Deep Free-Form Deformation Network for Object-Mask Registration

Haoyang Zhang
ANU, Data61 CSIRO
Canberra, Australia
haoyang.zhang@anu.edu.au

Xuming He
ShanghaiTech University
Shanghai, China
hexm@shanghaitech.edu.cn

Abstract

This paper addresses the problem of object-mask registration, which aligns a shape mask to a target object instance. Prior work typically formulate the problem as an object segmentation task with mask prior, which is challenging to solve. In this work, we take a transformation based approach that predicts a 2D non-rigid spatial transform and warps the shape mask onto the target object. In particular, we propose a deep spatial transformer network that learns free-form deformations (FFDs) to non-rigidly warp the shape mask based on a multi-level dual mask feature pooling strategy. The FFD transforms are based on B-splines and parameterized by the offsets of predefined control points, which are differentiable. Therefore, we are able to train the entire network in an end-to-end manner based on L_2 matching loss. We evaluate our FFD network on a challenging object-mask alignment task, which aims to refine a set of object segment proposals, and our approach achieves the state-of-the-art performance on the Cityscapes, the PASCAL VOC and the MSCOCO datasets.

1. Introduction

Aligning a shape mask to object instances is a commonly used strategy in segmenting objects from background or inferring shape deformation of individual objects, and has wide applications in semantic instance segmentation [34], object proposal generation [14] and visual object tracking [19], *etc*. While it can be viewed as a special case of image registration problem [39], such object-mask alignment task is more challenging as the mask lacks internal structure for finding the dense correspondence between the target object and itself.

Most existing approaches address this problem by formulating it as an object segmentation task, in which the shape mask is used as an initialization, such as contour matching [5], or an instance shape prior for binary object segmentation [23, 31]. However, the resulting segmentation task is usually equally challenging, and does not provide shape alignment between mask and object.

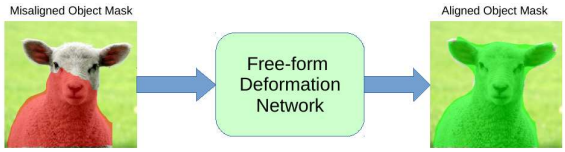


Figure 1. An illustration of the object-mask alignment problem and the transformation implemented by the deep free-form deformation network.

An alternative, and sometimes more natural approach to the object-mask alignment problem is to predict a 2D spatial transformation that registers mask onto the target object, as shown in Figure 1. Such a transformation-based strategy has several advantages in practice. First, the problem of predicting 2D transforms is typically simpler due to the fact that the common transformation families, such as affine or TPS [27], have fewer degrees of freedom and thus the output of prediction lies in a lower dimensional space. Second, for slightly mis-aligned mask and object, transforming binary masks is more efficient than recomputing the segmentation or doing image registration. Finally, the predicted transformation allows us to infer the detailed shape deformation of an instance relative to its canonical shape mask.

In this paper, we propose a deep learning approach to address the object-mask alignment problem. Given an input image containing the target object and an initial mask, our approach learns a non-rigid 2D transform that warps the mask onto the target object. To achieve this, we design a novel spatial transformer network that predicts a free-form deformation (FFD) [33] transform and applies the non-rigid transform to the input mask to generate a better alignment between the mask and object.

Specifically, we build a deep convolutional neural network consisting of two modules. The first module computes the convolutional feature maps from the input image, and extracts a feature representation of the image region covered by the mask. To encode the shape information of the initial mask, and the image cues around object, we develop a multi-level dual mask feature pooling method to capture the

misalignment between the mask and object. Based on the multi-level features, the second network module predicts a FFD transform parameterized by the offsets of predefined control points through regression. It then applies the B-spline based FFD transform to the initial mask based on a grid generator and a bilinear sampler, which produces the final warped object mask.As these two network modules are differentiable, we train the entire deformation network in an end-to-end fashion using L_2 matching loss.

We evaluate our FFD network on a challenging object-mask alignment task, in which we aim to refine a set of object segment proposals generated from state-of-the-art methods. Our results show that we achieve sizable improvements in Average Recall on the Cityscapes, the PASCAL VOC and the MSCOCO datasets for different initial proposal methods, which validates the efficacy of our deep FFD network.

2. Related Work

Image registration is a long-standing problem in computer vision and medical image analysis, which is typically applied to two images and aims to find dense or sparse correspondence between them based on similar local structures [39, 6]. The registration geometrically aligns two images (the reference and moving images), through gradually minimizing the difference between the images [27, 1]. In this work, however, we learn to predict the underlying deformations between a binary shape mask and its ground-truth object region, which is more challenging than the standard image registration task.

Our work is inspired by the B-spline FFD model [25], which is a powerful pool for modelling local and non-rigid deformations. It has been widely used in medical image registration [33] and shape registration [16]. The basic idea of the FFD model is to deform an object by manipulating an underlying mesh of control points. The control points act as parameters of the FFD model and determine the deformations being modelled. In our work, we use the FFD model to encode the transformation between the object mask and its ground-truth object region, for its flexibility and differentiable property.

The object-mask alignment can be formulated as an object segmentation problem and solved by a variety of semantic segmentation techniques (*e.g.*, [2, 37]). Early work on level-set based segmentation start from an initial contour and iteratively evolve the contour towards the target object by minimizing a functional energy function [5]. More recent approaches tend to use initial masks as a prior in inferring object segmentation. The masks can be transferred from similar images with object annotations [23, 24, 21], object shape prior [31] or discriminatively trained Exemplar-SVMs [34, 14]. However, it usually remains challenging to solve the corresponding segmentation problem. In this work, we take an alternative perspective and learn a non-rigid transformation to warp the mask onto object.

Learning deep regression networks to align objects has been explored in a variety of problem settings. In [36], the authors propose a deep deformation network for efficient object landmark localization. [20] introduces a warpnet to match images of objects, from which it builds single-view reconstruction. The spatial transformer network (STN) [18] learns a parametric transform to recover the canonical view of objects for better classification accuracy. Our method is built on top of the STN and mainly addresses the novel task of aligning a mask to object.

Object segment proposal generation is an important step for semantic instance segmentation task. One strategy is to generate object bounding boxes first based on handcrafted features [38] or deep networks [32] followed by object segmentation. Alternatively, grouping-based methods use mid-level image cues to generate and rank multiple segment candidates [3, 17, 35, 22, 30]. Recent approaches to proposing object segments learn a deep network that directly predicts object masks from the input image. In particular, DeepMask [28] builds a two-branch deep network, jointly producing a binary mask and an objectness score for every patch in an image. Dai *et al*. [8] propose a multi-task network cascade for instance segmentation, in which the first two stages generate generic bounding box proposals as well as an object mask for each bounding box. Only a few attempt to improve the quality of object segment proposals: recent work of SharpMask [29] builds a refinement network on top of the DeepMask net to obtain better boundary alignment. Our method, in contrast, explicitly learns a non-rigid spatial transform network to warp any initial object candidate towards its nearest object.

3. Deep Free-Form Deformation Network

We aim to generate an object segmentation by aligning an initial mask to its target object in an input image. To this end, we take the transformation-based strategy that learns a 2D spatial transformer to warp the initial mask to the target object. In this section, we introduce a deep convolutional neural network that first predicts a non-rigid transformation and then applies the transform to the initial mask to produce the aligned object mask. Our network is fully differentiable and can be trained in an end-to-end fashion.

More specifically, our network consists of two modules: the first computes convolutional feature maps and extracts multi-level features to capture the misalignment between the mask and object, while the second module predicts the non-rigid transformation and warps the initial mask. Figure 2 illustrates the overview of our network structure. We now describe each module of our system in detail.

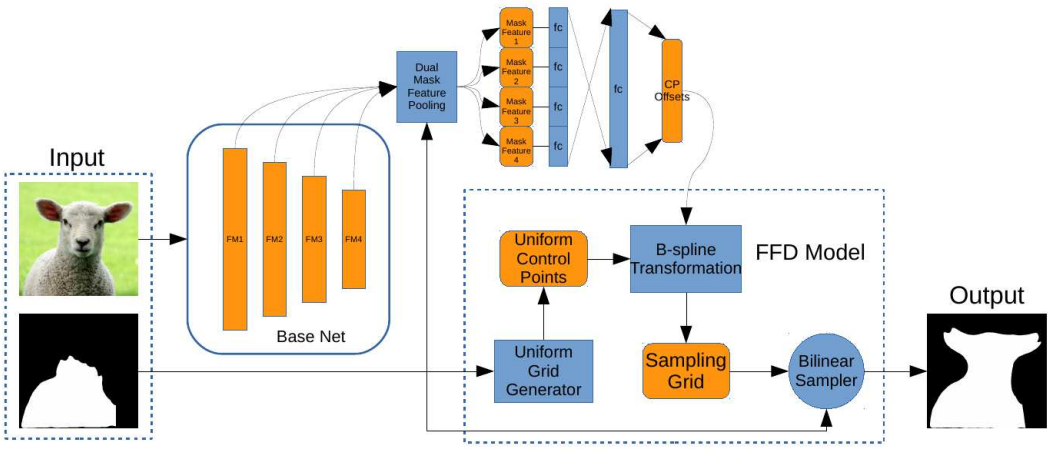


Figure 2. An overview of our deep FFD network for object-mask alignment. The entire network consists of two modules: the first computes the convolutional feature maps and extracts mask features using dual mask pooling, while the second predicts the FFD transform and warps the input mask onto the target object.

3.1. Convolutional Features and Mask Pooling

Our first network module uses a base convolutional neural network (CNN) to compute the convolutional feature maps of the input image. To capture the misalignment between the initial mask and its target object, we introduce a dual mask feature pooling scheme to extract multi-level features from the feature maps. In particular, this scheme enables us to capture the mask shape information and the spatial context cue around the object region that can guide the network to predict the spatial warping.

Our pooling layer takes as input a set of convolutional feature maps and an object mask, and generates an object-mask descriptor. Its design is inspired by the standard RoI pooling [11] and the convolutional feature masking [7] methods. Specifically, we form a tight bounding box enclosing the mask as well as a larger box by expanding the tight box in its height and width directions by 1.6 times. We first do weighted RoI pooling in the tight box, where the output of the standard RoI pooling in each cell is weighted by the overlap ratio between the cell and the mask. This generates the first type of mask features, encoding the shape and the convolutional features covered by the mask. We then perform the standard RoI pooling in the larger bounding box. This second type of features captures the spatial context cue of the mask and the target object. The final object-mask descriptor is formed by concatenating the two types of pooled mask features. Note that different from the RoI pooling in object detection [11], we compute the mask feature pooling on all convolution feature maps generated by the base network (as shown in Figure 2), which allows us to capture both local and global cues for predicting the transformation. Figure 3 illustrates the dual mask feature pooling process for a single level of feature maps.

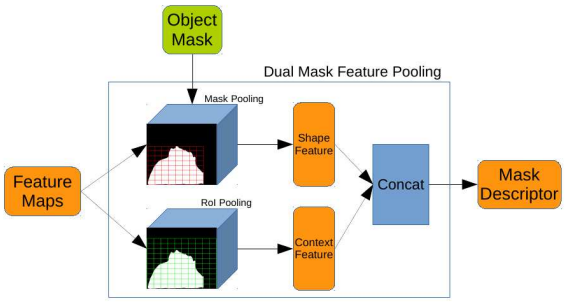


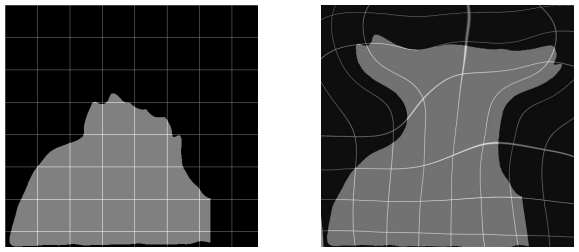
Figure 3. The dual mask feature pooling pipeline in our FFD network. Here only a single level of convolutional maps is shown. Note that we use much finer grid partition than the standard RoI pooling.

3.2. Free-Form Deformation Transformer

Given the object-mask descriptor, our second network module predicts a 2D spatial transform to warp the initial mask onto the target object. As the mask can have arbitrary shapes, we adopt a rich family of spatial transforms, which is capable of representing any non-rigid warping in image, referred to as free-form deformation (FFD) [33].

The FFD defines a family of non-rigid spatial transformations based on a mesh of control points. By shifting the control points and interpolating the dense deformation based on B-splines [25], it provides a flexible tool to describe the non-rigid transformation between the mask and object. Figure 4 shows an example of the deformation process.

Formally, let Φ be a 2-D mesh of control points and $T : (x, y) \mapsto (x', y')$ be a pointwise transformation of any location (x, y) in target image F to the location (x', y') in



(a) Uniformly spaced control points

(b) Deformation by shifting the control points

Figure 4. Illustration of FFD defined on a binary mask. Left is the original mask with uniformly spaced control points; Right is the deformed mask with displaced control points.

the source image R. Given a mesh of control points $\phi_{i,j}$ with uniform spacing δ pixels, the non-rigid transformation T by B-spline functions is defined by

$$T_{(x,y)} = \sum_{l=0}^{3}\sum_{m=0}^{3} B_l(u)B_m(v)\phi_{i+l,j+m} \tag{1}$$

where $i = \lfloor x/\delta \rfloor - 1$, $j = \lfloor y/\delta \rfloor - 1$, $u = x/\delta - \lfloor x/\delta \rfloor$, $v = y/\delta - \lfloor y/\delta \rfloor$, and B_l represents the l-th basis function of cubic B-splines [25]:

$$B_0(u) = (1-u)^3/6, \quad B_1(u) = (3u^3 - 6u^2 + 4)/6$$
$$B_2(u) = (-3u^3 + 3u^2 + 3u + 1)/6, \quad B_3(u) = u^3/6$$

From Equation (1), we note that the B-spline based FFD is locally controlled as each control point $\phi_{i,j}$ affects only its $4\delta \times 4\delta$ neighborhood. This indicates that the FFD can describe highly local transformation, which is required for capturing the complex non-rigid deformations between the mask and object. Additionally, the degree of non-rigid deformations can be controlled by changing the resolution of the mesh of control points Φ. A larger spacing of control points allows modelling of global and coarse deformations, while a small spacing of control points allows modelling of local and fine-grained deformations.

By shifting the locations of the control points from the uniform grid $\phi_{i,j}$ to $\phi_{i,j} + \Delta\phi_{i,j}$, the B-spline based FFD generates a non-rigid transformation as follows:

$$T_{(x,y)} = \sum_{l=0}^{3}\sum_{m=0}^{3} B_l(u)B_m(v)(\phi_{i+l,j+m} + \Delta\phi_{i+l,j+m}) \tag{2}$$

In this work, we parameterize the FFD by the offsets of its control points $\{\Delta\phi_{i,j}\}$, and our second network module first regresses the control point offsets from the object-mask descriptor. To achieve scale-invariance, we normalize the offsets by the size of the initial mask. Our transform regressor module consists of 3 fully connected (fc) layers and its outputs are the offset vectors of every control point.

To obtain the warped mask, we follow a similar strategy as the Spatial Transformer Network [18]. Given the predicted offsets, we compute the dense transformation according to Equation (2). The transform T then generates a sampling grid G, which is a set of points where the initial mask should be sampled in order to produce the warped mask. Next, a bilinear sampling layer takes the sampling grid and the initial mask as inputs and produces the final warped mask. We refer the reader to [18] for more details about the bilinear sampling process, especially the back propagation of the loss through the sampling mechanism.

We note that for the FFD transformer network, the gradients of loss L with respect to $\Delta\phi_{i,j}$ can be computed by:

$$\begin{aligned}\frac{\partial L}{\partial \Delta\phi_{i,j}} &= \frac{\partial L}{\partial G}\cdot\frac{\partial G}{\partial \Delta\phi_{i,j}} \\ &= \frac{\partial L}{\partial G}\cdot\sum_{l=0}^{3}\sum_{m=0}^{3} B_l(u)B_m(v)\end{aligned} \tag{3}$$

where $\frac{\partial L}{\partial G}$ is the gradients of loss L with respect to the sampling grid G. This equation shows that given $\frac{\partial L}{\partial G}$, $\frac{\partial L}{\partial \Delta\phi_{i,j}}$ can be computed efficiently by convolution, with the filter weights as $B_l(u)B_m(v)$ and the stride being the spacing of control points δ. The differentiable property of the FFD transformer network allows loss gradients to flow back to the feature maps, which enables us to train the network in an end-to-end fashion.

3.3. Network Details and Training

Network Architecture. We use ResNet-101 [13] pre-trained on the ImageNet dataset [9] for image classification task as our base net to learn the feature representation. We remove all the layers on top of *res4b22_bracnch2a_relu*, as the output from these layers are not used in our system.

For the mask feature pooling, we select a 30×30 grid for computing the feature on the feature maps output from layer *conv1_relu (64 channels)* and layer *res2c_relu (256 channels)*, and a 20×20 grid for layer *res3b3_branch2a_relu (128 channels)* and layer *res4b22_bracnch2a_relu (256 channels)*. We discover that the high resolution of the pooling grid is important for training the network, as the non-rigid transformations to be learned by the network are highly complex, which need quite discriminative and fine features to represent them.

As the mask features pooled from different layers are of different spatial sizes and channel depths, we first fully connect each set of them into a low dimensional output of size 128 and then concatenate all the outputs together to form a feature vector of size 512. Next are another two fc layers for predicting the offsets of the control points. The weight sizes of these two *fc* layers are 512×512 and

$512 \times 2 \times 13 \times 13$ respectively, which means the resolution of the mesh of control points is 13×13 in our experiments. All the fc layers except the last one are followed by a ReLU layer and a dropout layer.

Training Examples. To build the set of training examples, we select those segment proposals who have an IoU with the ground truth greater than 0.5 as the training samples. Specifically, for a qualified segment proposal, we crop it with a larger box whose size is $1.6\times$ to the tight box that encloses the segment in terms of height and width, so that the cropped region can cover more of the ground-truth object mask. We also use this large box to crop corresponding ground-truth mask as this region's ground truth.

Learning Details. We train the network to simply minimize the L_2 loss between the candidate's mask and the ground truth's, which we find is robust and effective. We adopt an image-centric training policy [11]. In our system, the mini-batch size is 1 and for every image we randomly sampled 128 training segments. Except the ResNet layers, the extra fc layers are initialized randomly from Gaussian distribution. We train the network for 10 epochs using a momentum of 0.9 and weight decay of 0.002. The learning rate we use for each epoch gradually decreases from 10^{-4} to 10^{-7} evenly in the log space.

4. Experiments

We apply our FFD network to the segment proposal refinement task in which we intend to improve a set of object segment proposals generated from state-of-the-art methods. We evaluate the performance of our approach on three public datasets: Cityscapes [4], PASCAL VOC 2012 [10, 12] and MSCOCO [26],

4.1. Evaluation Metrics and Protocols

For performance evaluation, we compute the average recall (AR) [15] between IoU 0.5 and 0.95 for a fixed number of segment proposals. The AR metric describes the overall quality of object proposals and has been shown to correlate highly with the detection accuracy in [15]. Additionally, we report the recall versus IoU threshold for different number of proposals.

On the Cityscapes dataset, we split the training set into two subsets: one for training (2,614 images) and the other for validation (361 images taken at Tubingen, Ulm and Zurich). We report our results on the original validation set (500 images) for evaluation as the test server does not provide the evaluation for segment proposals. For the PASCAL VOC dataset, we train our network on the training set (5,623 images) and evaluate on the validation set (5,732 images). We use the instance-level segmentation annotations from [12]. For the MSCOCO dataset, we follow the same protocol as in SharpMask [29].

Method	AR@10	AR@100	AR@1000
MNC-r	0.052	0.131	0.180
MNC	0.041	0.102	0.136
SharpMask-r	**0.103**	**0.175**	**0.215**
SharpMask	0.085	0.141	0.171
DeepMask	0.082	0.138	0.164
MCG	0.016	0.046	0.091

Table 1. Quantitative results of segment proposal refinement on **Cityscapes**: AR at different number of proposals (10, 100 and 1,000).

To demonstrate the generality of our method, we conduct our Cityscapes and PASCAL VOC experiments with two different sets of initial object segments, which are generated from the state-of-the-art segment proposal generation methods, SharpMask [29] and MNC [8], respectively. For each type of initial segments, we train our model from scratch with a set of selected segment proposals from the initial pool. However, when training the network with SharpMask proposals on the PASCAL VOC, we find that it is difficult for the network to converge, which might be due to much fewer training segments and their sparse spatial distribution. So for that case, we fine-tune the network that has been trained for MNC proposals on the PASCAL VOC. On the MSCOCO, we only report our experiment with the SharpMask proposals.

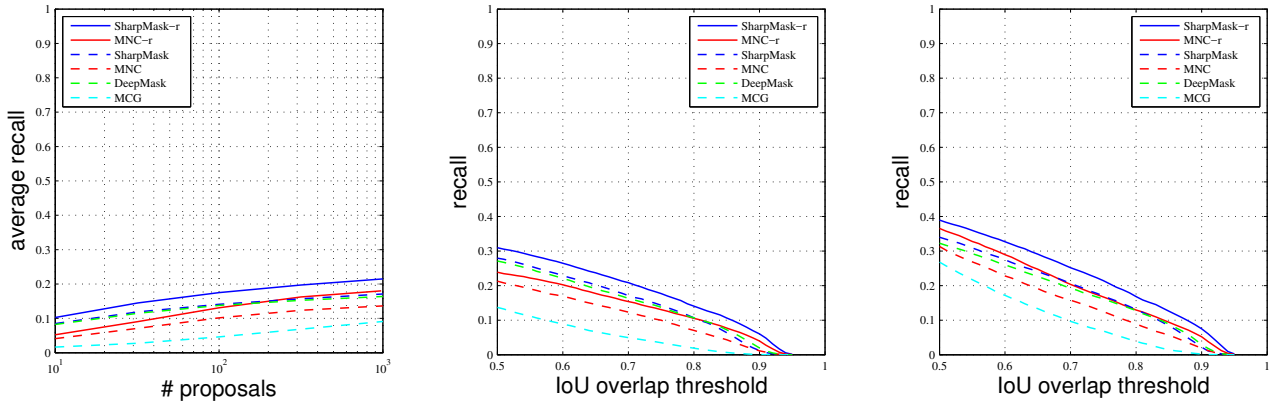
4.2. Results

4.2.1 Cityscapes

In Fgure 5(a), we first report the AR performances of the refined segment proposals (**MNC-r** and **SharpMask-r**), and compare the performance of our method against the original proposal methods as well as other baselines (DeepMask [28] and MCG [30]) on the Cityscapes. We can see that our FFD network can improve the quality of the initial segment proposals by a significant margin. Specifically, with 1,000 proposals, our FFD network increases the AR of MNC and SharpMask from 0.136 to 0.180 (32.4% improvement) and from 0.171 to 0.215 (25.7% improvement), respectively. More detailed quantitative results are shown in Table 1.

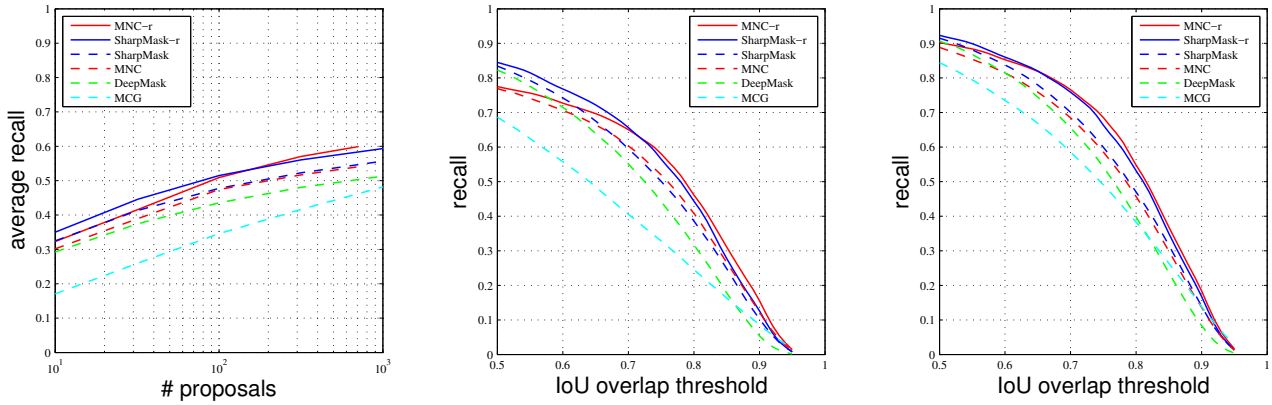
Figure 5(b) and 5(c) show the recall versus IoU threshold with 100 and 1,000 proposals respectively. They demonstrate that our method can improve the proposals with different segmentation qualities on the Cityscapes dataset.

We further report some qualitative results in Figure 8. These examples show that our FFD network is capable of predicting non-rigid deformations for both local and global warping, and produces better segmentation for the target objects with different scales and classes.



(a) Average recall (b) Recall @ 100 proposals (c) Recall @ 1000 proposals

Figure 5. Segment proposal refinement results on **Cityscapes**: (a) AR vs. number of proposals; (b) and (c) recall vs. IoU threshold with different number of proposals.



(a) Average recall (b) Recall @ 100 proposals (c) Recall @ 1000 proposals

Figure 6. Segment proposal refinement results on **PASCAL VOC**: (a) AR vs. number of proposals; (b) and (c) recall vs. IoU threshold with different number of proposals.

Method	AR@10	AR@100	AR@1000
MNC-r	0.323	0.509	**0.599**
MNC	0.302	0.474	0.541
SharpMask-r	**0.350**	**0.515**	0.594
SharpMask	0.325	0.477	0.557
DeepMask	0.293	0.436	0.513
MCG	0.171	0.346	0.481

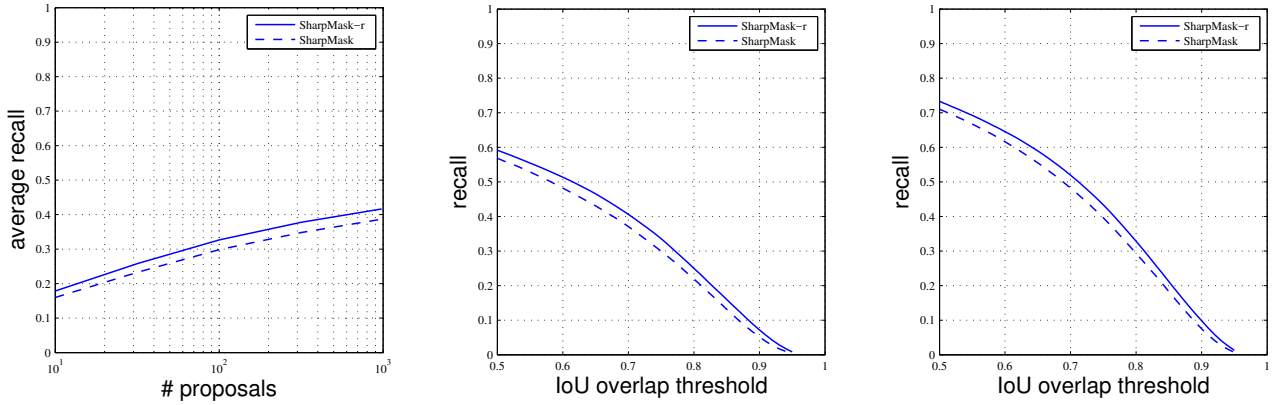
Table 2. Quantitative results of segment proposal refinement on **PASCAL VOC** : AR at different number of proposals (10, 100 and 1,000).

4.2.2 PACAL VOC

We compare the AR performances of our method with other baselines on the PASCAL VOC in Figure 6(a). It can be seen that our FFD network further improves the quality of the segment proposals generated from both state-of-the-art approaches. In particular, with 1,000 proposals, our FFD network increases the AR of MNC and SharpMask by 10.52% (from 0.542 to 0.599) and 6.64% (from 0.557 to 0.594). More detailed quantitative results are shown in Table 2. This demonstrates that our approach generalizes well to other types of datasets.

Figure 6(b) and 6(c) show the recall versus IoU threshold with 100 and 1,000 proposals respectively. We can see that the refined proposals have better quality, as with high IoU thresholds, *e.g.* 0.7, 0.8 and 0.9, the refined proposals have much higher recall than the initial proposals.

Additionally, we include some qualitative examples in Figure 9, which show that our FFD network produces a wide range of refinements on object shapes. Some of these results have a slightly better boundary alignment, while the others achieve large improvements over the initial segments.



(a) Average recall (b) Recall @ 100 proposals (c) Recall @ 1000 proposals

Figure 7. Segment proposal refinement results on **MSCOCO**: (a) AR vs. number of proposals; (b) and (c) recall vs. IoU threshold with different number of proposals.

Method	AR@10	AR@100	AR@1000
SharpMask-r	**0.179**	**0.327**	**0.416**
SharpMask	0.160	0.298	0.387

Table 3. Quantitative results of segment proposal refinement on **MSCOCO** : AR at different number of proposals (10, 100 and 1,000).

IoU Interval	[0.5, 0.6)	[0.6, 0.7)	[0.7, 0.8)	[0.8, 0.9)
mean PGIoU	0.548	0.648	0.749	0.849
mean RGIoU	0.665	0.737	0.796	0.861
Gain	21.35%	13.7%	6.28%	1.41%

Table 4. Statistics for the improvements in the quality of MNC proposals with different initial IoU scores on **PASCAL VOC**. The 'mean PGIoU' denotes the average IoU score of the original proposals, while the 'mean RGIoU' is the average IoU score of the warped proposals.

4.2.3 MSCOCO

Figure 7(a) demonstrates the AR improvement for SharpMask proposals on the MSCOCO, while Tabel 3 shows more detailed quantitative results. With 1,000 proposals, our approach improve the AR by 7.49% (from 0.387 to 0.416). Figure 7(b) and 7(c) show the recall versus IoU threshold with 100 and 1,000 proposals respectively. It is clear that our method can achieve consistent improvements on MSCOCO, and this demonstrates that our approach is able to scale up to a larger number of object classes.

4.3. Ablation Study

In order to get more insight into our FFD network, we analyze the IoU improvements for MNC segment proposals with different IoU scores on the PASCAL VOC. We divide the initial proposal set into 4 groups, which correspond to the IoU intervals of [0.5, 0.6), [0.6, 0.7), [0.7, 0.8) and [0.8, 0.9). We then compute the mean IoU improvements for each group after aligning the initial masks to their object regions through the FFD network. The results are shown in Table 4, from which we can see that our FFD network is more effective in modeling relatively coarse transformations than capturing fine-level local deformations. Encoding such fine-level misalignment between the object mask and its groundtruth might require richer features and denser control points.

We have also tried to learn a backward transformation that transforms the groundtruth mask to the proposal mask. Interestingly, we discover that the backward transformation is much easier to learn, which can be explored further in future work.

5. Conclusion

In this paper, we address the problem of object-mask registration and aim to align a shape mask to a target object instance. To this end, we take a transformation based approach that predicts a 2D non-rigid spatial transform and warps the shape mask onto the target object. In particular, we propose a deep spatial transformer network that learns free-form deformations (FFDs) to non-rigidly warp the shape mask based on a multi-level dual mask feature pooling strategy. Our network is fully differentiable and thus can be trained in an end-to-end manner. We evaluate our FFD network on the task of refining a set of object segment proposals, and our approach achieves the state-of-the-art performance on the Cityscapes, the PASCAL VOC and the MSCOCO datasets.

Acknowledgment Data61 is part of the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which is the federal government agency for scientific research in Australia. The Tesla K40 used for this research was donated by the NVIDIA Corporation.

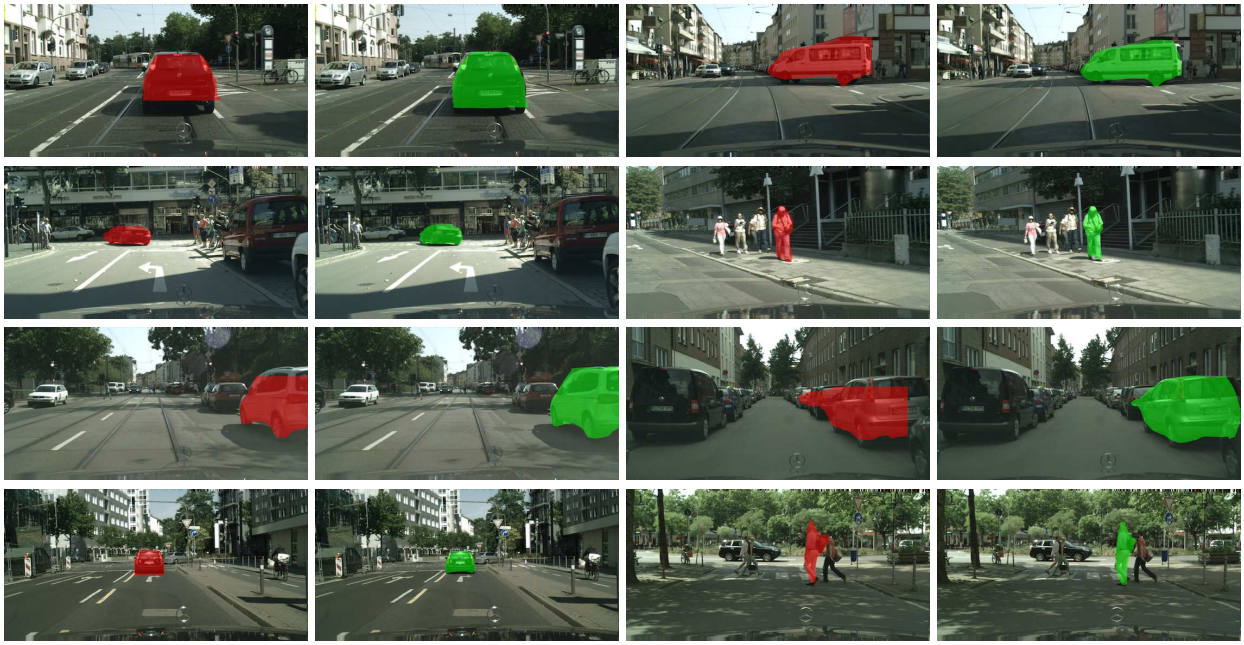


Figure 8. Qualitative results on **Cityscapes**. Red: original object mask. Green: aligned mask.

Figure 9. Qualitative results on **PASCAl VOC**. Red: original object mask. Green: aligned mask.

References

[1] J. Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 2007. 2

[2] J. Carreira, R. Caseiro, J. P. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. 2

[3] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *TPAMI*, 2012. 2

[4] 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 scene understanding. In *CVPR*, 2016. 5

[5] D. Cremers, F. Tischhäuser, J. Weickert, and C. Schnörr. Diffusion snakes: Introducing statistical shape knowledge into the mumford-shah functional. *IJCV*, 2002. 1, 2

[6] W. R. Crum, T. Hartkens, and D. Hill. Non-rigid image registration: theory and practice. *The British Journal of Radiology*, 2014. 2

[7] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 2015. 3

[8] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016. 2, 5

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 5

[11] R. Girshick. Fast R-CNN. In *ICCV*, 2015. 3, 5

[12] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 5

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[14] X. He and S. Gould. An exemplar-based crf for multi-instance object segmentation. In *CVPR*, 2014. 1, 2

[15] J. Hosang, R. Benenson, P. Dollr, and B. Schiele. What makes for effective detection proposals? *TPAMI*, 2015. 5

[16] X. Huang, N. Paragios, and D. N. Metaxas. Shape registration in implicit spaces using information theory and free form deformations. *TPAMI*, 2006. 2

[17] A. Humayun, F. Li, and J. M. Rehg. Rigor: Reusing inference in graph cuts for generating object regions. In *CVPR*, 2014. 2

[18] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 2, 4

[19] V. Jampani, R. Gadde, and P. V. Gehler. Video propagation networks. In *CVPR*, 2017. 1

[20] A. Kanazawa, D. W. Jacobs, and M. Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *CVPR*, 2016. 2

[21] J. Kim and K. Grauman. Shape sharing for object segmentation. In *ECCV*, 2012. 2

[22] P. Krähenbühl and V. Koltun. Geodesic object proposals. In *ECCV*. 2014. 2

[23] D. Kuettel and V. Ferrari. Figure-ground segmentation by transferring window masks. In *CVPR*, 2012. 1, 2

[24] D. Kuettel, M. Guillaumin, and V. Ferrari. Segmentation propagation in imagenet. In *ECCV*, 2012. 2

[25] S. Lee, G. Wolberg, and S. Y. Shin. Scattered data interpolation with multilevel b-splines. *IEEE transactions on visualization and computer graphics*, 1997. 2, 3, 4

[26] 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*, 2014. 5

[27] C. R. Meyer, J. L. Boes, B. Kim, P. H. Bland, K. R. Zasadny, P. V. Kison, K. Koral, K. A. Frey, and R. L. Wahl. Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin-plate spline warped geometric deformations. *Medical image analysis*, 1997. 1, 2

[28] P. O. Pinheiro, R. Collobert, and P. Dollar. Learning to segment object candidates. In *NIPS*, 2015. 2, 5

[29] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollar. Learning to refine object segments. In *ECCV*, 2016. 2, 5

[30] J. Ponttuset, P. Arbelaez, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *TPAMI*, 2015. 2, 5

[31] A. I. Popa and C. Sminchisescu. Parametric image segmentation of humans with structural shape priors. In *ACCV*, 2016. 1, 2

[32] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2

[33] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 1999. 1, 2, 3

[34] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013. 1, 2

[35] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 2013. 2

[36] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In *ECCV*, 2016. 2

[37] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2016. 2

[38] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*. 2014. 2

[39] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and vision computing*, 2003. 1, 2