# 11. One-Shot Viedo Object Segmentation

논문 : https://cvlsegmentation.github.io/osvos/

요약

- 동영상의 첫번째 프레임에 segmentation할 object를 표시해서 전체 프레임에서 object segmentation을 하는 기법을 제안
- ImageNet으로 CNN 모델 학습 → DAVIS Train set로 CNN 모델에 object segmentation task를 학습 → DAVIS Test data의 1 프레임으로 fine-tuning 후 나머지 프레임에 대한 정확도 계산

---

Figure 1. Example result of our technique: The segmentation of the first frame (red) is used to learn the model of the specific object to track, which is segmented in the rest of the frames independently (green). One every 20 frames shown of 90 in total.

## Contribution

- 동영상 중 1 프레임에 대한 segement 정보가 있으면 동영상에서 동일한 물체를 찾아낼 수 있음
- 각 프레임을 독립적으로 연산
    - Temporal consistency 기반의 기법들은 급변하는 물체를 찾아내기 힘들어함
    - 논문의 제안 기법은 프레임별로 독립적인 연산을 수행하므로 Occlusion 등에 robust함
    - 속도와 성능간 Trade-off가 자유로워서 사용자의 선택의 폭이 넓어짐
        - annotated frame을 추가해서 성능향상을 시킬 수 있음

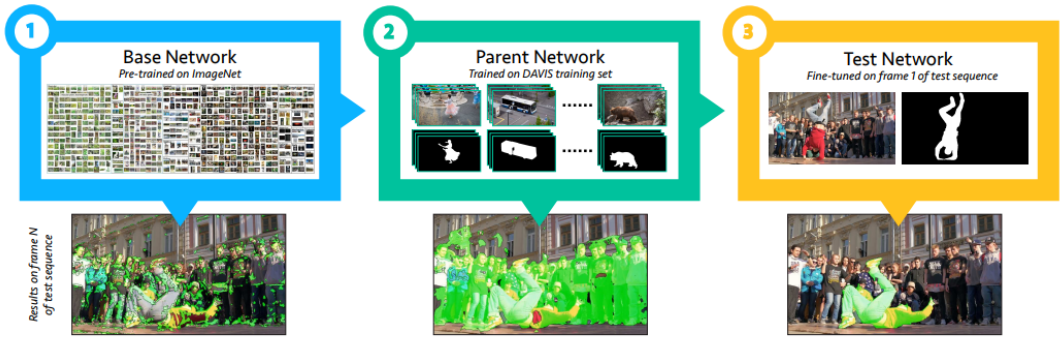



Figure 2. **Overview of OSVOS**: (1) We start with a pre-trained base CNN for image labeling on ImageNet; its results in terms of segmentation, although conform with some image features, are not useful. (2) We then train a *parent network* on the training set of DAVIS; the segmentation results improve but are not focused on an specific object yet. (3) By fine-tuning on a segmentation example for the specific target object in a single frame, the network rapidly focuses on that target.

## 학습 시나리오 (그림 2)

- 두 단계의 학습 방법
    - off-line : 배경으로부터 물체를 분리하는 방법 학습
    - on-line : 분리된 물체 중 특정한 물체를 구별하도록 학습
- Pixel-wise cross entropy loss를 사용
    - 프레임 별로 binary label의 imbalance를 해결하기 위해 식 1로 Loss 수정

$$\mathcal{L}(\mathbf{W}) = -\sum_j y_j \log P(y_j=1|X;\mathbf{W}) + (1-y_j)\log(1-P(y_j=1|X;\mathbf{W}))$$

$$= -\sum_{j\in Y_+} \log P(y_j=1|X;\mathbf{W}) - \sum_{j\in Y_-} \log P(y_j=0|X;\mathbf{W})$$

$$\mathcal{L}_{mod} = -\beta \sum_{j \in Y_+} \log P\left(y_j = 1 | X\right) - (1-\beta) \sum_{j \in Y_-} \log P\left(y_j = 0 | X\right) \quad (1)$$

$$\beta = |Y_-| / |Y|$$

- 그림 2 : Foreground Branch에 사용할 Model 학습 순서
  - Offline training
    - Base Network ( 그림 2의 (1) ): ImageNet과 같은 데이터셋으로 Classification task를 학습한 좋은 pre-train 모델을 그냥 사용
      - Base Network의 feature extractor Image Classification에 사용하는 feature representation을 사용
    - Parent Network ( 그림 2의 (2) ): DAVIS 데이터셋으로 binary mask를 학습
      - 배경과 물체를 구분할 수 있는 방법을 학습
      - 학습에는 식 1을 Loss로 사용
  - Online training/testing
    - Test set의 첫번째 frame을 가지고 parent network를 fine-tuning
      - fine-tuning을 몇번하는지에 따라 성능차이가 있음 (그림 3)
      - 480p frame (480 x 854) 학습에 102 ms 소요
    - 이후 나머지 frame을 입력으로 inference

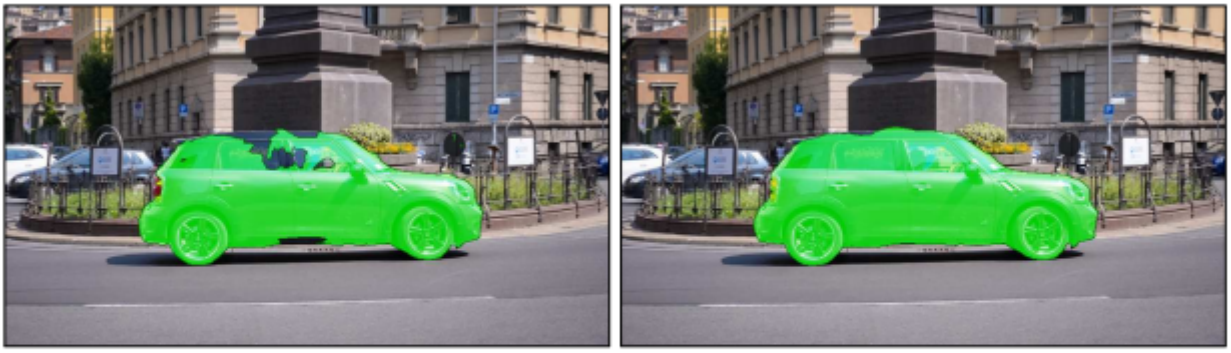


Figure 3. **Qualitative evolution of the fine tuning**: Results at 10 seconds and 1 minute per sequence.

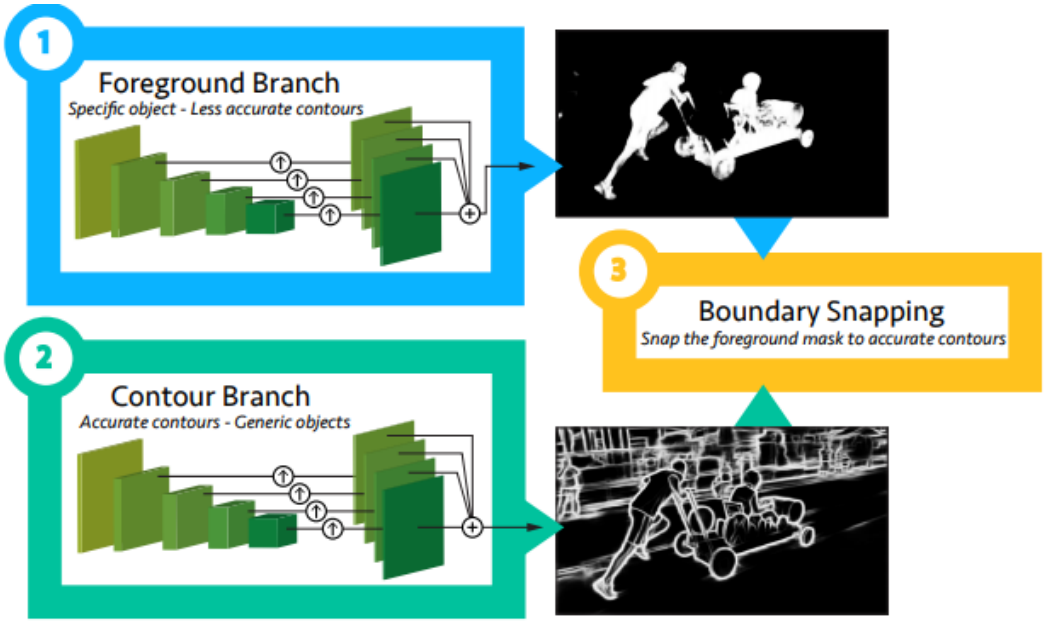



Figure 4. **Two-stream FCN architecture**: The main foreground branch (1) is complemented by a contour branch (2) which improves the localization of the boundaries (3).

## Contour snapping 그림 4의 (2)

- Image Classification task로 학습한 모델은 물체가 어디있던지 동일한 Class로 예측하기 때문에 윤곽선이 부정확하게 나옴
- 윤곽선 보정을 위해 Contour Branch로 보정하는 기법을 제안
    - Foreground branch와 동일한 모델 구조로 contours를 학습시키기
    - offline training만 하고 online에선 inference만 진행
    - PASCAL-Context dataset으로 contour 학습
- boundary snapping step (그림 4의 (3) )
    - Foreground Branch와 Contour Branch에서 얻은 결과값을 합치는 작업으로 두단계에 걸쳐 최종 출력값을 계산
    - a) Superpixel snapping
        - low strength의 threshold로 설정한 UCM(Ultrametric Contour Map)으로 계산된 contour (그림 4의 (2) 에 정렬되는 superpixel을 계산
          superpixel : perceptually simentic pixel들을 모아서 그룹화한 것
    - b) Contour recovery
        - foreground mask와 superpixel의 contour를 일치시킴
        - 일정 이상 일치하는 영역의 superpixel을 최종 mask로 사용

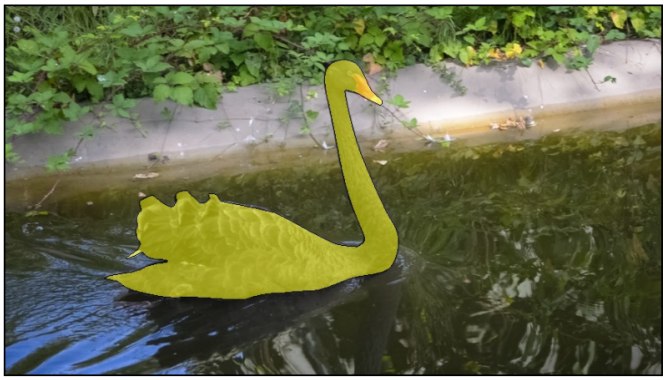
PASCAL-Context dataset 예시

Superpixel의 예시

---

## 실험

Frame: 00001



Ground Truth

| ID | Description |
|---|---|
| BC | *Background Clutter*. The back- and foreground regions around the object boundaries have similar colors ($\chi^2$ over histograms). |
| DEF | *Deformation*. Object undergoes complex, non-rigid deformations. |
| MB | *Motion Blur*. Object has fuzzy boundaries due to fast motion. |
| FM | *Fast-Motion*. The average, per-frame object motion, computed as centroids Euclidean distance, is larger than $\tau_{fm} = 20$ pixels. |
| LR | *Low Resolution*. The ratio between the average object bounding-box area and the image area is smaller than $t_{lr} = 0.1$. |
| OCC | *Occlusion*. Object becomes partially or fully occluded. |
| OV | *Out-of-view*. Object is partially clipped by the image boundaries. |
| SV | *Scale-Variation*. The area ratio among any pair of bounding-boxes enclosing the target object is smaller than $\tau_{sv} = 0.5$. |
| AC | *Appearance Change*. Noticeable appearance variation, due to illumination changes and relative camera-object rotation. |
| EA | *Edge Ambiguity*. Unreliable edge detection. The average ground-truth edge probability (using [11]) is smaller than $\tau_e = 0.5$. |
| CS | *Camera-Shake*. Footage displays non-negligible vibrations. |
| HO | *Heterogeneus Object*. Object regions have distinct colors. |
| IO | *Interacting Objects*. The target object is an ensemble of multiple, spatially-connected objects (*e.g.* mother with stroller). |
| DB | *Dynamic Background*. Background regions move or deform. |
| SC | *Shape Complexity*. The object has complex boundaries such as thin parts and holes. |

Table 1: List of video attributes and corresponding description. We extend the annotations of [50] (*top*) with a complementary set of attributes relevant to video *object* segmentation (*bottom*). We refer the reader to the supplementary material for the list of attributes for each in video in the dataset, and corresponding visual examples.

- DAVIS dataset 사용
  - Full-HD video에 frame 별 pixel level의 object segmentation이 되어있는 데이터
  - object segmentation label 외에 attribute도 제공됨
- 평가 지표
  - $\mathcal{J}$ : region similarity in terms of intersection over union (IoU)
  - $\mathcal{F}$ : contour accuracy

- $\mathcal{T}$ : temporal instability of the masks

## 실험

| | Measure | Ours | -BS | | -PN-BS | | -OS-BS | | -PN-OS-BS | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mathcal{J}$ | Mean $\mathcal{M}\uparrow$ | **79.8** | 77.4 | *2.4* | 64.6 | *15.2* | 52.5 | *27.3* | 17.6 | *62.2* |
| | Recall $\mathcal{O}\uparrow$ | **93.6** | 91.0 | *2.6* | 70.5 | *23.2* | 57.7 | *35.9* | 2.3 | *91.3* |
| | Decay $\mathcal{D}\downarrow$ | 14.9 | 17.4 | *2.5* | 27.8 | *13.0* | **−1.9** | *16.7* | 1.8 | *13.1* |
| $\mathcal{F}$ | Mean $\mathcal{M}\uparrow$ | **80.6** | 78.1 | *2.5* | 66.7 | *13.9* | 47.7 | *32.9* | 20.3 | *60.4* |
| | Recall $\mathcal{O}\uparrow$ | **92.6** | 92.0 | *0.6* | 74.4 | *18.3* | 47.9 | *44.7* | 2.4 | *90.2* |
| | Decay $\mathcal{D}\downarrow$ | 15.0 | 19.4 | *4.5* | 26.4 | *11.4* | **0.6** | *14.3* | 2.4 | *12.6* |
| $\mathcal{T}$ | Mean $\mathcal{M}\downarrow$ | 37.6 | **33.5** | *4.0* | 60.9 | *23.3* | 53.8 | *16.2* | 46.0 | *8.4* |

Table 1. **Ablation study on DAVIS**: Comparison of OSVOS against downgraded versions without some of its components.

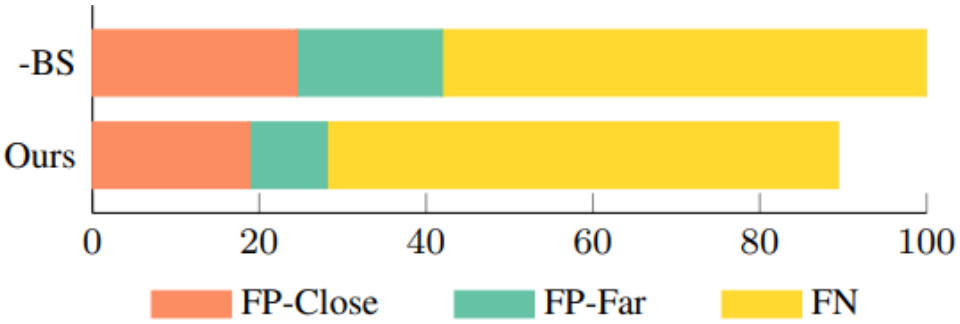



Figure 5. **Error analysis of our method**: Errors divided into False Positives (FP-Close and FP-Far) and False Negatives (FN). Values are total error pixels relative to the error in the -BS case.

- Ablation Study on DAVIS (테이블 1, 그림 5)
  - BS : boundary sanpping
  - PN : pre-training the parent network on DAVIS
  - OS : one-shot learning on the specific sequence
  - 그림 5
    - 전체 에러의 비율 분석
    - FP-Close : 20 pixel 이내의 차이로 False Positive
    - FP-Far : 20 pixel 이상의 차이로 False Positive
    - FN : False Negative

| | Measure | Semi-Supervised | | | | | | | | Unsupervised | | | | | | | Bounds | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ours | OFL | BVS | FCP | JMP | HVS | SEA | TSP | FST | NLC | MSG | KEY | CVOS | TRC | SAL | COB\|SP | COB | MCG |
| $\mathcal{J}$ | Mean $\mathcal{M}\uparrow$ | **79.8** | 68.0 | 60.0 | 58.4 | 57.0 | 54.6 | 50.4 | 31.9 | **55.8** | 55.1 | 53.3 | 49.8 | 48.2 | 47.3 | 39.3 | **86.5** | 79.3 | 70.7 |
| | Recall $\mathcal{O}\uparrow$ | **93.6** | 75.6 | 66.9 | 71.5 | 62.6 | 61.4 | 53.1 | 30.0 | **64.9** | 55.8 | 61.6 | 59.1 | 54.0 | 49.3 | 30.0 | **96.5** | 94.4 | 91.7 |
| | Decay $\mathcal{D}\downarrow$ | 14.9 | 26.4 | 28.9 | **−2.0** | 39.4 | 23.6 | 36.4 | 38.1 | **0.0** | 12.6 | 2.4 | 14.1 | 10.5 | 8.3 | 6.9 | 2.8 | 3.2 | **1.3** |
| $\mathcal{F}$ | Mean $\mathcal{M}\uparrow$ | **80.6** | 63.4 | 58.8 | 49.2 | 53.1 | 52.9 | 48.0 | 29.7 | 51.1 | **52.3** | 50.8 | 42.7 | 44.7 | 44.1 | 34.4 | **87.1** | 75.7 | 62.9 |
| | Recall $\mathcal{O}\uparrow$ | **92.6** | 70.4 | 67.9 | 49.5 | 54.2 | 61.0 | 46.3 | 23.0 | 51.6 | 51.9 | **60.0** | 37.5 | 52.6 | 43.6 | 15.4 | **92.4** | 88.5 | 76.7 |
| | Decay $\mathcal{D}\downarrow$ | 15.0 | 27.2 | 21.3 | **−1.1** | 38.4 | 22.7 | 34.5 | 35.7 | **2.9** | 11.4 | 5.1 | 10.6 | 11.7 | 12.9 | 4.3 | 2.3 | 3.9 | **1.9** |
| $\mathcal{T}$ | Mean $\mathcal{M}\downarrow$ | 37.6 | 21.7 | 34.5 | 29.6 | 15.3 | 35.0 | **14.9** | 41.2 | 34.3 | 41.4 | 29.1 | 25.2 | **24.4** | 37.6 | 64.1 | **27.4** | 44.1 | 69.8 |

Table 2. **DAVIS Validation**: OSVOS versus the state of the art, and practical bounds.

| Attr | Ours | | OFL | | BVS | | FCP | | JMP | | HVS | | SEA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AC | **80.6** | *-1.2* | 56.6 | *17.6* | 48.6 | *17.6* | 52.8 | *8.6* | 52.4 | *7.0* | 41.4 | *20.4* | 43.2 | *11.1* |
| DB | **74.3** | *6.5* | 44.3 | *27.9* | 31.9 | *33.0* | 53.4 | *5.9* | 40.7 | *19.1* | 42.9 | *13.9* | 31.1 | *22.7* |
| FM | **76.5** | *5.1* | 49.6 | *28.2* | 44.8 | *23.3* | 50.7 | *11.9* | 45.2 | *18.0* | 34.5 | *31.0* | 30.9 | *30.1* |
| MB | **73.7** | *11.0* | 55.5 | *22.8* | 53.7 | *11.5* | 50.9 | *13.6* | 50.9 | *11.1* | 42.3 | *22.5* | 39.3 | *20.3* |
| OCC | **77.2** | *3.7* | 67.3 | *1.0* | 67.3 | *-10.4* | 49.2 | *13.2* | 45.1 | *16.9* | 48.7 | *8.5* | 38.2 | *17.5* |

Table 3. **Attribute-based performance**: Quality of the techniques on sequences with a certain attribute and the gain with respect to this quality in the sequences without it (in italics and smaller font). See DAVIS [37] for the meaning of the acronyms.

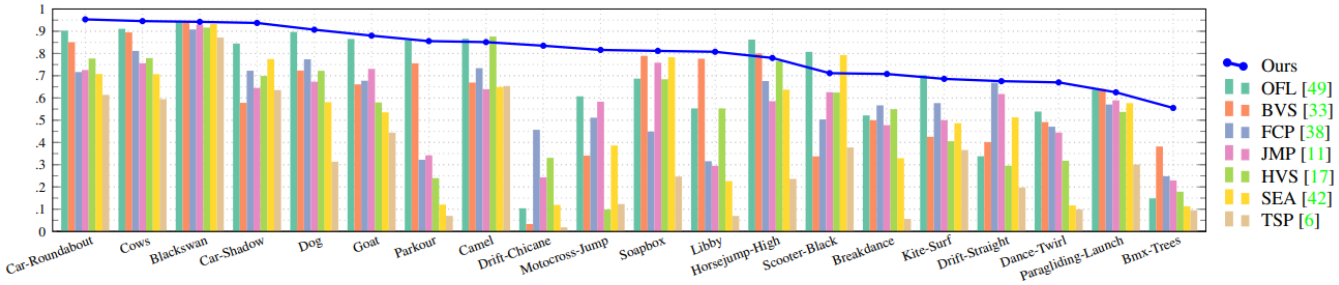



Figure 6. **DAVIS Validation**: Per-sequence results of region similarity ($\mathcal{J}$).

- 다른 기법들과 성능비교 on DAVIS (표 2, 3, 그림 6)
  - 표 2 : DAVIS Validation set 결과
  - 표 3 : 추가 속성이 주어졌을 때의 정확도 비교
    - AC : Appearance Change
    - DB : Dynamic Background
    - FM : Fast-Motion
    - MB : Motion Blur
    - OCC : Occlusion
  - 그림 6 : class 별 정확도 비교

| Training data | 100 | 200 | 600 | 1000 | 2079 |
|---|---|---|---|---|---|
| Quality ($\mathcal{J}$) | 74.6 | 76.9 | 77.2 | 77.3 | 77.4 |

Table 4. **Amount of training data**: Region similarity ($\mathcal{J}$) as a function of the number of training images. Full DAVIS is 2079.

- Parent Network 학습에 사용할 이미지 개수에 따른 정확도 변화(표 4)
  - 데이터를 넣을 수록 정확도가 상승하다가 한계점이 있는것을 볼 수 있음

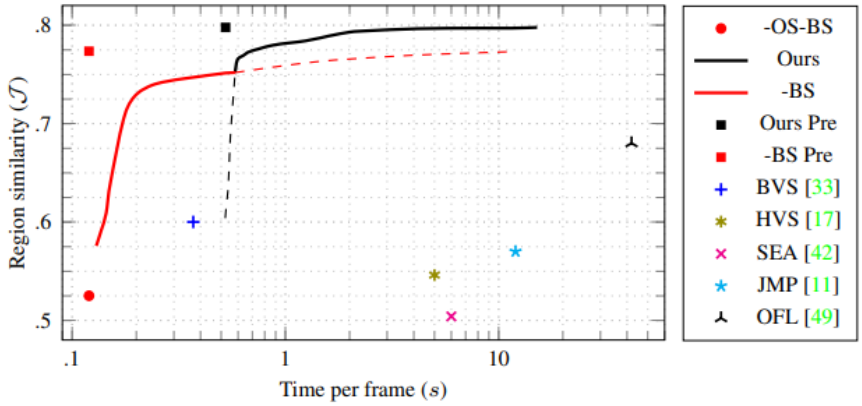



Figure 8. **Quality versus timing**: Region similarity with respect to the processing time per frame.

- **Timing (그림 8) : annotated frame의 fine-tuning 시간 투자에 따른 정확도 변화**
  - 실험에 사용한 fine-tuning 시나리오
    - 제안기법(Ours)과 -BS의 평균 1회 fine-tuning 시간 측정
    - 횟수를 늘려가면서 정확도 확인
    - 1회 fine-tuning 시간만큼 늘려가면서 그래프를 그림
  - 10초 fine-tuning으로 specific object에 대한 segmentation accuracy가 80%가까이 나오는 것을 확인

- **Refinement of result (표 5, 그림 9)**
  - fine-tuning에 사용하는 Annotated frame 개수에 따른 정확도 변화 관찰
  - 0개의 경우 parent network를 바로 적용했을 때의 결과 (fine-tune 하지 않음)

| Annotations | 0 | 1 | 2 | 3 | 4 | 5 | All |
|---|---|---|---|---|---|---|---|
| Quality ($\mathcal{J}$) | 58.5 | 79.8 | 84.6 | 85.9 | 86.9 | 87.5 | 88.7 |

Table 5. **Progressive refinement**: Quality achieved with respect to the number of annotated frames OSVOS trains from.

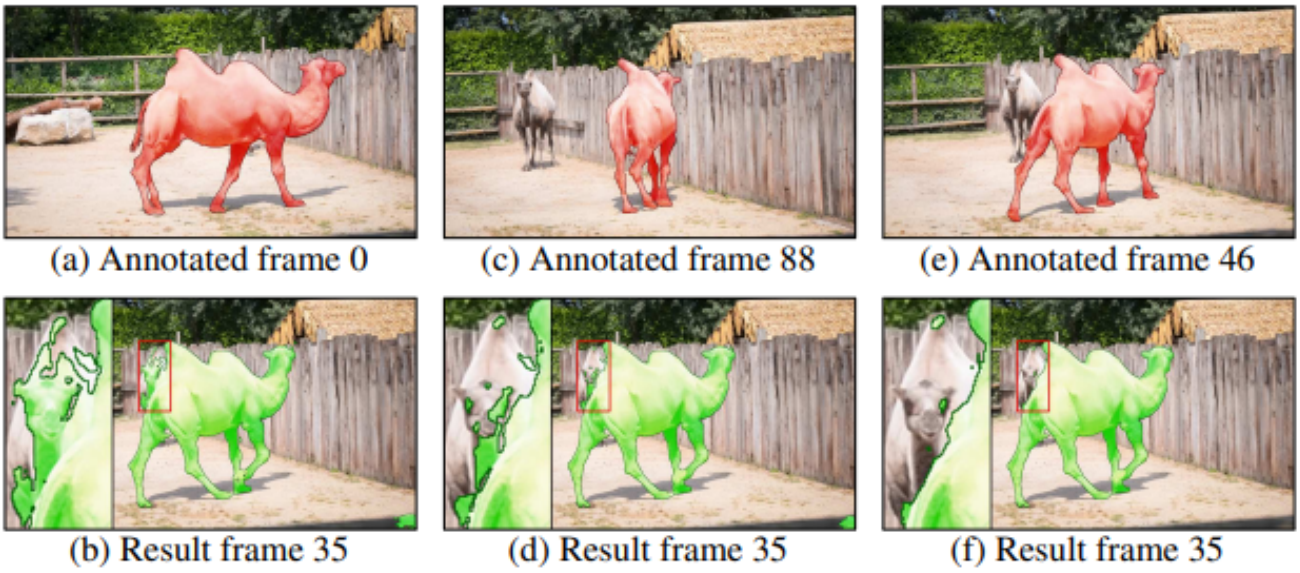



Figure 9. **Qualitative incremental results**: The segmentation on frame 35 improves after frames 0, 88, and 46 are annotated.

- Evaluation as a tracker (표 6)
  - Visual Object Racking (VOT)에 대한 다른 기법과의 정확도 비교
  - Overlap : Bounding box 겹침 정도

| Overlap | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|
| Ours | **78.2** | **72.2** | **65.8** | **59.4** | **49.6** |
| MDNET [32] | 66.4 | 57.8 | 43.4 | 29.5 | 14.7 |

Table 6. **Evaluation as a tracker**: Percentage of bounding boxes that match with the ground truth at different levels of overlap.

- Results on Youtube-Objects (표 7)
  - Youtube-Objects 데이터셋에서의 다른 기법과의 성능비교
  - state of the art인 OFL과 비등비등

| Category | Ours | OFL | JFS | BVS | SCF | AFS | FST | HBT | LTV |
|---|---|---|---|---|---|---|---|---|---|
| Aeroplane | 88.2 | **89.9** | 89.0 | 86.8 | 86.3 | 79.9 | 70.9 | 73.6 | 13.7 |
| Bird | **85.7** | 84.2 | 81.6 | 80.9 | 81.0 | 78.4 | 70.6 | 56.1 | 12.2 |
| Boat | **77.5** | 74.0 | 74.2 | 65.1 | 68.6 | 60.1 | 42.5 | 57.8 | 10.8 |
| Car | 79.6 | **80.9** | 70.9 | 68.7 | 69.4 | 64.4 | 65.2 | 33.9 | 23.7 |
| Cat | **70.8** | 68.3 | 67.7 | 55.9 | 58.9 | 50.4 | 52.1 | 30.5 | 18.6 |
| Cow | 77.8 | **79.8** | 79.1 | 69.9 | 68.6 | 65.7 | 44.5 | 41.8 | 16.3 |
| Dog | **81.3** | 76.6 | 70.3 | 68.5 | 61.8 | 54.2 | 65.3 | 36.8 | 18.0 |
| Horse | **72.8** | 72.6 | 67.8 | 58.9 | 54.0 | 50.8 | 53.5 | 44.3 | 11.5 |
| Motorbike | 73.5 | **73.7** | 61.5 | 60.5 | 60.9 | 58.3 | 44.2 | 48.9 | 10.6 |
| Train | 75.7 | 76.3 | **78.2** | 65.2 | 66.3 | 62.4 | 29.6 | 39.2 | 19.6 |
| Mean | **78.3** | 77.6 | 74.0 | 68.0 | 67.6 | 62.5 | 53.8 | 46.3 | 15.5 |

Table 7. **Youtube-Objects evaluation**: Per-category mean intersection over union ($\mathcal{J}$).