Progressive Semantic Segmentation Supplementary Material

Chuong Huynh[1] Anh Tuan Tran[1,2] Khoa Luu[1,3] Minh Hoai[1,4]
[1]VinAI Research, Hanoi, Vietnam, [2]VinUniversity, Hanoi, Vietnam
[3]University of Arkansas, Fayetteville, Arkansas, USA
[4]Stony Brook University, Stony Brook, NY 11790, USA
{v.chuonghm,v.anhtt152,v.khoal,v.hoainm}@vinai.io

Input	mIoU(%)	mBA(%)	Time(s)	Mem.(MB)
256×128	63.23	43.31	0.02	1575
512×256	65.03	47.85	0.03	3585
1024×512	73.18	58.22	0.05	9993

Table 1: Performance of HRNetV2W18+OCR with different input sizes on Cityscapes

1. Cityscapes

1.1. Implementation details

For the PointRend, we used the feature maps at the last stage of HRNet to refine initial segmentation. DenseCRF ran in 50 steps with pairwise bilateral, sxy of 256, srgb of 13, and compat of 1. For DeepLabV3+, Resnet-50 was used as the backbone with an output stride of 8.

1.2. Compare MagNet with one-run models

Table 1 shows the results of single backbone HRNetV2W18+OCR with different input sizes that can be fitted into the memory of an RTX2080Ti 11GB. While MagNet consumes only ∼2GB memory and can improve the mIoU from 66.91% to 67.57%, the one-run model with a higher resolution of 512×256 uses ∼3.5GB memory and yields a lower mIoU. With the input size of 1024×512, the one-run model has higher accuracy but it requires ∼10GB.

1.3. Additional results

1.3.1 Quantitative results

Table 2 provides more quantitative results on the Cityscapes dataset. The mean boundary accuracy (mBA) which has been mentioned in CascadePSP is also used to evaluate the performance models. Compare to SegFix, our model yields about 2% improvement. In class-specific IoU of 19 categories in the Cityscapes, our MagNet settings outperform on 16 classes with large margins.

1.3.2 Qualitative results

Fig. 1 and Fig. 2 display more results produced by our MagNet framework and other methods. These images are crops

Method	mBA (%)	Class-IoU (%)																		
		road	side-walk	build-ing	wall	fence	pole	traffic light	traffic sign	vege-tation	terrain	sky	person	rider	car	truck	bus	train	motor-cycle	bi-cycle
Patching	47.0	95.1	68.4	85.0	26.2	33.3	**42.9**	46.1	59.4	**89.2**	52.1	91.3	61.2	16.8	86.6	14.4	32.7	9.0	22.1	59.9
Downsample	43.3	96.6	75.8	87.2	44.8	45.4	35.6	43.1	55.0	87.2	55.7	90.1	64.6	40.9	87.4	63.6	74.1	61.9	36.2	56.5
DenseCRF	43.3	96.7	75.9	87.1	44.9	45.3	32.3	40.5	54.7	86.8	55.8	90.2	64.3	40.6	87.6	64.0	74.3	62.1	36.2	56.7
DGF	43.8	96.6	76.0	87.3	44.9	45.4	35.6	43.1	55.3	87.2	55.7	90.6	64.9	40.9	87.6	63.6	74.1	62.0	36.2	56.6
PointRend	45.3	97.1	78.4	88.2	**48.3**	45.0	42.7	**49.2**	**60.2**	88.6	**58.2**	90.9	67.1	43.1	89.6	**65.8**	69.5	41.6	39.2	**60.7**
SegFix	**47.7**	**97.2**	**78.6**	**88.5**	46.9	**46.5**	38.8	47.7	59.2	88.5	57.6	**92.5**	**69.9**	**44.7**	**89.9**	64.4	**75.8**	**63.3**	**41.1**	60.0
MagNet-Fast	49.0	96.7	78.4	89.4	47.5	48.6	**49.4**	**53.1**	64.8	90.0	**56.3**	**92.3**	**72.7**	46.1	91.4	67.1	58.5	**65.4**	44.0	59.9
MagNet	**49.5**	**98.3**	**80.7**	**89.8**	**48.0**	**51.6**	**49.4**	44.0	**66.1**	**90.5**	55.7	90.0	72.2	**47.1**	**91.8**	**67.4**	**78.0**	56.4	**45.2**	**61.8**
Improvement	1.8	1.1	2.1	1.3	-0.3	5.1	6.5	3.9	5.9	1.3	-1.9	-0.2	2.8	2.4	1.9	1.6	2.2	2.1	4.1	1.1

Table 2: **Mean boundary accuracy and IoU for some specific categories on Cityscapes.** The best result of our method is highlighted in red while the best of previous methods is in blue color.

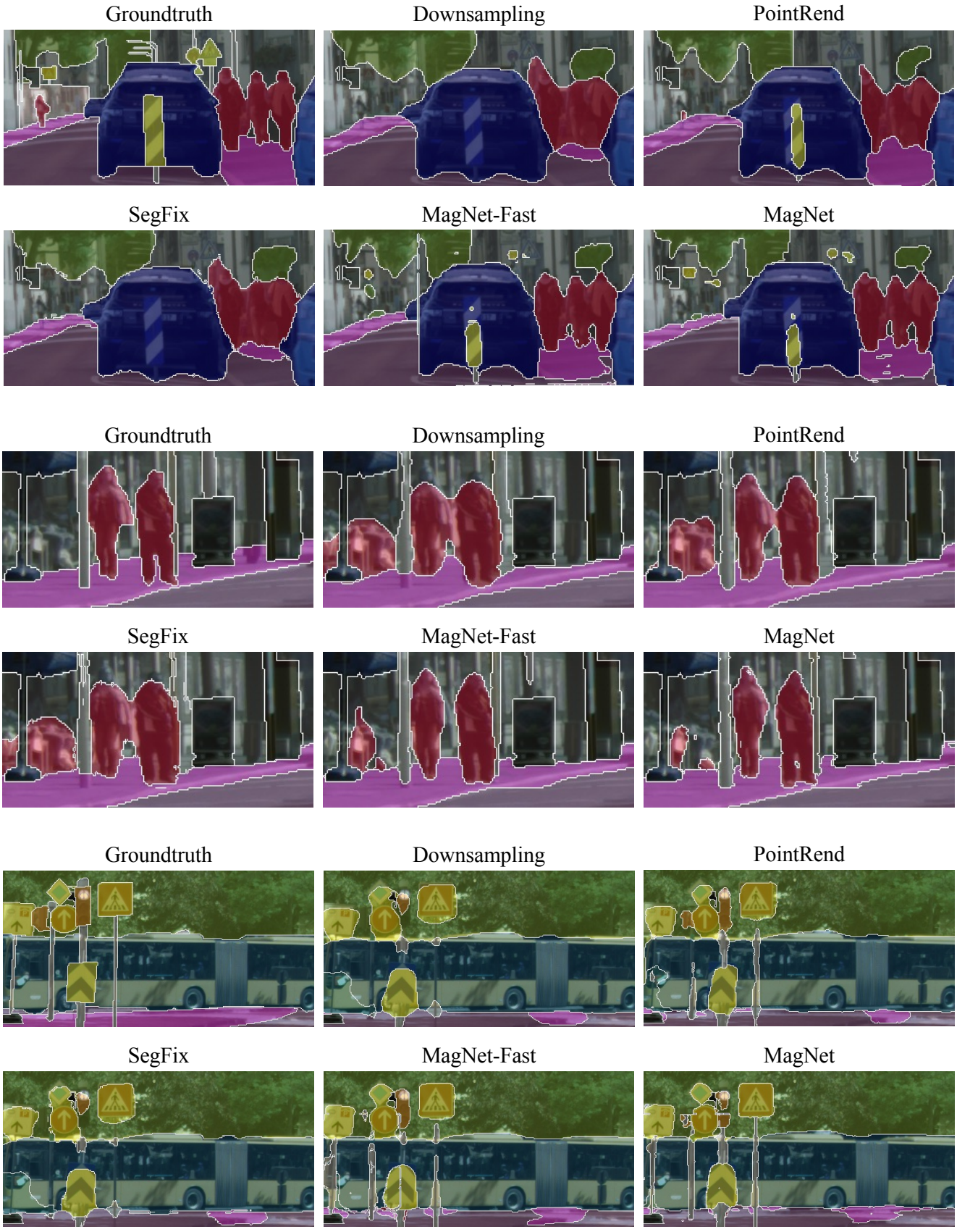

Figure 1: Our MagNet outperforms the other methods on the Cityscapes dataset. The MagNet framework successfully recognized small details. (Best view in color)

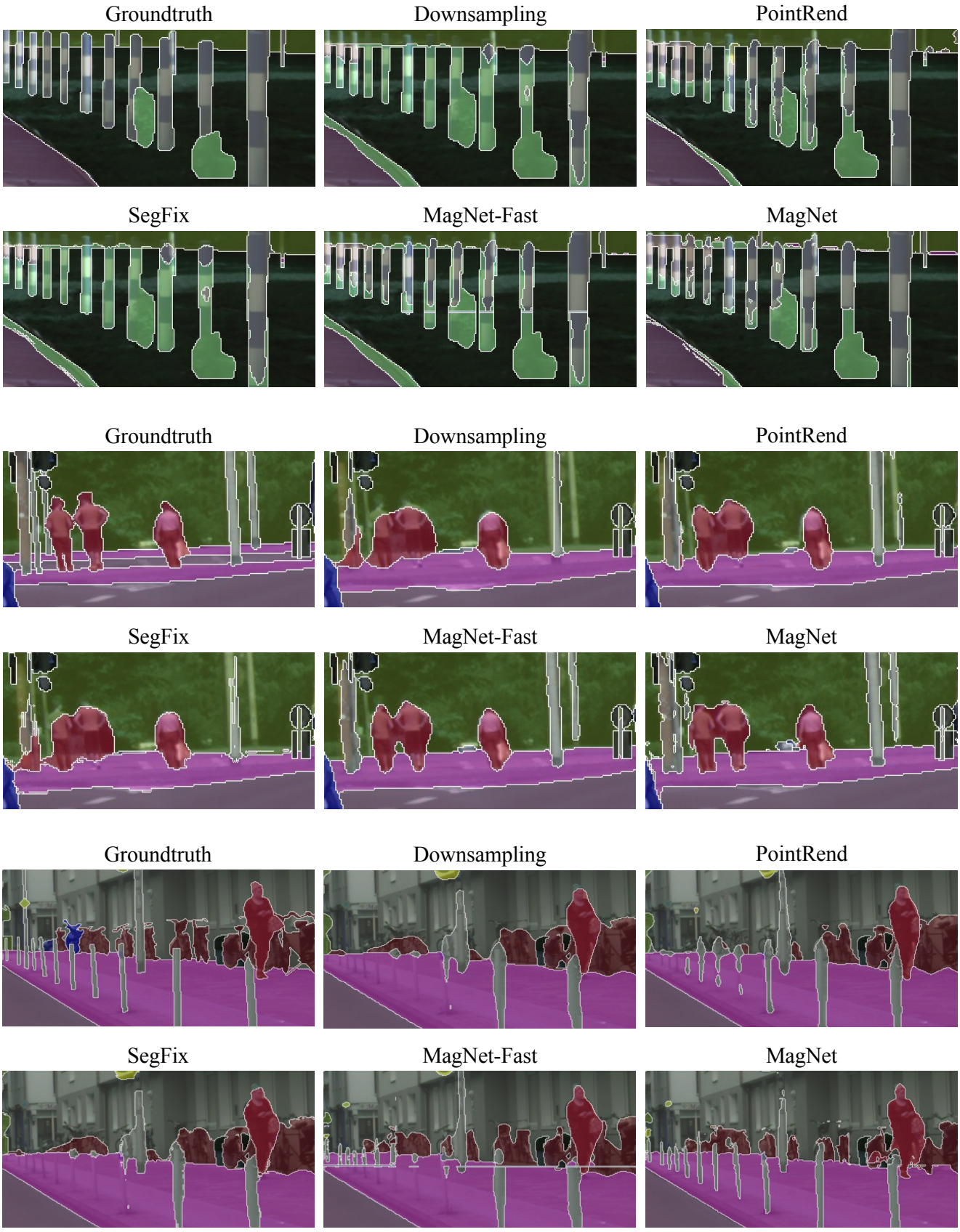

Figure 2: (cont.) Our MagNet outperforms the other methods on the Cityscapes dataset. The MagNet framework successfully recognized small details. (Best view in color)

from the Cityscapes dataset. With the initial segmentation of downsampled images, while PointRend can add a little number of details to the segmentation map, SegFix fails to add new tiny objects because this method focuses on boundary refinement only. Our MagNet produces finer prediction when it uses global context from downsampling and local details from patch processing.

2. DeepGlobe

2.1. Implementation details

The configuration of DeepLabV3+ and DenseCRF are the same as in the experiments on the Cityscapes dataset. With PointRend, the feature map of *conv2* was used to improve the segmentation map.

2.2. Additional results

Fig. 3 and Fig. 4 show additional results of state-of-the-art methods and our framework on the DeepGlobe dataset. As can be observed, our framework combines the segmentation of downsampling and patch processing in an accurate prediction and outperforms other methods.

3. Gleason

3.1. Implementation details

The same configuration was still applied to DenseCRF for the experiments on this dataset. With PointRend, the feature maps from the backbone Resnet-101 of PSPNet were used to refine the output.

3.2. Qualitative results

Fig. 5 and Fig. 6 illustrate some good predictions of our MagNet in the comparison with DenseCRF. Overall, our framework can fix errors of coarse segmentation better than the old approach. Patch processing cannot segment well because of the lack of global information.

3.3. Failed cases

Although working well in most of cases, our MagNet still fails to handle extreme cases that have large errors in coarse segmentation. Fig. 7 depicts some examples of these failed cases. Similar to DenseCRF, our network cannot fix those major errors completely.

4. Inrial Aerial

We also do one more experiment to compare the MagNet framework with GLNet [5] on the foreground-background segmentation task. Inrial Aerial [1], which contains 180 satellite images of resolution 5000 pixels, is used for this experiment. Each image is associated with a binary segmentation mask for the building locations in the image. There is a class imbalance between the building class and the background class. We trained and evaluated MagNet on this dataset with the same train, validation, and test splits used by GLNet [5], which have 127, 27, and 27 images respectively.

Method	mIoU(%)
Downsample	51.29
Patch processing	86.04
GLNet (reported in [5])	71.20
GLNet (our implementation)	67.73
MagNet	87.01

Table 3: Results on Inria Aerial. MagNet outperforms the other methods including GLNet. All methods use the same Resnet50-FPN backbone and input size of 536×536.

Because there is no implementation of GLNet on this dataset, we need to implement it by ourselves but cannot get the number reported in the paper. As can be observed in the Table 3, with this dataset, the patch processing approach can achieve very high accuracy and our MagNet adds 1% more to the IoU of patch processing. The low accuracy of GLNet can be explained by the domination of the downsampling branch.

References

[1] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229. IEEE, 2017. 4

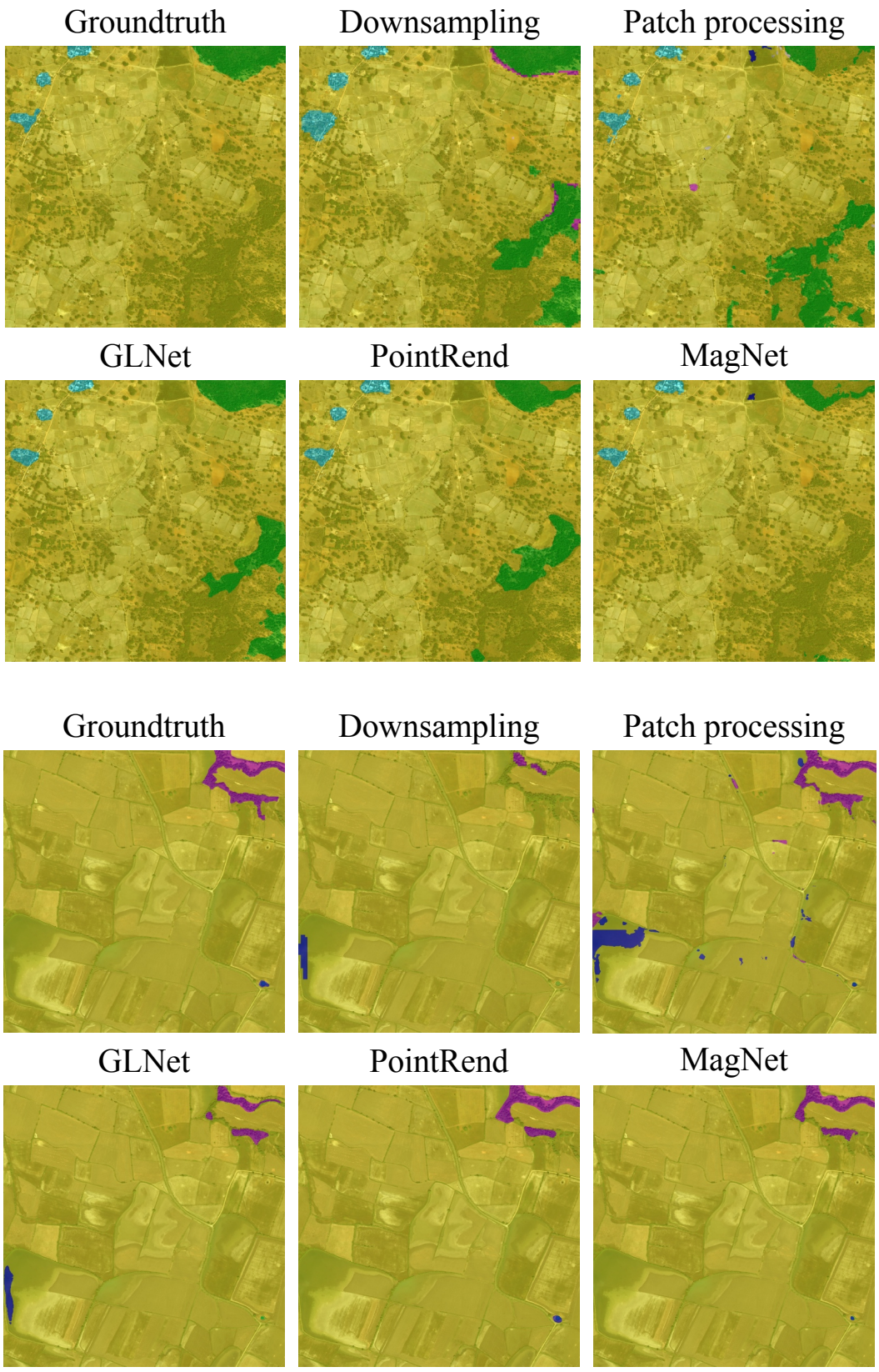

Figure 3: Some more results of MagNet on DeepGlobe dataset. Comparing to other state-of-the-art methods, our framework predicts more accurate segmentation. (Best view in color)

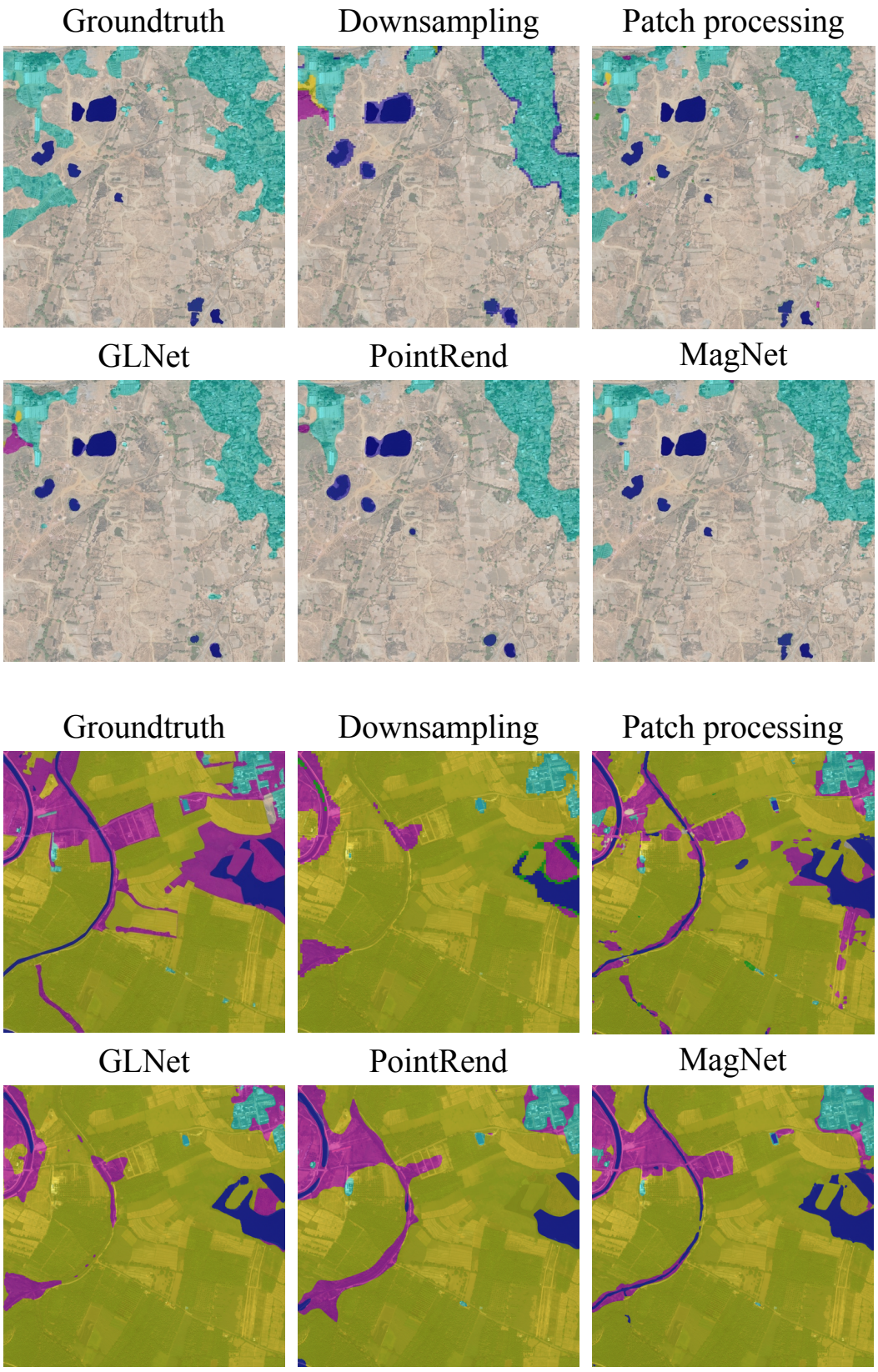

Figure 4: (cont.) Some more results of MagNet on DeepGlobe dataset. Comparing to other state-of-the-art methods, our framework predicts more accurate segmentation. (Best view in color)

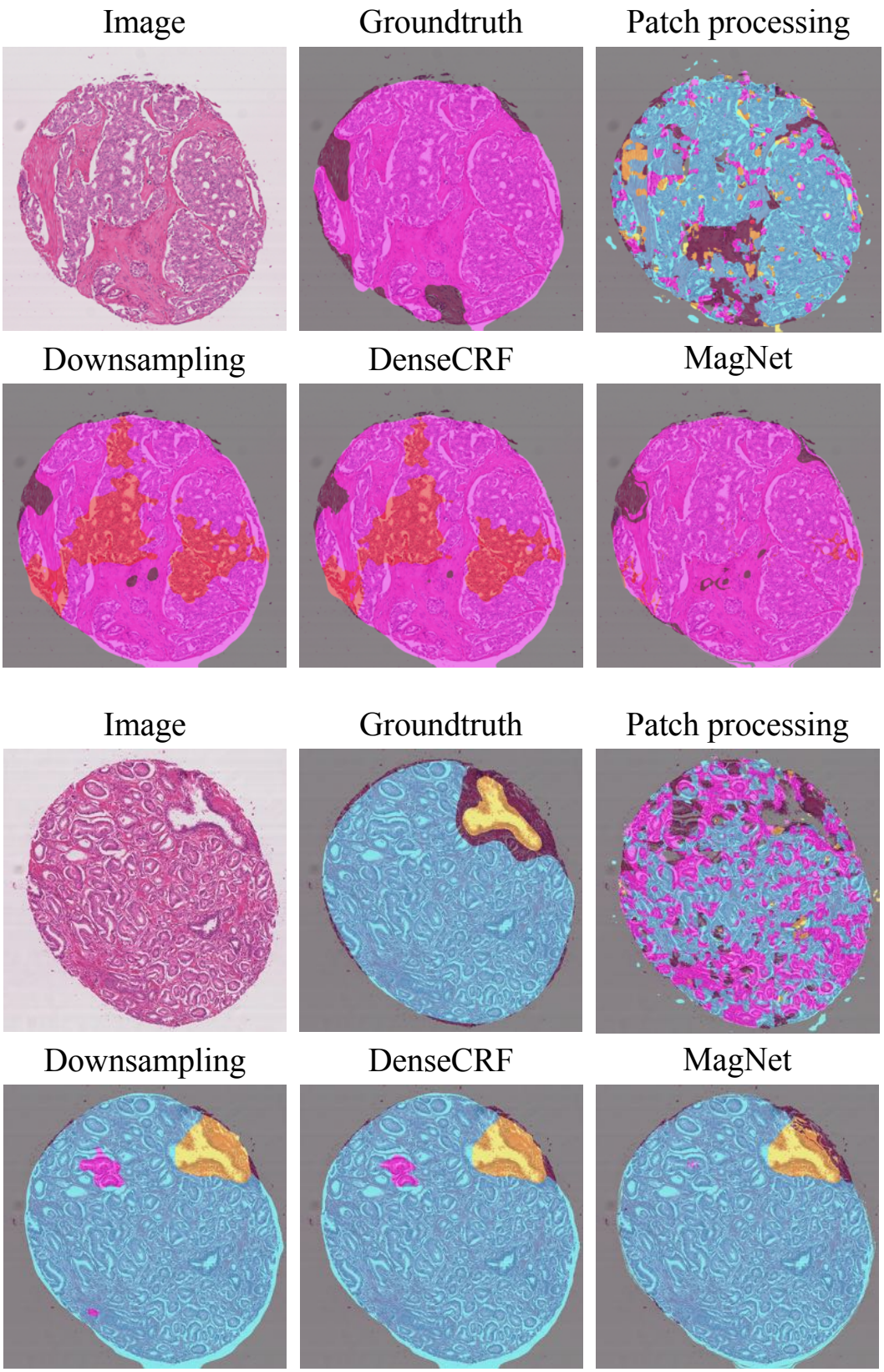

Figure 5: Some visualizations of our framework and other methods on Gleason. The MagNet can improve the coarse prediction by fixing some wrong classified regions. (Best view in color and zoom-in)

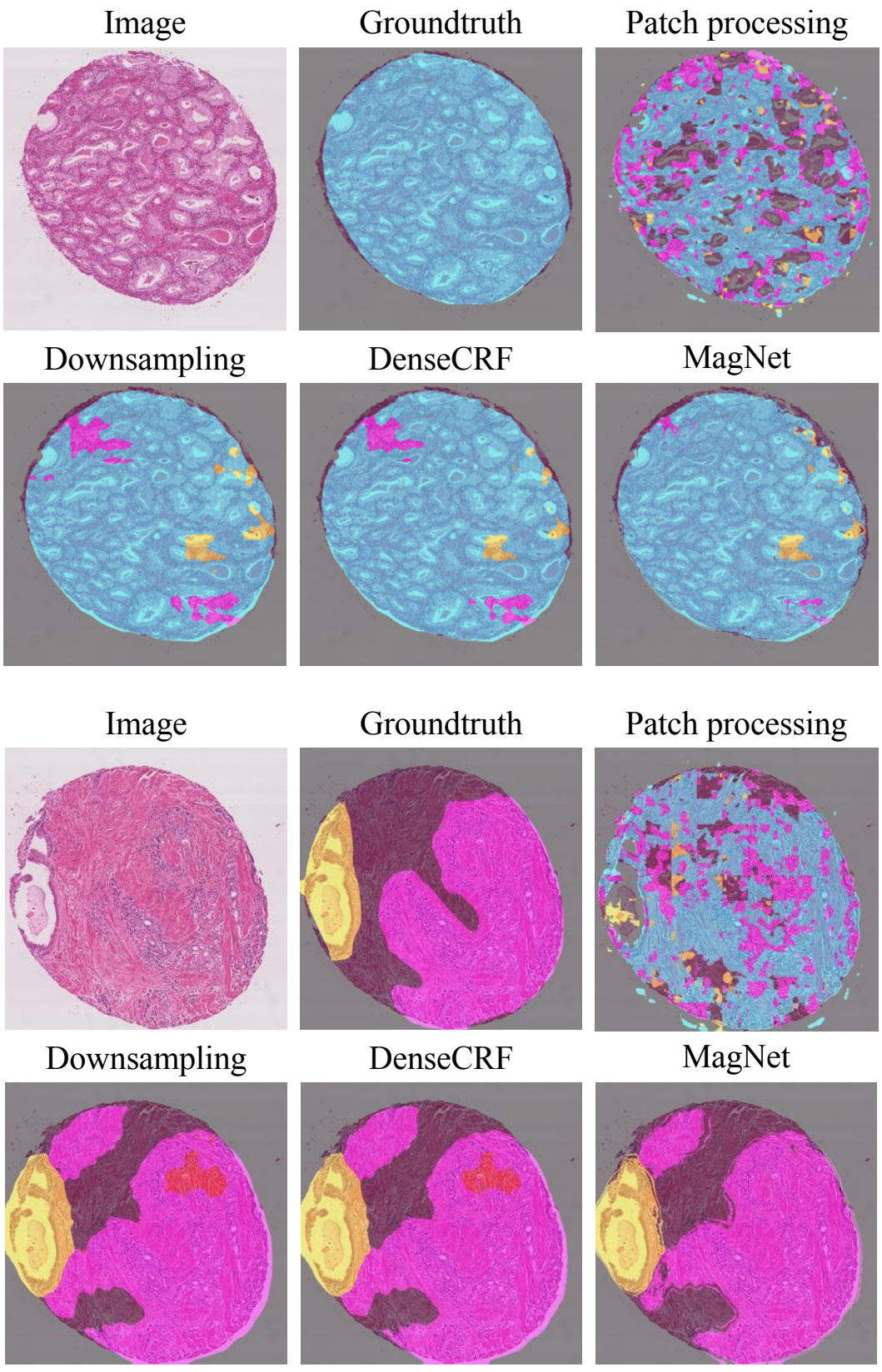

Figure 6: (cont.) Some visualizations of our framework and other methods on Gleason. The MagNet can improve the coarse prediction by fixing some wrong classified regions. (Best view in color and zoom-in)

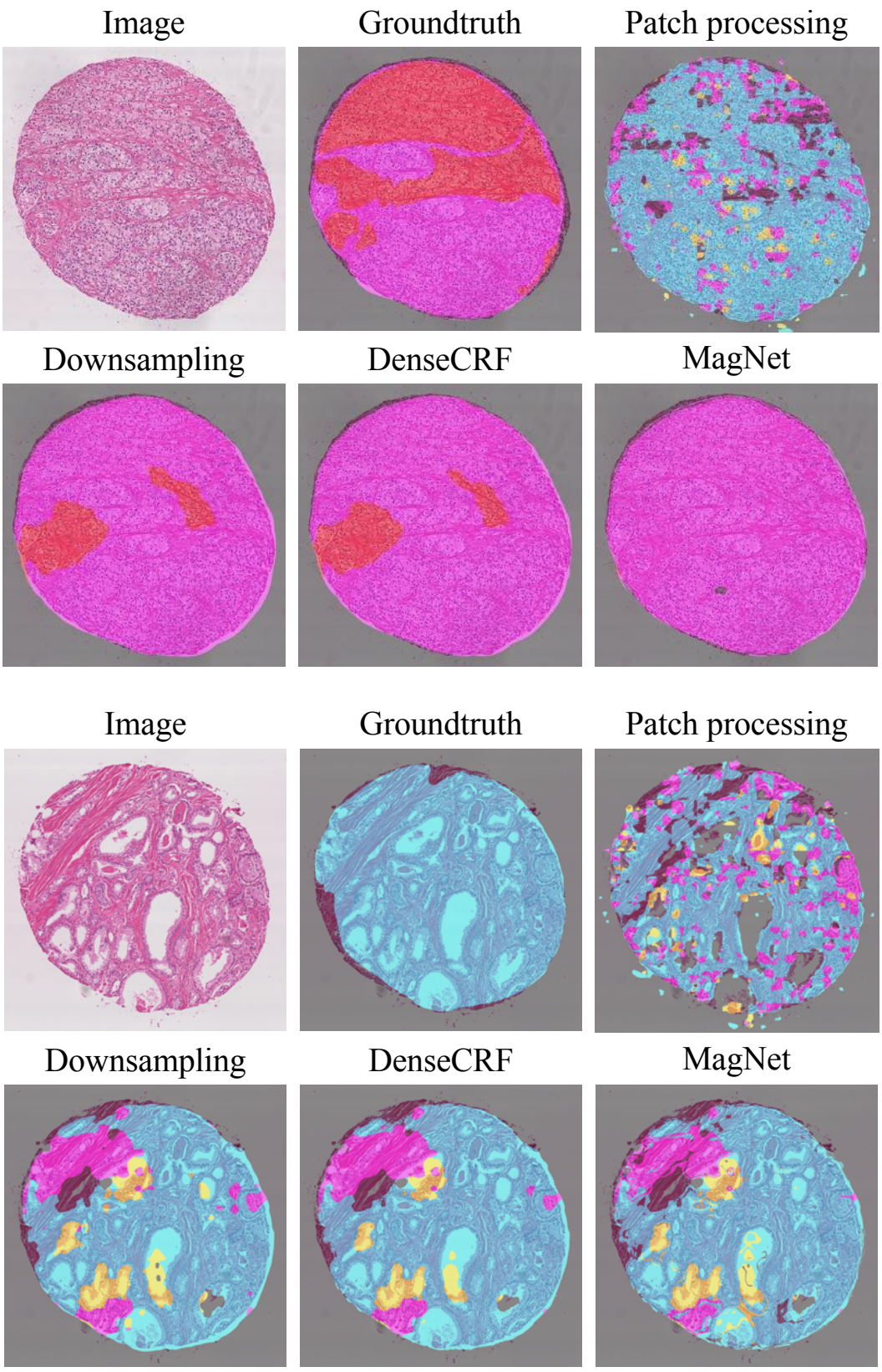

Figure 7: Some failures of our MagNet on Gleason dataset. Some major errors of the coarse segmentation cannot be fixed by our framework. (Best view in color)