Rethinking Interactive Image Segmentation: Feature Space Annotation

Jordão Bragantini, Alexandre Falcão, Laurent Najman

Rethinking Interactive Image Segmentation: Feature Space Annotation

Jordão Bragantini[1,2] *, Alexandre X. Falcão[1], Laurent Najman[2]
[1]University of Campinas, Laboratory of Image Data Science, Brazil
[2]Université Gustave Eiffel, LIGM, Equipe A3SI, ESIEE, France
jordao.bragantini@gmail.com, afalcao@ic.unicamp.br, laurent.najman@esiee.fr

Abstract

Despite the progress of interactive image segmentation methods, high-quality pixel-level annotation is still time-consuming and laborious — a bottleneck for several deep learning applications. We take a step back to propose interactive and simultaneous segment annotation from multiple images guided by feature space projection and optimized by metric learning as the labeling progresses. This strategy is in stark contrast to existing interactive segmentation methodologies, which perform annotation in the image domain. We show that our approach can surpass the accuracy of state-of-the-art methods in foreground segmentation datasets: iCoSeg, DAVIS, and Rooftop. Moreover, it achieves 91.5% accuracy in a known semantic segmentation dataset, Cityscapes, being 74.75 times faster than the original annotation procedure. The appendix presents additional qualitative results. Code and video demonstration will be released upon publication.

1 Introduction

Convolutional Neural Networks (CNNs) can achieve excellent results on image classification, image segmentation, pose detection, and other images/video-related tasks [1, 2, 3, 4, 5], at the cost of an enormous amount of high-quality annotated data and processing power. Thus interactive image segmentation with reduced user effort is of primary interest to create such datasets for the training of CNNs. Concerning image segmentation tasks, the annotations are pixel-wise labels, usually defined by interactive image segmentation methods [6, 7, 8, 9, 10, 11] or by specifying polygons in the object boundaries [12].

Recent interactive image segmentation methods based on deep learning can significantly reduce user effort by performing object delineation from a few clicks, sometimes in a single user interaction [13, 14]. However, such deep neural networks do not take user input as hard constraints and so cannot provide enough user control. Novel methods can circumvent this issue by refining the neural network's weights while enforcing the correct results on the annotated pixels [15, 16, 17]. Their results are remarkable for foreground segmentation. Still, in complex cases, segmentation may be unsatisfactory even by extensive user effort.

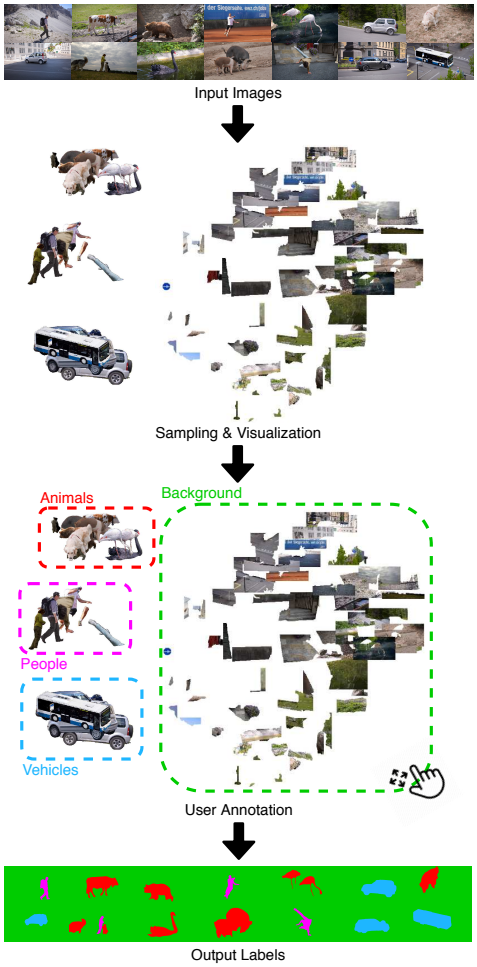


Figure 1: Our approach to interactive image segmentation: candidate segments are sampled from the dataset and presented in groups of similar examples to the user, who annotates multiple segments in a single interaction.

*Work done during a internship at LIGM.

The big picture in today's image annotation tasks is that thousands of images with multiple objects require user interaction. While they might not share the same visual appearance, their semantics are most likely related. Hence, thousands of clicks to obtain thousands of segments with similar contexts do not sound as appealing as before.

This work presents a scheme for interactive large-scale image annotation that allows user labeling of many similar segments at once. It starts by defining segments from multiple images and computing their features with a neural network pre-trained from another domain. User annotation is based on feature space projection, Figure 1. As it progresses, the similarities between segments are updated with metric learning, increasing the discrimination among classes, and further reducing the labeling burden.

Our goal is to demonstrate that other forms of human-machine interaction, notably feature space interaction, can benefit the interactive image segmentation paradigm.

This work is organized as follow: Section 2 reviews related works. Our approach, with a new way to conceive user interaction, is presented in Section 3. We evaluate our work, with an ablation study, in Section 4, and additional experiments using state-of-the-art interactive segmentation methods and the budget reported to produce the original ground-truth. The methods are evaluated on several datasets: iCoSeg [18], DAVIS [19], Rooftop [20], and Cityscapes [21].

2 Related Works

In this work, we address the problem of assigning a label (i.e. class) to every pixel in a collection of images, denoted as image annotation in the remaining of the paper. This is related to the foreground (i.e. region) segmentation microtasks, where the region of interest has no specific class assigned to it, and the delineation of the object is of primary interest — in standard image annotation procedures, this is the step preceding label assignment [12].

Current deep interactive segmentation methods, from click [13, 14, 22, 17, 15, 16], bounding-box [22, 23] to polygon-based approaches [24, 25, 26] address this microtask of segmenting and then labeling each region individually to generate annotated data.

A minority of methods segment multiple objects jointly; to our knowledge, in deep learning, this has been employed only once [27]; in classical methods, a hand-full could do this efficiently [28, 29, 10].

Fluid annotation [30] proposes a unified human-machine interface to perform the complete image annotation; the user annotation process starts from the predictions of an existing model, requiring user interaction only where the model lacks accuracy, further reducing the annotation effort. The user decides which action it will perform at any moment without employing active learning (AL). Hence, the assumption is that the user will take actions that will decrease the annotation budget the most.

This approach falls in the Visual Interactive Labeling (VIAL) [31] framework, where the user interface should empower the users, allowing them to decide the optimal move to perform the task efficiently. Extensive experiments [32] have shown that this paradigm is as competitive as AL and obtains superior performance when starting with a small amount of annotated data.

Inside the VIAL paradigm, feature space projection has been employed for user guidance in semi-supervised label propagation [33, 34] and for object detection in remote-sensing [35]. However, its use for image segmentation has not been explored yet.

Other relevant works and their relationship with our methodology are further discussed in the next section.

3 Proposed Method

Our methodology allows the user to annotate multiple classes jointly and requires lower effort when the data are redundant by allowing multiple images to be annotated at once. Multiple image annotation is done by extracting candidate regions (i.e. segments) and presenting samples of the same class simultaneously to the user; multiple examples can be annotated in a single interaction. The candidate segments are displayed closer together according to their visual similarity [36] for the label assignment of multiple segments at once. Hence, user interaction in the image domain is only employed when necessary — not to assign labels but to fix incorrect segments. Since, this action is the one that requires the most user effort and has been the target of several weakly-supervised methodologies [37, 38, 39, 40] that try to avoid it altogether. Therefore, this approach is based on the following pillars:

- The segment annotation problem should be evaluated as a single task [30]. While dividing the problem into microtasks is useful to facilitate the user and machine interaction, they should not be treated independently since the final goal is the complete image annotation.
- The human is the protagonist in the process, as described in the VIAL process [31], deciding which action minimizes user effort for image annotation while the machine assists in well-defined microtasks.
- The annotation in the image domain is burdensome; thus, it should be avoided, but not neglected, since perfect segmentation is still an unrealistic assumption.
- The machine should assist the user initially, even when no annotated examples are present [32], and as the annotation progresses, labeling should get easier because more information is provided.

3.1 Overview

The proposed methodology is summarized in Figure 2 and each component is described in the subsequent sections.

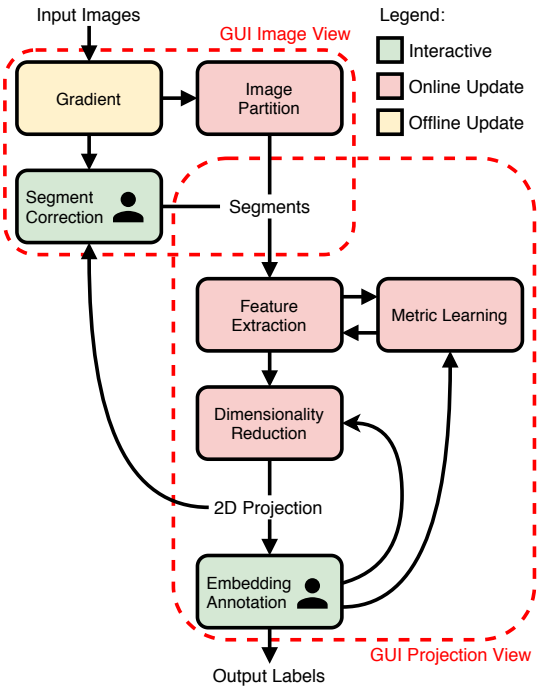


Figure 2: The proposed feature space annotation pipeline.

The user interface is composed of two primary components, the Projection View and the Image View. Red contours in Figure 2 delineate which functionalities are present in these widgets. The former is concerned with displaying the segments arranged in a canvas (as shown in Figure 1), enabling the user to interact with it: assigning labels to clusters, focusing on cluttered regions, and selecting samples for segmentation inspection and correction in the image domain.

Image View displays the image containing a segment selected in the canvas. It is highlighted to allow fast component recognition among the other segments' contours. Samples already labeled are colored by class. This widget allows further user interaction to fix erroneous delineation, as further discussed in Section 3.7. Segment selection also works backwardly, when a region is selected in the Image View, its feature space neighborhood is focused on the projection canvas, accelerating the search for relevant clusters by providing a mapping from the image domain to the projection canvas.

The colored rectangles in Figure 2 represent data processing stages: yellow represents fixed operations that are not updated during user interaction, red elements are updated as the user annotation progresses, and the greens are the user interaction modules. Arrows show how the data flows in the pipeline.

The pipeline works as follows, starting from a collection of images, their gradients are computed to partition each image into segments, which will be the units processed and annotated in the next stages.

Since we wish to cluster together similar segments, we must define a similarity criterion. Therefore, for each segment, we obtain their deep representation (i.e. features). Their Euclidean distances are used to express this information — they are more dissimilar as they are further apart in the feature space.

The next step concerns the notion of similarity between segments as presented and perceived by the user. We propose communicating this information to the user by displaying samples with similar examples in the same neighborhood. Hence, the segments' features are used to project them into the 2D plane while preserving, as best as possible, their relative feature space distances.

The user labeling process is executed in the 2D canvas by defining a bounding-box and assigning the selected label to the segments inside it. As the labeling progresses, their deep representation is updated using metric learning, improving class separability, enhancing the 2D embedding, thus, reducing the annotation effort. We refer to [31] for a review in visual interactive labeling and [36] for interactive dimensionality reduction systems.

This pipeline relies only upon the assumption that it is possible to find meaningful candidate segments from a set of images and extract discriminant features from them to cluster together similar segments. Even though these problems are not solved yet, existing methods can satisfy these requirements, as they are validated in our ablation studies (Section 4.3).

3.2 Gradient and Image Partition

Gradient computation and watershed-based image partition are operations of the first step of the pipeline, obtaining initial candidate segments. In the ideal scenario, the desired regions are represented by a single connected component, requiring no further user interaction besides labeling.

However, obtaining meaningful regions is a challenging and unsolved problem. Desired segments vary from application to application. On some occasions, users wish to segment humans and vehicles in a scene, while in the same image, other users may desire to segment the clothes and billboards. Thus, the proposed approach has to be class agnostic and enables the user to obtain different segment categories without effort.

The usual approach computes several solutions (e.g. Multiscale Combinatorial Grouping (MCG) [41]) and employs a selection policy to obtain the desired segments. However, it does not guarantee disjoint regions, and it gen-

erates thousands of candidates, further complicating the annotation process.

Given that segments should be disjoint, and they are also task dependent, we chose to employ hierarchical segmentation techniques for this stage. Most of the relevant literature for this problem aims at obtaining the best gradient (edge saliency) to compute segmentation.

In Convolutional Oriented Boundaries [42], a CNN predicts multiple boundaries in multiple scales and orientations and are combined into an ultrametric contour map to perform the hierarchical segmentation – Refer to [43, 44] to review the duality between contours and hierarchies.

Recently, various CNN architectures for edge estimation were developed [45, 46, 47, 48, 49], where the multi-scale representation is fused into a single output image. In such methods, boundaries that appear in finer scales present lower values than the clearest ones.

While segmentation can be computed from the gradient information directly, performing it in learned features has shown to be competitive. Isaacs et al. [50] propose a novel approach to use deep features to improve class agnostic segmentation, learning a new mapping for the pixels values, without considering any class information and further separating their representation apart.

Other tasks also employ edge estimation to enhance their performance, notably PoolNet [51] switches between saliency object prediction and edge estimation in the training loop with the same architecture to obtain saliency with greater boundary adherence. We noticed that this approach produces less irrelevant boundaries for image segmentation; thus, we chose this method for our experimental setup.

We opted to employ the flexible hierarchical watershed framework [52, 53] for delineating candidate segments on the gradient image estimated by PoolNet [51] architecture. The hierarchical watershed allows manipulation of the region merging criterion, granting the ability to rapidly update the segments' delineation. Besides, hierarchical segmentation lets the user update segments without much effort (e.g. , obtaining a more refined segmentation by reducing the threshold, but as a trade-off, the number of components increases). Further, a watershed algorithm can also interactively correct the delineated segments (Section 3.7).

Starting from a group of N images without annotations, $\{I_1, I_2, \ldots, I_N\}$, their gradient images are computed. For each gradient G_i, $i \in [1, N]$, its watershed hierarchy is built and disjoint segments $\{S_{i,1}, \ldots, S_{i,n_i}\}$ are obtained by thresholding the hierarchy. The required parameters (threashold and hierarchy criterion) are robust and easy to be defined by visual inspection on a few images. More details about that are presented in Section 3.7.

3.3 Feature Extraction

Before presenting the regions arranged by similarity, a feature space representation where dissonant samples are separated must be computed. For that, we refer to CNN architectures for image classification tasks, without their fully connected layers [54] used for image classification.

Each segment is treated individually; we crop a rectangle around the segment in the original image, considering an additional border to not impair the network's receptive field. In this rectangle, pixels that do not belong to the segment are zeroed out. Otherwise, segments belonging to the same image would present similar representations. The segment images are then resized to 224×224 and forwarded through the network, which outputs a high-dimensional representation, $\phi_{i,j}$. In this instance $\phi_{i,j} \in \mathbb{R}^{2048}$, for each $S_{i,j}$. We noticed that processing the segment images without resizing them did not produce significant benefits and restricted the use of large batches' efficient inference.

Since our focus is on image annotation, where labeled data might not be readily available, feature extraction starts without fine-tuning. It is only optimized as the labeling progresses. Any CNN architecture can be employed, but performance is crucial. We use the High-Resolution Network (HRNet) [5] architecture, pre-trained on ImageNet, but without fully connected layers. It is publicly available with multiple depths, and its performance is superior to other established works for image classification, such as ResNet [1]. During the development of this work, EfficientNet [55] was proposed, significantly improving the classification performance while using comparable computing resources. It was not employed in our experimental setup, but it might improve our results.

3.4 Dimensionality Reduction

Dimensionality reduction aims at reducing a feature space from a higher to a lower dimension with similar characteristics. Some methods enforce the global structure (e.g. PCA); other approaches, such as non-linear methods, focus on local consistency, penalizing neighborhood disagreement between the higher and lower dimensional spaces.

In some applications, the reduction aims to preserve the characteristics of the original features; in our case, we wish to facilitate the annotation as much as possible. Hence, a reduction that groups similar segments and segregates dissonant examples is more beneficial than another one that preserves the original information.

The t-SNE [56] algorithm is the most used technique for non-linear dimensionality reduction. It projects the data into a lower-dimensional space while minimizing the divergence between the higher- and lower-dimensional neighborhood distributions. However, we employed a more recent approach, known as UMAP [57], for the following

reasons: the projection is computed faster, samples can be added without fitting the whole data, its parameters seems to provide more flexibility to choose the projection scattering — enforcing local or global coherence — and most importantly, it allows to use labeled data to enforce consistency between samples of the same class while still allowing unlabeled data to be inserted.

Note that dimensionality reduction is critical to the whole pipeline because it arranges the data to be presented to the user, where most of the interaction will occur.

The 2D embedding can produce artifacts, displaying distinct segments clustered together due to the trade-off between global and local consistency even though they might be distant in the higher-dimensional feature space. Therefore, the user can select a subset of samples and interact with their local projection in a pop-up window, where the projection parameter is tuned to enforce local consistency. The locally preserving embedding (Figure 3) separates the selected cluttered segments (in pink) into groups of similar objects (tennis court, big households, small households, etc.), making it easier for label assignment.

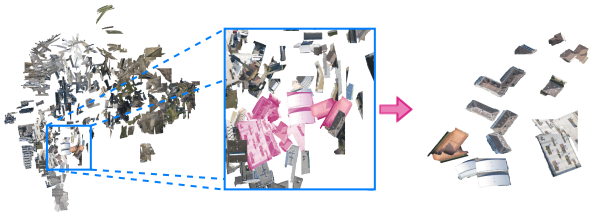


Figure 3: Local re-projection example: Global projection with a region highlighted in blue; A subset of segments is selected by the user, in pink; Their local embedding is computed for a simpler annotation.

On images, CNN's features obtain remarkable results consistent with the human notion of similarity between objects. Considering that an annotator evaluates images visually, sample projection is our preferred approach to inform the user about possible clusters, as explained in the next section. Other visualizations must be explored for other kinds of data, such as sound or text, where the user would have difficulty to visually exploit the notion of similarity [31, 36].

3.5 Embedding Annotation

Each segment is displayed on their 2-dimensional coordinate, as described in the previous section. To annotate a set of segments, the user selects a bounding-box around them in the canvas, assigning the designed label. Hence, each $S_{i,j}$ inside the defined box is assigned to a label $L_{i,j}$. Finally, to obtain annotated masks in the image domain, the label $L_{i,j}$ of a segment $S_{i,j}$ is mapped to its pixels in I_i, thus resulting in an image segmentation (pixel annotation).

Due to several reasons, such as spurious segments or over-segmented objects, a region could be indistinguishable. Hence, when a single segment is selected in the projection, its image is displayed in the Image View as presented earlier. This action also works backwardly, the user can navigate over the images, visualize the current segments, and upon selection, the segments are focused in the projection view. Thus, avoiding the effort of searching individual samples in the segment scattering.

Additional care is necessary when presenting a large number of images, mainly if each one contains several objects — the number of segments displayed on the canvas may impair the user's ability to distinguish their respective classes for annotation. Therefore, only a subset of the data is shown to the user initially. Additional batches are provided as requested while the labeling and the embedding progress, reducing the annotation burden.

3.6 Metric Learning

The proposed pipeline is not specific to any objects' class and does not require pre-training, but as the annotation progresses, the available labels can be employed to reduce user effort — less effort is necessary when the clusters are homogeneous and not spread apart.

For that, we employ a *metric learning* algorithm. Figure 4 shows an example of how metric learning can make clusters of a same class more compact and better separate clusters from distinct classes in our application.

Initially, metric learning methods were concerned with finding a metric where some distance-based (or similarity) classification [58, 59] and clustering [60] would be optimal, in the sense that samples from the same class should be closer together than adversary examples. Given some regularity conditions, learning this new metric is equivalent to embedding the data into a new space.

The metric learning objective functions can be grossly divided into two main varieties, soft assignment and triplet-based techniques. The former, as proposed in NCA [59], maximizes a soft-neighborhood assignment computed through the soft-max function over the negative distance between the data points, penalizing label disagreement of immediate neighbors more than samples further apart. Triplet-based methods [58] select two examples from the same class and minimize their distance while pushing away a third one from a different class when it violates a threshold given the pair distance. Thus, avoiding unnecessary changes when a neighborhood belongs to a single class.

More recently, these methods began to focus mostly on improving embedding through neural networks rather than on the metric centered approach. Notably, in computer vision, it gained traction in image retrieval tasks with improvements in triplet mining methods [61, 62] and novel

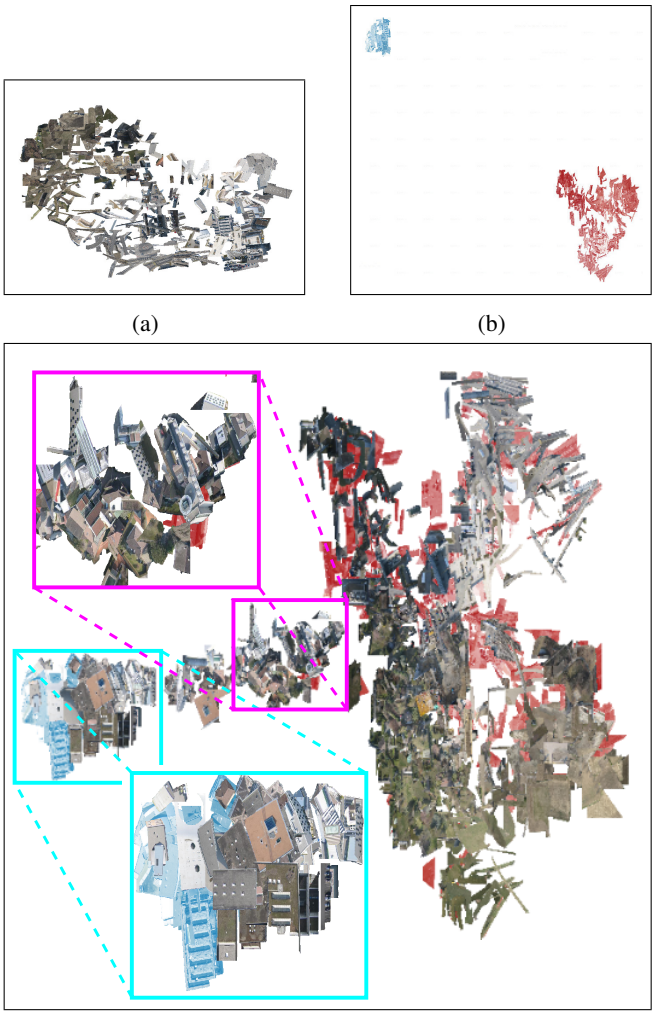


Figure 4: Example of metric learning in the Rooftop dataset: (a) Segment arrangement from an initial batch (10 images). (b) Displacement after labeling and employing metric learning, foreground (rooftops) in blue and background in red. (c) Projection with additional unlabeled data (plus 20 images). Most rooftops' segments are clustered together (the cyan box). The magenta box indicates clusters from mixed classes and spurious segments. They suggest where labels are required the most for a next iteration of labeling and metric learning. The remaining clusters can be easily annotated.

loss functions [63, 64].

In our pipeline, the original large-margin loss [58] was employed using our previously mentioned feature extractor network due to its excellent performance with only a single additional parameter. We follow here Musgrave et al. [65], which showed that some novel methods are prone to overfitting and require more laborious parameter tuning.

3.7 Segment Correction

Since segment delineation is not guaranteed to be perfect, component correction is crucial, especially for producing ground-truth data, where pixel-level accuracy is of uttermost importance.

Segments with multiple objects (under segmentation) are corrected by positive and negative clicks, splitting the segment into two new regions according to the user's positive and negative cues.

Given an under-segmented region $S_{i,j}$, we define an undirected graph $\mathcal{G} = (V, E, w)$ where the vertices V are the pixels in $S_{i,j}$, the edges connect 4-neighbors, constrained to be inside $S_{i,j}$, and each edge $(p, q) \in E$ is weighted by $w(p, q) = (G_{i,j}(p) + G_{i,j}(q))/2$. A segment partition is obtained by the image foresting transform [29] algorithm for the labeled watershed operator given two sets of clicks C_{pos} and C_{neg} as defined by the user. This operation offers full control over segmentation, is fast, and improves segment delineation.

Further, the user can change the hierarchical criterion for watershed segmentation, preventing interactive segment correction in multiple images. For example, in an image overcrowded with irrelevant small objects, the user can bias the hierarchy to partition larger objects by defining the hierarchy ordering according to the objects' area in the image domain, filtering out the spurious segments.

Therefore, the Image View interface allows inspection of multiple hierarchical segmentation criteria and their result for given a threshold. The segments are recomputed upon user confirmation, maintaining the labels of unchanged segments. Novel segments go through the pipeline for feature extraction and projection into the canvas.

4 Experiments

This section starts by describing the datasets chosen for the experiments and the implementation details for our approach. Since our method partitions the images into segments and solves the simultaneous annotation of multiple objects by interactive labeling of similar segments' clusters, we present an ablation study of two ideal scenarios to evaluate each main step individually. In the first experiment, the images are partitioned into perfect segments constrained to each object's mask, the rest of the pipeline is executed as proposed, evaluating the feature space annotation efficiency. Next, we assess the image partition by assuming optimal labeling, thus, only investigating the initial segmentation performance. Hence, the ablation study evaluates the limitations of the initial unsupervised hierarchical segmentation techniques, description with metric learning, and projection. Subsequently, we describe the state-of-the-art methods used for comparison and present the quantitative

and qualitative results.

Note that our goal goes beyond showing that the proposed method can outperform others. We are pointing a research direction that exploits new ways of human-machine interaction for more effective data annotation.

Typically, a robot user executes the deep interactive segmentation experiments; in contrast, our study is conducted by a volunteer with experience in interactive image segmentation. Thus we are taking into account the effort required to locate and identify objects of interest.

4.1 Datasets

We selected image datasets from video segmentation, co-segmentation, and semantic segmentation tasks, in which the objects of interest are to some extent related.

1. *CMU-Cornell iCoSeg* [18]: It contains 643 natural images divided into 38 groups. These groups contain famous landmarks, animals, and other images with distinct foregrounds. Within a group, the images have the same foreground and background but are seen from different point-of-views.
2. *DAVIS* [19]: It is a video segmentation dataset containing 50 different sequences. Following the same procedure as in [15], multiple objects in each frame were treated as a single one, and the same subset of 345 frames (10 % of the total) was employed.
3. *Rooftop* [20]: It is a remote sensing dataset with 63 images, divided into 32 images for training and 31 for testing, and in total containing 390 instances of disjoint rooftop polygons.
4. *Cityscapes* [21]: It is a semantic segmentation dataset for autonomous driving research. It contains video frames from 27 cities divided into 2975 images for training, 500 for validation, and 1525 for testing. The dataset contains 30 classes (e.g. roads, cars, trucks, poles), we evaluated using only the 19 default classes.

4.2 Implementation details

We implemented a user interface in Qt for Python. To segment images into components, we used Higra [66, 67] and the image gradients generated with PoolNet [68]. We computed gradients over four scales, 0.5, 1, 1.5, and 2, and averaged their output to obtain a final gradient image. For the remaining operations, including the baselines, we used NumPy [69], the PyTorch Metric Learning package [70] and the available implementations in PyTorch [71].

For segment description with metric learning, we used the publicly available *HRNet-W18-C-Small-v1* [72] configuration pre-trained on the ImageNet dataset. In the metric-learning stage, the Triplet-Loss margin is fixed at 0.05. At each call, the embedding is optimized through Stochastic Gradient Descent (SGD) with momentum of 0.8 and weight decay of 0.0005 over three epochs with 1000 triplets each. The learning rate starts at 0.1 and, at each epoch, it is divided by 10.

Dataset	Avg. IoU	Median IoU	Time (s)
iCoSeg	95.07	99.96	5.32
DAVIS	98.54	99.97	7.82
Rooftop	95.10	99.99	3.96

Table 1: Average IoU, median IoU, average total processing time in seconds per image.

We used UMAP [57] with 15 neighbors for feature projection and a minimum distance of 0.01 in the main canvas. The zoom-in canvas used UMAP with five neighbors and 0.1 minimum distance; when labels were available, the semi-supervised trade-off parameter was fixed at 0.5, penalizing intra-class and global consistencies equally.

4.3 Ablation Study

Our approach depends on two main independent steps: the image partition into segments and the interactive labeling of those regions. The inaccuracy of one of them would significantly deteriorate the performance of the feature space annotation for image segmentation labeling. Therefore, we present an ablation study that considers two ideal scenarios: (a) interactive projection labeling of perfect segments and (b) image partition into segments followed by optimal labeling.

In (a), the user annotates segments from a perfect image partition inside and outside the objects' masks. Hence, every segment will always belong to a single class. Table 1 reports the results. The Intersection over Union (IoU) distribution is heavily right-skewed, as noticed from the differences between average IoU and median IoU, indicating that most segments were labeled correctly. Visual inspection revealed that user annotation errors occurred only in small components. Table 1 also reports the total time (in seconds) spent annotating (user) and processing (machine), starting from the initial segment projection presented to the user. It indicates that feature space projection annotation with metric learning is effective for image segmentation annotation.

In (b), we measure the IoU of the watershed hierarchical cut using a fixed parameter — the threshold of 1000 with the volume criterion. The segments were then labeled by majority vote among the true labels of their pixels. Table 2 shows the quality of segmentation, which imposes the upper bound to the quality of the overall projection labeling procedure if no segment correction was performed.

For reference, Click Carving [73] reports an average IoU of 84.31 in the iCoSeg, dataset when selecting the optimal

Dataset	Avg. IoU	Median IoU
iCoSeg	84.15	91.86
DAVIS	82.46	88.50
Rooftop	75.14	76.77

Table 2: Automatic segmentation results with their respective dataset.

segment (i.e. highest IoU among proposals) from a pool of approximately 2000 segment proposals per image, produced with MCG [41]. In contrast, we obtain an equivalent performance of 84.15 with a fixed segmentation with disjoints candidates only. Moreover, with just a few segment corrections, our approach can beat other interactive segmentation methods, as it will be shown next. The appendix contains additional examples, displaying that the candidate segments are not over segmented.

4.4 Quantitative analysis using baselines

We used the reported IoU from baselines on Rooftop and iCoSeg datasets and evaluated state-of-the-art interactive segmentation method, f-BRS [74], on all three datasets. This method optimizes the feature representation after the ASPP layer minimizing the errors in the annotated pixels in the f-BRS-B configuration; the backbone used is from $Resnet101$, and it was trained on the SBD dataset [75]. The results are reported over the final segmentation mask, given a sequence of 3 and 5 clicks.

Tables 3–5 report the average IoU and the total time spent in annotation. Our baselines perform segmentation as microtasks, and so it is possible to measure the instance IoU across each disjoint segment in the Rooftop dataset. Note that both image and instance measurements are similar, as shown for f-BRS (Table 5).

For click-based methods, when it was not available, the interaction time was estimated as 2.4s for the initial click and 0.9s for additional clicks [76]. For methods based on bounding boxes, the interaction time is reported in between 7s and 42s [73]. We considered 7s. As observed, our approach can outperform the baselines in foreground segmentation with comparable time and superior accuracy. A few results are presented in Figure 5 and additional images are shown in the appendix.

4.5 Qualitative analysis for semantic segmentation

To validate the proposed approach in a domain-specific scenario, we evaluate its performance on Cityscapes [21]. Since the true labels of the test set are not available, we

Method	IoU	Time (s)
GrabCut	83.14	7
Poly-RNN++	72.46	7
Click Carving*	82.13	10.36
f-BRS (3 clicks)	79.82	4.2
f-BRS (5 clicks)	82.14	6
Ours	**84.29**	5.96

Table 3: iCoSeg dataset - Average IoU and time across images. *Click carving validated with only 20% of the dataset.

Method	IoU	Time (s)
f-BRS (3 clicks)	79.87	4.2
f-BRS (5 clicks)	82.44	6
Ours	**84.53**	8.74

Table 4: DAVIS dataset - Average IoU and time across images.

Method	Image IoU	Instance IoU	Time (s)
Poly-RNN++*	-	65.67	7
Curve-GCN*	-	66.78	7
Spline-GCN*	-	68.33	7
f-BRS (3 clicks)	62.57	65.00	4.2
f-BRS (5 clicks)	74.53	**74.74**	6
Ours	**77.28**	-	7.02

Table 5: Rooftop dataset - Average IoU, average time across instances. *Poly-RNN++ and Curve/Spline-GCN were fine-tuned on training set, accuracy reported for a subset of the test set only.

took the same approach as [24], by testing on the validation set. Further, the annotation quality was evaluated on 98 randomly chosen images (about 20% of the validation set).

PoolNet was optimized based on the boundaries of the training set's true labels for five epochs using the Adam optimizer, a learning rate of $5e^{-5}$, weight decay of $5e^{-4}$, and a batch of size 8. Predictions were performed on a single scale. The fine-tuned gradient ignores irrelevant boundaries, reducing over segmentation.

The original article reports an agreement (i.e. accuracy) between annotators of 96%. We obtained an agreement of 91.5% with the true labels of the validation set (Figure 6), while spending less than 1.5% of their time — i.e. our experiment took 1 hour and 58 minutes to annotate the 98 images, while producing the same amount of ground-truth data took approximately 6.1 days with other approaches, considering an average of 1.5 hour per image [21] — our method is then 74.75 times faster than the original procedure.

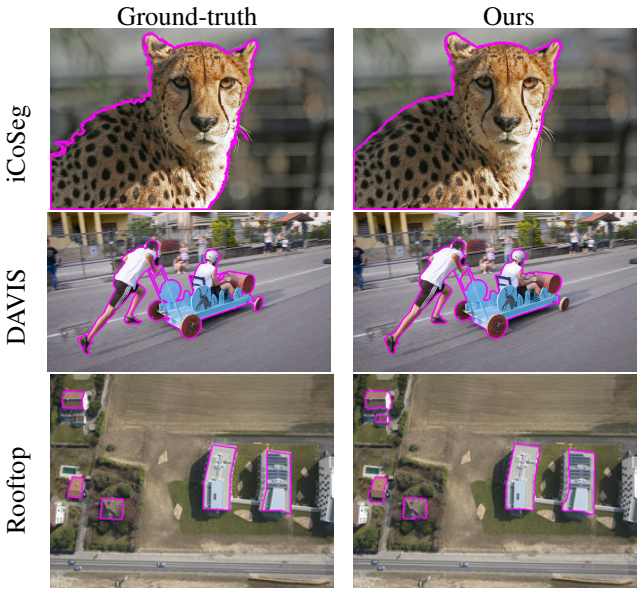


Figure 5: Resulting segmentation. Rows indicate dataset, the first column is the original image with ground-truth contour, and the second column presents our results.

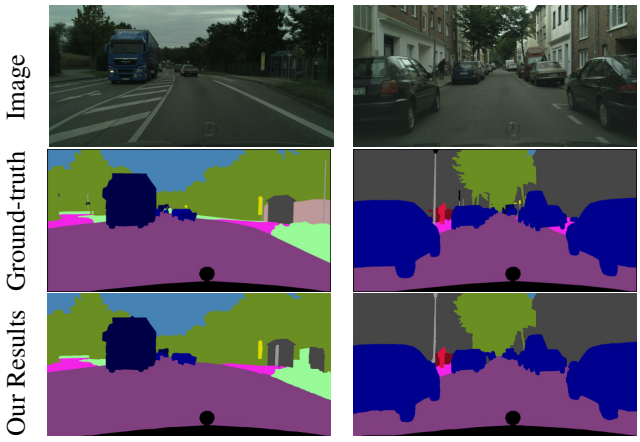


Figure 6: Cityscapes result, each column is a different image, row indicates which kind.

5 Conclusion

We presented a novel interactive image segmentation paradigm for simultaneous annotation of segments from multiple images in the projection space of their deep features. The feature embedding is optimized with metric learning as the labeling progress, reducing the annotation burden. Its feasibility is confirmed by surpassing the performance of state-of-the-art methods in multiple datasets. The proposed paradigm creates opportunities to investigate new applications and interactive machine learning mechanisms (e.g., active learning, semi-supervised labeling, weakly-supervised methods). We intend to improve segment delineation without over segmentation, given its incredible impact in the final results, as shown in the ablation study.

Acknowledgements

This work was supported by FAPESP research grants #2014/12236-1, #2019/11349-0, and #2019/21734-9.

Appendix

This appendix presents additional qualitative results of the methodology presented in the main article in the DAVIS [19], iCoSeg [18] and Rooftop [20].

A Ablation Study Qualitative Results

Figures 7–9 presents examples of the unsupervised segmentation obtained with fixed parameters defined by the user, watershed volume attribute with threshold 1000. The method is not sensitive to parameter choice even across multiple datasets with images with different characteristics and sizes. These segments were employed in the ablation study (b) of Section 4.3 of the main text.

Note that, during the interactive experiments, the user can adjust these parameters to correct the segmentation. However, this is undesirable because it requires additional interactions.

B Interactive Segmentation Qualitative Results

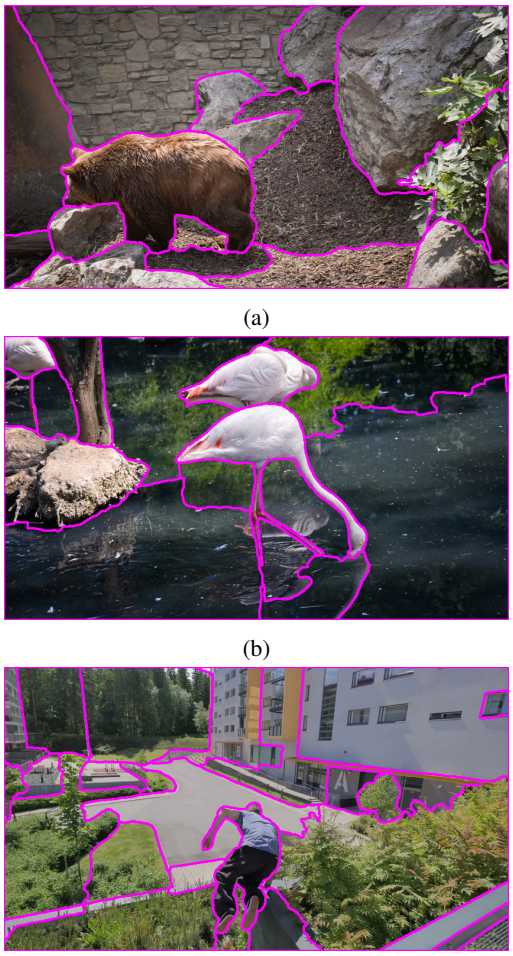
Figures 10–12 presents qualitative results from the experiments of Section 4.4 from the main article. The magenta contour delineates the objects' boundaries, the top rows are the ground-truth data, and the bottom rows are our results.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. 1, 4

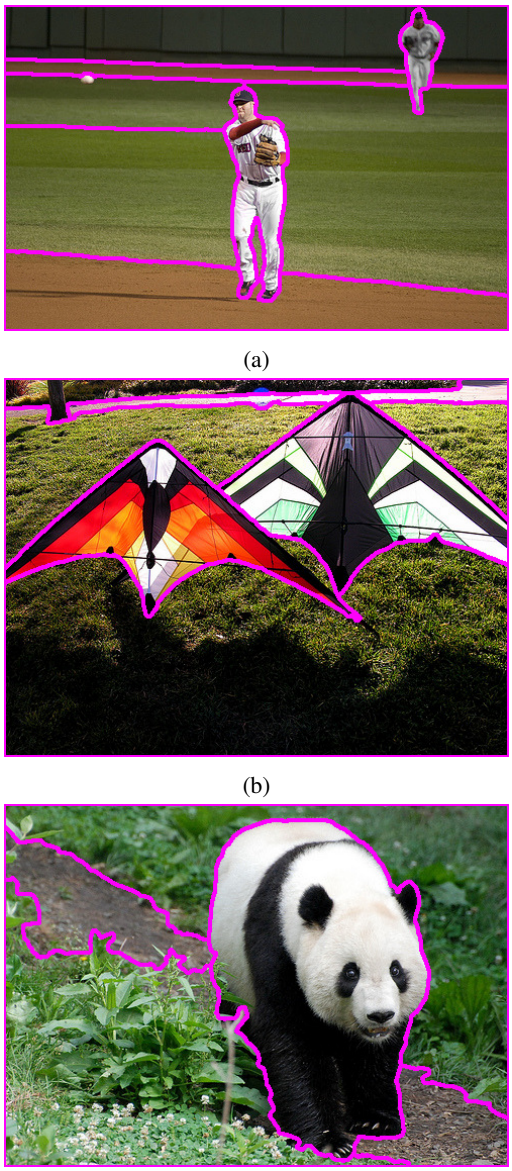
[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017. 1

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural



(a) (b) (c)

Figure 7: Candidate segments of DAVIS



(a) (b) (c)

Figure 8: Candidate segments of iCoSeg.

networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. 1

[4] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018. 1

[5] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 4

[6] E. N. Mortensen and W. A. Barrett, "Intelligent scissors for image composition," in *Proceedings of the 22nd Annual Conference On Computer Graphics and Interactive Techniques*, pp. 191–198, 1995. 1

[7] A. X. Falcão, J. K. Udupa, S. Samarasekera, S. Sharma, B. E. Hirsch, and R. A. Lotufo, "User-steered image segmentation paradigms: Live wire and live lane," *Graphical models and image processing*, vol. 60, no. 4, pp. 233–260, 1998. 1

[8] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004. 1

[9] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut:

(a)

(b)

(c)

Figure 9: Candidate segments of Rooftop

Interactive foreground extraction using iterated graph cuts," in *ACM Trans. on Graphics*, vol. 23, pp. 309–314, 2004. 1

[10] L. Grady, "Random walks for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1768–1783, 2006. 1, 2

[11] M. Tang, L. Gorelick, O. Veksler, and Y. Boykov, "Grabcut in one cut," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 1769–1776, 2013. 1

[12] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008. 1, 2

[13] N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang, "Deep interactive object selection," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 373–381, 2016. 1, 2

[14] Z. Li, Q. Chen, and V. Koltun, "Interactive image segmentation with latent diversity," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 577–585, 2018. 1, 2

[15] W.-D. Jang and C.-S. Kim, "Interactive image segmentation via backpropagating refinement scheme," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 5297–5306, 2019. 1, 2, 7

[16] K. Sofiiuk, I. Petrov, O. Barinova, and A. Konushin, "f-brs: Rethinking backpropagating refinement for interactive segmentation," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 8623–8632, 2020. 1, 2

[17] T. Kontogianni, M. Gygli, J. Uijlings, and V. Ferrari, "Continuous adaptation for interactive object segmentation by learning from corrections," in *IEEE European Conference on Computer Vision*, pp. 579–596, Springer, 2020. 1, 2

[18] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "Interactively co-segmentating topically related images with intelligent scribble guidance," *International Journal of Computer Vision*, vol. 93, no. 3, pp. 273–292, 2011. 2, 7, 9

[19] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 724–732, 2016. 2, 7, 9

[20] X. Sun, C. M. Christoudias, and P. Fua, "Free-shape polygonal object localization," in *IEEE European Conference on Computer Vision*, pp. 317–332, Springer, 2014. 2, 7, 9

[21] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban

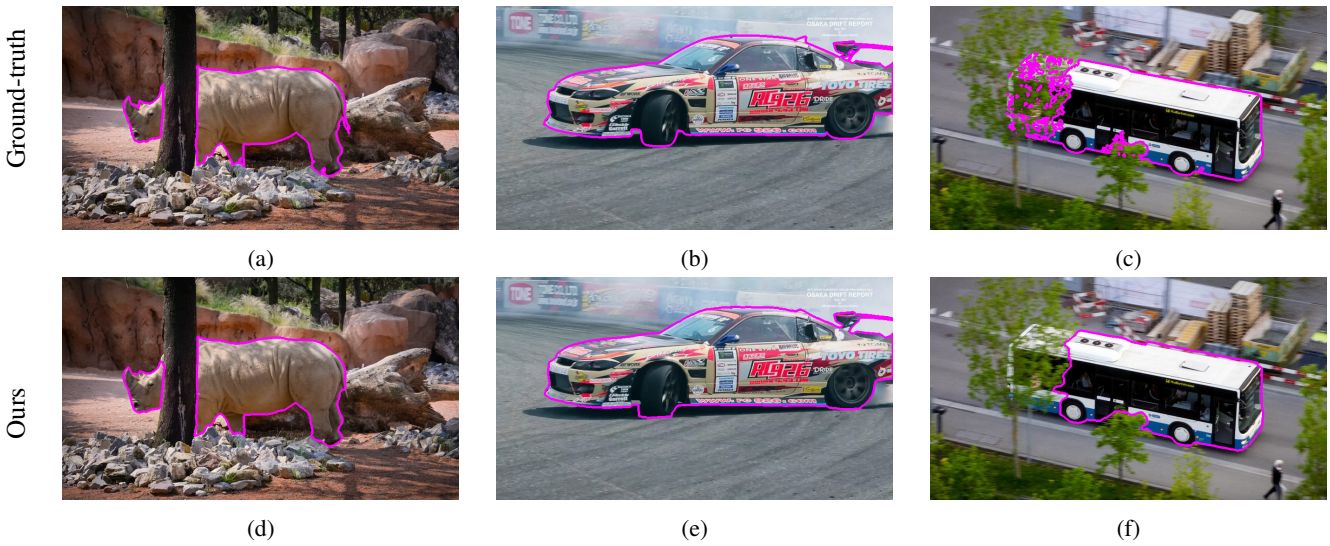


Figure 10: The top row indicates the ground-truth, and the bottom row shows the results from the proposed methodology in the DAVIS dataset; The magenta boundaries define regions with foreground labels.

Figure 11: The top row indicates the ground-truth, and the bottom row shows the results from the proposed methodology in the iCoSeg dataset; The magenta boundaries define regions with foreground labels.

scene understanding," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 3213–3223, 2016. 2, 7, 8

[22] K.-K. Maninis and et al., "Deep extreme cut: From extreme points to object segmentation," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, 2018. 2

[23] R. Benenson, S. Popov, and V. Ferrari, "Large-scale interactive object segmentation with human annotators," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 11700–11709, 2019. 2

[24] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a polygon-rnn," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 5230–5238, 2017. 2, 8

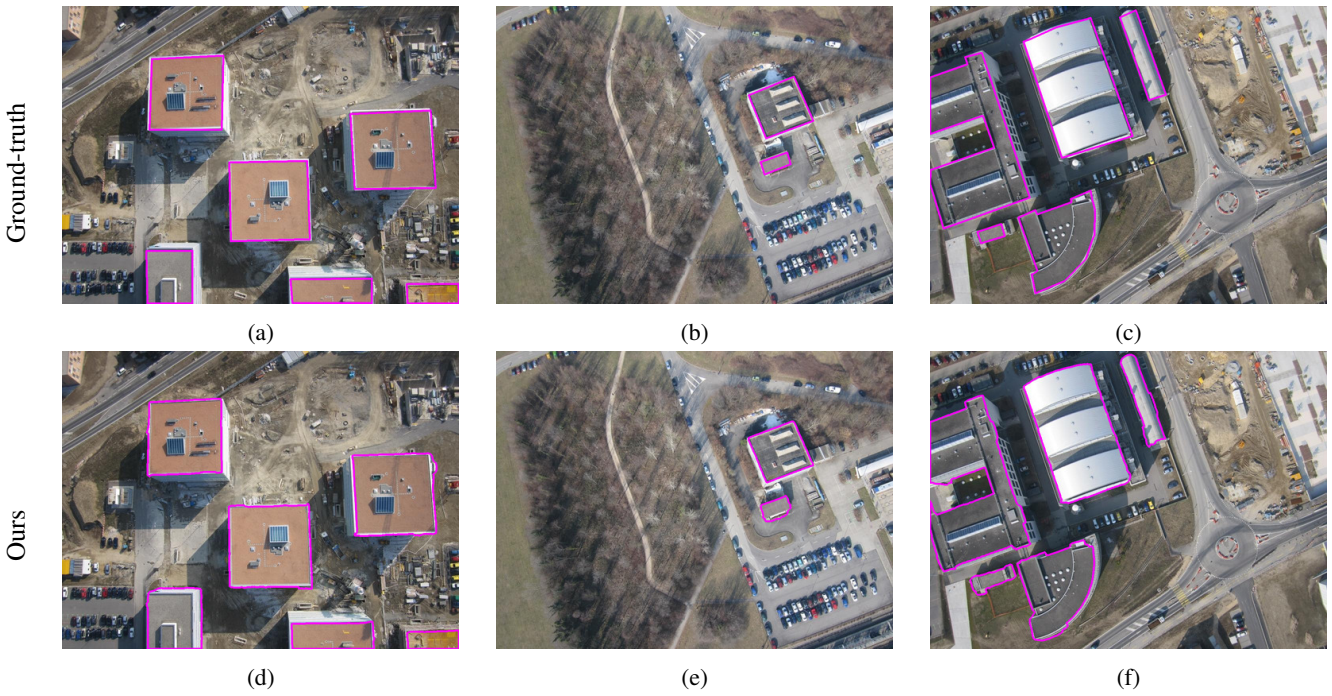


Figure 12: The top row indicates the ground-truth, and the bottom row shows the results from the proposed methodology in the Rooftop dataset; The magenta boundaries define regions with rooftop labels.

[25] D. Acuna, H. Ling, A. Kar, and S. Fidler, "Efficient interactive annotation of segmentation datasets with polygon-rnn++," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 859–868, 2018. 2

[26] H. Ling, J. Gao, A. Kar, W. Chen, and S. Fidler, "Fast interactive object annotation with curve-gcn," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 5257–5266, 2019. 2

[27] E. Agustsson, J. R. Uijlings, and V. Ferrari, "Interactive full image segmentation by considering all regions jointly," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 11622–11631, 2019. 2

[28] C. Couprie, L. Grady, L. Najman, and H. Talbot, "Power watershed: A unifying graph-based optimization framework," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1384–1399, 2011. 2

[29] A. X. Falcão, J. Stolfi, and R. A. de Lotufo, "The image foresting transform: Theory, algorithms, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 19–29, 2004. 2, 6

[30] M. Andriluka, J. R. Uijlings, and V. Ferrari, "Fluid annotation: a human-machine collaboration interface for full image annotation," in *Proceedings of the 26th ACM international conference on Multimedia*, pp. 1957–1966, 2018. 2

[31] J. Bernard, M. Zeppelzauer, M. Sedlmair, and W. Aigner, "Vial: a unified process for visual interactive labeling," *The Visual Computer*, vol. 34, no. 9, pp. 1189–1207, 2018. 2, 3, 5

[32] J. Bernard, M. Hutter, M. Zeppelzauer, D. Fellner, and M. Sedlmair, "Comparing visual-interactive labeling with active learning: An experimental study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 298–308, 2017. 2

[33] B. C. Benato, A. C. Telea, and A. X. Falcão, "Semi-supervised learning with interactive label propagation guided by feature space projections," in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 392–399, IEEE, 2018. 2

[34] B. C. Benato, J. F. Gomes, A. C. Telea, and A. X. Falcão, "Semi-automatic data annotation guided by feature space projection," *Pattern Recognition*, vol. 109, p. 107612, 2019. 2

[35] J. E. Vargas-Muñoz, P. Zhou, A. X. Falcão, and D. Tuia, "Interactive coconut tree annotation using

feature space projections," in *IEEE International Geoscience and Remote Sensing Symposium*, pp. 5718–5721, IEEE, 2019. 2

[36] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, "Visual interaction with dimensionality reduction: A structured literature analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 241–250, 2016. 2, 3, 5

[37] N. Araslanov and S. Roth, "Single-stage semantic segmentation from image labels," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 4253–4262, 2020. 2

[38] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 3791–3800, 2018. 2

[39] Y. Zhu, Y. Zhou, H. Xu, Q. Ye, D. Doermann, and J. Jiao, "Learning instance activation maps for weakly supervised instance segmentation," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 3116–3125, 2019. 2

[40] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Deepco3: Deep instance co-segmentation by co-peak search and co-saliency detection," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 8846–8855, 2019. 2

[41] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 128–140, 2016. 3, 8

[42] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Convolutional oriented boundaries: From image segmentation to high-level tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 819–833, 2017. 4

[43] L. Najman and M. Schmitt, "Geodesic saliency of watershed contours and hierarchical segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 12, pp. 1163–1173, 1996. 4

[44] L. Najman, "On the equivalence between hierarchical segmentations and ultrametric watersheds," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 3, pp. 231–247, 2011. 4

[45] S. Xie and Z. Tu, "Holistically-nested edge detection," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 1395–1403, 2015. 4

[46] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 3000–3009, 2017. 4

[47] R. Deng, C. Shen, S. Liu, H. Wang, and X. Liu, "Learning to predict crisp boundaries," in *IEEE European Conference on Computer Vision*, pp. 562–578, 2018. 4

[48] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bi-directional cascade network for perceptual edge detection," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 3828–3837, 2019. 4

[49] D. Linsley, J. Kim, A. Ashok, and T. Serre, "Recurrent neural circuits for contour detection," in *International Conference on Learning Representations*, 2019. 4

[50] O. Isaacs, O. Shayer, and M. Lindenbaum, "Enhancing generic segmentation with learned region representations," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 12946–12955, 2020. 4

[51] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 3917–3926, 2019. 4

[52] J. Cousty, L. Najman, Y. Kenmochi, and S. Guimarães, "Hierarchical segmentations with graphs: quasi-flat zones, minimum spanning trees, and saliency maps," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 4, pp. 479–502, 2018. 4

[53] J. Cousty and L. Najman, "Incremental algorithm for hierarchical minimum spanning forests and saliency of watershed cuts," in *Mathematical Morphology and Its Applications to Signal and Image Processing*, pp. 272–283, Springer, 2011. 4

[54] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015. 4

[55] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, pp. 6105–6114, 2019. 4

[56] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008. 4

[57] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018. 4, 7

[58] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification.," *Journal of Machine Learning Research*, vol. 10, no. 2, 2009. 5, 6

[59] J. Goldberger, G. E. Hinton, S. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," *Advances in Neural Information Processing Systems*, vol. 17, pp. 513–520, 2004. 5

[60] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Advances in Neural Information Processing Systems*, pp. 521–528, 2003. 5

[61] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017. 5

[62] W. Zheng, Z. Chen, J. Lu, and J. Zhou, "Hardness-aware deep metric learning," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 72–81, 2019. 5

[63] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 2840–2848, 2017. 6

[64] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, and R. Jin, "Softtriple loss: Deep metric learning without triplet sampling," in *IEEE Conferencee on Computer Vision and Pattern Recognition*, pp. 6450–6458, 2019. 6

[65] K. Musgrave, S. Belongie, and S.-N. Lim, "A metric learning reality check," in *IEEE European Conference on Computer Vision*, Springer, 2020. 6

[66] B. Perret, G. Chierchia, J. Cousty, S. Guimarães, Y. Kenmochi, and L. Najman, "Higra: Hierarchical graph analysis," *SoftwareX*, vol. 10, p. 100335, 2019. 7

[67] ""Higra: Hierarchical graph analysis" package.." https://higra.readthedocs.io/en/stable/. Last Accessed: 2020-11-06. 7

[68] "Repository of "a simple pooling-based design for real-time salient object detection"." https://github.com/backseason/PoolNet. Last Accessed: 2020-11-06. 7

[69] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020. 7

[70] K. Musgrave, S. Belongie, and S.-N. Lim, "PyTorch metric learning," 2020. 7

[71] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, pp. 8026–8037, 2019. 7

[72] "Repository of "high-resolution networks (HRNets for image classification"." https://github.com/HRNet/HRNet-Image-Classification. Last Accessed: 2020-11-06. 7

[73] S. D. Jain and K. Grauman, "Click carving: Interactive object segmentation in images and videos with point clicks," *International Journal of Computer Vision*, vol. 127, no. 9, pp. 1321–1344, 2019. 7, 8

[74] "Repository of "f-BRS: Rethinking backpropagating refinement for interactive segmentation repository"." https://github.com/saic-vul/fbrs_interactive_segmentation. Last Accessed: 2020-11-06. 8

[75] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *IEEE International Conference on Computer Vision*, pp. 991–998, IEEE, 2011. 8

[76] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *IEEE European Conference on Computer Vision*, pp. 549–565, Springer, 2016. 8