Segmentation with Fovea-magnifying Warped Crops

William Whitehead
Department of Electrical and Computer Engineering
University of California, Los Angeles
williamow@ucla.edu

Abstract

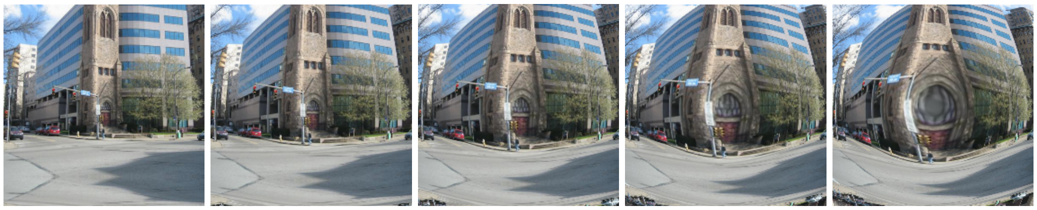
When deep neural networks are used to segment images, they are trained to provide different classifications for minimally translated inputs. Soft transitions across semantic edges in published results suggest that learning to provide different classifications at boundaries is difficult for neural networks, which leads us to speculate that the translational invariance of convolutional neural networks is not ideal for semantic segmentation. This speculation compels us to individually classify each pixel in a scene using warped crops as the classifier input. Here warping is a transformation of a regular crop (rectangular subset of an image) that magnifies the fovea (center that is being classified) and pushes the rest of the scene to the periphery. Warping effectively decreases translational invariance by making the crops of adjacent pixels appear more distant. We segment using warped crops and find that warped crops are more accurate than regular crops in all cases, sometimes dramatically. Class prediction heatmaps and a translation sensitivity analysis show that reducing translational invariance is an important aspect of warped crops. Both translational sensitivity and warping show promise as future research subjects.

1 Introduction

Background Semantic scene segmentation entails labeling every pixel in an image by its semantic meaning. Doing so requires both finding the context of a scene (such as identifying a cyclist on a road) and identifying what each pixel belongs to (the cyclist or the road). Accuracy records for semantic segmentation tasks are held by fully-convolutional network (FCN) variants (7; 24; 26; 28). These top performers tend to improve results by building context awareness into their network architectures. While they directly output a dense set of semantic predictions for each pixel in an image, their function can still be thought of as classification. The simplest FCNs are identical to applying a convolutional neural network (CNN) classifier to a translating input (11; 17); the task of the network becomes identifying the center of a crop (4). This task is in direct contrast to the expectation that CNNs produce translationally invariant features; consider the edge of a segmentation region, where the neural network is asked to produce different classifications for nearly identical inputs. Examples of soft-edge class prediction heatmaps (6; 8), use of CRFs to compensate for said heatmaps (6; 8), and missed edges in segmentations (17; 28) suggest that translational invariance can indeed be problematic for segmentation. Note that some work (3) has shown that CNNs already show low translational invariance; we interpret their observed variance as noise, whereas we are interested in the variance of useful information.

Introducing warped crops Under the assumption that the translational invariance of deep neural networks is problematic for segmentation tasks, we speculate that making it easier to differentiate adjacent pixels will have a positive effect on segmentation performance. Instead of focusing on the design of the convolutional network, crop warping changes the image input to the network such that

Preprint. Under review.



A=1, B=0 A=0.75, B=0.25 A=0.5, B=0.5 A=0.25, B=0.75 A=0, B=1

Figure 1: Crops at the left are unwarped, and gradually become more warped towards the right. A and B parameters describe the warping, discussed later in section 2.1.

adjacent pixels look significantly different from each other. This property can only be achieved in the context of pixel-classifying CNNs. The warping we study here is analogous to cortical magnification, a phenomena of natural vision that has been studied extensively (5; 20). The effect is radial, spanning outward from the center of the image crop. The center is magnified; shapes near the center of the crop appear larger than physically identical shapes near the edge. A large fraction of pixels in the crop are focused on one spot in a scene, but unlike reducing the field of view of the crop, the same contextual information is still squeezed into the periphery of the crop. We will sometimes refer to crops that are not warped as being flat, in the sense that their spatial representation is uniform.

1.1 Potential benefits of warped crops

Decreasing translational invariance With the foveal details emphasized by warped crops, the input to the CNN will change more dramatically as the crop's focus point shifts across the scene, seen in figure 2. This property most directly solves the problem of translational invariance in section 1. Without warping, most of an image does not change under translation with objects having the same shape, size, and position relative to each other. In contrast, when a crop is warped different objects will appear enlarged as they near the point of focus. As the crop focus translates and different pixels come into the center of magnification, the neural network's response should also change quickly enabling more precise differentiation between pixels. In section 4, we perform a sensitivity analysis to check if crop warping actually decreases translational invariance and makes the network's output change faster.

Similarity to cortical magnification Cortical magnification refers to the foveal region of the eye having a higher density of photoreceptors than the periphery (20). Mimicking cortical magnification seemed desirable since it is easy to identify objects and edges at the focus of vision where photoreceptor density is high. Peripheral information is important but does not need to have great detail (23).

Weighting central pixels While a regular image contains both the local and contextual information needed for accurate segmentation, there is little evidence that CNNs are capable of considering the local area being segmented differently from the scene as a whole. Warped crops take the guesswork out of the distinction by making the local region in the center obviously different than the rest of the image. Local (center, foveal) pixels naturally affect a greater fraction of both the CNN input and output spaces since warping has magnified them.

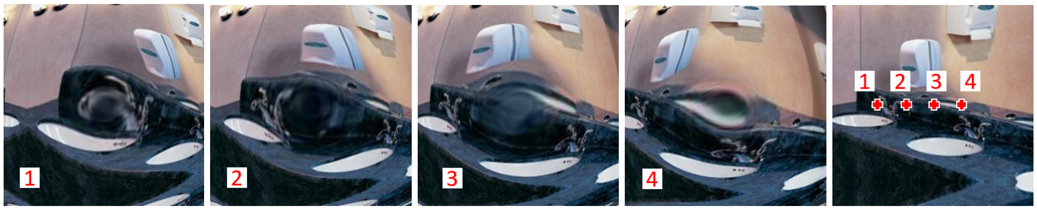


Figure 2: Example of warped crops as the crop focus is translated across an image; rightmost frame shows the path of the crop focus.

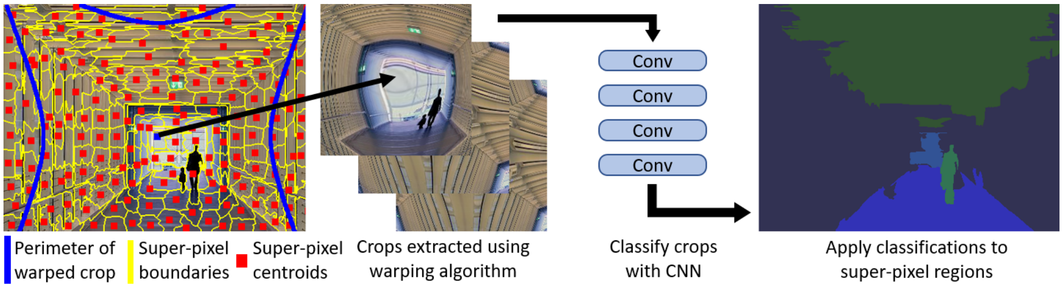


Figure 3: Process flow for segmenting with crop warping.

Scale invariance As a bonus effect, warped crops may beneficially promote scale invariance at the same time that it decreases translational invariance. Consider that a significant portion of the source scene is represented by a smaller sliver at the perimeter of a warped crop. Thus an increase or decrease of the field of view is largely absorbed into the periphery of the crop, leaving the magnified center relatively unchanged.

1.2 Related work

Translational invariance in other words While this paper primarily questions segmentation through the lens of translational invariance, the fundamental problem being solved is one that has attracted numerous solutions already. The difficulties of segmentation are often framed as the need to include both contextual and local information. As we translate across an image we want both some constant contextual information and rapidly changing local information. This understanding has motivated many modifications to the basic structure of an FCN. Early FCN developments such as connections that skipped from shallower layers towards the prediction layer are justified by this thinking (17; 19). The idea was that deeper layers would contain more contextual, high level information and that the skip connections would bring more detailed information deeper into the network, so that the final classification layers would have both types of information available. Recent successful FCN designs (24; 26), are also driven by this idea of context vs. detail.

Fovea inspired This paper is also not the first to treat the fovea of an image different from the periphery. Several works have proposed splitting scenes down two pathways, one ‘peripheral’ and one ‘foveal’, where the foveal pathway is treated to processing at a higher resolution and smaller field of view (16; 25). Motivation for this seems to come from the idea that some parts of a scene simply exist at a different scale than other parts, and thus deserve to be magnified. Additional related work has modified crops before being processed by a neural network just as we do, although not in the context of segmentation. In particular, (22; 23) have tried to understand the relative importance of the foveal and peripheral regions by darkening one region or the other, transforming images into a log-polar form, and covering either the fovea or periphery of images. The work most mechanically similar to this paper has used blurring to emphasize a central image region in the context of replicating human visual attention (2).

2 Methods

Segmentation pipeline Crop warping fits into a relatively straightforward segmentation procedure where a neural network classifies one point in a scene at a time, such as in (4). The scene is divided into numerous super-pixel regions which are then each classified. We divide scenes into super-pixels using either SLIC or a simple grid. In the extreme case, all pixels in an image are segmented individually. To classify a region, a point within the region (the centroid) is selected as a focus point or center of another image, the crop. A crop centered on that point is classified by a CNN; the CNN classification is then applied as the segmentation result for all scene pixels included in that region. What we are most interested in here is the step of creating crops and how this step can affect segmentation performance.

2.1 Formulating Warped Crops

Spatial mapping It is easiest to model the crop warping in polar coordinates. We consider r_c the radius of a point in the crop, measured from the center of the crop, and r_s the radius of a point in the scene, measured from the focus point in the scene. Creating a crop is done by transferring information from the scene to the crop as in equation 1 where C is the crop image and S is the scene image.

$$C[r_c, \theta] = S[r_s, \theta] \tag{1}$$

Field of view scaling To complete the spatial mapping r_s needs to be related to r_c. First the warping characteristics need to be isolated from the scale of the pixel maps. Both the characteristic warping relation, f_c, and its input are normalized to the range $[0, 1]$. Values s_s and s_c represent the pixel size of the scene and the crop, respectively; they are calculated as the mean of the image's width and height. An extra variable, fov, is introduced to adjust the field of view of the crop and unless otherwise noted is set to 0.8.

$$r_s = f_c({}^{r_c}\!/_{s_c}) * fov * s_s \tag{2}$$

Warp characterization All that is left now is the definition of f_c, which defines the shape of a crop. It could take any form, but a reasonable f_c is a continuous, monotonically increasing function from $f_c(0) = 0$ to $f_c(1) = 1$. The simplest function satisfying these requirements is $f_c(r_o) = r_o$. This represents a regular, unwarped crop. The family of warping functions used for the rest of the paper is equation 3 with the limitation that $A + B = 1$.

$$f_c(r_o) = Ar_o + Br_o^2 \tag{3}$$

It is important to interpret the derivative of the warping function. The derivative is proportional to image pixels per crop pixel. Thus when the derivative is small, such as at small radii, space in the scene is spread out in the crop and when the derivative is large, the scene is compressed into the crop.

Several other features are added to the cropping algorithm. When a pixel in a crop maps to a pixel in the scene that is outside of the scene boundary, the scene is reflected to provide data for that pixel. This was found to yield better results than simply leaving the areas outside the crop a uniform or random value. Rotation and horizontal stretching are also added into the mapping from crop pixel to scene pixel, which simply serves as native data augmentation during training.

2.2 Convolutional neural network specification

Network design In all cases pretrained mobilenet v2 (21) and resnet50 v2 (13) Tensorflow networks (1) are used. They are modified for segmentation datasets by replacing the final pooling and fully connected layers with a series of three fully-connected layers. This choice of ending was found to perform better than several other layer endings in one set of training trials. The native input dimensions of 224x224 pixels was kept.

Training hyperparameters Each classification network was trained on 40k batches of 75 crops each. The crops in each batch were selected randomly from the dataset; for each crop, first an image was taken from a queue, and then the crop center was chosen at random from the entire space of the image. The extraction of crops from scenes was done concurrently with training by several parallel worker processes. Momentum optimization with momentum 0.99 was used; the learning schedule is the same power decay used in (8) with an initial rate of 0.001. Loss was softmax with cross entropy. These hyperparameters were chosen by a few iterations of hand adjustment. The memory-saving gradient computation from (9) was helpful during training.

Reference FCN We also trained an FCN network based on mobilenet v2 to check against the pixel classification method. Ideally the classifier networks with unwarped crops would yield a similar segmentation accuracy as this FCN network. The FCN has the same final computation layers as the classifiers. The FCN was trained on 20k batches of 7 images. The input was scaled to mimic the field of view seen by the crop-classifying networks. The FCN was trained and evaluated with an output stride of four, which was then bilinearly upsampled.

Code All experiments were run with TensorFlow (1). Code can be downloaded from https://drive.google.com/file/d/1eCu3eX2pFs-CHnCc5eJ7zg3GJvfQ8-n6/view.

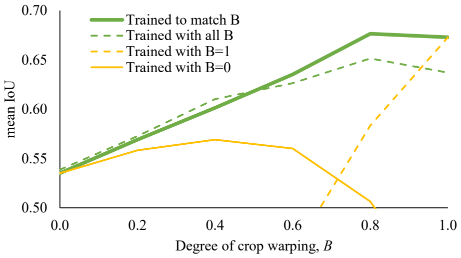


Figure 4: mIoU of Cityscapes validation data, as a function of the warping parameter B.

Table 1: Results on Cityscapes validation data.

Method	CNN base	mIoU
FCN-8s (10)	VGG 16	0.653
By crops B=0 (ours)	**Mobilenet v2**	**0.531**
By crops B=1 (ours)	**Mobilenet v2**	**0.686**
DeepLab (8)	Resnet-101	0.714
DPC (7)	Xception	0.809

3 Segmentation results

The success of crop warping depends on the setting it is evaluated in. We will see that warped crops only slightly improve benchmark scores for the ADE20K dataset (27) but drastically improve benchmark scores on the Cityscapes dataset (10). Qualitatively, the class prediction heatmaps for both datasets show better boundary identification. The different degrees of success in the benchmark scores likely depends on the the challenges of each dataset. The ADE20K dataset has 150 classes and achieving a good benchmark score does not require accurately segmenting many small features. In contrast, Cityscapes only requires identification of 19 classes with many small regions in its images - exactly the sort of issues that crop warping should perform best at. All accuracy data points are an evaluation of a single trained network, and the split between training and validation data is determined by the dataset authors.

3.1 Cityscapes dataset benchmark scores

The Cityscapes dataset primarily uses the mean intersection over union (mIoU) metric (10). The mIoU calculation for Cityscapes typically looks at 19 object and region classes. Since the objects are often small, accurate segmentation for a whole image can only be achieved by classifying a huge number of small super-pixels. Because of this computational difficulty, sampling was used to estimate the mIoU Cityscapes scores instead of fully segmenting the entire validation dataset. Pixels were sampled at random from the validation dataset as during training, and the presented mIoU scores were calculated from the results of these samples. Each mIoU score represents 37,500 sample points. Cityscapes mIoU are shown in table 1 alongside results from state-of-the-art FCNs. For this dataset, using warped crops made a significant improvement over using flat crops.

Figure 4 shows the dependence of the cityscapes mIoU on the warping parameter B, with different curves representing networks with different training conditions. When the network is trained and evaluated on the same B, maximum accuracy is reached by $B = 0.8$ and then levels off. Figure 4 also includes curves for networks that were trained with $B = 0$ or $B = 1$, but then evaluated on the full range of B values. Unsurprisingly, the networks do not perform well when evaluated with warping that is different than what they were trained on. A network that has been trained to identify crops with all values of B is able to do so, but does not perform as well as the networks trained and evaluated with a single B value.

3.2 ADE20K dataset benchmark scores

The common metrics for evaluating the ADE20K dataset, specified in (27), are pixel accuracy and the mIoU between predictions and ground truths. On both these metrics warped crops give better results than unwarped crops, although only by small margins. The pixel accuracies and mIoU scores for validation datasets are in the typical range for the ADE20K dataset but fall short of state-of-the-art scores, shown in table 2. For the crop classifier segmentation accuracies reported in table 2, the entire ADE20K validation dataset of 2000 images were segmented by classifying super-pixels as outlined in section 2. The segmentation scores depend on the super pixel segmentations, improving as the number of segmentations increases. For evaluation on ADE20K, roughly 1000 super-pixels per image

Table 2: Performance on ADE20K validation data

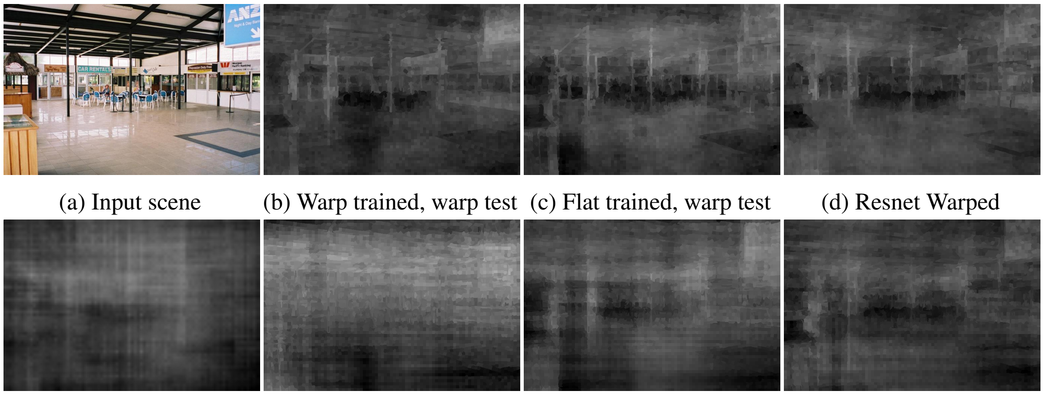
		Unwarped Acc. $B = 0$		Warped Acc. $B = 1$	
Method	CNN base	Pixel acc.	mean IoU	Pixel acc.	mean IOU
By crops (ours)	**mobilenet v2**	**0.741**	**0.292**	**0.753**	**0.312**
By crops (ours)	Resnet v2 50	0.734	0.285	0.749	0.308
FCN (ours)	mobilenet v2	0.726	0.258	N/A	N/A
FCN-8s (27)	VGG 16	0.713	0.294	N/A	N/A
FCN (24)	Resnet 50	0.746	0.344	N/A	N/A
EncNet (24)	Resnet 50	0.797	0.411	N/A	N/A

were used, which produces satisfactory results but was too slow for real-time use at 2.5 seconds per image using mobilenet v2, running on an Nvidia GTX 1080 GPU.

3.3 Class prediction heatmaps

These class prediction heatmaps are from one-hot predictions before the softmax operation. Brighter regions indicate a higher probability of the given object class. With warping the heatmaps show sharper edges between regions. Figure 5 shows heatmaps for the ADE20K "column, pillar" class, created by various neural networks for the same image. Both our FCN segmentation and flat crop segmentations have a difficult time localizing the pillars since they are small in the image. The heatmaps for these methods do not show sharp delineations where the columns are, presumably because a crop centered on the column does not look much different than a crop centered adjacent to the column. When warped crops are used, the heatmap for columns becomes much more differentiated and the shape of the columns can be seen clearly. Heatmaps are also shown for networks trained on flat crops but evaluated on warped crops, and vice versa. Even the network trained to identify flat crops produces sharper segmentations when evaluated with warped crops, even though this setup results in worse scores. Heatmaps produced with a Resnet architecture show that the benefits of warped crops are seen across network designs, but not all to the same degree.

Improved delineation between regions is not limited to small objects such as the columns in figure 5. large regions such as buildings or sky also show sharper edges when warped crops are used. Figure 6 shows a sampling of heatmaps and segmentation results that show the delineating power of warped crops on both the ADE20K and Cityscapes datasets.



(a) Input scene (b) Warp trained, warp test (c) Flat trained, warp test (d) Resnet Warped

(e) Mobilenet FCN (f) Warp trained, flat test (g) Flat trained, flat test (h) Resnet flat

Figure 5: Column class heatmaps from warped and unwarped crops. The bottom row shows results without warping and heatmaps in the top row show heatmaps produced with warped crops. Mobilenet results are labeled for the warping used during network training (- trained) and at evaluation (- test).

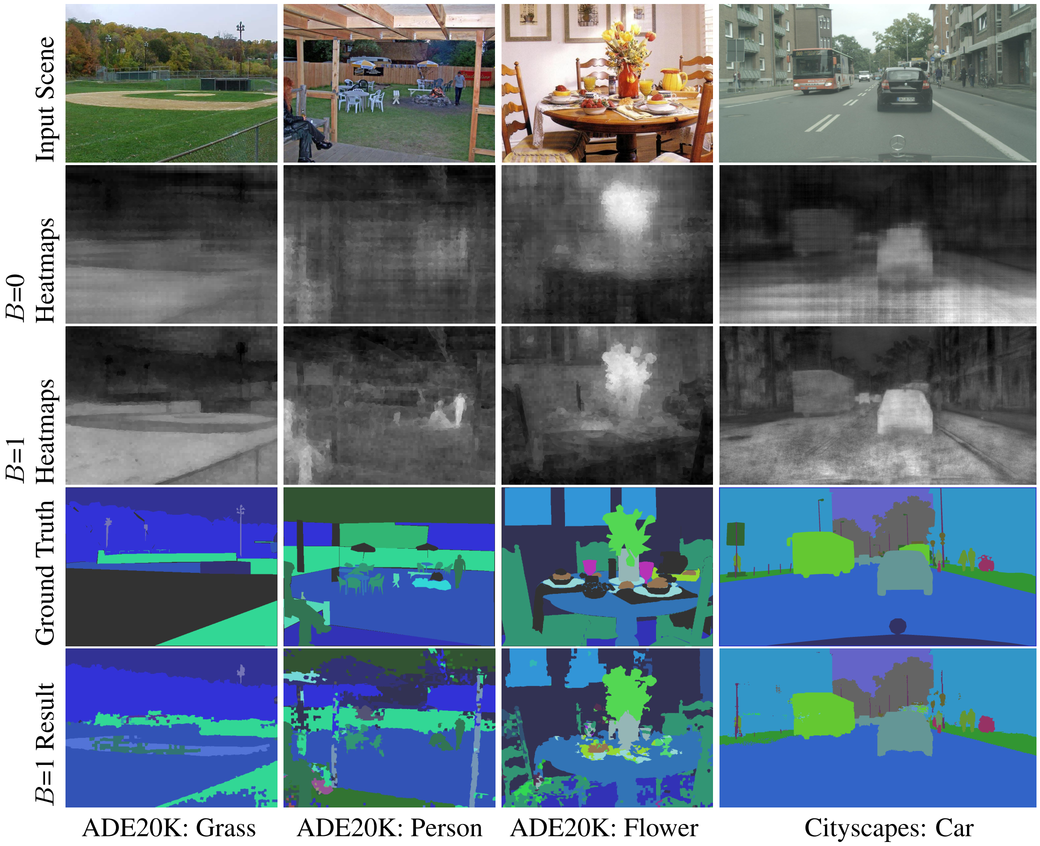


Figure 6: Select heatmaps and segmentation results. Classes listed along the bottom denote which heatmap is shown.

4 Sensitivity analysis

Motivation There are a couple reasons we would like to perform a sensitivity analysis. The first reason is to check if translational invariance does decrease when we use warped crops. The second reason for this sensitivity analysis is to identify alternative quantitative descriptions of network behavior. Comparing networks across a range of alternative metrics would let us determine if crop warping is in some way equivalent to existing methods. If not we would be able to say that crop warping can be combined with other methods to yield a new state-of-the-art.

Sensitivity formulation To quantify the network sensitivity to transformations we evaluate the logits produced for two inputs and compare the resulting feature activations. The difference between the two sets of activations serves as a proxy for sensitivity. Near identical techniques have been introduced by (12; 15) to measure translational invariance. Others have performed similar analysis but have looked at classification accuracy as a metric instead of feature activations (3; 18). The differences are normalized by their expected value so that a sensitivity of one indicates that the crops being compared appear unrelated to each other to the CNN. The sensitivity calculation is shown in equation 4 where S is the sensitivity of a transformation T and f_{CNN} is the output of a convolutional neural network operating on the input image x.

$$S(T) = \left\langle \frac{f_{CNN}(x_o) - f_{CNN}(T(x_o))}{\langle f_{CNN}(x_m) - f_{CNN}(T(x_n))\rangle} \right\rangle \tag{4}$$

Translational sensitivity To further test the claim that decreasing translational invariance corresponds to better segmentations, we measure the sensitivity of network feature maps to translation. In the measure of sensitivity defined in equation 4 the transformation operation T_n is translation that

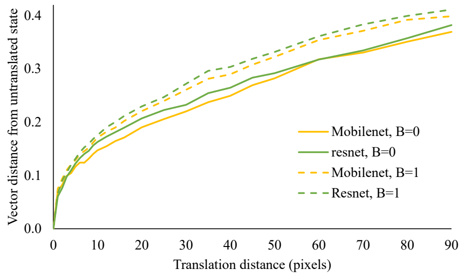


Figure 7: Sensitivity of the final feature layer to input translation.

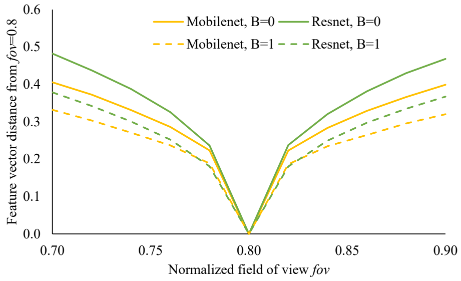


Figure 8: Sensitivity of the final feature layer to the scale of the input image.

moves the focus point of the crop n pixels away from its original location. Sensitivity S is plotted in figure 7 against translation distance for various networks with and without warped inputs. The averages in equation 4 are computed across 1500 sample pairs from the ADE20K validation data set. Higher sensitivity is less invariance. Both Mobilenet and Resnet behave as proposed, demonstrating greater sensitivity to translation with warped crops than with flat crops.

Scale Sensitivity It would be good if CNN evaluation of warped crops is less sensitive to changes in scale. In the proposed sensitivity measurement equation 4 the transformation T changes the field of view parameter fov from equation 2. The native field of view is taken to be 0.8, and T varies fov between 0.7 and 0.9. The results for various networks with various warping is shown in figure 8. As hoped, both Mobilenet and Resnet appear less sensitive to changes in scale when warped crops are used than when flat crops are used. Unlike translational invariance which can be easily identified in heatmaps, we do not have another reason to suspect that scale invariance is an important effect. The lack of secondary confirmation means that it is more difficult to verify scale sensitivity as a legitimate measure, but it also means that this analysis is powerful for showing a phenomena that is otherwise not an obvious mechanism of crop warping.

5 Discussion

Summary We have shown how to warp an image to magnify a specific spot while retaining peripheral information. Segmenting using crops warped in this manner produce more accurate segmentations than flat crops across two datasets and two CNN architectures. The improvement on the Cityscapes benchmark is particularly dramatic. Warped crops also produce heatmaps with sharp delineations between regions that we have not seen from any other convolution-based segmentation. A translational sensitivity analysis quantifies the refined edges in the heatmaps and confirms that the observed performance gains are related to the translational invariance of the system. However, this methodology both requires excessive computation and does not match state of the art benchmark scores, which makes it hard to promote segmenting with warped crops as a practical method. Instead, validating our ideas about spatial warping and translational invariance for semantic segmentation is the most valuable finding, discussed in detail below.

Future of warping While we warp space to magnify the center of a crop for classifying a single pixel, warping space in a scene as a whole is an avenue worth exploring. Light warping to magnify certain regions in an image could be used as a pre-processing step for FCN segmentation. For instance, the finest regions in the Cityscapes dataset are all long the horizon of the image, and this region could be magnified relative to other regions in the scene. This plot is reminiscent of the dual-pathway method in (16), but does not involve adjusting the neural architecture. The location and degree of warping could also be designed as adaptive, learnable parameters akin the innovation of spatial transformer networks (14). Our results suggest that these methods could yield improvements, and since the FCN architectures would not require modification, any improvement would be on top of the state-of-the-art with little computational cost.

Future of translational invariance Our results have also shown that accuracy can be improved by effectively reducing translational invariance. That there should be less translational invariance is a particularly intriguing result since most work related to translational invariance finds that more is better. The context information emphasized in (24) essentially increased translational invariance by providing identical features across the entire image, and also improved segmentation results. That leaves us with both decreasing and increasing translational invariance as methods of improving semantic segmentation. Rather than having contradictory statements, we hope that this is a window into a more sophisticated model of translational invariance and segmentation. For various FCN designs we would like to see how activations can be placed into a spectrum from translationally invariant to translationally sensitive.

Acknowledgments

This work was completed as part of the UCLA Internet Research Initiative (IRI) program. Thanks to Jeff Burke of the UCLA School of Theater, Film and Television for discussion and guidance.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Ana Filipa Almeida, Rui Figueiredo, Alexandre Bernardino, et al. Deep networks for human visual attention: A hybrid model using foveal vision. In *ROBOT 2017: Third Iberian Robotics Conference*, pages 117–128. Springer International Publishing, Cham, 2018.

[3] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *CoRR*, abs/1805.12177, 2018.

[4] Clemens-Alexander Brust, Sven Sickert, Marcel Simon, et al. Convolutional patch networks with spatial prior for road detection and urban scene understanding. In *VISAPP*. 2015.

[5] Marisa Carrasco and Karen S. Frieder. Cortical magnification neutralizes the eccentricity effect in visual search. *Vision research*, 37:63–82, 02 1997.

[6] Siddhartha Chandra and Iasonas Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *ECCV*. 2016.

[7] Liang-Chieh Chen, Maxwell Collins, Yukun Zhu, et al. Searching for efficient multi-scale architectures for dense image prediction. In *Advances in Neural Information Processing Systems 31*, pages 8699–8710. Curran Associates, Inc., 2018.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, April 2018.

[9] Tianqi Chen, Bing Xu, Chiyuan Zhang, et al. Training deep nets with sublinear memory cost. *CoRR*, abs/1604.06174, 2016.

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, et al. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[11] Alessandro Giusti, Dan C. Cireşan, Jonathan Masci, et al. Fast image scanning with deep max-pooling convolutional neural networks. In *2013 IEEE International Conference on Image Processing*, pages 4034–4038. Sep. 2013.

[12] Ian Goodfellow, Honglak Lee, Quoc V. Le, et al. Measuring invariances in deep networks. In *Advances in Neural Information Processing Systems 22*, pages 646–654. Curran Associates, Inc., 2009.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, pages 630–645. Springer International Publishing, Cham, 2016.

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems 28*, pages 2017–2025. Curran Associates, Inc., 2015.

[15] Eric Kauderer-Abrams. Quantifying translation-invariance in convolutional neural networks. *CoRR*, abs/1801.01450, 2018.

[16] Xin Li, Zequn Jie, Wei Wang, et al. Foveanet: Perspective-aware urban scene parsing. *CoRR*, abs/1708.02421, 2017.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440. June 2015.

[18] Erik Rodner, Marcel Simon, Robert B. Fisher, et al. Fine-grained recognition in the noisy wild: Sensitivity analysis of convolutional neural networks approaches. *CoRR*, abs/1610.06756, 2016.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, Cham, 2015.

[20] Jyrki. Rovamo and Veijo. Virsu. An estimation and application of the human cortical magnification factor. *Experimental Brain Research*, 37(3):495–510, Nov 1979.

[21] Mark B. Sandler, Andrew G. Howard, Menglong Zhu, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[22] Panqu Wang and Garrison W. Cottrell. Central and peripheral vision for scene recognition: A neurocomputational modeling exploration. *Journal of Vision*, 17(9), June 2017.

[23] Kevin Wu, Eric Wu, and Gabriel Kreiman. Learning scene gist with convolutional neural networks to improve object recognition. pages 1–6. 03 2018.

[24] Hang Zhang, Kristin Dana, Jianping Shi, et al. Context encoding for semantic segmentation. In *CVPR*. 2018.

[25] Xiaodan Zhang, Xinbo Gao, Wen Lu, et al. A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction. *CoRR*, abs/1812.07989, 2018.

[26] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, et al. Pyramid scene parsing network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[27] Bolei Zhou, Hang Zhao, Xavier Puig, et al. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[28] Yueqing Zhuang, Fan Yang, Li Tao, et al. Dense relation network: Learning consistent and context-aware representation for semantic image segmentation. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3698–3702. Oct 2018.