.: Multiclass semantic video segmentation with object-level active inference. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1, 3, 4, 8
3. Liu, C., Yuen, J., Torralba, A.: Nonparametric scene parsing via label transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2011) 2368–2382 1, 3
4. Liu, X., Zhao, Y., Zhu, S.C.: Single-view 3d scene parsing by attributed grammar. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1
5. Zhang, C., Wang, L., Yang, R.: Semantic segmentation of urban scenes using dense depth maps. In: Proceedings of the 11th European Conference on Computer Vision. (2010) 1
6. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1, 6, 10
7. Tighe, J., Lazebnik, S.: Superparsing: Scalable nonparametric image parsing with superpixels. International Journal of Computer Vision **101** (2013) 329–349 1, 3, 10
8. Tang, K., Sukthankar, R., Yagnik, J., Fei-Fei, L.: Discriminative segment annotation in weakly labeled video. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2013) 1, 2, 3, 9, 13
9. Wang, L., Hua, G., Sukthankar, R., Xue, J., Zheng, N.: Video object discovery and co-segmentation with extremely weak supervision. In: Proceedings of the 13th European Conference on Computer Vision. (2014) 1
10. Zhang, Y., Chen, X., Li, J., Wang, C., Xia, C.: Semantic object segmentation via detection in weakly labeled video. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1, 3
11. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph-based video segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2010) 2
12. Xu, C., Xiong, C., Corso, J.J.: Streaming hierarchical video segmentation. In: Proceedings of the 12th European Conference on Computer Vision. (2012) 2, 10
13. Liu, X., Tao, D., Song, M., Ruan, Y., Chen, C., Bu, J.: Weakly supervised multiclass video segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2014) 2, 3, 4, 7, 9, 11, 12, 13
14. Hartmann, G., Grundmann, M., Hoffman, J., Tsai, D., Kwatra, V., Madani, O., Vijayanarasimhan, S., Essa, I., Rehg, J., Sukthankar, R.: Weakly supervised learning of object segmentations from web-scale video. In: Proceedings of the 12th European Conference on Computer Vision Workshop. (2012) 2, 3, 9
15. Chen, A.Y., Corso, J.J.: Propagating multi-class pixel labels throughout video frames. In: Proceedings of Western New York Image Processing Workshop. (2010) 2, 9, 13
16. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: Proceedings of IEEE International Conference on Computer Vision. (2011) 2

17. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: Proceedings of IEEE International Conference on Computer Vision. (2013) 2
18. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: Proceedings of IEEE International Conference on Computer Vision. (2013) 2, 3
19. Jain, S.D., Grauman, K.: Supervoxel-consistent foreground propagation in video. In: Proceedings of the 13th European Conference on Computer Vision. (2014) 2, 3
20. Wen, L., Du, D., Lei, Z., Li, S.Z., Yang, M.H.: Jots: Joint online tracking and segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2
21. Nagaraja, N.S., Schmidt, F., Brox, T.: Video segmentation with just a few strokes. In: Proceedings of IEEE International Conference on Computer Vision. (2015) 2
22. Tsai, Y.H., Yang, M.H., Black, M.J.: Video segmentation via object flow. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2, 3
23. Tsai, Y.H., Zhong, G., Yang, M.H.: Semantic co-segmentation in videos. In: Proceedings of the 14th European Conference on Computer Vision. (2016) 2, 5, 7
24. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: Proceedings of the 11th European Conference on Computer Vision. (2010) 3
25. Rubio, J.C., Serrat, J., López, A.: Video co-segmentation. In: Proceedings of the 11th Asian Conference on Computer Vision. (2012) 3
26. Chiu, W.C., Fritz, M.: Multi-class video co-segmentation with a generative multi-video model. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2013) 3
27. Fu, H., Xu, D., Zhang, B., Lin, S.: Object-based multiple foreground video co-segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2014) 3
28. Guo, J., Cheong, L.F., Tan, R.T., Zhou, S.Z.: Consistent foreground co-segmentation. In: Proceedings of the 12th Asian Conference on Computer Vision. (2014) 3
29. Zhang, D., Javed, O., Shah, M.: Video object co-segmentation by regulated maximum weight cliques. In: Proceedings of the 13th European Conference on Computer Vision. (2014) 3
30. Socher, R., Lin, C.C., Manning, C., Ng, A.Y.: Parsing natural scenes and natural language with recursive neural networks. In: Proceedings of the 28th International Conference on Machine Learning. (2011) 3
31. Munoz, D., Bagnell, J.A., Hebert, M.: Stacked hierarchical labeling. In: Proceedings of the 11th European Conference on Computer Vision. (2010) 3
32. Yang, J., Price, B., Cohen, S., Yang, M.H.: Context driven scene parsing with attention to rare classes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2014) 3
33. Xu, J., Schwing, A.G., Urtasun, R.: Tell me what you see and I will show you where it is. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2014) 3
34. Chen, X., Jain, A., Davis, L.S.: Object co-labeling in multiple images. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision. (2014) 3
35. Galvão, R.D.: Uncapacitated facility location problems: contributions. Pesquisa Operacional **24** (2004) 7–38 4

36. Lazic, N., Givoni, I., Frey, B., Aarabi, P.: Floss: Facility location for subspace segmentation. In: Proceedings of IEEE International Conference on Computer Vision. (2009) 4
37. Zhu, F., Jiang, Z., Shao, L.: Submodular object recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2014) 5, 6
38. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2007) 7
39. Yang, F., Jiang, Z., Davis, L.S.: Submodular reranking with multiple feature modalities for image retrieval. In: Proceedings of the 12th Asian Conference on Computer Vision. (2014) 8
40. Siva, P., Russell, C., Xiang, T.: In defence of negative mining for annotating weakly labelled data. In: Proceedings of the 12th European Conference on Computer Vision. (2012) 9, 12
41. Vezhnevets, A., Ferrari, V., Buhmann, J.M.: Weakly supervised semantic segmentation with a multi-image model. In: Proceedings of IEEE International Conference on Computer Vision. (2011) 9
42. Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Yang, M.H.: Sky is not the limit: Semantic-aware sky replacement. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH) (2016) 10