2019 IEEE Winter Conference on Applications of Computer Vision

Semantic Instance Meets Salient Object: Study on Video Semantic Salient Instance Segmentation*

Trung-Nghia Le[†]
le.trung.nghia.437@m.kyushu-u.ac.jp
Graduate University for Advanced Studies (SOKENDAI), Japan

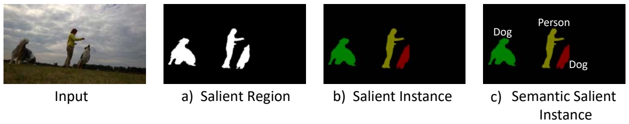
Akihiro Sugimoto
sugimoto@nii.ac.jp
National Institute of Informatics, Japan

Abstract

Focusing on only semantic instances that only salient in a scene gains more benefits for robot navigation and self-driving cars than looking at all objects in the whole scene. This paper pushes the envelope on salient regions in a video to decompose them into semantically meaningful components, namely, semantic salient instances. We provide the baseline for the new task of video semantic salient instance segmentation (VSSIS), that is, Semantic Instance - Salient Object (SISO) framework. The SISO framework is simple yet efficient, leveraging advantages of two different segmentation tasks, i.e. *semantic instance segmentation and salient object segmentation to eventually fuse them for the final result. In SISO, we introduce a sequential fusion by looking at overlapping pixels between semantic instances and salient regions to have non-overlapping instances one by one. We also introduce a recurrent instance propagation to refine the shapes and semantic meanings of instances, and an identity tracking to maintain both the identity and the semantic meaning of instances over the entire video. Experimental results demonstrated the effectiveness of our SISO baseline, which can handle occlusions in videos. In addition, to tackle the task of VSSIS, we augment the DAVIS-2017 benchmark dataset by assigning semantic ground-truth for salient instance labels, obtaining SEmantic Salient Instance Video (SESIV) dataset. Our SESIV dataset consists of 84 high-quality video sequences with pixel-wisely per-frame ground-truth labels.*

1. Introduction

Recent advances in salient object segmentation (SOS) in videos using CNN [22, 27, 29, 50] have demonstrated impressive performance in accuracy. Such SOS methods [22, 27, 29, 50] focus on only localizing the region of interest by labeling "salient" or "non-salient" to each pixel in the video frame. The localized salient region, however, may involve multiple (interacting) objects (Fig. 1 a), which is a more reasonable scenario in the real-world scenes. Therefore, localized salient regions should be decomposed into conceptually meaningful components (Fig. 1 b), called salient instances [26], for better understanding of videos. Furthermore, attaching a semantic label to each salient instance (Fig. 1 c) will widen the range of applications of SOS even to autonomous driving [54] and robotic interaction [52]. Nevertheless, segmenting semantic salient instances is not yet addressed in the literature.



Input
a) Salient Region
b) Salient Instance
c) Semantic Salient Instance

Figure 1: Segmentation levels of salient objects. The input video frame is followed by different levels of label annotation. Our work focuses on segmenting semantic salient instances (most right).

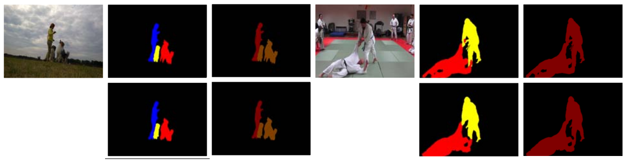


Figure 2: Examples obtained by our method on the SESIV dataset. From left to right, the original video frame is followed by instance label and semantic label. The first and second rows show ground-truth labels, and segmented results, respectively.

To achieve this semantic-instance level segmentation of salient regions, we aim to jointly identify individual instances in the segmented salient regions and categorize these salient instances (Fig. 1 c). We refer this problem to *semantic salient instance segmentation*, which aims to identify *only individual prominent foreground object* classes.

[*] This work was in part supported by CART of the Ministry of Land, Infrastructure, Transport and Tourism of Japan, JST CREST (Grant No. JPMJCR14D1), and Grant-in-Aid for Scientific Research (Grant No. 16H02851) of the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

[†] Trung-Nghia Le is now with Kyushu University, Japan.

978-1-7281-1975-5/19/$31.00 ©2019 IEEE
DOI 10.1109/WACV.2019.00194

The problem is even more challenging on the videos because instances need to be tracked over the entire video to maintain their identifications even if they are occluded at some frames. We remark that in this paper, an instance in a video is defined to be salient if it appears in the first video frame and stands out for more than 50% duration of the video in total.

Many computer vision tasks such as action localization [15], action recognition [8], object relation detection [17], or video captioning [46], focus on dominant objects in the scene to avoid the expensive computational cost. Narrowing down dominant objects further using semantic salient instances is more appropriate in real application scenarios. Indeed, for autonomous robots or self-driving car, it is sufficient to focus on only a few useful semantic instances on the street view such as pedestrians or cars with high performance in accuracy and processing time instead of looking at all semantic object classes in the whole scene.

There are many methods proposed for each task of semantic instance segmentation (SIS) [9, 16, 34] and SOS [29, 33, 50]. However, to the best of our knowledge, no work exists on semantic salient instance segmentation in images or videos to the date. Li *et al*. [26] very recently proposed a method for salient instance segmentation for images but do not deal with semantic level of segmentation.

On the other hand, the CNN-based approach to SOS requires a large number of training samples. As illustrated in Table 1, several benchmark datasets for various tasks of SOS have been provided [1, 2, 6, 7, 12, 28, 29, 37, 44, 47, 49, 51, 53]. The dataset quality is improved over the time in terms of the number of samples and the detailed annotation. Though some datasets for salient instance segmentation are recently available (*e.g*. SOI dataset [26] for images and SegTrack2 dataset [25] for videos), they do not have sufficient numbers of samples to train deep networks. For semantic salient instance segmentation, to the best of our knowledge, no dataset having a sufficient number of samples for training is available to the date.

The overall contribution of this paper is three-fold:

First, we address the new task of **video semantic salient instance segmentation (VSSIS)** and analyze in-depth challenges of the problem. Finding semantic salient instances in videos is a useful task and it can be an interesting problem for the community. Existing work individually performs SIS or SOS, but no work can jointly perform these two tasks, which is considered as the new task of VSSIS.

Second, we introduce the baseline for VSSIS, called **S**emantic **I**nstance - **S**alient **O**bject (**SISO**). SISO is a simple yet efficient two-stream framework leveraging advantages of two different segmentation tasks, *i.e*. SIS and SOS, through combining outputs of two streams. SISO possesses three key features: sequential fusion, recurrent instance propagation, and identity tracking. The sequential fusion frame-wisely fuses the outputs of the two streams.

Table 1: Datasets for salient object segmentation tasks.

Task	**Image**	**Video**
Salient Object Segmentation	MSRA [7], CSSD [53], Judd-A [1], THUR [6], HKU-IS [28], XPIE [51], DUTS [47]	SegTrack [44], DAVIS-2016 [37], 10-Clips [12], FBMS [2], ViSal [49], VOS [29]
Salient Instance Segmentation	SOI [26]	SegTrack2 [25]
Semantic Salient Instance Segmentation	None	**Our proposed SESIV**

Using our introduced instance merging order and frame-confidence, the salient region obtained from the SOS stream is decomposed into non-overlapping salient instances one by one. The recurrent instance propagation recovers unsegmented semantic salient instances by recurrently propagating instances in frames with high frame-confidence to ones in frames with low frame-confidence. Identity tracking, on the other hand, maintains the consistency of instance identities and semantic labels over the entire video where identity propagation is for short-term consistency and re-identification is for long-term consistency. We also comprehensively evaluate the performance of the proposed baseline and deeply analyze results to show promising avenues for future research.

Third, we provide a dataset, **SE**mantic **S**alient **I**nstance **V**ideo (**SESIV**) dataset [1]accompanied with complementary metrics specifically designed for the task of VSSIS. The SESIV dataset consists of 84 high-quality video sequences with various densely annotated, pixel-accurate and per-frame ground-truth labels for different segmentation tasks. Our SESIV annotations are built on top of existing DAVIS-2017 annotations [38]. From pixel-wise instance-level labels of the DAVIS-2017 dataset, we identify salient instances and assign a semantic label to each instance. We emphasize that this is the very first dataset for VSSIS. Figure 2 shows some example results obtained by the SISO baseline in the SESIV dataset. We believe that our introduced SESIV dataset and metrics raise interest to the community and promote further research on VSSIS.

2. Related Work

Semantic instance segmentation (SIS) is the task of unifying object detection and semantic segmentation. It has been intensively studied in recent years where the segmentation based approach or the proposal based approach is employed. The segmentation based approach [19, 24, 31, 54] generally adopts two-stage processing: segmentation first and then instance clustering. The proposal based approach [4, 9, 16, 30, 34], on the other hand, predicts bounding-boxes first and then parses the bounding-boxes to obtain mask regions [9] or exploits object detection mod-

[1] The SESIV annotations and evaluation scripts are publicly available at https://sites.google.com/view/ltnghia/research/sesiv

els (*e.g.*, Faster R-CNN [40] or R-FCN [10]) to classify mask regions [4, 16, 30, 34]. Among these methods, Mask R-CNN [16] achieves the state-of-the-art performance, and recent work [13, 34] is based on Mask R-CNN's architecture. To the best of our knowledge, no work exists that deals with video semantic instance segmentation. We thus use the strategy of frame-by-frame segmentation followed by instance linkage over the entire video.

Recent video salient object segmentation (VSOS) methods are based on the convolutional neural network (CNN) [22, 23, 27, 29, 50] and have demonstrated superior results over early work utilizing only hand-crafted features [21, 35, 39, 48, 49, 55]. These CNN based methods are classified into two approaches: segmentation based approach and end-to-end saliency inference approach. The segmentation based approach first segments each frame of a video into regions and uses deep features extracted from each region for saliency inference [23]. The end-to-end saliency inference approach, on the other hand, uses fully convolutional networks (FCNs) [22, 27, 29, 50] to utilize optical flow [27, 29, 50] or 3D kernels [22]. The end-to-end saliency inference approach achieves better performance than the segmentation based one, and using 3D kernels can deal with more frames than optical flow to incorporate temporal information. We thus employ [22] as the SOS stream in SISO.

3. Semantic Salient Instance Video Dataset

3.1. Overview

To promote VSSIS, a publicly available dataset with pixel-wise ground-truth annotation is mandatory. We thus construct the **SE**mantic **S**alient **I**nstance **V**ideo (SESIV) dataset. We emphasize that no other dataset is publicly available for VSSIS. Figure 3 illustrates examples from our SESIV dataset with their corresponding ground-truth labels.

The proposed SESIV dataset consists of 84 videos with 185 semantic salient instances categorized into 29 classes. The training set consists of 58 videos (with 136 instances and 27 categories), and the testing set consists of 26 videos (with 49 instances and 14 categories). For each video frame, we provide various ground-truth labels (*i.e.*, saliency label, instance label, and semantic label, as exampled in Fig. 3). We remark that SESIV annotations are built on top instance-level ground-truth labels of the DAVIS-2017 dataset [38].

3.2. Dataset Construction

To build the dataset, we used 90 videos in the DAVIS-2017 dataset [38], which has pixel-wise instance-level ground-truth. This dataset is designed for semi-supervised instance segmentation where instances are indicated in the first frame of the video regardless of whether they are salient. Therefore, instance labels in the DAVIS-2017 dataset are annotated regardless of whether they are salient or non-salient, and they do not contain any semantic labels.

In order to construct annotations for the task of VSSIS, we need to identify salient instances and assign a semantic label to each salient instance to create the SESIV dataset. Figure 4 illustrates the flowchart of constructing the SESIV dataset. We first manually eliminated non-salient instances and kept only salient instances (Fig. 4 (a)). Then, we annotated semantic labels to the instances to have semantic salient instances using 29 among 80 categories of the MS-COCO dataset [32] (Fig. 4 (b)). They are *person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, bird, cat, dog, horse, sheep, cow, elephant, bear, backpack, snowboard, sports ball, kite, skateboard, surfboard, tennis racket, chair, tv, remote, cell phone,* and *clock*. The famous MS-COCO dataset is the largest large-scale dataset for object detection/segmentation to the date, thus synchronizing our SESIV with MS-COCO has advantages for training models. After that, we merged unlabeled salient instances into their neighboring one so that the merged instance can be labeled. For example, "mask" instance is merged into "clothes" instance to obtain a new instance that is annotated with person (Fig. 4 (c)). Finally, we discarded six videos, namely, *camel, goat, gold-fish, pigs, rhino,* and *varanus-cage* as in Fig. 4 (d), because these videos do not have any labeled semantic salient instances.

3.3. Dataset Description

The SESIV dataset consists of 84 videos, and the average length of the 84 videos is 68 frames. We note that 28% of the videos have from 71 to 80 frames. We also note that the challenge of the SESIV dataset is enhanced due to the same properties as the DAVIS-2017 dataset [38]. They are *background clutter, dynamic background, deformation, appearance change, shape complexity, small instance, occlusion, out of view, motion blur,* and *fast motion*.

We here analyze in-depth two other properties that are specifically designed for VSSIS:

- Number of semantic salient instances.
- Number of categories used for semantic annotation.

We present the distribution of these two properties over the SESIV dataset in Fig. 5. Each video has the maximum of 8 semantic salient instances. Most videos have from 1 to 3 semantic salient instances: 37% of the videos have one instance, 35% do two instances, and 18% do three instances (Fig. 5 (a)). Each video has the maximum of 4 categories. 54% of the videos have only one category while 39% do two categories (Fig. 5 (b)). A large number of videos have a single instance (37%) or two instances from different categories (25%) (Fig. 5 (c)).

It is also noteworthy that instances can disappear in several frames in a video due to, for example, *full occlusion* or *out of view*. 17.9% of the videos have instances that disappear in their some frames. They are, for example, *bmx-bumps, color-run, dog-gooses, drone, surf,* and *walking*.

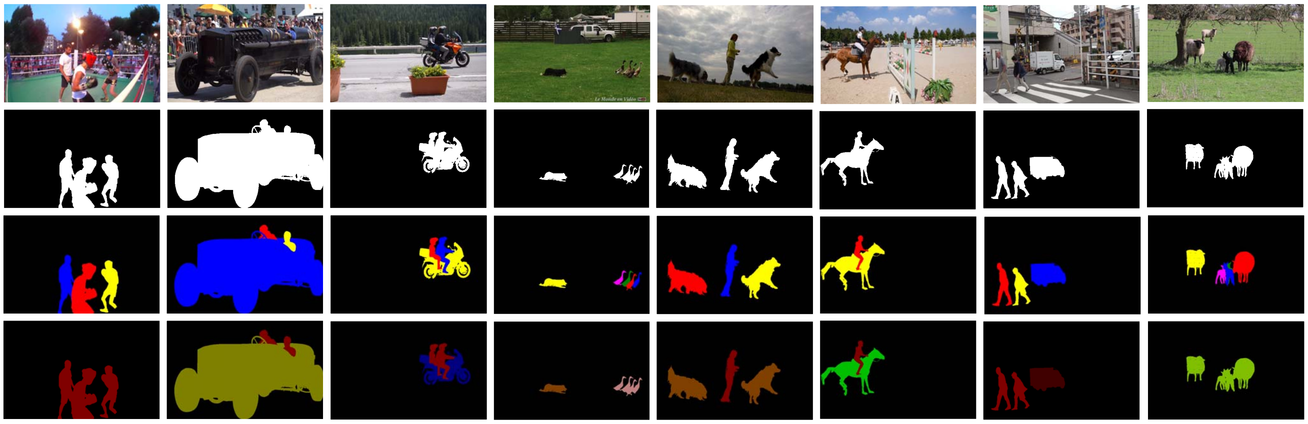


Figure 3: Samples from the SESIV dataset. From top to bottom, video frame is followed by saliency ground-truth, instance ground-truth, and semantic label ground-truth.

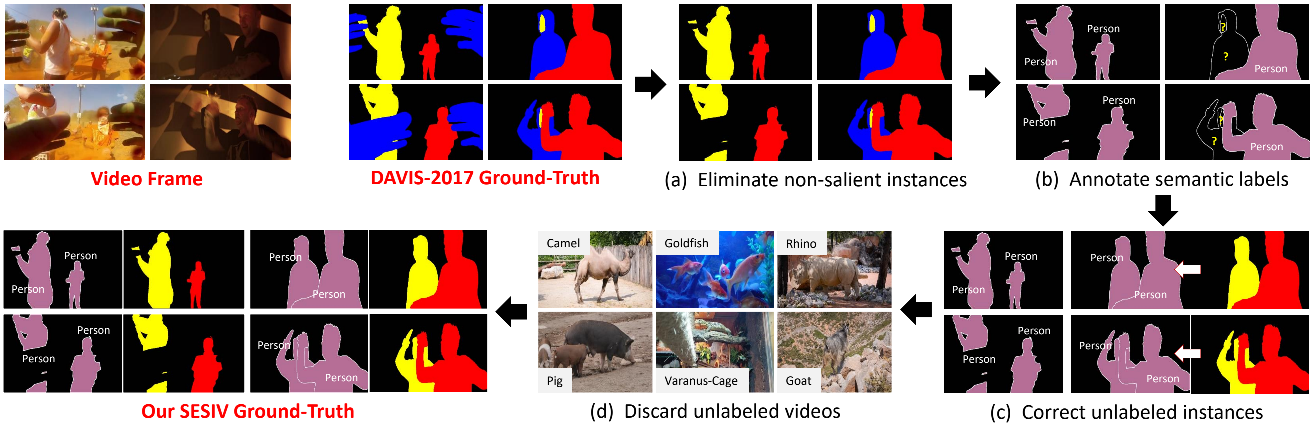


Figure 4: SESIV dataset construction.

4. Proposed Baseline

4.1. Overview

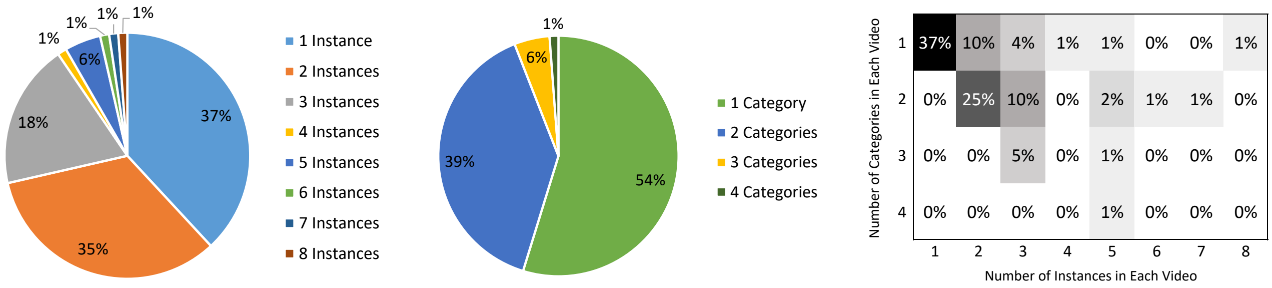
The most straightforward approach to VSSIS is to segment individual instances frame-by-frame and then combine them to obtain final results. However, this approach does not guarantee consistency of labels over frames due to frame-by-frame processing. Furthermore, this approach faces the problem that instances overlap with each other.

To overcome such issues, we propose a **S**emantic **I**nstance - **S**alient **O**bject (SISO) baseline, consisting of two streams (one for SIS, and the other for SOS), where salient instances in the current frame are propagated to those in subsequent frames to maintain consistency of their labels, in terms of identity and semantic, even if instances disappear in some frames. Therefore, SISO is able to deal with a varying number of salient semantic instances and is scalable to the length of videos.

Figure 6 illustrates pipeline of SISO. Two streams (*e.g*. SIS and SOS) of SISO work on both spatial and temporal domains. Outputs of streams are fused to remove non-salient instances, producing a pixel-wisely labeled instance map. We remark that both salient region mask and semantic instances are spatially refined before the fusion, using boundary snapping method [3, 20, 22], improving accuracy of the final result. Finally, the identity tracking maintains the consistency of instance labels over the entire video.

SIS Stream. No existing work can deal with SIS in videos. We thus use the strategy of frame-by-frame segmentation followed by instance linkage over the entire video. Particularly, semantic instances segmented at each frame are temporally propagated over the entire video using the recurrent instance propagation to improve the accuracy of instance shapes. In section 5.3, we evaluate the performance of various network architectures implemented in the SIS stream.

SOS Stream. We employ the 3D FCN model proposed by Le *et al*. [22] as the SOS stream, thus computed saliency map implicitly contains temporal information. The saliency map is then binarized to have salient region mask using, for example, the adaptive threshold $\theta = \mu + \eta$ where μ and η



(a) Number of instances.

(b) Number of categories.

(c) Number of instances and categories.

Figure 5: Statistics of videos over the SESIV dataset based on the numbers of instances/categories in each video.

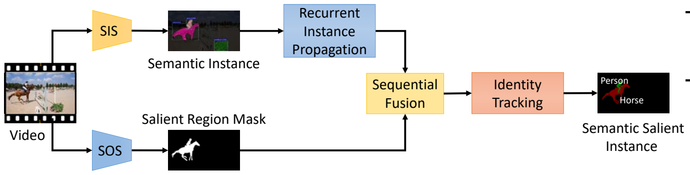


Figure 6: Pipeline of proposed SISO baseline. Yellow, orange, and blue blocks are for spatial, temporal, and spatiotemporal computation, respectively. Both two streams can work on the spatiotemporal domain.

are the mean value and the standard deviation of pixel-wise saliency values over the frame.

4.2. Sequential Fusion

When fusing semantic instances with the salient region mask for the semantic salient instance map, dealing with the areas where different instances overlap with each other becomes a crucial issue.

Almost all multi-instance segmentation methods ignore such areas and randomly merge instances [5, 36, 42, 43, 45]. Though Le *et al*. [20] proposed to merge instances depending on the order based on their topological relationships, their method requires the ground-truth label of the first video frame to learn the order. We here propose a novel sequential fusion that does not require any ground-truth label. We compute the merging order in each frame using the salient region mask from the SOS stream.

Algorithm 1 describes our proposed sequential fusion to select a set of instances (hereafter, referred to confident-instances) to compute a fusion map. We select the instance that overlaps the salient region mask best where we use IOU [14] to compute the overlapping area between the instance and the mask. We set semantic label of the selected instance to each pixel in its corresponding region of the fusion map. We then remove the overlapping area from the salient region mask. Next, we select the instance from other remaining instances that overlaps the remaining salient region mask best and then remove the overlapping area. We iterate this procedure until no instance exists inside the remaining salient region mask. In our experiments, when the IOU score for an instance is less than $\theta = 0.1$, we regarded the instance is not present in the salient region mask.

Algorithm 1: Sequential fusion and confidence computation.

```
input     : salient region mask $M$; set of instance identities $I$, where each instance $i \in I$ has region $R_i$, category $C_i$, and classification score $S_i^{(\mathrm{cls})}$
output    : fusion map $FM$; frame-confidence $FC$, set of confident-instance identities $J$, where each instance $j \in J$ has new region $\widetilde{R}_j$
parameter: threshold $\theta$
initialize: $FM \leftarrow [0]^{h \times w}$; $S^{(\mathrm{conf})} \leftarrow 0$; $J \leftarrow \emptyset$;
repeat
    $S^{(\mathrm{seg})} \leftarrow [0]^{|I|}$; // set of segmentation scores
    for $i \in I$ do
        $S_i^{(\mathrm{seg})} \leftarrow$ IOU $(R_i, M)$;
        if $S_i^{(\mathrm{seg})} == 0$ then
            $S_i^{(\mathrm{cls})} \leftarrow \emptyset$; $S_i^{(\mathrm{seg})} \leftarrow \emptyset$; $R_i \leftarrow \emptyset$; $C_i \leftarrow \emptyset$;
        end if
    end for
    $j \leftarrow \arg\max_i S_i^{(\mathrm{seg})}$;
    $\widetilde{R}_j \leftarrow R_j \cap M$;
    Pixels in $FM$ corresponding to $\widetilde{R}_j \leftarrow C_j$;
    $M \leftarrow M \backslash \widetilde{R}_j$; $J \leftarrow J \cup \{j\}$;
    $S^{(\mathrm{conf})} \leftarrow S^{(\mathrm{conf})} +$ CS$(S_j^{(\mathrm{seg})}, S_j^{(\mathrm{cls})})$;
    $S_j^{(\mathrm{cls})} \leftarrow \emptyset$; $S_j^{(\mathrm{seg})} \leftarrow \emptyset$; $R_j \leftarrow \emptyset$; $C_j \leftarrow \emptyset$;
until max $S^{(\mathrm{seg})} \leq \theta$ or $S^{(\mathrm{seg})} == \emptyset$;
$FC \leftarrow \frac{S^{(\mathrm{conf})}}{|J|}$
```

We also compute the frame-confidence for each frame by averaging the confidence scores of all the semantic salient instances in the frame. The confident score of a semantic salient instance, denoted by $\mathbf{CS}(\cdot)$, is computed as a trade-off between the IOU score and the classification accuracy: $\mathbf{CS} = \frac{(1+\beta^2)S^{(\mathrm{seg})}S^{(\mathrm{cls})}}{\beta^2 S^{(\mathrm{seg})} + S^{(\mathrm{cls})}}$, where $S^{(\mathrm{seg})}$ is the segmentation IOU score of the instance and $S^{(\mathrm{cls})}$ is the classification accuracy score of the instance. We remark that in our experiments we set $\beta^2 = 0.3$ so that the segmentation score $S^{(\mathrm{seg})}$ is more weighted.

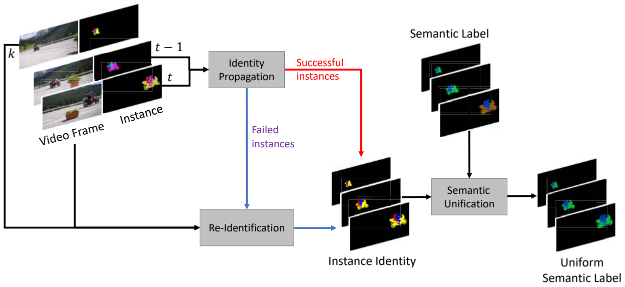


Figure 7: Flowchart of identity tracking module. Instance identities from a video frame are propagated to its new frames by flow warping. When an instance is occluded or out of frame, it is re-identified in next frames by using the feature extracted at its key-frame k. The consistency of identities and semantic labels of instances is maintained over the entire video.

4.3. Recurrent Instance Propagation

Some semantic instances may not be segmented due to severely deformed appearances caused by object motion and/or camera motion. To recover such missing semantic instances, we introduce the recurrent instance propagation where instances are recurrently propagated to neighboring frames. More specifically, We propagate instances in frames with high frame-confidences to those in frames with low frame-confidences.

Video frames are first sorted in the descending order based on their frame-confidences computed by Algorithm 1. We sequentially update confident-instances and frame-confidences of all video frames in this order. If a video frame has larger frame-confidence than its adjacent frames, instances of the frame are propagated to the next frame and the previous frame using flow wrapping/inverse flow wrapping where the flow is computed using FlowNet2 [18]. The propagated instances are then integrated to instances already segmented in the target frame. After that, we recompute frame-confidence and confident-instances of the target frame. If the frame-confidence increases, we update the frame-confidence and confident-instances of the target frame. After updating frame-confidences of all the video frames, the average confidence of the video is computed by averaging all frame-confidences. This propagation is recurrently executed until the average confidence of the video converges. We remark that we empirically observe that semantic salient instances are effectively propagated after around five iterations.

4.4. Identity Tracking

Since the semantic label of an instance is attached frame-by-frame, how to maintain the consistency of the label over the entire video is critical. To enable SISO to maintain this consistency, we introduce the identity tracking where the identities of instances are propagated over frames to maintain short-term consistency and they are re-identified and unified for long-term consistency. With this identity tracking, the identities of instances are consistently tracked over the entire video even if the instances disappear (or are occluded) and re-appear in some frames in the video. Fig. 7 depicts the flowchart of our proposed identity tracking.

4.4.1 Identity Propagation

We initialize identities of instances in the first frame. The identity propagation propagates the identifies of instances in a given frame to its next frame using flow warping. We then check how each propagated instance overlaps with instance already segmented in the target frame. Namely, for a propagated instance, we compute IOU [14] scores between the instance and each of the instances already segmented in the target frame. We then update the identity of the instance having the largest IOU score so that it is the same with the identity of the propagated instance. If none of the instances in the target frame achieves $\theta = 0.7$ of the IOU score, we regard that the propagated instance is out of frame or occluded in the target frame. Re-identification is required for such an instance.

We note that any instance at the target frame that is not propagated from the previous frame is regarded as a new instance and annotated with a new identity.

4.4.2 Re-identification

We employ instance search [41] for re-identifying instance identity, where we use feature of an instance of interest in a previous frame to detect the instance in future frames.

Given an instance of interest to be re-identified in a target frame, we first select its key-frame from previous frames and then extract a query feature from the bounding-box around the region of the instance in the key-frame. After that, we apply Faster R-CNN [40] to the target frame to generate region proposals and extract features from each of proposed regions. We then select the proposed region that is most similar to the instance based on cosine similarity between the query feature and the feature extracted from each region. Next, we compute IOU [14] between the selected proposed region and each region of all instances already segmented in the target frame. If the largest IOU score is larger than the threshold $\theta = 0.7$, the corresponding instance is updated with identity of the instance of interest.

For instance i, the key-frame is selected as follows. The instance may have multiple separated regions in a frame. We thus compute the average area of connected regions of the instance i in a frame t: $S_{i,t}^{(\text{area})} = \frac{area_{i,t}}{n_{i,t}}$, where $area_{i,t}$ is the area where instance i exists at frame t, and $n_{i,t}$ denotes the number of separated regions of instance i at frame t. The key-frame of the instance i is given by $\arg\max_{t} S_{i,t}^{(\text{area})}$.

Table 2: Results on SESIV. The best results are shown in blue.

Method	$\mathcal{JS}$	$\mathcal{FS}$
Mask R-CNN$_{\text{org}}$ [16]	0.41	0.43
Mask R-CNN$_{\text{prop}}$	15.65	16.70
Mask R-CNN$_{\text{SISO}}$	41.71	42.59
MNC$_{\text{org}}$ [9]	0.72	0.72
MNC$_{\text{prop}}$	20.36	18.97
MNC$_{\text{SISO}}$	31.07	31.08

4.4.3 Semantic Unification

For a semantic salient instance and a category, we first compute the summation over the entire video of the classification scores that the instance belongs to the category. We then choose for the instance the semantic label of the category that achieves the maximum value among all the categories. In this way, the semantic labels attached to salient instances are unified over the entire video.

5. Experiments

5.1. Implementation Details

For the SOS stream, we employed DSRFCN3D [22], using the public pre-trained model on video saliency datasets [22] (without any fine-tuning).

For the SIS stream, we employed Mask R-CNN [16] and MNC [9] to evaluate the performance of the proposed SISO on various network architectures. We used public pre-trained models without any fine-tuning (Mask R-CNN is pre-trained on the MS-COCO dataset [32], and MNC is pre-trained on the VOC Pascal dataset [11]). We remark that we used only semantic instances whose classification scores are larger than 0.7; we eliminated the other instances. We also remark that to evaluate MNC, we converted semantic ground-truth labels of the MS-COCO to their corresponding categories of the VOC Pascal and used only convertible semantic salient instances.

We implemented optical flow [18], instance search [41], and SIS models [9, 16] with python, VSOS model [22] and other modules with Matlab. All experiments were conducted on a computer with a Core i7 3.6GHz processor, 32GB of RAM, and GTX1080Ti GPU.

5.2. Evaluation Criteria

To evaluate performances, we introduce semantic region similarity and semantic contour accuracy defined as follows. Let m and g be binary masks of the predicted instance and the ground-truth instance. The semantic region similarity $\mathcal{JS}$ and the semantic contour accuracy $\mathcal{FS}$ are

$$\mathcal{JS}(m,g) = \delta_{id(m),id(g)}\delta_{sl(m),sl(g)}\mathcal{J}(m,g), \quad (1)$$

$$\mathcal{FS}(m,g) = \delta_{id(m),id(g)}\delta_{sl(m),sl(g)}\mathcal{F}(m,g), \quad (2)$$

where $\mathcal{J}(\cdot)$ and $\mathcal{F}(\cdot)$ are region similarity [11] and contour accuracy [37]. δ denotes the Kronecker delta, and $id(m)$ and $sl(m)$ are the identity and the semantic label of instance m, respectively. Remark that we compare the similarity of two instances only if they have the same identity and the same semantic label. We note that region similarity is the intersection over the union of the estimated segmentation and the ground-truth mask while contour accuracy is a trade-off between the contour-based precision and recall.

Table 3: Effectiveness of confident instance utilization. The best results are shown in blue.

Method or Metric	$SISO_a$	$SISO_b$	$SISO_c$
Sequential Fusion		✓	✓
Recurrent Instance Propagation			✓
$\mathcal{JS}$	36.57	37.43	41.71
$\mathcal{FS}$	39.59	40.32	42.59

Table 4: Effectiveness of identity tracking. The best results are shown in blue.

Method or Metric	$SISO_\alpha$	$SISO_\beta$	$SISO_\gamma$
Identity Propagation		✓	✓
Re-Identification			✓
$\mathcal{JS}$	0.95	33.74	41.71
$\mathcal{FS}$	1.02	34.56	42.59

Similar to [38], we first evaluate each instance and then take the average over the dataset. More precisely, letting V be a set of videos in the dataset, and $\mathcal{M} \in \{\mathcal{JS}, \mathcal{FS}\}$ be a given metric, the performance $\mathcal{M}(V)$ over V is defined by

$$\mathcal{M}(V) = \frac{1}{|I_V|}\sum_{i\in I_V}\frac{1}{|F_{v(i)}|}\sum_{f\in F_{v(i)}}\mathcal{M}(m_i^f, g_i^f), \quad (3)$$

where I_V is the set of annotated instances in V, $v(i) \in V$ is the sequence in which the instance $i \in I_V$ appears, and F_v is the set of frames in sequence v. m_i^f and g_i^f are respectively the predicted region and the ground-truth of instance i in frame f.

We remark that we matched identities of predicted instances at the first frame with those of the ground-truth by maximizing IOU scores between the predicted instances and the ground-truth. This avoids the identity permutation problem in the evaluation.

5.3. Results of SISO Instances

We emphasize that SISO is the first work for VSSIS, meaning that no state-of-the-art method is available for comparison. We thus evaluated the performance of various network architectures implemented in the SIS stream. Each method M, where $M = \{$Mask R-CNN, MNC$\}$, is employed with three different settings: M_{org} is the original

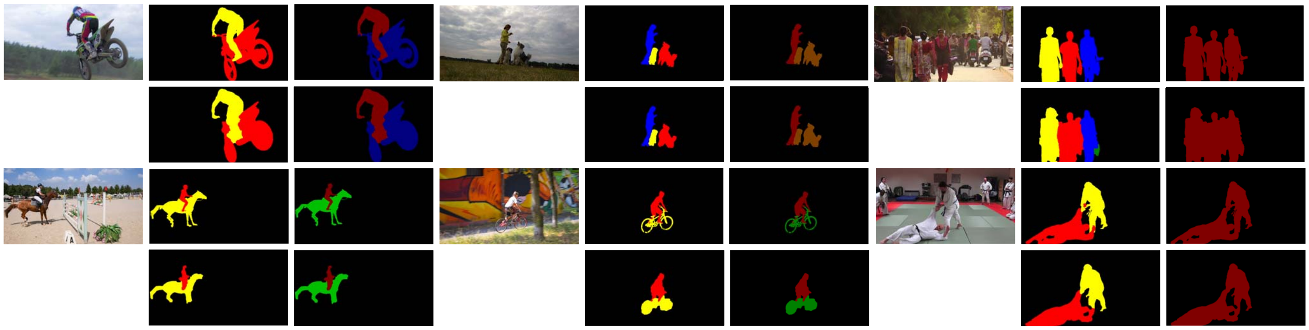


Figure 8: Visualization of some results by our method on the SESIV dataset. From left to right, original video frame is followed by instance label and semantic label, respectively. The top row indicates ground-truth labels and the bottom row shows results by our method.

model (we applied this frame-by-frame for videos), M_{prop} is the model incorporating our identity propagation module (this is just to simply exploit temporal information), and M_{SISO} is the model incorporated in our proposed SISO.

The quantitative results are shown in Table 2, indicating that SISO significantly outperforms the other settings for any SIS method on all metrics. This suggests that SISO is capable of eliminating non-salient instances and maintaining consistent identities of instances over the entire video. We also note that the setting M_{org} achieves the worst performances. This is because it is a frame-by-frame method and does not take into account temporal information. Figure 8 is the visualization of a few examples obtained by Mask R-CNN$_{\text{SISO}}$. We see that our method handles complex instances with background clutter, giving accurate and consistent segmentation.

5.4. Ablation Studies

To demonstrate the effectiveness of components in SISO, i.e., sequential fusion, recurrent instance propagation, and identity tracking, we performed experiments under controlled settings and compared results. We note that we used Mask R-CNN$_{\text{SISO}}$ for these experiments because we see that Mask R-CNN$_{\text{SISO}}$ performed better than MNC$_{\text{SISO}}$.

5.4.1 Effectiveness of Confident-Instance Utilization

As shown in Section 4, confident-instances are utilized in the sequential fusion and the recurrent instance propagation modules. To evaluate the effectiveness of confident-instances, we performed experiments under three different controlled settings: merging instances in the random order without using any confident-instance (denoted by $SISO_a$), using the sequential fusion only (denoted by $SISO_b$), and using both the sequential fusion and the recurrent instance propagation (denoted by $SISO_c$). Table 3 shows their results, indicating that (1) merging instances based on introduced confident-instances ($SISO_b$) achieves better performance than in the random order ($SISO_a$), and that (2) utilizing confident instances ($SISO_b$ and $SISO_c$) performs better than not using confident-instances ($SISO_a$). In particular, our complete method $SISO_c$ performs best.

5.4.2 Effectiveness of Identity Tracking

To evaluate the effectiveness of the identity tracking module, we performed experiments under three different controlled settings: not tracking any instances (denoted by $SISO_\alpha$), using the identity propagation only (denoted by $SISO_\beta$), and using both the identity propagation and the re-identification (denoted by $SISO_\gamma$). Table 4 shows their results and indicates that our complete method $SISO_\gamma$ exhibits outperformance against the other settings on all the metrics. In particular, the outperformance over $SISO_\alpha$ is significant. We also observe that using both the identity propagation and the re-identification ($SISO_\gamma$) brings more gains than using the identity propagation only ($SISO_\beta$). This suggests that the identity tracking contributes to maintain consistent identities of instances over the entire video.

6. Conclusion

We addressed a new task of video semantic salient instance segmentation (VSSIS), and proposed the first baseline for for VSSIS, called Semantic Instance - Salient Object (SISO). SISO is a simple yet efficient framework that jointly performs semantic instance segmentation and salient object segmentation in videos. Furthermore, SISO is capable of eliminating non-salient instances and maintaining consistency of both identities and semantic labels for salient instance over the entire video thanks to our introduced sequential fusion, recurrent instance propagation, and identity tracking. To address the task of VSSIS, we provided a new dataset SESIV consisting of 84 video sequences with pixelwisely annotated per-frame ground-truth labels.

Besides extending the quantity of the dataset, developing a way to directly segment salient instances from videos is left for future work.

References

[1] A. Borji, M. M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE TIP*, 24(12):5706–5722, 2015.

[2] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, pages 282–295, 2010.

[3] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017.

[4] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *CVPR*, 2018.

[5] J. Cheng, S. Liu, Y.-H. Tsai, W.-C. Hung, S. Gupta, J. Gu, J. Kautz, S. Wang, and M.-H. Yang. Learning to segment instances in videos with spatial propagation network. *CVPR Workshop*, 2017.

[6] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu. Salient shape: group saliency in image collections. *The Visual Computer*, 30(4):443–453, 2014.

[7] M.-M. Cheng, G.-X. Zhang, N. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011.

[8] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid. Potion: Pose motion representation for action recognition. In *CVPR*, pages 1–10, 2018.

[9] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016.

[10] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, pages 379–387, 2016.

[11] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[12] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato. Saliency-based video segmentation with graph cuts and sequentially updated priors. In *ICME*, pages 638–641, 2009.

[13] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran. Detect-and-Track: Efficient Pose Estimation in Videos. In *CVPR*, 2018.

[14] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, pages 297–312, 2014.

[15] J. He, Z. Deng, M. S. Ibrahim, and G. Mori. Generic tubelet proposals for action localization. In *WACV*, pages 343–351, March 2018.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988, 2017.

[17] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. Relation networks for object detection. In *CVPR*, June 2018.

[18] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017.

[19] A. Kirillov, E. Levinkov, B. Andres, B. Savchynskyy, and C. Rother. Instancecut: from edges to instances with multicut. In *CVPR*, volume 3, page 9, 2017.

[20] T.-N. Le, K.-T. Nguyen, M.-H. Nguyen-Phan, T.-V. Ton, T.-A. Nguyen, X.-S. Trinh, Q.-H. Dinh, V.-T. Nguyen, A.-D. Duong, A. Sugimoto, T. V. Nguyen, and M.-T. Tran. Instance re-identification flow for video object segmentation. *CVPR Workshop*, 2017.

[21] T.-N. Le and A. Sugimoto. Contrast based hierarchical spatial-temporal saliency for video. In *PSIVT*, pages 734–748, 2015.

[22] T.-N. Le and A. Sugimoto. Deeply supervised 3d recurrent fcn for salient object detection in videos. In *BMVC*, 2017.

[23] T.-N. Le and A. Sugimoto. Spatiotemporal utilization of feep features for video saliency detection. In *ICME Workshop*, 2017.

[24] E. Levinkov, J. Uhrig, S. Tang, M. Omran, E. Insafutdinov, A. Kirillov, C. Rother, T. Brox, B. Schiele, and B. Andres. Joint graph decomposition & node labeling: Problem, algorithms, applications. In *CVPR*, 2017.

[25] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, pages 2192–2199, 2013.

[26] G. Li, Y. Xie, L. Lin, and Y. Yu. Instance-level salient object segmentation. In *CVPR*, 2017.

[27] G. Li, Y. Xie, T. Wei, K. Wang, and L. Lin. Flow guided recurrent neural encoder for video salient object detection. In *CVPR*, 2018.

[28] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015.

[29] J. Li, C. Xia, and X. Chen. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE TIP*, 27(1):349–364, 2018.

[30] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, 2017.

[31] X. Liang, L. Lin, Y. Wei, X. Shen, J. Yang, and S. Yan. Proposal-free network for instance-level semantic object segmentation. *IEEE TPAMI*, pages 1–1, 2017.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[33] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, pages 678–686, 2016.

[34] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018.

[35] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE TPAMI*, 33(2):353–367, 2011.

[36] A. Newswanger and C. Xu. One-shot video object segmentation with iterative online fine-tuning. *CVPR Workshop*, 2017.

[37] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016.

[38] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[39] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä. Segmenting salient objects from images and videos. In *ECCV*, pages 366–379, 2010.

[40] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[41] A. Salvador, X. G. i Nieto, F. Marqus, and S. Satoh. Faster r-cnn features for instance search. In *CVPR Workshop*, pages 394–401, 2016.

[42] A. Shaban, A. Firl, A. Humayun, J. Yuan, X. Wang, P. Lei, N. Dhanda, B. Boots, J. M. Rehg, and F. Li. Multiple-instance video segmentation with sequence-specific object proposals. *CVPR Workshop*, 2017.

[43] G. Sharir, E. Smolyansky, and I. Friedman. Video object segmentation using tracked object proposals. *CVPR Workshop*, 2017.

[44] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg. Motion coherent tracking using multi-label mrf optimization. *IJCV*, 100(2):190–202, 2012.

[45] P. Voigtlaender and B. Leibe. Online adaptation of convolutional neural networks for the 2017 davis challenge on video object segmentation. *CVPR Workshop*, 2017.

[46] B. Wang, L. Ma, W. Zhang, and W. Liu. Reconstruction network for video captioning. In *CVPR*, June 2018.

[47] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017.

[48] W. Wang, J. Shen, and F. Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, pages 3395–3402, June 2015.

[49] W. Wang, J. Shen, and L. Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE TIP*, 24(11):4185–4196, 2015.

[50] W. Wang, J. Shen, and L. Shao. Video salient object detection via fully convolutional networks. *IEEE TIP*, 27(1):38–49, Jan 2018.

[51] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang. What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors. In *CVPR*, pages 4399–4407, 2017.

[52] N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang. Deep interactive object selection. In *CVPR*, pages 373–381, 2016.

[53] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013.

[54] Z. Zhang, S. Fidler, and R. Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *CVPR*, pages 669–677, 2016.

[55] F. Zhou, S. B. Kang, and M. Cohen. Time-mapping using space-time saliency. In *CVPR*, pages 3358–3365, June 2014.