Semantic Image Segmentation

Falong Shen

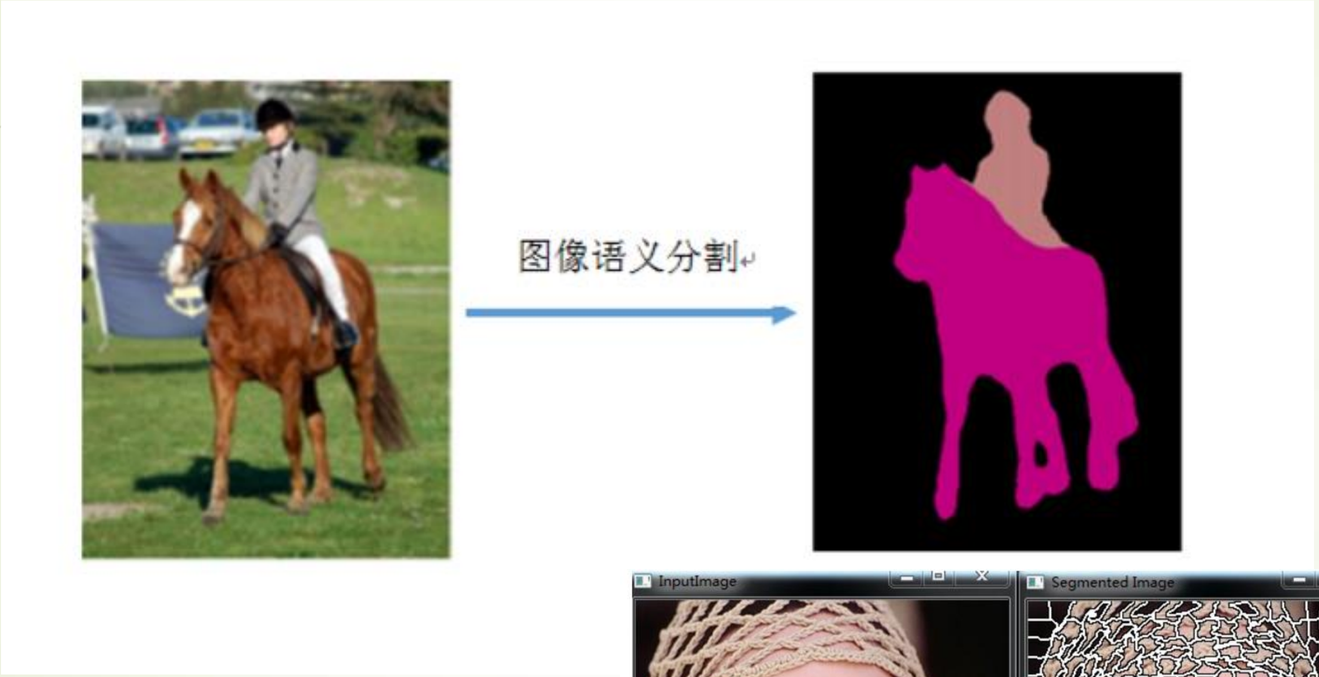
08/20/2017

Outline

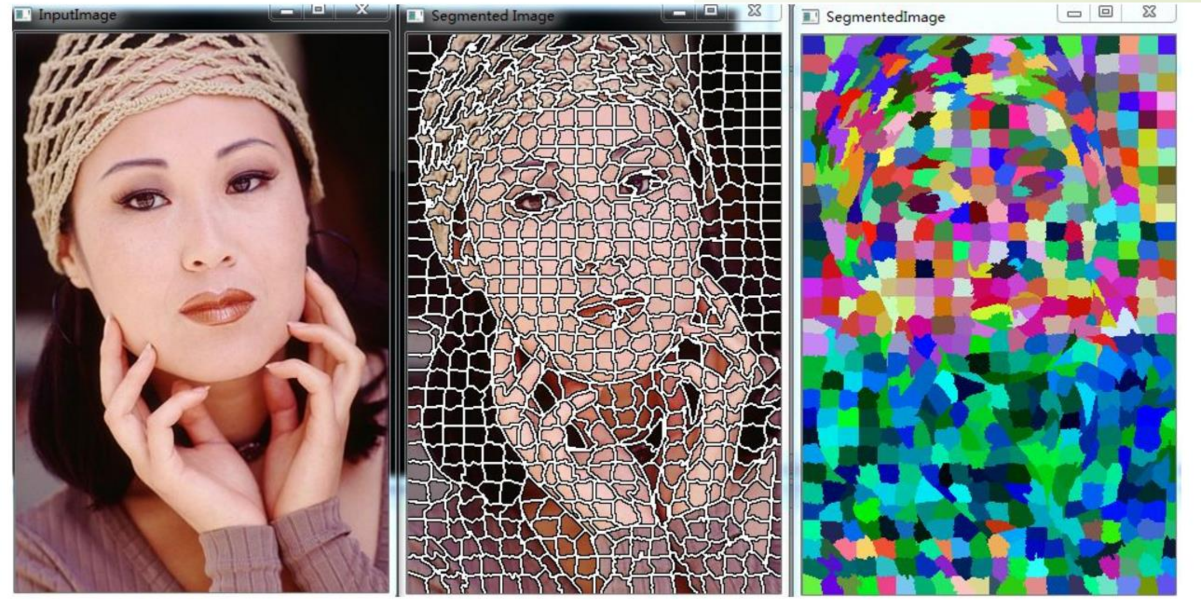
- Introduction to semantic image segmentation
- Related Works
  - FCN
  - Deeplab
- Proposed Methods
  - High order context: MAP
  - Guidance CRF: delineate the object boundary
- Experiments
  - Top performance on segmentation datasets: pascal voc, cityscapes, imageNet…
  - Ablative Studies
  - Fast and accurate segmentation
- Conclusion

Introduction to semantic image segmentation

Task:



Old methods:

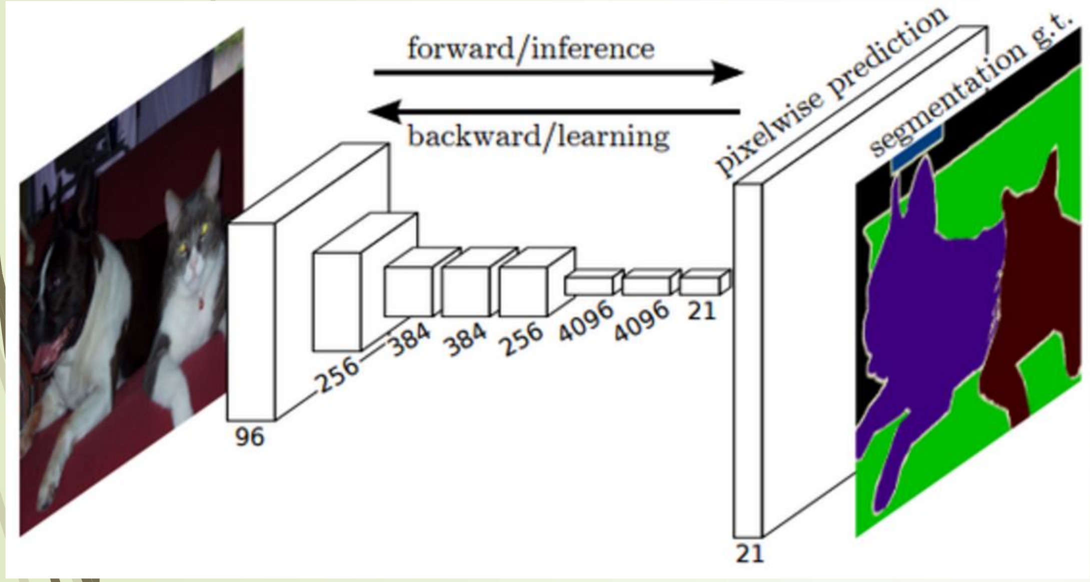


Datasets

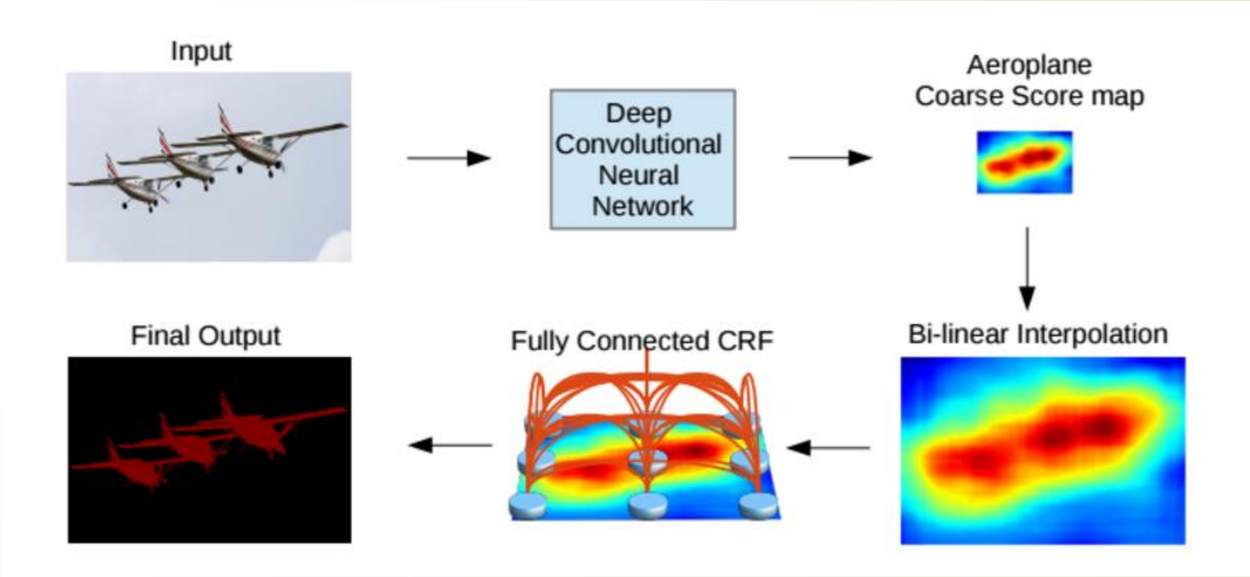
- Pascal VOC 2012
  - ~300x500 arbitrary image
  - 21 categories: plane, person, bird, bike…
  - ~10,000 pixel label images. (inconsistent label strategy for some categories)
- Cityscapes
  - 1024x2048 urban street image
  - 19 categories: person, car, bus, sky…
  - 3475 pixel label images.
- MIT Scene Parsing
  - Arbitrary image
  - 150 categories: person, sky, road, grass
  - ~20,000 pixel label images

Related Works

Fully convolutional neural networks



Deeplab (dilated CNN + bilateral CRF)



Proposed Methods

- Structured Patch Prediction
- High order context: MAP
- Guidance CRF: delineate the object boundary

Structured Patch Prediction

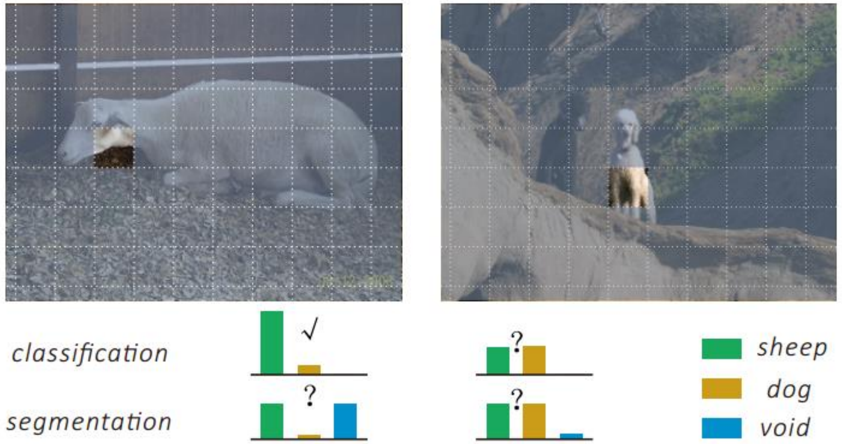


Figure 1: The belief-frequency ambiguity when transferring model from classification to segmentation. The right image is a hard example and both models produce a confusing prediction. The left image is an easy example, the segmentation model still produces a confusing prediction in order to make spatial prediction.

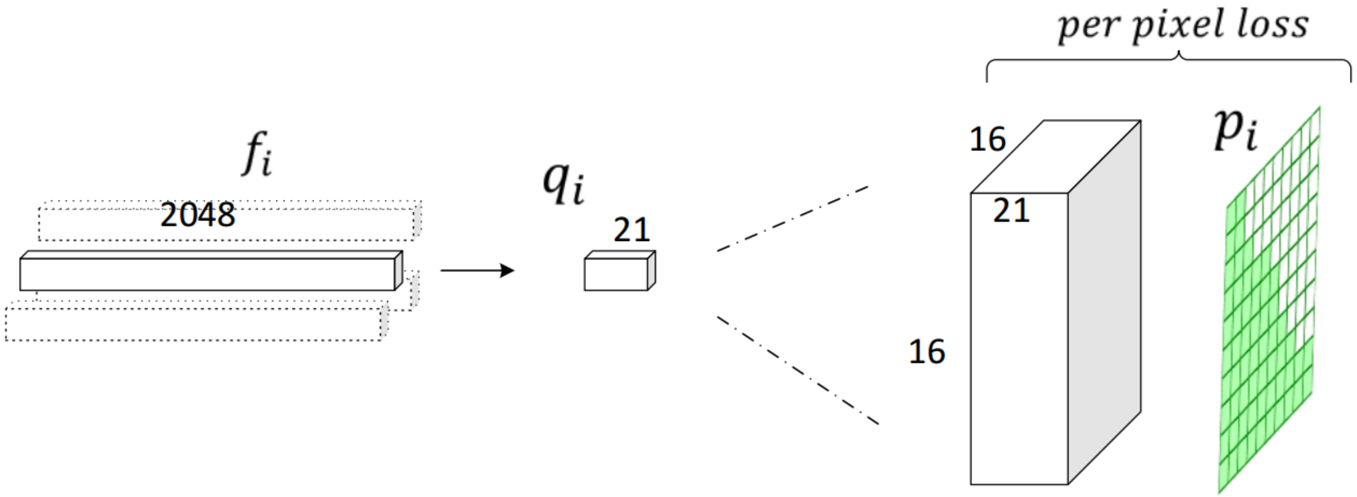
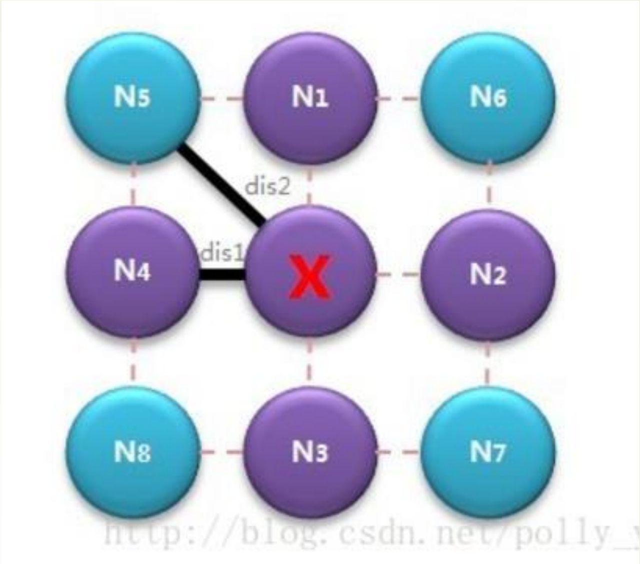


Figure 2: The 2048-D feature vector goes through a 21-D bottle neck before up-sampling to 16×16, which leads to heavily information loss.

Markov random field



$$\max_X P(X|I)$$

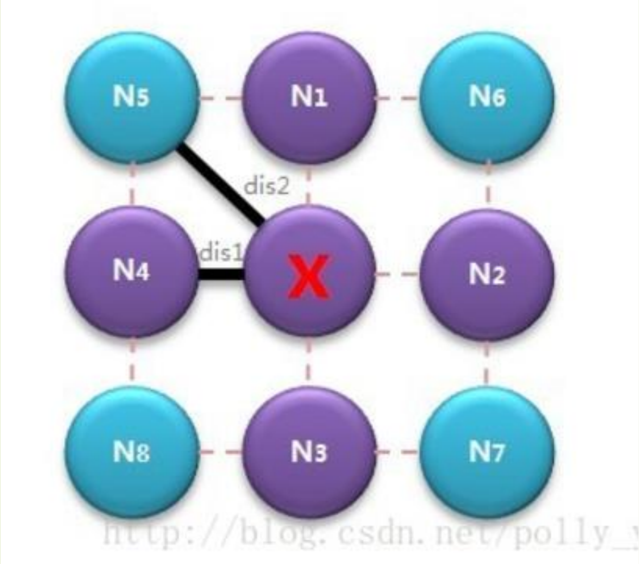
- MRF = Undirected graph:
  - $P(X_i|X_{G\backslash i}) = P(X_i|X_{N_i})$
  - Independent without edge
- MRF ↔ Gibbs distribution
  - $P(X) \propto \prod_c \phi_c(X_c)$

Proof:

1. Gibbs distribution → MRF
   Very Easy. Del the common factor is OK.

2. MRF → Gibbs distribution
   Construct the potential function ϕ for each cliques (connected sub-graph).

Second order MRF



$$\max_X \sum_i \phi_i(x_i) + \sum_{i,j} \psi_{ij}(x_i, x_j)$$

$$\max_X P(X|I)$$

$$P(X) \propto \prod_c \phi_c\,(X_c)$$

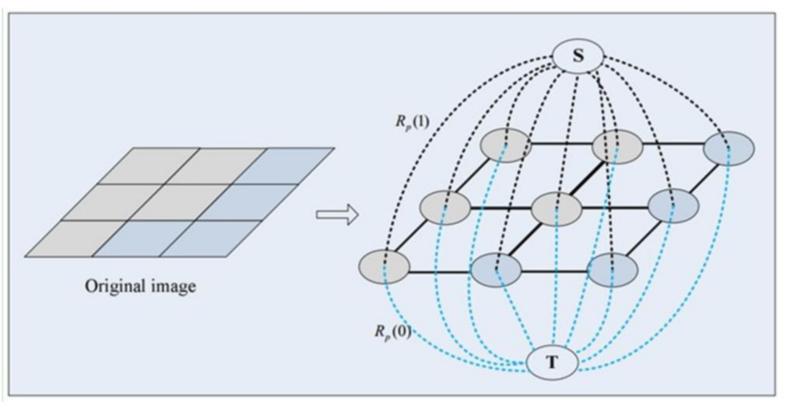
Maximum a posteriori (MAP)

$$\max_X \sum_i \phi_i(x_i) + \sum_{i,j} \psi_{ij}(x_i, x_j)$$

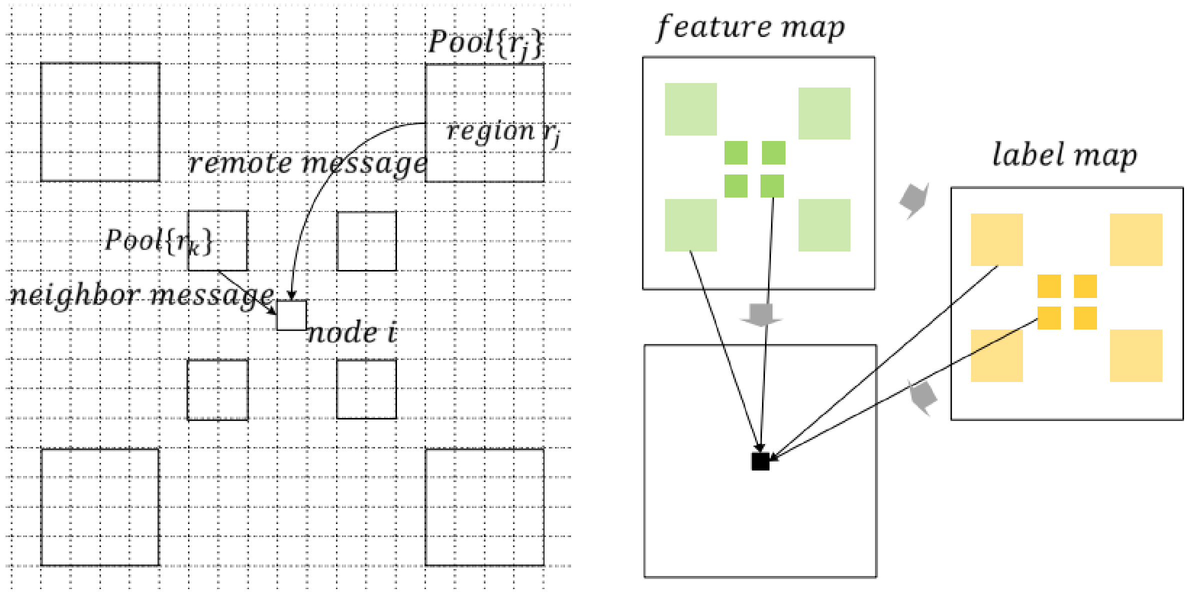
Solution 1. Mean field

$$\min_Q \mathrm{KL}\left(\exp\left(\sum_i \phi_i(x_i) + \sum_{i,j} \psi_{ij}(x_i, x_j)\right), \qquad \prod_i Q_i(x_i) \right)$$

Solution 2. Graph cut (two states)



High order context



(a) (b)

Figure 3: Illustration of context CRF. (a) We exploit a quite large field (28 × 28 on the feature map) to collect context information. The messages from neighbor regions and remote regions are pooled with different size in order to avoid over-fitting. (b) Both feature map and score map are exploited to produce messages.

Guidance CRF

Algorithm 1 Guidance CRF

Forward

input: Down-sampled Guidance image I, segmentation score map ϕ^u, compatibility matrix μ, weight parameter λ,maximum iteration k_{max}, $k = 0$, $\phi^0 = \phi^u$.

while $k < k_{max}$

1. $q^k(x_i) = \frac{1}{Z_i} \exp[-\phi^k(x_i)]$. ▷ Softmax
2. $g^k(x_i) = \sum_j w_{ij}(I) q^k(x_j)$ ▷ Guided filtering
3. $m^k(x_i) = \sum \mu(x_i, x_j) g^k(x_j)$ ▷ Compatibility transform
4. $\phi_i^k(x_i) = \phi_i^u(x_i) - \lambda m^k(x_i)$ ▷ Local update
5. $k = k + 1$

endwhile

output: marginal potential ϕ^b

Experiments

Table 1: Results on Pascal VOC 2012 *test* set and Cityscapes *test* set. Measured by the mean IoU (%). Both of our submitted models are fine-tuned from Resnet-101 and exploit MS-COCO.

Method	PasVOC12	CityScapes
DPN[23]	77.5	66.8
Dilation10[33]	-	67.1
Adelaide_context[19]	77.8	71.6
Adelaide_VeryDeep[31]	79.1	-
LRR_4x[7]	79.3	71.8
DeepLab-v2[4]	79.7	70.4
CentraleSupelec Deep G-CRF[1]	80.2	-
SegModel	82.5	79.2

Table 2: Results on ADE20K *val* set and *test* set. Measured by the average of mean IoU and pixel accuracy (%). Our models are trained on ADE20K *train* set, without resorting to MS-COCO or Place365. The performance on the *val* set is evaluated by a single model.

Method	*val*	*test*
CRFasRNN[35]	-	47.0
ACRV-Adelaide[19]	-	53.3
Hikvision	60.4	53.4
CASIA_IVA	-	54.3
SegModel	61.2	54.5
360+MCG-ICT-CAS_SP	-	55.6
Adelaide[31]	-	56.7
SenseCUSceneParsing[34]	63.1	57.2

Ablative study

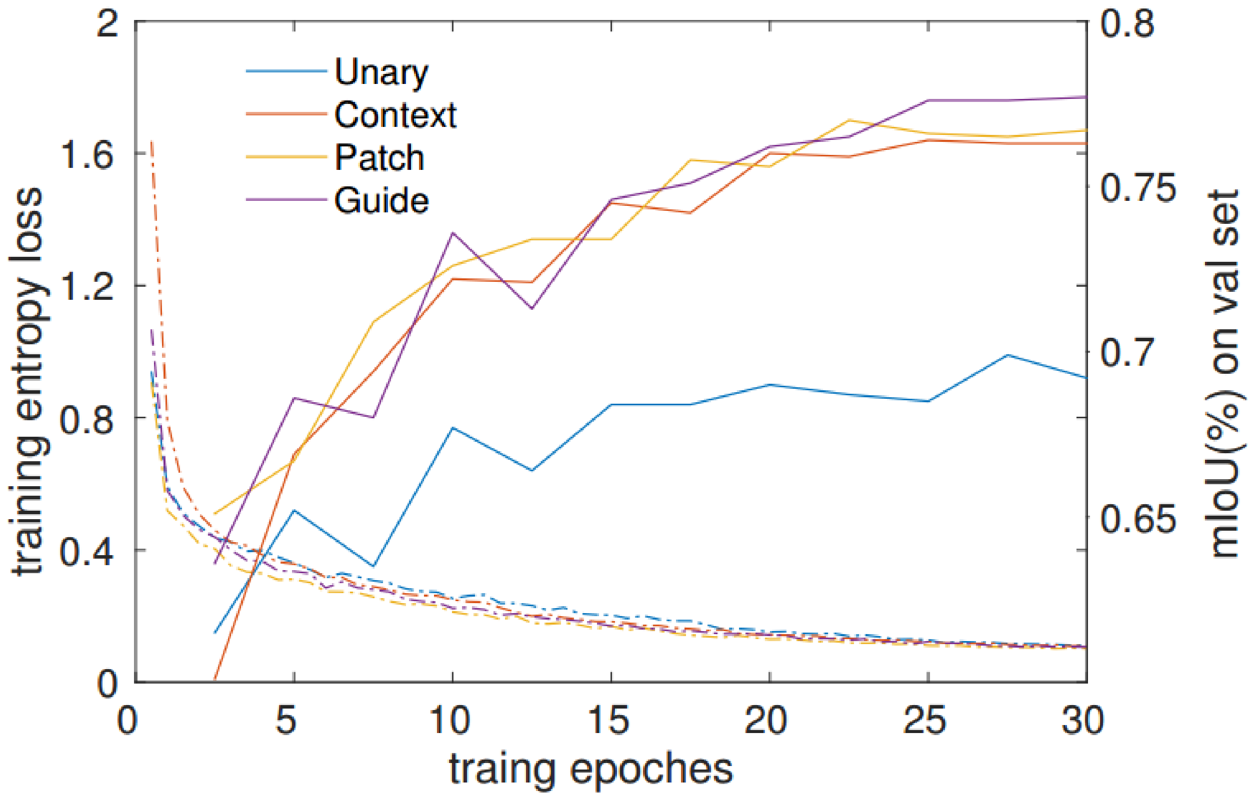


Figure 7: Training curves.

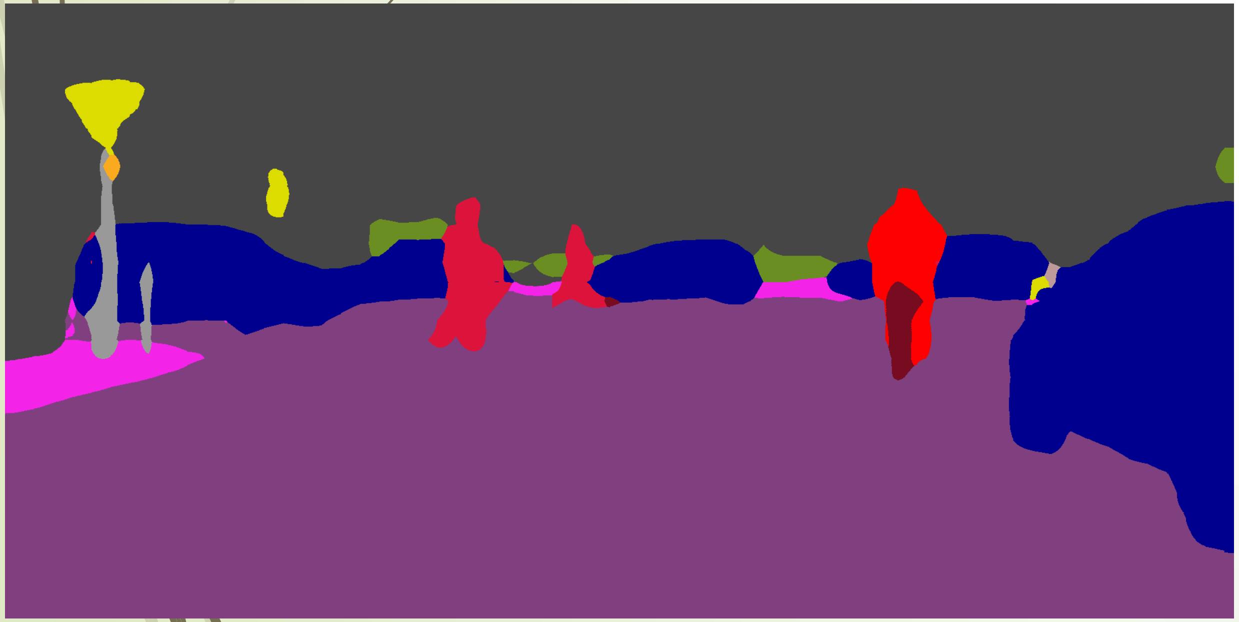
part	time(ms)
Unary	43.7
Context	49.8
Patch	50.0
Guide	54.4

Table 4: Inference time for a 500 × 300 color image.

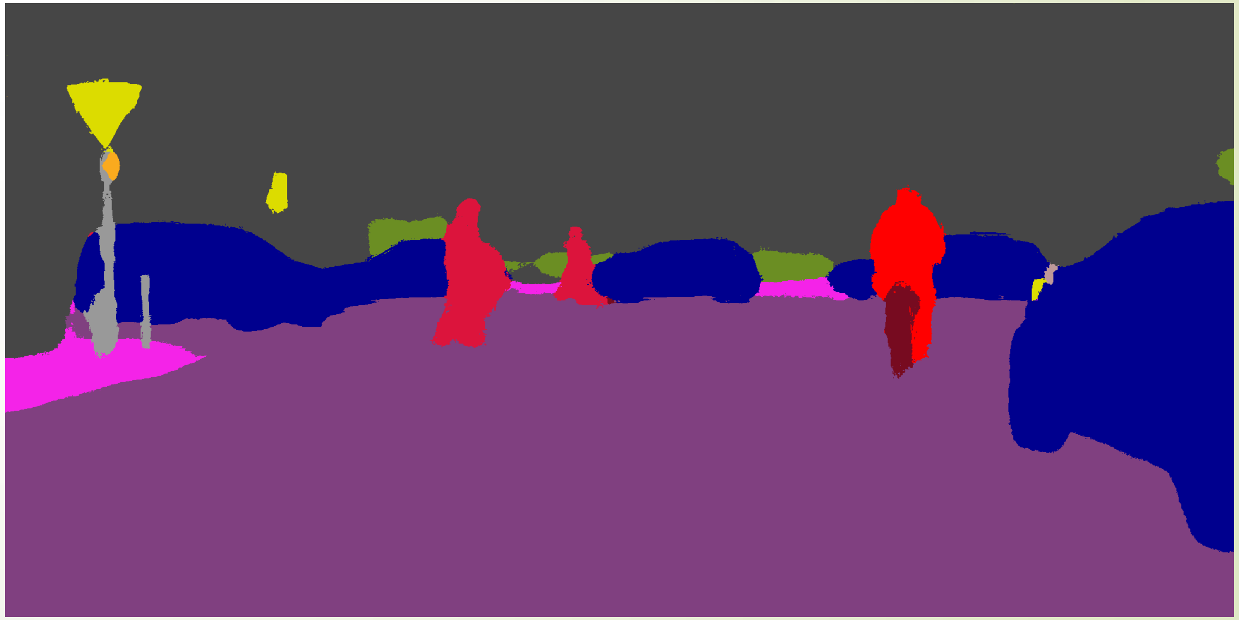
Fast segmentation

- Down-sample input image from 1024x2048 → 256x512
- Feed into FCN
- Up-sample the scoremap and align the object boundary with guidance CRF.

The total process costs about 60ms on a Pascal Titan X with fp32.

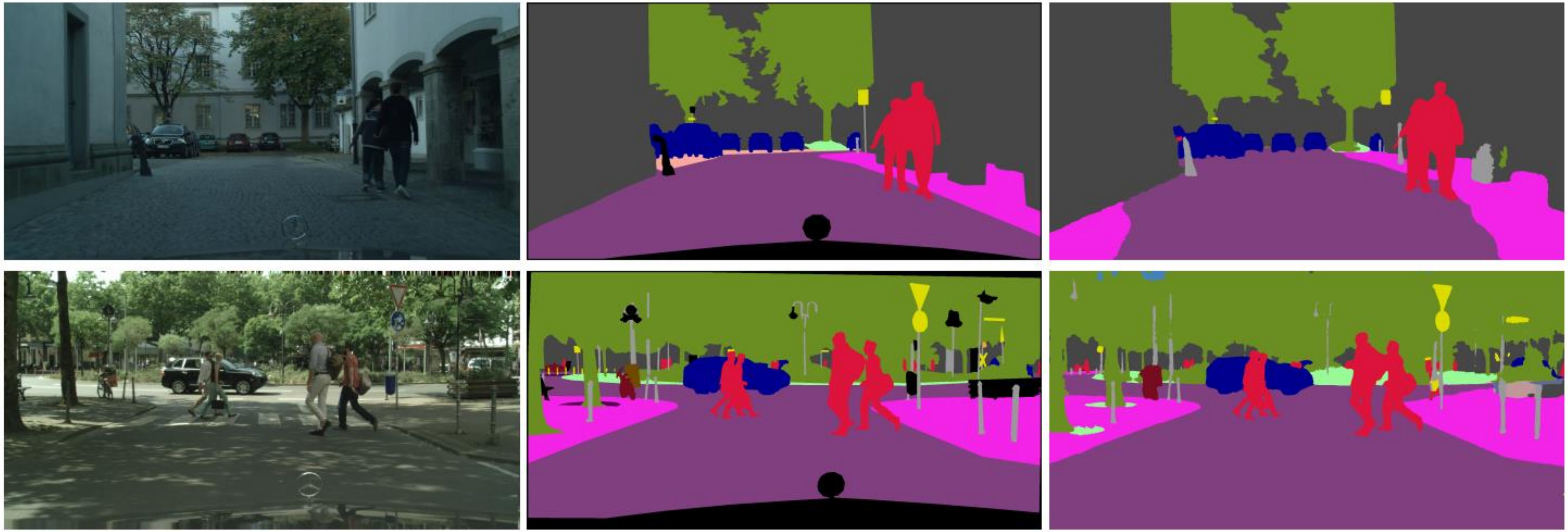


Before CRF



After CRF

Experiments



(a) Input (b) Truth (c) Prediction

Figure 5: Some visual results of Cityscapes *val* set. It costs about 0.5s for a 2048 $\times$ 1024 color image.

Conclusion

- The dominant framework of semantic segmentation is FCN + CRF.
- The base model is important to train a good segmentation model.
  - Good classification model are Not always good segmentation model.
  - Very important to get rid of over-fitting.
- Our segmentation model is fast and accurate. It is a good choice to use our SegModel for semantic image segmentation.