VRT-Net: Real-Time Scene Parsing via Variable Resolution Transform

Jogendra Nath Kundu* Gaurav Singh Rajput* R. Venkatesh Babu
Video Analytics Lab, CDS, Indian Institute of Science, Bangalore, India
jogendrak@iisc.ac.in, gauravrajput@iisc.ac.in, venky@iisc.ac.in

Abstract

Urban scene parsing is a basic requirement for various autonomous navigation systems especially in self-driving. Most of the available approaches employ generic image parsing architectures designed for segmentation of object focused scene captured in indoor setups. However, images captured in car-mounted cameras exhibit an extreme effect of perspective geometry, causing a significant scale disparity between near and farther objects. Recognizing this, we formalize a unique Variable Resolution Transform (VRT) technique motivated from the foveal magnification in human eye. Following this, we design a Fovea Estimation Network (FEN) which is trained to estimate a single most convenient fixation location along with the associated magnification factor, best suited for a given input image. The proposed framework is designed to enable its usage as a wrapper over the available real-time scene parsing models, thereby demonstrating a superior trade-off between speed and quality as compared to the prior state-of-the-arts.

1. Introduction

Urban scene parsing is a fundamental task which can enhance the ability of a wide range of consequent artificial decision making systems, more importantly in areas of autonomous navigation (e.g. self-driving). Scene parsing, also regarded as semantic segmentation, aims to assign category label to each individual pixel for a given RGB image, usually captured by car mounted cameras. Understanding images at pixel-level granularity [9, 8] provides a much richer representation to perform effective reasoning for various decision making tasks requiring varied level of local precision or global context information.

Acknowledging the significance of this problem a wide range of deep convolutional neural network (CNN) based solutions have been proposed [26, 5, 4], which has pushed the benchmark performances to a significant extent. However, in the race of improving segmentation accuracy such approaches employ sophisticated deep architectures [26, 5] which inevitably increases the computational cost to a significant extent in terms of both number of operations and parameter size. Recognizing this trade-off, there has been a diversion to realize efficient scene-parsing solution wherein

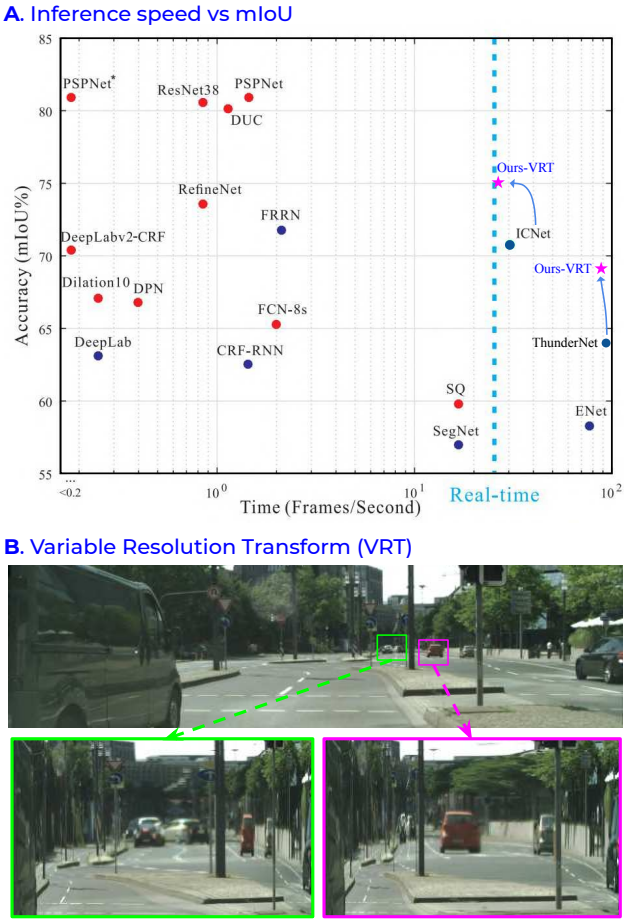

Figure 1. **A.** Real time segmentation mIoU scores on Cityscapes test set. Other Methods include: PSPNet [26], ResNet38 [22], DUC [21], RefineNet [11], FRRN [18], DeepLabv2-CRF [5], Dilation10 [24], DPN [14], FCN-8s [15], DeepLab [4], CRF-RNN [27], SQ [20], ENet [17], SegNet [1], ICNet [25], ThunderNet [23] and Ours. **B.** An illustration of the proposed transformation applied at two fovea regions to cater the issue of scale disparity.

*equal contribution

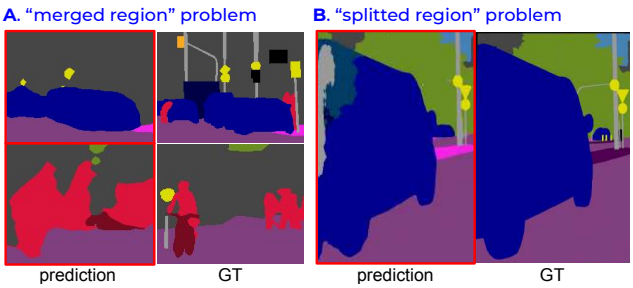

Figure 2. Illustration of the "merged region" and "splitted region" problem for farther objects and nearer objects respectively.

the objective is to improve the inference speed while attempting to sustain a comparable accuracy [25, 17]. This is a prime requirement towards practical deployment of scene-parsing models on embedded systems with limited compute power (see Figure 1A). Pushing the state-of-the-art performance with a good trade-off between speed and quality remains an open research problem (see Figure 1A).

One of the key differences between urban road scenes [6, 3] against indoor scene segmentation problems [28] can be attributed to the extreme effect of perspective geometry as a result of unobstructed vision in outdoor scenes. This effect is also common ego-centric video feeds while navigating in outdoor environments. The perspective projection from actual scene onto the image plane creates a significant scale disparity between near and farther objects, though they are of the same size in reality. Available scene parsing solutions do not employ explicit modifications to cater this problem. Similar architecture characteristics is used for both object-focused [28, 13] and urban road scene segmentation [6]. As a result such approaches yield poor performance on distant objects as compared to the nearer ones. We also notice that the nearby objects are usually distributed on the peripheral region specifically in the car mounted camera feeds [10]. This implies an uneven distribution of information content across the spatial map, with high density crucial information concentrated at the central region (see Figure 1B).

While closely analyzing the failure cases of parsing urban road scenes, we notice two crucial shortcomings (see Figure 2). **a)** The predictions of available parsing models exhibit "splitted region" problem, wherein segmentation of a large-scale object (a single segmentation class) situated in the image periphery splits into island-like regions of varied semantic categories. **b)** On the contrary, the predictions of parsing models exhibit "merged region" problem especially for farther objects concentrated in the central image region in absence of distinctive details. For example, a rider on vehicle (bicycle or motorcycle) is entirely segmented as either a rider or the vehicle causing parsing errors (see Figure 2A).

Our contributions. Motivated by the above discussion, we propose a novel scene parsing solution based on a carefully devised variable resolution transform (VRT) formalization. We aim to undo the effect of perspective geometry, which is the root cause behind both the failure scenarios discussed above (i.e. "splitted region" and "merged region"). The proposed variable resolution transform aims to enhance the scale of farthest objects, while simultaneously reducing scale of the peripheral ones. We draw motivation from a key aspect of modeling image representation in the human eye, while attending a particular object in the full visual scene [16]. In such a scenario, the representation of the region close to the fixated point is mapped onto the fovea with a considerably greater spatial resolution as compared to the representation of the unattended peripheral areas. We model this transformation using 2 distinct set parameters which are named as a) the fovea point, i.e. the spatial location of the fixation in the image plane, and b) the foveal magnification factor, separately along the height and width of the image plane. Besides this, we propose a novel Fovea Estimation Network (FEN) which is trained to predict a single most convenient fixation location alongside the associated magnification factors along both the spatial directions best suited for a given urban scene image. In absence of the ground-truth, we propose various strategies to obtain an estimate of the most convenient fovea parameters which would yield a significant improvement in the final parsing performance. After obtaining the predictions for both the natural image and the transformed image through a common parsing network, a perspective-aware aggregation of these predictions is performed using a spatial weight-map to weigh the predictions based on their reliability at individual pixel locations. We evaluate the proposed VRT-Net framework as a wrapper over the available real-time semantic segmentation approaches, such as ICNet [25] and ThunderNet [23] on two benchmark datasets, Cityscapes [6] and CamVid [3]. In summary our prime contributions are:

- We formalize a unique variable resolution transformation technique based on the foveal magnification in human eye while fixating a particular point in a scene.
- We design a Fovea Estimation Network by utilizing the initial layer of an off-the-shelf real-time scene parsing model, where the additional parameters are trained to output a single set of most convenient VRT parameters maintaining an optimal computational overhead.
- The proposed framework can be used as a wrapper over the available real-time scene parsing models followed by an end-to-end fine-tuning, where the final prediction is obtained as a perspective-aware aggregation of parsing results from both the natural and the transformed image.
- The proposed framework demonstrates a superior trade-off between speed and quality as compared to the prior state-of-the-art approaches.

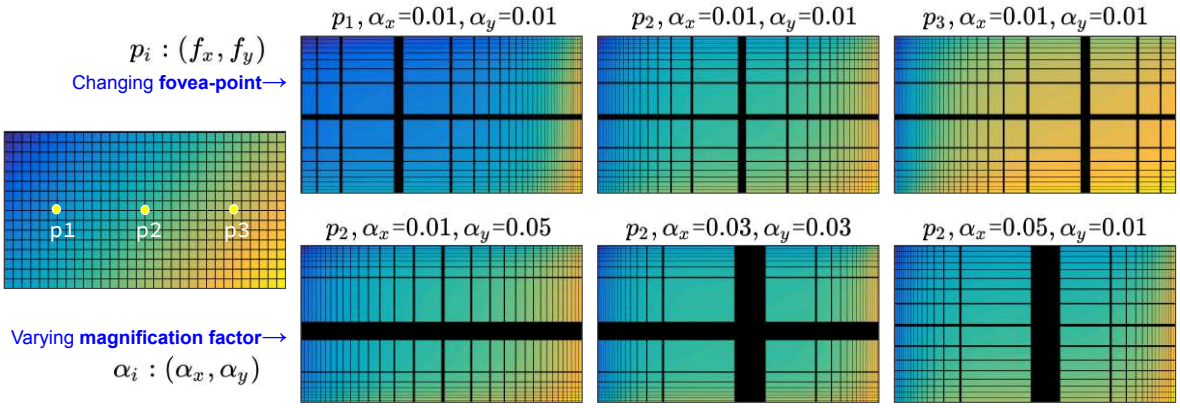

Figure 3. An exemplar grid image on left to demonstrate the effect of the proposed Variable Resolution Transformation for varying fovea point (top row) and magnification factors (bottom row).

2. Related Work

With the availability of large-scale labelled datasets for scene segmentation, there has been a tremendous progress towards developing efficient scene parsing solutions. Among them, the Fully Convolutional Network (FCN) [15] and DeepLab models [4] have achieved have pushed the benchmark performances to a significant extent via efficient realization of deep convolutions architectures. Several approaches employ conditional random fields [12, 27] to further enhance the parsing performance. An efficient fusion of multi-scale representations has proven to be another effective technique for outdoor scene parsing as a solution to the huge spatial scale variance. (ICNet) [25] is a parsing model that uses multi-resolution branches of the images i.e. high, medium and low for parsing. Images at different resolutions are parsed separately with deep and shallow convolution Networks. Final results are produced after merging each of these feature using a cascade feature fusion box. Pyramid Scene Parsing Network (PSPNet) [26] employs a different variant of multi-scale fusion where the features obtained from the image are pooled at different scales via a pyramid pooling module.

As a different approach certain recent works propose to explicitly address the extreme effect of perspective geometry, specifically for outdoor scene parsing. FoveaNet [10] propose to fist obtain a fovea region which is used to obtain a zoomed image cropped around the fovea. Both the original and the cropped images are parsed using a shared scene paring CNN. And finally the results are fused to obtain the final segmentation output. In our proposed method, we avoid cropping the localized fovea region by performing Variable resolution transform directly on the full image and use a common architecture to parse both image and the fovea-localized region. This is more effective in undoing the perspective issue as it preserves the full image context thereby leading to an improved parsing performance.

3. Proposed Architecture

Our proposed architecture process the images through three important stages which are pre-processing, scene-parsing and post-processing. Pre-processing stage identifies the fovea point(s) in the input image using Depth Estimation Technique [7] or a Grid-Search technique, and then performs Variable Resolution Transform using the found fovea point. This stage is responsible for addressing changes in the perspective view. Scene-parsing stage then parses the input image and the transformed image to produce their respective segmentation maps. Finally, post-processing stage merges the two segmentation maps to produce the final result. Here we explain the involved components in detail.

3.1. Variable Resolution Transform (VRT)

Variable Resolution Transform (VRT) [2] is a technique that is used in image compression applications. In this case, we aim to formalize a nonlinear spatial transformation which allows us to magnify the regions close to a fixation point while compressing the resolution towards the periphery. This transformation is parameterized by, a) spatial location of the fovea point represented as $p_i : (f_x, f_y)$, and b) magnification factor $\alpha_i : (\alpha_x, \alpha_y)$. Here, the magnification factor decides the extent of spatial magnification, separately along both the spatial dimensions (see Figure 3). Let, (x, y) denotes spatial-index of the original image in a $H \times W$ lattice. And, $(\hat{x}, \hat{y})$ denotes the spatial-index of the transformed image. The distance of an arbitrary pixel location (x, y) from the fovea-point (f_x, f_y) is represented as,

$$d_x = x - f_x, \qquad d_y = y - f_y \tag{1}$$

A logarithmic nonlinear transformation defines the mapping, $[(x, y) \rightarrow (\hat{x}, \hat{y})]$, as realized in the following equations. $\hat{x} = f_x + dv_x, \quad \hat{y} = f_y + dv_y$ where; (2)

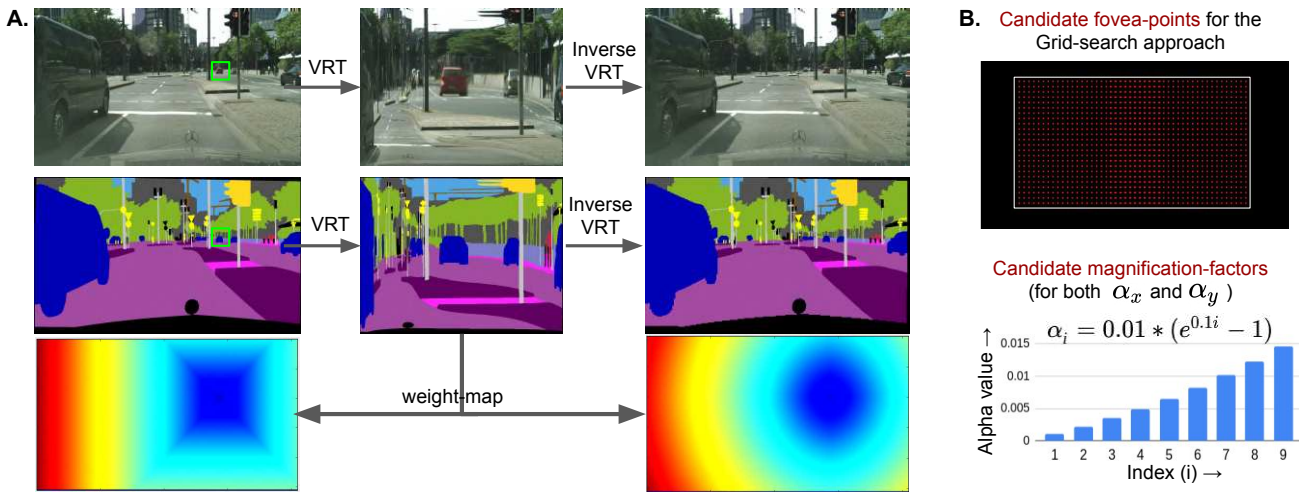

Figure 4. **A.** Variable Resolution Transform(VRT) applied on a raw image and its segmentation map. Below we show the spatial weight-maps obtained using Chebyshev (left) and L2 (right) distance in equation 6. We have used L2 distance in our computation. **B.** The figure depicts the process of selecting candidate fovea-points and magnification factors for the Grid-search approach discussed in Sec. 3.2b.

$$\begin{aligned} dv_x &= \ln\left(d_x * \alpha_x + 1\right) * sf_x \\ dv_y &= \ln\left(d_y * \alpha_y + 1\right) * sf_y \end{aligned} \tag{3}$$

Here, sf_x and sf_y control scaling of the transformed image. Moreover, sf_x and sf_y are set to maintain a fixed predefined spatial size of the transformed output image, i.e.

$$\begin{aligned} sf_x &= dv_x^{\max} / (\ln\left(d_x^{\max} * \alpha_x + 1\right)) \\ sf_y &= dv_y^{\max} / (\ln\left(d_y^{\max} * \alpha_y + 1\right)) \end{aligned} \tag{4}$$

The above equations convey how a pixel at a distance (d_x, d_y) from the fovea in the original image is moved to a distance (dv_x, dv_y) form the fovea in the transformed image. The above defined spatial transformation is also invertible (i.e. $[(\hat{x}, \hat{y}) \rightarrow (x, y)]$ using following equations,

$$\begin{aligned} d_x &= (\exp\left(dv_x / sf_x\right) - 1) / \alpha_x \\ d_y &= (\exp\left(dv_x / sf_y\right) - 1) / \alpha_y \end{aligned} \tag{5}$$

We obtain a spatial weight map, M in the original image space, depicting spatial information loss in a cyclic reconstruction of the original image (see Figure 4).

$$\begin{aligned} \hat{M}(\hat{x}, \hat{y}) &= dist(\ 0,\ \Delta[(\hat{x}, \hat{y}) \rightarrow (x, y)]^{(\hat{x}, \hat{y})}) \\ M(x, y) &= \hat{M}([(x, y) \rightarrow (\hat{x}, \hat{y})]^{(x, y)}) \end{aligned} \tag{6}$$

3.2. Fovea Point Estimation

Identification of fovea point is central towards addressing the issue of perspective geometry. We propose to use a Fovea Estimation Network (FEN) for this purpose. Given an input image FEN estimates both the fovea point (f_x, f_y) and the magnification factor (α_x, α_y). However, in absence of a reliable ground-truth for the above transformation parameters, we propose two different strategies to prepare supervised training samples for the FEN.

a) Depth-based fovea estimation. In first strategy we fix the magnification factor to some empirically chosen appropriate value, i.e. $(\alpha_x, \alpha_y) = (0.004, 0.004)$, thereby keeping fovea point as the only parameter to be estimated. We found the above value as the mode over a set of candidate values (see Figure 4), when tested for an improved parsing performance on a subset of training dataset. We rely on a monocular depth estimation network [7] to obtain a depth map for each image in the training dataset. Following this, the fovea point is obtained as a pixel-location with the maximum depth value. For cases where the maximum depth value is obtained at multiple pixel-locations, we perform a weighted (i.e. depth-value) mean of the pixel locations to realize the fovea-point. We train the FEN network by treating these fovea-points as ground-truth for the corresponding input image. However, fixing the magnification factor limits the capability of the proposed variable transformation technique. Different images require different level of magnification or zooming to reliably parse the distant objects. Further, the choice of maximum depth as the fovea-point is not optimal. In certain scenarios, the maximum depth is attended on background regions such as tree, road, or sky, instead of attending a distance object of interest such as car, pedestrian, rider, etc.

b) Fovea estimation via grid-search. Acknowledging the above shortcoming, we argue that the fovea-point and the magnification-factors must be selected in a way which would surely improve the final segmentation performance. Following this analogy, we adopt a greedy approach to obtain the most reliable ground-truth for the transformation

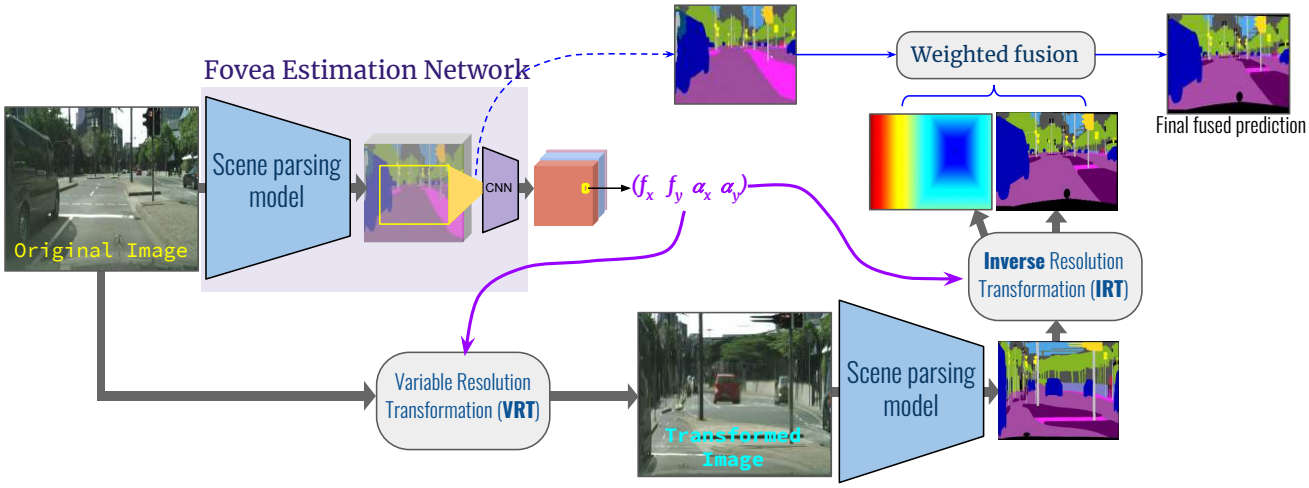

Figure 5. An overview of the proposed framework. In top branch, the raw image is taken as input to the Fovea Estimation Network (FEN) to obtain the parameters required for the transformation. We reuse the Scene parsing model in FEN which also outputs a segmentation map of the original input image. In another branch, the transformation module (VRT) uses the estimated parameters to obtain a transformed image. This image is passed through the scene parsing model whose output is transformed back to the original image space via a inverse transformation (IRT). Both the segmentation estimates are merged using the spatial weight map to obtain the final segmentation result.

parameters. Given a pre-trained scene parsing model, we setup the full-framework (see Figure 5) using pre-selected values for $(f_x, f_y, \alpha_x, \alpha_y)$ by skipping the fovea estimation network. We compute mIoU score with respect to the ground-truth to greedily select the best suited values for $(f_x, f_y, \alpha_x, \alpha_y)$, for each training samples.

Though grid-search is adopted to prepare the training samples, we formalize a strategy to reduce the search space. Given a input image of size 2048×1024, we first search for the best fovea-point out of a predefined set of candidate locations as shown in Figure 3B. The fovea-point candidate set is a 48×24 grid close to the central region of the image, following the general nature of perspective effect. Each point is spaced 32 pixels both vertically and horizontally. We consider each candidate point as the fovea, while fixing the magnification factors as $(\alpha_x, \alpha_y) = (0.004, 0.004)$. We select the point with maximum mIoU as the fovea-point ground-truth. After fixing the fovea-point, we search for the best magnification factor which further improves the mIoU metric. We first vary α_x over the 9 candidate values (see Figure 3B) while fixing α_y as 0.004. Following this, we fix both fovea-point and α_x to select the best suited α_y among the 9 candidate values. This is performed for each image in the training sample to obtain the corresponding ground-truth tuple, $(f_x, f_y, \alpha_x, \alpha_y)$.

Training FEN. Training samples for FEN consists of an input image along with four labels namely f_x, f_y for fovea point and α_x, α_y for magnification factor. In FEN, we have a feature extractor network followed by multiple CNN layers and finally an output layer. We use a pre-trained scene parsing model (ICNet) as feature extraction network whose last layer logit is of size of size 256×512×19 (19 channels for 19 classes of cityscapes). This output is cropped from the center (see Figure 5) according to the position of the candidate fovea-locations shown in Fig 4B. The cropped spatial maps of size 200×312×19 is upsampled to 192×384×19 before passing it as input to a shallow CNN (3 convolution layers with kernels of size 3x3 and output channels 64, 64, 19 respectively) to estimate the transformation parameters. In the last layer (19 channels), we use the channels 1-9 to predict α_x as multi-class classification over the 9 candidates via *softmax* nonlinearity. Similarly, channels 10-18 is dedicated to predict α_y via another *softmax* nonlinearity. Here, the last channel (spatial size: 48×24) predicts probability of the spatial location to be selected as a fovea-point via *sigmoid* nonlinearity.

Before formalizing the loss function to train the newly introduced shallow CNN, we define a spatial weight map as, $W^{(x,y)} = \exp(-\sqrt{|f_x - x|^2 + |f_y - y|^2}/\beta)$. Here, β controls temperature of the spatial weight map. This softly allows the spatial coordinates close to the (f_x, f_y) to have a better estimate of the magnification factor as enforced by the following loss function.

$$\mathcal{L} = \sum_{(x,y)} (W^{(x,y)} * (\mathcal{L}_{CE}(z^{(x,y)}_{[1:9]}, \alpha_x) + \mathcal{L}_{CE}(z^{(x,y)}_{[10:18]}, \alpha_y)) + \lambda \mathcal{L}_{BCE}(z^{(x,y)}_{[19]}, \mathbb{1}_{(f_x,f_y)}(x,y)))$$

In the above equation, z represents the last layer output after applying the respective non-linearities. $\mathcal{L}_{CE}$ and $\mathcal{L}_{BCE}$ represent multi-class cross-entropy and binary cross-entropy objectives respectively. Here, $\mathbb{1}$ represents the indicator function which is active when (x, y) matches with

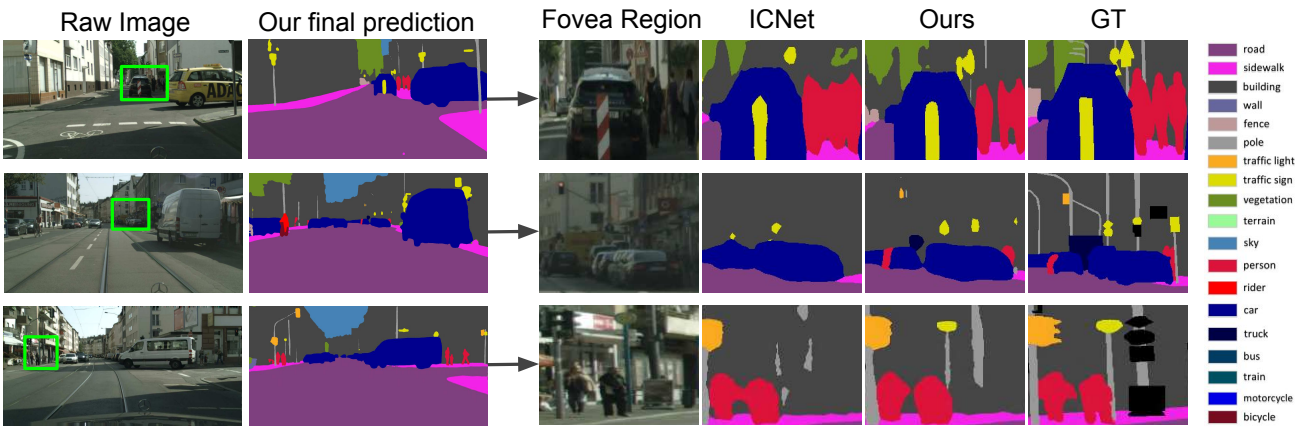

Figure 6. Parsing results for images from Cityscapes validation indicating the fovea region (first row, in green box), the final merged segmentation results through the proposed VRT-Net and magnified segmentation performance on a patch centered at the fovea point.

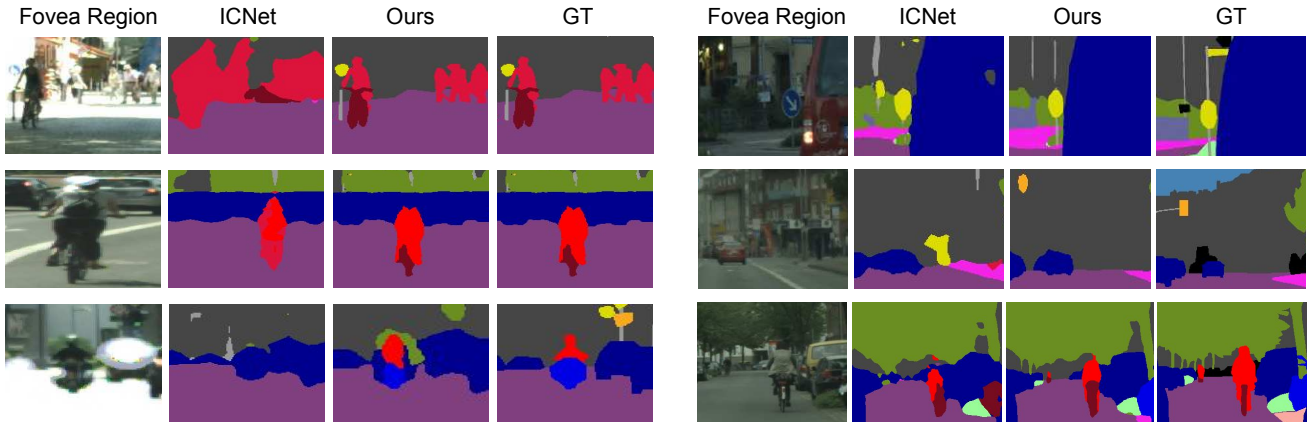

Figure 7. Qualitative comparison of parsing performance specifically on recognizing important small-scale objects such as traffic sign, pedestrian etc. Note that, the prediction results obtained from ICNet [25] often misses the crucial traffic sign objects by miss-classifying it as one of the background categories.

the ground-truth (f_x, f_y), thereby providing a label for the binary cross-entropy.

3.3. Proposed scene parsing framework

Our proposed architecture is shown in Figure 5. Our model works in three stages. Firstly, given a raw image as an input it finds the fovea points (f_x, f_y) and the magnification factor (α_x, α_y) using the techniques discussed above. Following this, one branch performs VRT using the estimated $(f_x, f_y, \alpha_x, \alpha_y)$ to obtained the transformed image which is passed to the scene parsing model to obtain the corresponding segmentation map. We perform inverse resolution transform (IRT) on this segmentation output to obtain the segmentation map back into the original image space, denoted as S_t. Additionally, we obtain another segmentation output, S_o directly from the original image which is already computed as an intermediate representation of the Fovea estimation network (see Figure 5). Note that, the transformation module also outputs a spatial weight map, M which is later used to fuse both the segmentation outputs to obtain the final segmentation result, S_f. We compute it as,

$$S_f = (1 - M) * S_t + M * S_o$$

As the transformed image undoes the effect of perspective projection, the segmentation estimate for regions close to the fovea is more reliable in the corresponding output, i.e. S_t. And conversely, segmentation estimate, S_o is comparatively more reliable at the peripheral regions. The fused result leverages advantages of both the estimation to realize an improved parsing performance.

4. Experiments and Results

To evaluate the efficacy of the proposed scence parsing framework, we perform experiments on two widely used datasets, viz. Cityscapes and CamVid.

Table 1. Validation results (per class) on Cityscapes dataset.

Model	road	sidewalk	building	wall	fence	pole	tr. light	tr. sign	veg.	terrain	sky	person	rider	car	truck	bus	train	mcycle	bicycle
Trained Depth Info	97.2	81.5	90.2	40.0	47.2	56.6	64.5	74.2	91.0	57.4	93.1	78.3	58.9	92.1	65.2	80.0	58.1	66.2	74.8
Trained Grid Search	97.8	82.8	91.2	45.3	52.1	60.9	68.4	76.8	92.4	59.8	94.5	80.9	65.8	94.2	70.4	85.0	70.4	70.9	78.1

Table 2. Comparison of Test results on Cityscapes dataset. In last 3 rows, *Grid-Search* and *Depth-Info* denote the adopted fovea estimation approach as discussed in Section 3.2.

Method	mIoU (%)	Time (ms)	Frame (fps)
SegNet [1]	57.0	60	16.7
ENet [17]	58.3	13	76.9
SQ [20]	59.8	60	16.7
CRF-RNN [27]	62.5	700	1.4
DeepLab [4]	63.1	4000	0.25
FCN-8S [15]	65.3	500	2
Dilation10 [24]	67.1	4000	0.25
ThunderNet [23]	64.0	10.4	96.2
ICNet [25]	70.6	33	30.3
VRT-ThunderNet (Grid-Search)	69.1	11	90.9
VRT-ICNet (Depth-Info)	73.1	34	29.4
VRT-ICNet (Grid-Search)	75.2	34	29.4

Table 3. Comparison of Test results on CamVid dataset. In last 2 rows, *Grid-Search* and *Depth-Info* denote the adopted fovea estimation approach as discussed in Section 3.2.

Method	mIoU (%)	Time (ms)	Frame (fps)
SegNet [1]	46.4	217	4.6
DPN [14]	60.1	830	1.2
DeepLab [4]	61.1	203	4.9
Dilation10 [24]	65.3	227	4.4
ICNet [25]	67.1	36	27.8
Our-VRT-ICNet (Depth-Info)	69.7	37	27.1
Our-VRT-ICNet (Grid-search)	71.7	37	27.1

4.1. Results on Cityscapes dataset:

Cityscapes [6] dataset is a widely used dataset for benchmarking semantic segmentation performance. The dataset contains diverse stereo video sequences recorded in street scenes from 50 different cities, with high-quality pixel-level dense annotations for 5000 frames. These images are further divided into 2975 for training, 500 for validation and 1,525 for testing. They consider 19 classes for evaluation, which include the objects such as a car, pedestrian, cycle, sidewalk, buildings, etc. We have used both ICNet and ThunderNet as the baseline model for the scene parsing network. A monocular RGB image is given as an input to the model which outputs a pixel-level prediction map representing the semantic category of each pixel. We use the standard mean Intersection over Union ($mIoU$) metric to evaluate efficacy of the proposed framework against the prior approaches. IoU is defined as true positives divided by the sum of true positives (tp), false positives (fp) and false negatives (fn) for a given category and *mIoU* is the mean over all the categories.

We also show per-class mIoU score on the standard validation set of Cityscapes in Table 1. Figure 6 depicts a qualitative comparison of the proposed approach against ICNet. In Figure 6, the area marked on the raw images with the green colored box is the region where fovea point is located. On the right in *fovea region* column, we show the zoomed view around the fovea point. Using the fovea point and the magnification factor α, our model is able to focus on distant small object thereby delivering an improved parsing performance. For example, our model is able to distinguish between the rider class and vehicle (e.g. motorcycle) class which is not distinguished by ICNet. Evaluation results on Cityscapes test set are shown in the Table 2.

4.2. Results on CamVid dataset:

Cambridge-Driving labeled video data (CamVid) is a real-world dataset consisting of images taken using a car mounted camera. The standard CamVid [3] dataset contains images extracted from high resolution 10 min video sequences. A small-fraction of pixels are labelled as void in original dataset. To have a fair comparison, we use the dataset split proposed by Sturgess *et al.* [19]. They partitions the dataset into 367, 100 and 233 images for training, validation and testing respectively. Total 11 categories were used for evaluation. Table 3 contains quantitative comparison of the proposed framework against the prior arts.

5. Conclusion

We proposed a novel approach to improve urban scene parsing performance by undoing the effect of perspective projection. We design an efficient Fovea estimation network, which is trained to predict the most convenient transformation parameters. Re-usage of base scene parsing model for fovea estimation enables us to achieve improved segmentation performance in an optimal computational overhead. Effectiveness of such transformations for other dense prediction tasks remains to be explored.

Acknowledgements. This project was partly supported by Indo-UK project (DST/INT/UK/P-179/2017), DST, India; Uchhatar Avishkar Yojana (UAY) project (IISC_010), MHRD, India; and a Wipro PhD Fellowship (Jogendra).

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[2] A. Basu and K. J. Wiebe. Videoconferencing using spatially varying sensing with multiple and moving foveae. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 2-Conference B: Computer Vision & Image Processing.(Cat. No. 94CH3440-5)*, pages 30–34. IEEE, 1994.

[3] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 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, 2018.

[6] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[7] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[8] J. N. Kundu, P. Krishna Uppala, A. Pahuja, and R. Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] J. N. Kundu, N. Lakkakula, and R. V. Babu. Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[10] X. Li, Z. Jie, W. Wang, C. Liu, J. Yang, X. Shen, Z. Lin, Q. Chen, S. Yan, and J. Feng. Foveanet: Perspective-aware urban scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 784–792, 2017.

[11] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[12] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[14] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE international conference on computer vision*, pages 1377–1385, 2015.

[15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[16] F. P. Ottes, J. A. Van Gisbergen, and J. J. Eggermont. Visuomotor fields of the superior colliculus: a quantitative model. *Vision research*, 26(6):857–873, 1986.

[17] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.

[18] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[19] P. Sturgess, K. Alahari, L. Ladicky, and P. H. Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC-British Machine Vision Conference*. BMVA, 2009.

[20] M. Treml, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich, et al. Speeding up semantic segmentation for autonomous driving. In *MLITS, NIPS Workshop*, volume 1, page 5, 2016.

[21] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1451–1460. IEEE, 2018.

[22] Z. Wu, C. Shen, and A. Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019.

[23] W. Xiang, H. Mao, and V. Athitsos. Thundernet: A turbo unified network for real-time semantic segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1789–1796. IEEE, 2019.

[24] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[25] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–420, 2018.

[26] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[27] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015.

[28] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.