OVeNet: Offset Vector Network for Semantic Segmentation
- Supplementary Material -

Stamatis Alexandropoulos
Princeton University *

Christos Sakaridis
ETH Zürich

Petros Maragos
National Technical University of Athens

1. Network Instantiation

OVeNet follows the same general architecture as [3]. Since it is built on HRNet [3], our network contains four stages and two heads, as shown in Table 1. The branch occurs on the 4th stage. The semantic head of our model is constructed from modularized blocks, which are repeated a specific number of times in each of the four stages. In particular, the blocks are repeated 1, 1, 4, and 3 times, respectively, following the same configuration as the original HRNet model. However, in the offset head, we had to make a modification to the number of block repetitions in order to address memory constraints. As a result, the blocks are repeated 2 times in the 4th stage of the offset head. Each modularized block in our network consists of 1, 2, 3, or 4 branches, depending on the stage in which it is located (1st, 2nd, 3rd, or 4th). Each branch is associated with a different resolution and is composed of four residual units and one multi-resolution fusion unit.

2. Additional Qualitative Results

We provide additional qualitative comparisons of OVeNet to its HRNet baseline for the three examined datasets: Cityscapes [1], ACDC [2] and ADE20K [5]. More specifically, we provide sets of successful segmentations in Fig. 1, 2 and 3 and sets of challenging and failure cases in Fig. 4, 5 and 6 correspondingly.

In Fig. 1 we depict the successful results on Cityscapes. We also present an additional comparison of the default instance of OVeNet built only on HRNet [3] with the instance of OVeNet built on HRNet+OCR [4]. As we can see, OVeNet (*HRNet*) reduces correctly the number of misclassified terrain pixels on the right side of the road in the 3^{rd} row of Fig. 1. On the other hand, OVeNet (*HRNet* + *OCR*) achieves a better prediction since it enlarges the sidewalk segment and it eliminates the terrain. Regarding the 6^{th} row, OVeNet (*HRNet*) expands the terrain on the right, while the HRNet + OCR-based model increases additionally the sidewalk segment in the background. Overall, the latter generally achieves better predictions than the former.

In Fig. 2 we observe the results of successful segmentations on ACDC, where HRNet [3] is compared to OVeNet built only on it. Specifically, in the 2^{nd} as well as the 7^{th} row of Fig. 2, OVeNet enhances the prediction made by HRNet in the sidewalk on the right. A similar result happens with the terrain segment in the 5^{th} row, as it covers correctly all the more space. Thus, our model surpasses baseline's performance on adverse conditions.

In Fig. 3 we depict the results of successful segmentations on ADE20K. To be more specific, in the 3^{rd} as well as the 4^{th} row, OVeNet enhances the prediction made by HRNet by enlarging some segments (river and house respectively). As a result, our model surpasses baseline's performance on everyday images.

As for Fig. 4, we provide some challenging cases on Cityscapes. To be more specific, we can see that in both rows, HRNet [3] results in a better prediction than OVeNet (*HRNet*) (e.g car on the left in the 1^{st} row, terrain on the right in the 2^{nd} one). On the other hand, OVeNet (*HRNet* + *OCR*) outperforms both the HRNet and HRNet-based model, leading to a better total outcome.

Regarding Fig. 5, we see some some failure cases of our model built on HRNet [3] on ACDC. Specifically, in both rows, there are more correctly predicted labels in the vegetation segment on the right outputted by HRNet than by OVeNet.

As for Fig. 6, we see some some false results predicted our model built on HRNet [3] on ADE20K. Specifically, in both set of images, there are more correctly predicted labels in the tree and wash machine segment respectively outputted by HRNet than by OVeNet.

All in all, we observe that OVeNet not only produces results that are very faithful to ground-truth annotations, but its predictions also surpass the predictions made by the initial HRNet in terms of quality. By correctly classifying several pixels which are misclassified by the HRNet baselines, OVeNet (*HRNet* + *OCR*) eliminates inconsistencies and enhances the shape and appearance of respective segments,

* This work was done when Stamatis Alexandropoulos was in the National Technical University of Athens.

Table 1. **The architecture of the OVeNet (main body)**. Within each cell of our network, there are three distinct components. The first component, represented by $[\cdot]$, refers to the residual unit. The second component is a numerical value that specifies the number of times the residual unit is repeated. The final component is another numerical value that indicates how many times the modularized block is repeated within the cell. In each residual unit, the variable C is used to represent the number of channels.

Head	Resolution	Stage 1	Stage 2	Stage 3	Stage 4
Semantic	$4\times$	$\begin{bmatrix} 1\times 1, 64 \\ 3\times 3, 64 \\ 1\times 1, 256 \end{bmatrix} \times 4\times 1$	$\begin{bmatrix} 3\times 3, C \\ 3\times 3, C \end{bmatrix} \times 4\times 1$	$\begin{bmatrix} 3\times 3, C \\ 3\times 3, C \end{bmatrix} \times 4\times 4$	$\begin{bmatrix} 3\times 3, C \\ 3\times 3, C \end{bmatrix} \times 4\times 3$
	$8\times$		$\begin{bmatrix} 3\times 3, 2C \\ 3\times 3, 2C \end{bmatrix} \times 4\times 1$	$\begin{bmatrix} 3\times 3, 2C \\ 3\times 3, 2C \end{bmatrix} \times 4\times 4$	$\begin{bmatrix} 3\times 3, 2C \\ 3\times 3, 2C \end{bmatrix} \times 4\times 3$
	$16\times$			$\begin{bmatrix} 3\times 3, 4C \\ 3\times 3, 4C \end{bmatrix} \times 4\times 4$	$\begin{bmatrix} 3\times 3, 4C \\ 3\times 3, 4C \end{bmatrix} \times 4\times 3$
	$32\times$				$\begin{bmatrix} 3\times 3, 8C \\ 3\times 3, 8C \end{bmatrix} \times 4\times 3$
Offset Vector	$4\times$				$\begin{bmatrix} 3\times 3, C \\ 3\times 3, C \end{bmatrix} \times 4\times 2$
	$8\times$				$\begin{bmatrix} 3\times 3, 2C \\ 3\times 3, 2C \end{bmatrix} \times 4\times 2$
	$16\times$				$\begin{bmatrix} 3\times 3, 4C \\ 3\times 3, 4C \end{bmatrix} \times 4\times 2$
	$32\times$				$\begin{bmatrix} 3\times 3, 8C \\ 3\times 3, 8C \end{bmatrix} \times 4\times 2$

resulting in more realistic outputs.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt 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. 1

[2] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 1

[3] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1, 3, 4, 5, 6

[4] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *CoRR*, abs/1909.11065, 2019. 1

[5] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442*, 2016. 1

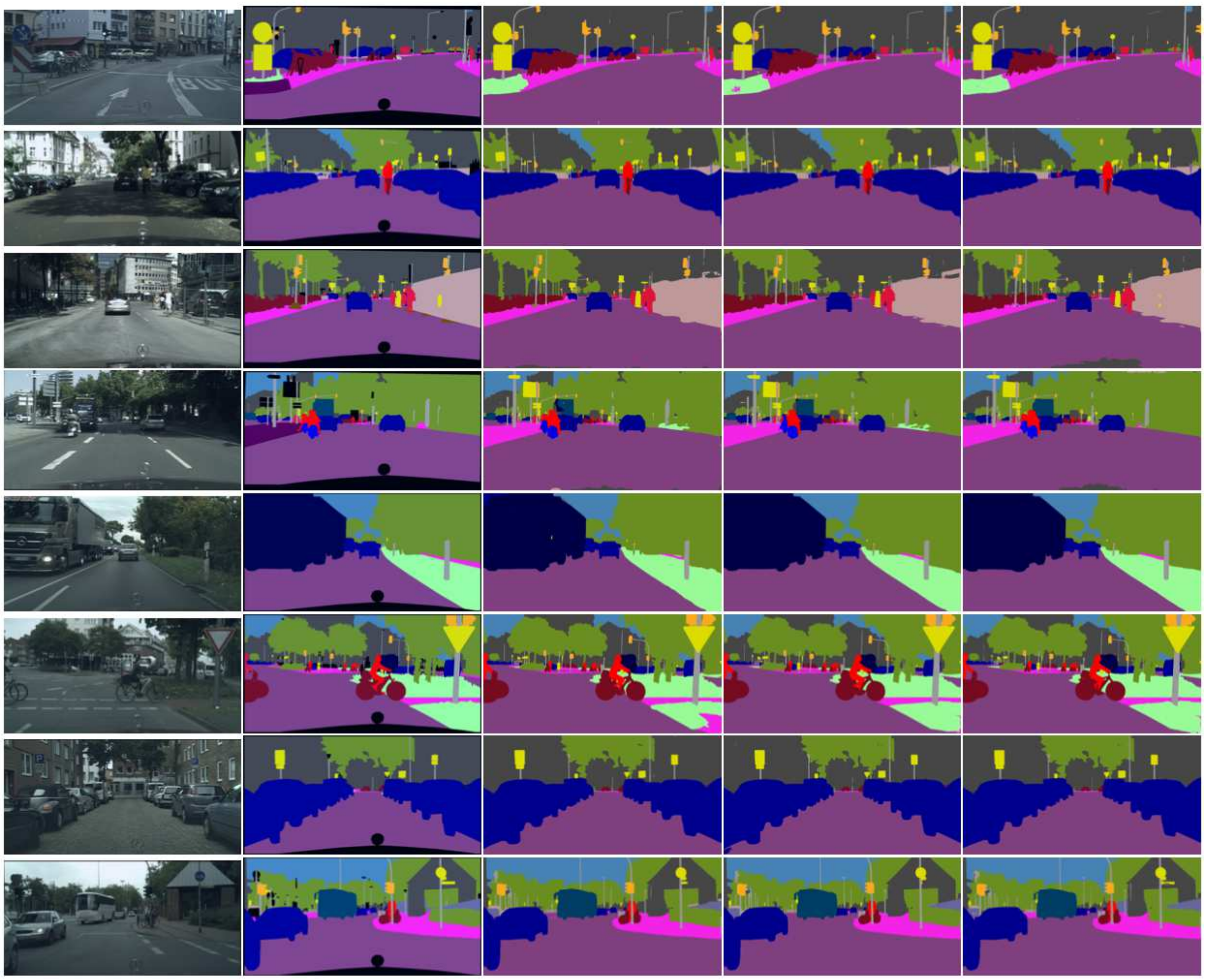

Figure 1. **Additional qualitative results of selected examples on Cityscapes**. From left to right: input image, ground-truth annotation, and prediction with HRNet [3], OVeNet (*HRNet*), and OVeNet (*HRNet+OCR*). Best viewed on a screen and zoomed in.

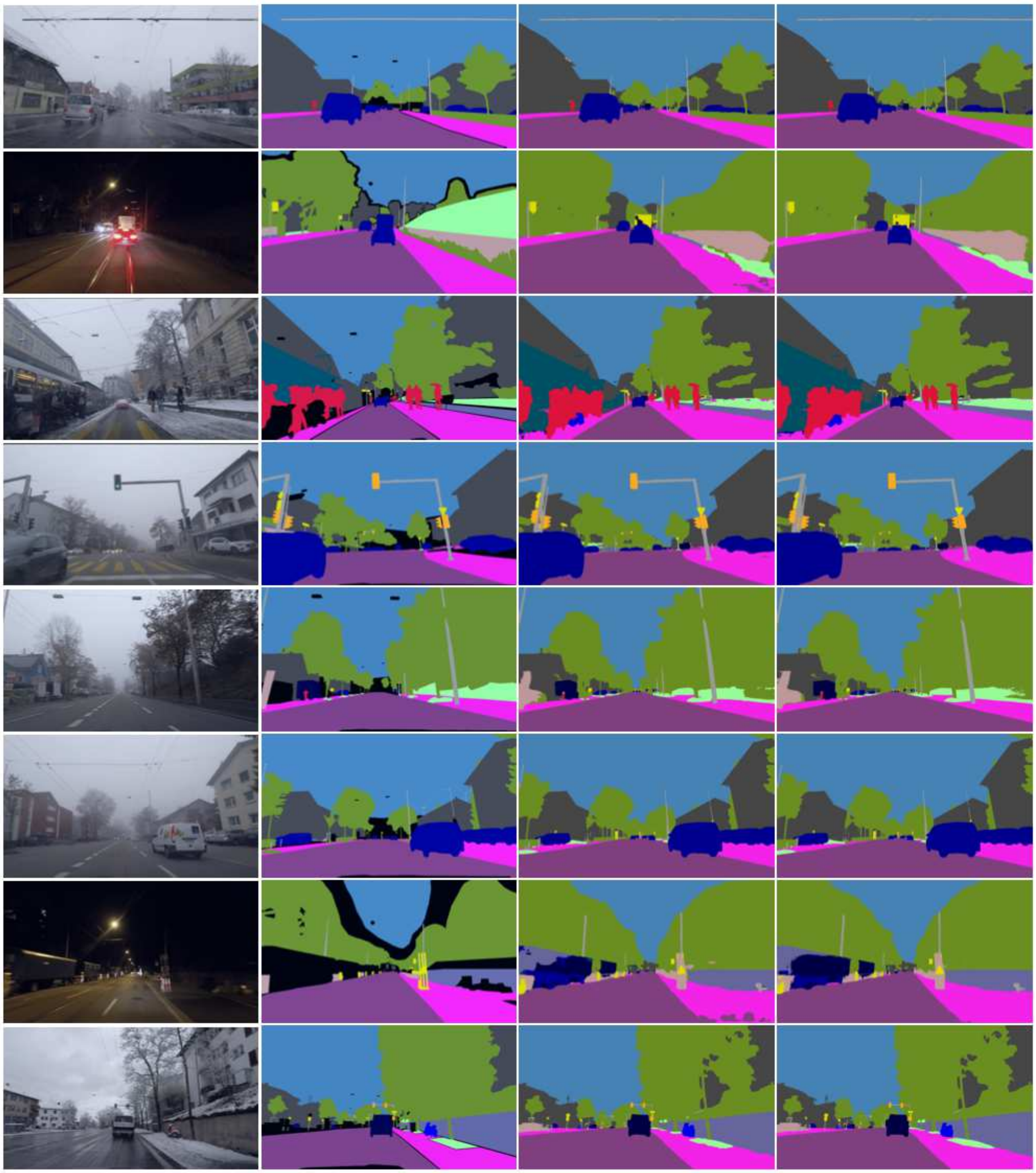

Figure 2. **Additional qualitative results of selected examples on ACDC**. From left to right: input image, ground-truth annotation, and prediction with HRNet [3] and OVeNet. Best viewed on a screen and zoomed in.

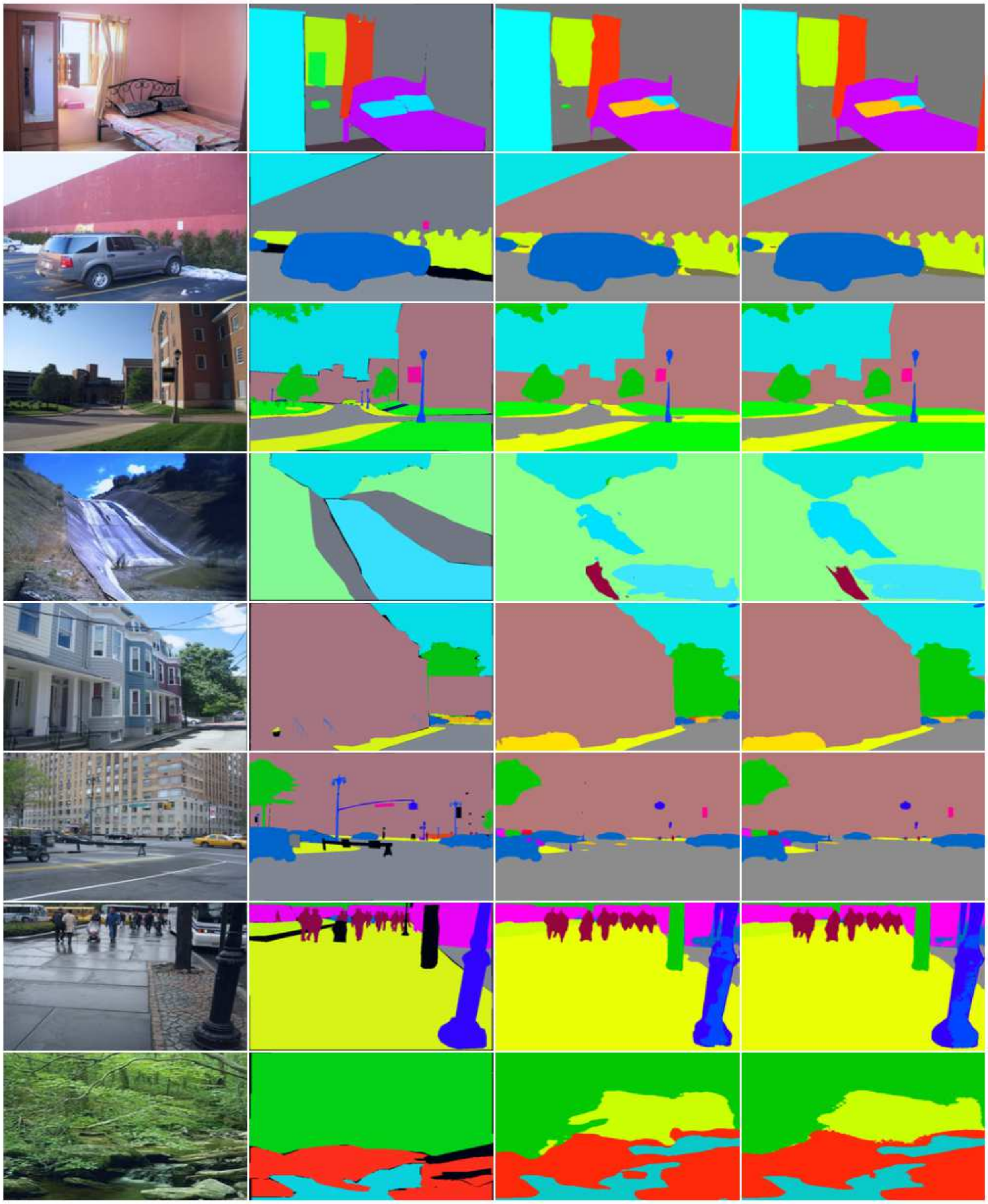

Figure 3. **Additional qualitative results of selected examples on ADE20K**. From left to right: input image, ground-truth annotation, and prediction with HRNet [3] and OVeNet. Best viewed on a screen and zoomed in.

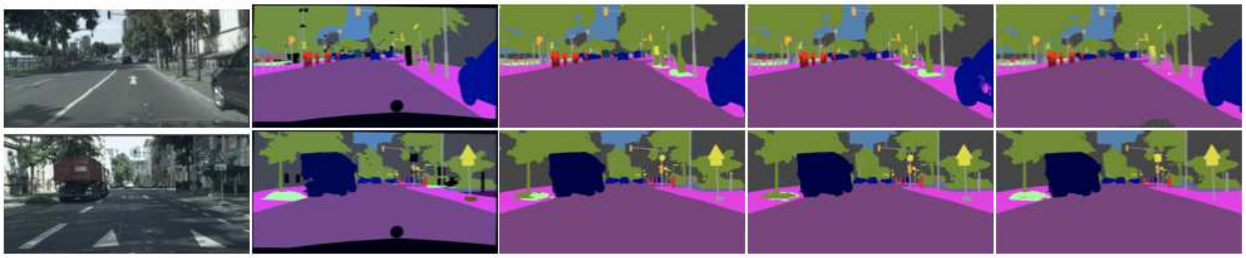

Figure 4. **Challenging cases on Cityscapes**. From left to right: input image, ground-truth annotation, and prediction with HRNet [3], OVeNet (*HRNet*), and OVeNet (*HRNet+OCR*). Best viewed on a screen and zoomed in.

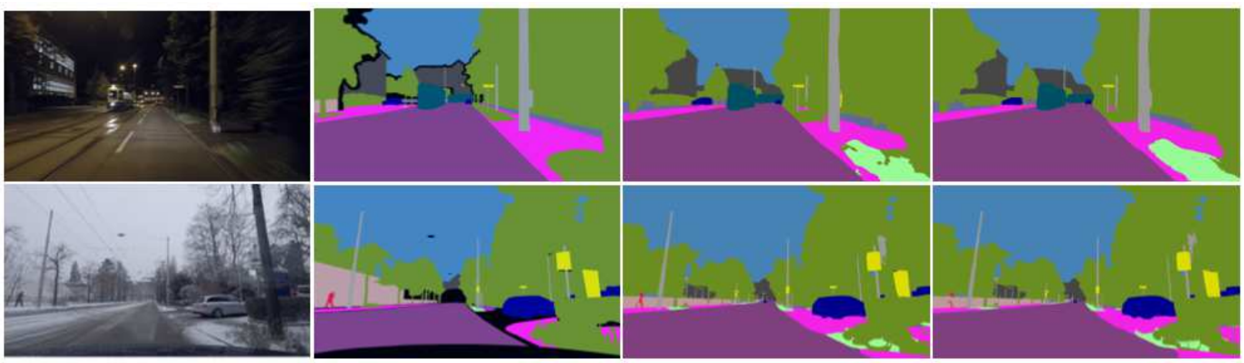

Figure 5. **Failure cases on ACDC**. From left to right: input image, ground-truth annotation, and prediction with HRNet [3] and OVeNet. Best viewed on a screen and zoomed in.

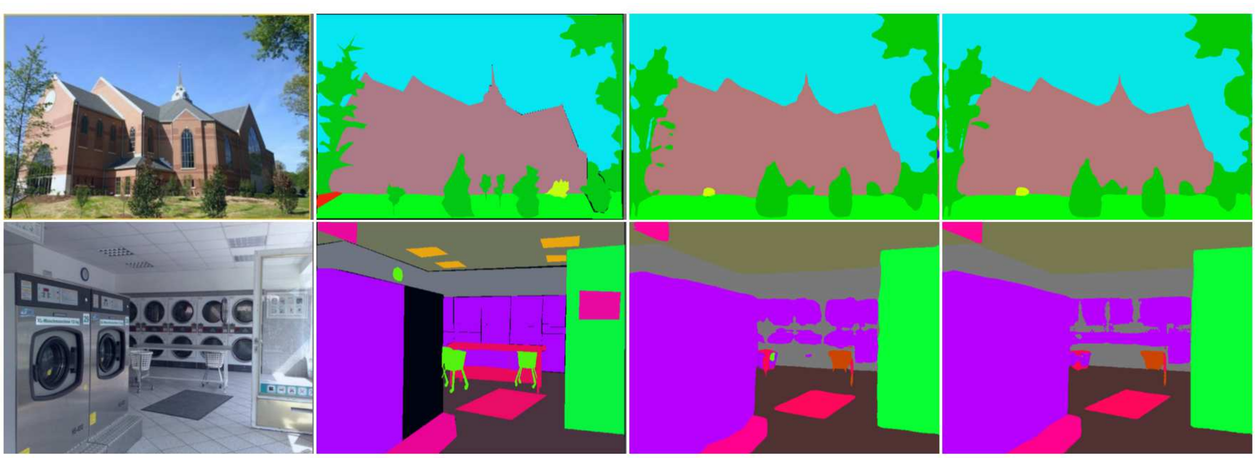

Figure 6. **Failure cases on ADE20K**. From left to right: input image, ground-truth annotation, and prediction with HRNet [3] and OVeNet. Best viewed on a screen and zoomed in.