Supplementary Materials for Heterogeneous Grid Convolution for Adaptive, Efficient, and Controllable Computation

The document explains 1) the full architectural specifications; 2) remaining details on the noise-canceling operation; 3) details on the differentiable clustering algorithm; 4) details on the importance map modulation with the application-specific attention maps; 5) datasets and metrics used in the experiments and 6) more results. Table 2 gives the list of hyper-parameters used in the experiments.

1. Full architectural specifications

Figure 1 provides the details of the four representative HG-CNN networks. For certain variants (specifically, HG-ResUNet18$^+$-Attn, HG-Orientation$^+$-Attn, and HG-ResUNet50-Attn), there exist some minor notes on the architecture and the following explains all the points.

HG-ResUNet18$^+$-Attn (road extraction): 1) The feature map from 4th residual stage of the encoder is skipped and concatenated with the output of HG-Conv modules on both of the encoder and decoder. (The concatenated features are further processed by 1×1 convolution layers.) 2) We reuse the same assignment matrix $\boldsymbol{S}$ at the decoder (i.e., the clustering is only performed on the encoder.).

HG-Orientation$^+$-Attn (road extraction): The major modification from the original model [1] is 1) hourglass module is replaced by multiple branches of small hourglass modules, 2) HG-Conv is applied on the encoder and decoders in each branch, and 3) active focus is introduced on the second stack of the refinement part.

HG-ResUNet50-Attn (salient object detection): The basic architecture is almost the same as HG-ResUNet18$^+$-Attn except several tiny differences as below; 1) The entry part is the same as the original ResNet (i.e., the stride of the first 7×7 convolution is 2, and the max-pooling layer follows the convolution), 2) the dilated convolution is used at the 3rd and 4th residual stages. 3) the auxiliary loss is eliminated when the active focus is not used (i.e., HG-ResUNet50).

Algorithm 1 Differentiable SLIC

Input: Input feature matrix $\boldsymbol{X}$; a set of cluster center coordinates $\boldsymbol{V}_0$; a set of input pixel coordinates $\boldsymbol{V}$; number of clusters m; and number of iterations N_{iter}

Output: Assignment matrix $\boldsymbol{S}$

$\boldsymbol{S}_0 \leftarrow \mathcal{F}_{nearest}(\boldsymbol{V}, \boldsymbol{V}_0)$ ▷ Initial assignment

for $t = 1$ **to** N_{iter} **do**

  $\bar{\boldsymbol{S}} \leftarrow \boldsymbol{Z}^{-1}\boldsymbol{S};\ Z_{jj} = \sum_i \boldsymbol{S}_{ij}$ ▷ Col-normalize

  $\boldsymbol{X}^c \leftarrow \bar{\boldsymbol{S}}^T \boldsymbol{X}$ ▷ Update cluster centers

  $\boldsymbol{S}_{ij} \leftarrow \exp\left\{-\| \boldsymbol{X}_i - \hat{\boldsymbol{X}}^c_j \|^2\right\}$ ▷ Update assignment

end for

$\boldsymbol{S}_{ij} \leftarrow \boldsymbol{S}_{ij}\, \mathcal{I}\left(\boldsymbol{S}_{ij} > 10^{-3}\right)$ ▷ Filtering

$\bar{\boldsymbol{S}} \leftarrow \boldsymbol{Z}^{-1}\boldsymbol{S};\ Z_{ii} = \sum_j \boldsymbol{S}_{ij}$ ▷ Row-normalize

return $\boldsymbol{S}$

2. Details on the noise canceling operation

The noise canceling operation has a few more post-processing steps on the group adjacency matrices $\hat{\boldsymbol{A}}^{\boldsymbol{\delta}}$. Firstly, small connection weights ($\hat{\boldsymbol{A}}^{\boldsymbol{\delta}}_{ij} < 10^{-7}$) are filtered out from the matrices (i.e., set to zero). Secondly, self-loop $\hat{\boldsymbol{A}}^{\circlearrowleft}$ is always reset to an identity matrix. Finally, the diagonal elements of the matrices $\hat{\boldsymbol{A}}^{\boldsymbol{\delta}}$ are set to zeros except for the self-loop adjacency matrix $\hat{\boldsymbol{A}}^{\circlearrowleft}$.

3. Details on the differentiable clustering

We use the differentiable SLIC algorithm [4] in our architecture with one enhancement. SLIC samples initial cluster centers by uniform sampling. We use an adaptive algorithm from a prior work [5] for further improvements. The complete process is given in Algorithm 1. The hyper-parameters of the algorithms are given in Table 2.

4. Details on the importance map modulation

Active focus modulates the importance map based on an application-specific attention map via simple weighted averaging. The section provides the full formula.

Let $\boldsymbol{C}_{imp}$ denotes the $(H \times W)$ importance map and

$\boldsymbol{C}_{attn}$ denotes the $(H \times W)$ attention map with its elements $(\boldsymbol{C}_{attn})_{ij}$ ranges from 0 to 1. The importance map is updated as follows:

$$\boldsymbol{C}_{focus} = \frac{1}{Z}\boldsymbol{C}_{imp} + \alpha \boldsymbol{C}_{attn}. \quad (1)$$

$Z = \max\{\boldsymbol{C}_{imp}[i,j]\}$ is a normalization coefficient. α(=10) is a weight coefficient for the attention map.

As introduced in the main paper, we define two types of attention maps: object-aware and uncertainty-aware. Let $\boldsymbol{P} = \{P_i | i = 1 \cdots K\}$ denote the K class prediction map of shape $(K \times H \times W)$. Object-aware active focus sets the attention map to P_k, where we are interested in an object in the k_{th} class.

$$\boldsymbol{C}_{attn} = \boldsymbol{P}_k. \quad (2)$$

This form of attention focuses the cluster centers around the target class k.

Uncertainty-aware active focus computes the attention map as the entropy of the probability map.

$$\boldsymbol{C}_{attn} = \frac{1}{\log K}\sum_k \{-\boldsymbol{P}_k \log \boldsymbol{P}_k\}. \quad (3)$$

The above attention focuses the cluster centers on the regions where the model is uncertain for the prediction.

5. Datasets and metrics

Semantic segmentation: To evaluate the proposed method, we use three semantic segmentation datasets, Cityscapes [2], ADE20K [16], and PASCAL-context [8]. For all the experiments, we use validation sets to evaluate the models.

• **Cityscapes** is a dataset for urban scene parsing, which contains 5,000 images of resolution $1,024 \times 2,048$ with fine pixel-wise annotations. The annotations have 30 classes. We use the major 19 classes by following a prior convention. The dataset has the training, validation, and testing sets with 2,975/500/1,525 images, respectively. Only fine annotations are used for training.

• **ADE20K** is a dataset for the ILSVRC2016 Scene Parsing Challenge, which contains more than 20K annotated images of natural scenes. The annotations are provided for 150 semantic categories such as objects and object-parts. The dataset has training, validation, and testing sets with 20K/2K/3K images, respectively.

• **PASCAL-context** is a scene parsing dataset with 59 classes and one background class, which consists of 4,998 training images and 5,105 validation images. Following previous works, we evaluated our models on the 59 classes and excluded background class.

Road extraction: We evaluate our method on DeepGlobe dataset [3], which consists of satellite images and corresponding pixel-wise road mask annotations. The images has 50 cm spatial resolution and pixel size of $1,024 \times 1,024$. Following [1], we split the dataset into training and validation with 4,696 and 1,530 images, where the performance is evaluated on road IoU and APLS metrics [10]. The APLS metric measures similarity between a predicted and a ground truth road graph.

Salient object detection: Following previous works [12, 15], we train our models on DUTS [11] dataset, and evaluate the models on ECSSD [13], PASCAL-S [7], DUT-OMRON [14], HKU-IS [6], SOD [9], and DUTS [11]. For evaluation metrics, mean absolute error (MAE) and maximum F-measure (maxF) are used as in prior works.

6. More results

Table 1 shows the per-class segmentation performance gain achieved by HG-Conv for Cityscapes dataset. We see that the HG-Conv performs well on both small and large objects.

Figures 2, 3, 4, 5, 6, and 7 show additional experimental results for the same problems discussed in the paper.

References

[1] Anil Batra, Suriya Singh, Guan Pang, Saikat Basu, and C V Jawahar Manohar. Improved Road Connectivity by Joint Learning of Orientation and Segmentation. *CVPR*, 2019. 1, 2

[2] 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. *CVPR*, 2016. 2

[3] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raska. DeepGlobe 2018: A challenge to parse the earth through satellite images. *CVPRW*, 2018. 2

[4] Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel Sampling Networks. *ECCV*, 2018. 1

[5] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image Segmentation As Rendering. *CVPR*, 2020. 1

[6] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. *CVPR*, 2015. 2

[7] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. *CVPR*, 2014. 2

[8] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam Gyu Cho, Seong Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. *CVPR*, 2014. 2

[9] Vida Movahedi and James H. Elder. Design and perceptual validation of performance measures for salient object segmentation. *CVPRW*, 2010. 2

Table 1. Per-class performance gain achieved by HG-conv for Cityscapes dataset.

Base network	road	s. walk	build.	wall	fence	pole	t-light	t-sign	veg	terrain	sky	person	rider	car	truck	bus	train	m-cycle	bicycle
ResNet101-Dilation	+0.2%	+1.2%	+0.3%	+11.0%	+1.6%	+0.8%	+0.3%	+0.2%	+0.2%	-0.5%	+0.2%	+0.4%	+2.5%	+0.3%	+10.0%	+7.6%	+27.6%	+1.2%	+0.6%
ResNet101-DCN	-0.2%	-0.8%	+0.3%	+12.0%	+2.4%	-0.1%	-0.3%	-0.1%	0.0%	-0.9%	0.0%	+0.4%	+0.6%	+0.4%	+14.9%	-0.3%	-7.3%	+0.9%	+0.1%
HRNet-W48	0.0%	+0.4%	0.0%	+1.3%	+0.4%	0.0%	0.0%	+0.5%	+0.1%	+1.2%	-0.1%	+0.1%	-0.2%	+0.5%	+9.6%	+1.4%	-0.8%	+0.3%	-0.3%

Table 2. List of hyperparameters used in the experiments.

Dataset	Base Network	HG-Conv	Hyperparemeters for HG-Conv				Other hyperparameters						
			#iter clustering	down ratio	sampling	aux coef	input size	train iter	batch size	optimizer	initial LR	weight decay	lr schedule
Cityscapes	ResNet	non-HG	-	-	-	0.4	713x713	60K	16	SGD	1.00E-02	1.00E-04	poly p=0.9
		HG (DCN)	3	1/64	top-k+random k=7, b=0.75	0.4	713x713	30K	16	Adam	1.00E-04	1.00E-04	poly p=0.9
		HG (Dilation)	3	1/64	top-k+random k=7, b=0.75	0.4	713x713	30K	16	Adam	1.00E-04	1.00E-04	poly p=0.9
	HRNetV2	non-HG	-	-	-	-	512x1024	484epochs	16	SGD	1.00E-02	5.00E-04	poly p=0.9
		HG	1	1/64	top-k+random k=7, b=0.75	0.1	512x1024	484epochs	16	SGD	1.00E-02	5.00E-04	poly p=0.9
ADE20K	ResNet	non-HG	-	-	-	0.4	473x473	125K	16	SGD	1.00E-02	1.00E-04	poly p=0.9
		HG (DCN)	3	1/64	top-k+random k=7, b=0.75	0.4	473x473	90K	16	Adam	1.00E-04	1.00E-04	poly p=0.9
		HG (Dilation)	3	1/64	top-k+random k=7, b=0.75	0.4	473x473	45K	16	Adam	1.00E-04	1.00E-04	poly p=0.9
	HRNetV2	non-HG	-	-	-	-	512x1024	120epochs	16	SGD	2.00E-02	1.00E-04	poly p=0.9
		HG	1	1/64	top-k+random k=7, b=0.75	0.1	512x1024	120epochs	16	SGD	2.00E-02	1.00E-04	poly p=0.9
PASCAL context	ResNet	non-HG	-	-	-	0.4	520x520	30K	16	SGD	1.00E-02	1.00E-04	poly p=0.9
		HG (DCN)	3	1/64	top-k+random k=7, b=0.75	0.4	520x520	30K	16	Adam	1.00E-04	1.00E-04	poly p=0.9
		HG (Dilation)	3	1/64	top-k+random k=7, b=0.75	0.4	520x520	30K	16	Adam	1.00E-04	1.00E-04	poly p=0.9
	HRNetV2	non-HG	-	-	-	-	512x1024	200epochs	16	SGD	1.00E-02	1.00E-04	poly p=0.9
		HG	1	1/64	top-k+random k=7, b=0.75	0.1	512x1024	200epochs	16	SGD	1.00E-02	1.00E-04	poly p=0.9
Road Extraction	ResUNet	non-HG	-	-	-	-	256x256	120epochs	32	Adam	1.00E-04	1.00E-04	poly p=0.9
		HG	1	1/16	top-k+random k=7, b=0.75	0.4	256x256	120epochs	32	Adam	1.00E-04	1.00E-04	poly p=0.9
	Orientation	non-HG	-	-	-	-	256x256	120epochs	32	SGD	1.00E-02	5.00E-04	step $\gamma = 0.1$ 60, 90, 110epochs
		HG	1	1/4 1/16 1/64	top-k+random k=4, b=0.75) k=16, b=0.75 k=64, b=0.75	-	256x256	120epochs	32	SGD	1.00E-02	5.00E-04	step $\gamma = 0.1$ 60, 90, 110epochs
Salient Object Detection	ResUNet	non-HG	-	-	-	-	352x352	24epochs	16	Adam	1.00E-04	1.00E-04	poly p=0.9
		HG	1	1/4	top-k+random k=4, b=0.75	0.4	352x352	24epochs	16	Adam	1.00E-04	1.00E-04	poly p=0.9

[10] Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. SpaceNet: A Remote Sensing Dataset and Challenge Series. *CoRR*, 2018. 2

[11] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. *CVPR*, 2017. 2

[12] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. *ICCV*, 2019. 2

[13] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. *CVPR*, 2013. 2

[14] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming Hsuan Yang. Saliency detection via graph-based manifold ranking. *CVPR*, 2013. 2

[15] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and Balance: A Simple Gated Network for Salient Object Detection. *ECCV*, 2020. 2

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. *CVPR*, 2017. 2

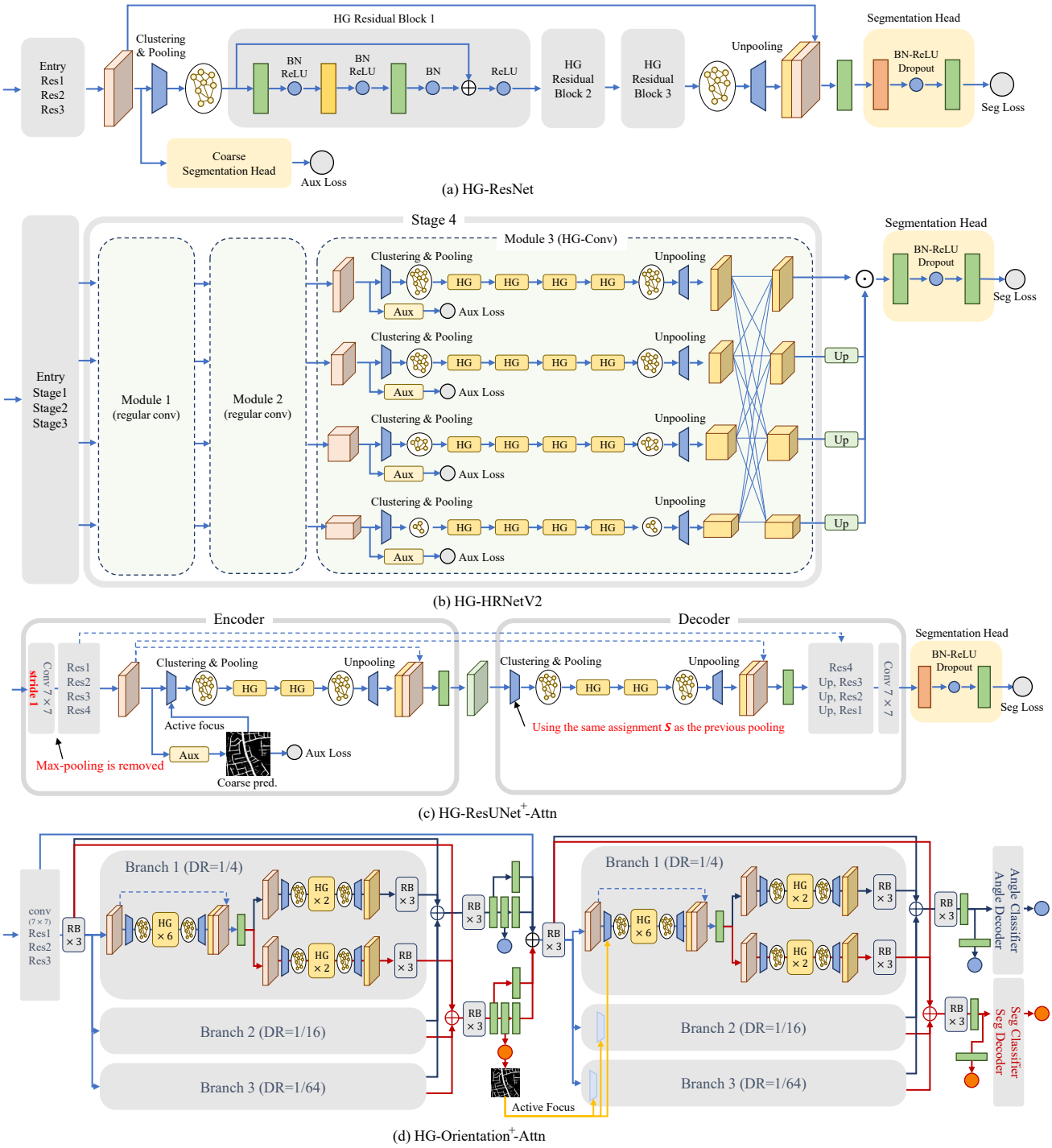


Figure 1. Architecture specifications for the four representative HG-CNN networks.

Input GT ResNet101-DCN HG-ResNet101-DCN

Figure 2. Prediction results of semantic segmentation.

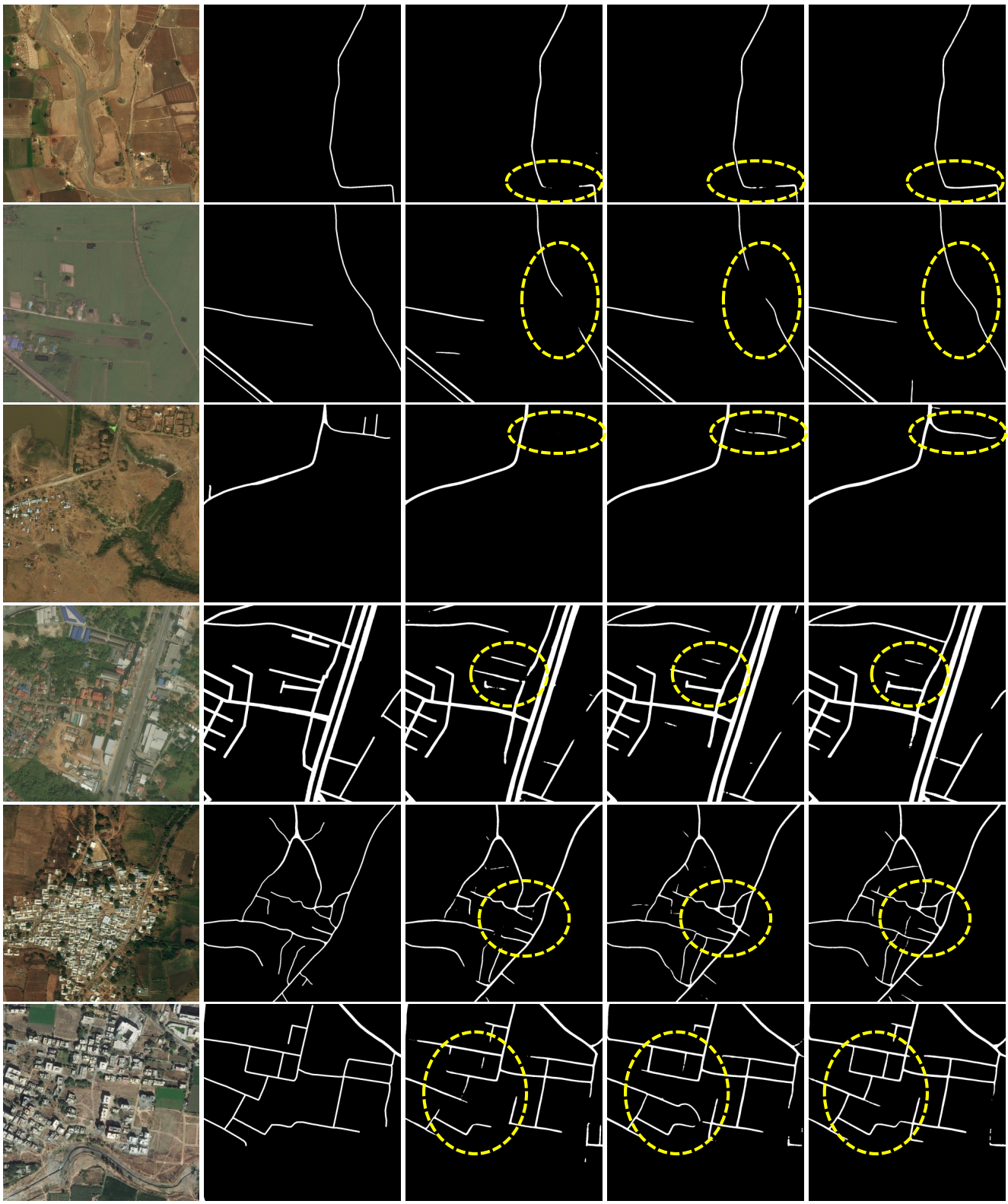


Figure 3. Prediction results of road extraction.

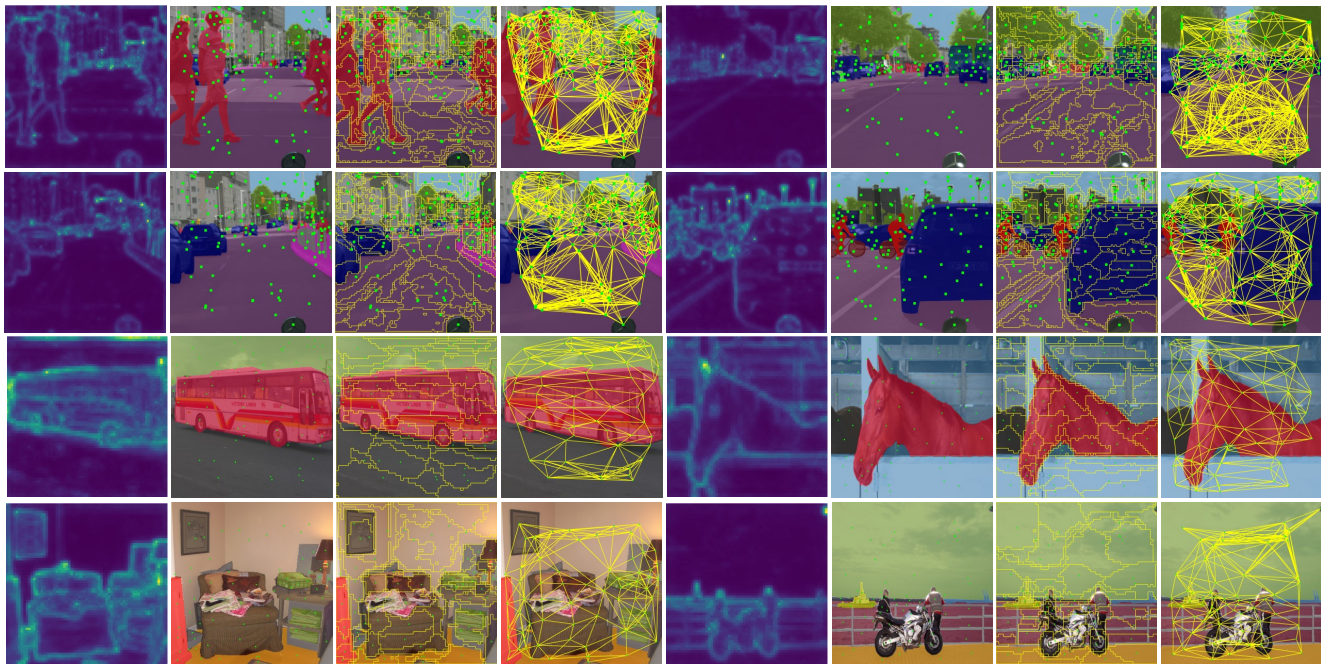


Figure 4. Visualization of HG-Conv on semantic segmentation. From left to right, the figures show, importance map, cluster center allocation, clustering result, and adjacency connection (same order below). The adjacency connection shows the summed up connection for all the direction-wise adjacency.

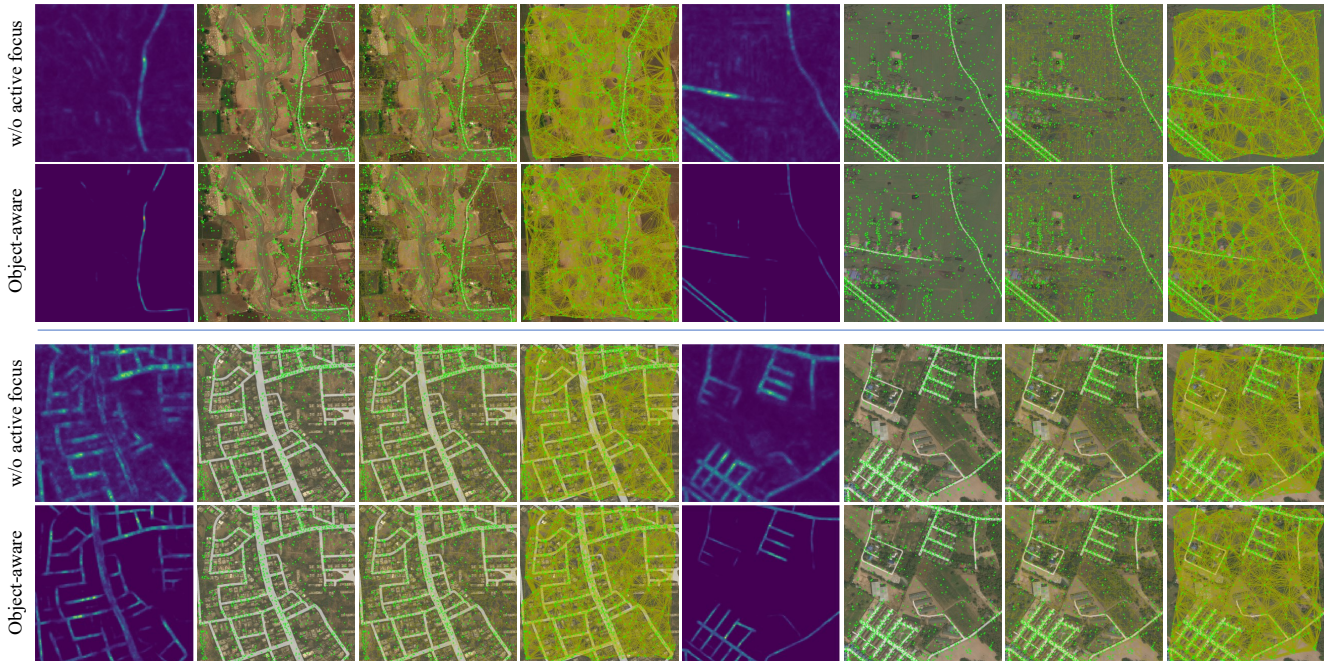
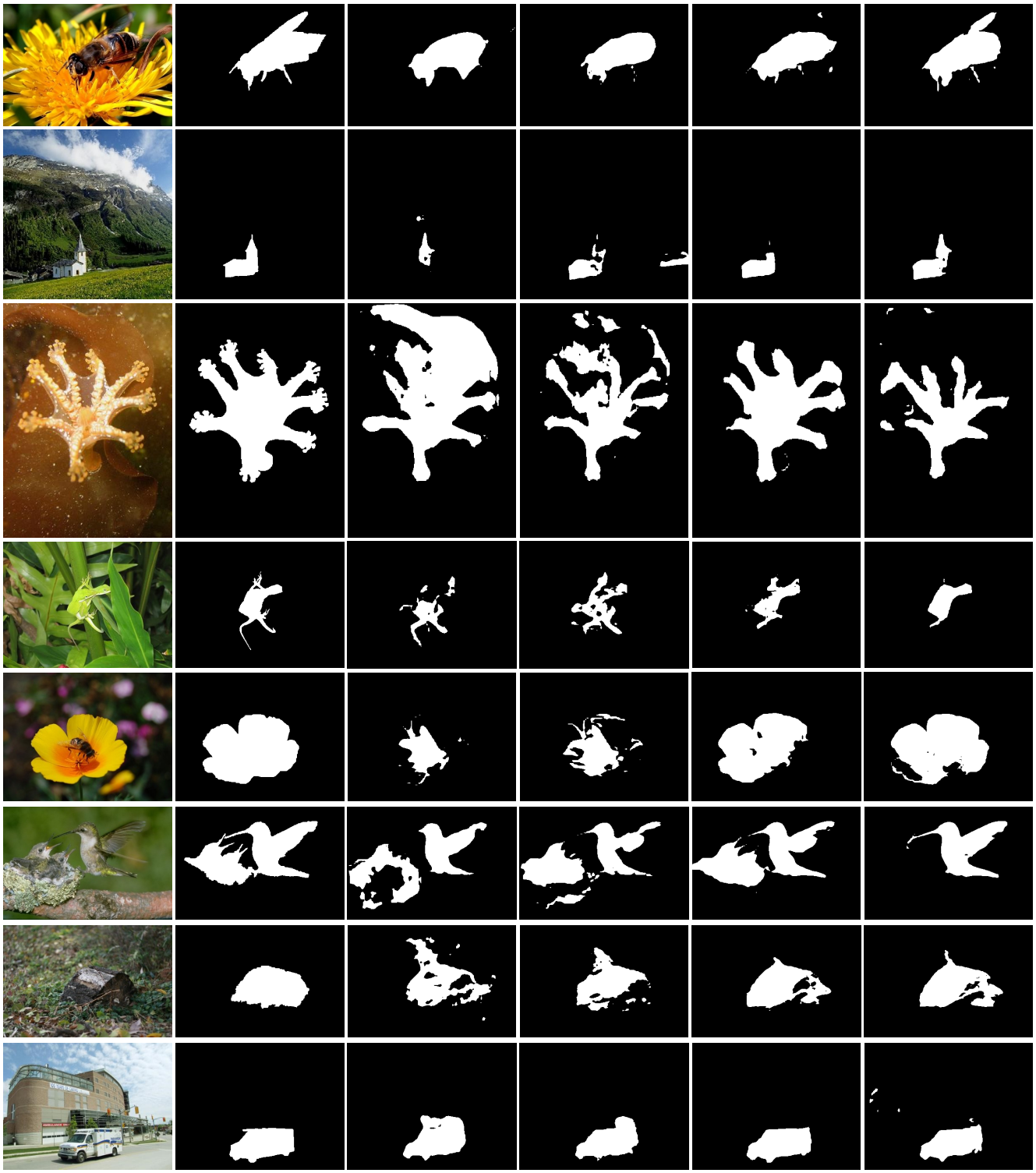


Figure 5. Visualization of HG-Conv on road extraction. From left to right, the figures show, importance map, cluster center allocation, clustering result, and adjacency connection (same order below). The adjacency connection shows the summed up connection for all the direction-wise adjacency.



Input | GT | ResUNet50 | HG-ResUNet50 | HG-ResUNet50-Attn (object-aware) | HG-ResUNet50-Attn (uncertainty-aware)

Figure 6. Prediction results of salient object detection

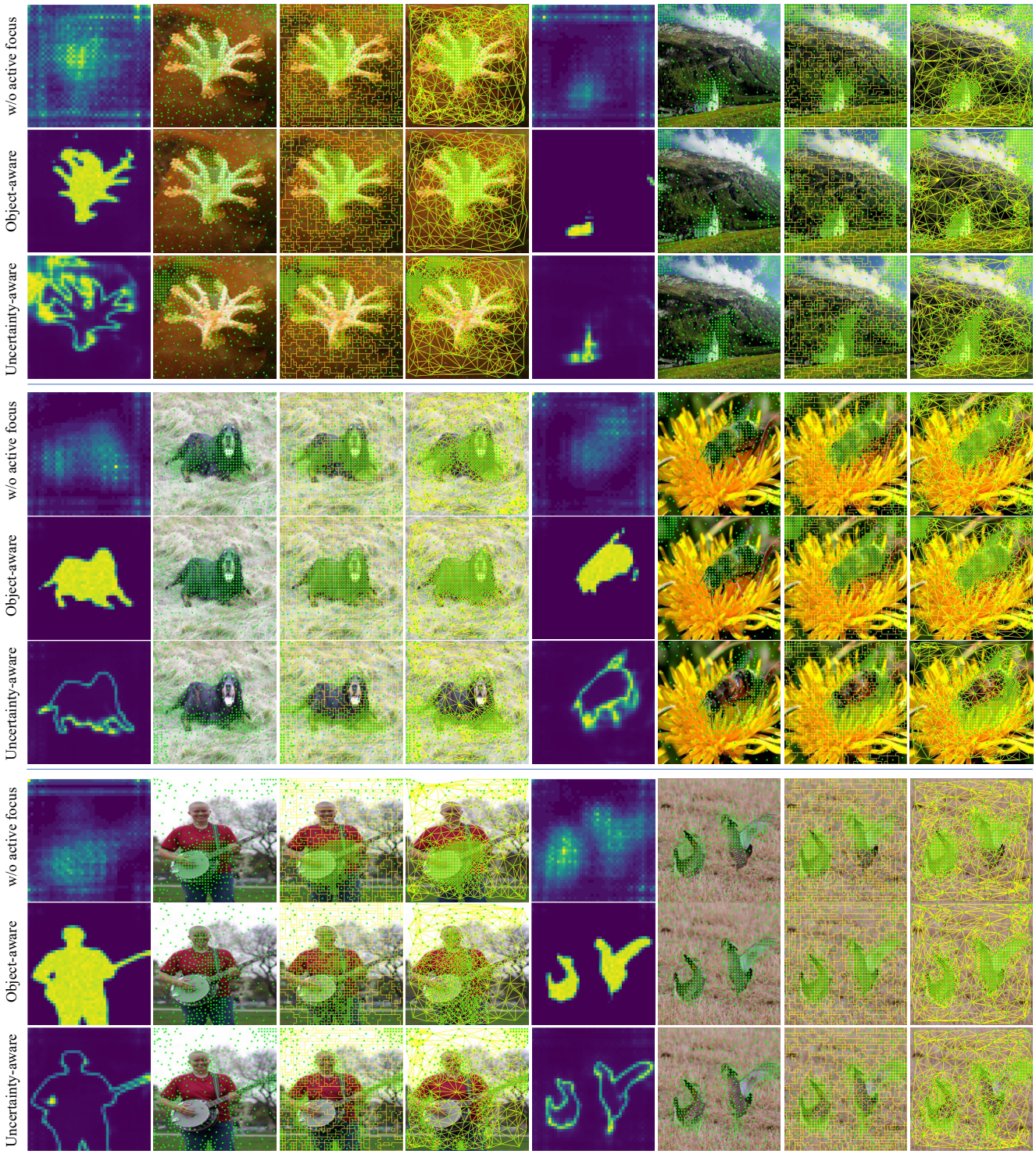


Figure 7. Visualization of HG-Conv on salient object detection. From left to right, the figures show, importance map, cluster center allocation, clustering result, and adjacency connection (same order below). The adjacency connection shows the summed up connection for all the direction-wise adjacency.