1. The results on the PASCAL VOC 2012 *test* set. They are shown in Tab. 1. With DeepLabv3 and DeepLabv3+, the improvements are 2.00% and 1.93% respectively. Models trained with RMI show better generalization ability on *test* set than on *val* set. The results are supposed to be in Tab.2 in the main paper. However, our submissions are stuck in the official server (too much submissions or some other reasons). You can check the link of DeepLabv3&CE (`http://host.robots.ox.ac.uk:8080/anonymous/ERHL10.html`) to verify our words – the submission date is 2019-05-23 and we receive the results after 4 days (the time we received the reminder mail). Some earlier attempts on *val* set are also stuck. Furthermore, the links of DeepLabv3&RMI and DeepLabv3+&RMI are `http://host.robots.ox.ac.uk:8080/anonymous/2RBVFL.html` and `http://host.robots.ox.ac.uk:8080/anonymous/SC5YIQ.html`.

Table 1: Per-class results on the PASCAL VOC 2012 *test* set.

Method		backg.	aero.	bike	bird	boat	bottle	bus	car	cat	chair	cow	d.table	dog	horse	mbike	person	p.plant	sheep	sofa	train	tv	mIoU (%)
DeepLabv3	CE	94.10	79.58	41.16	84.67	67.68	75.09	87.69	87.40	92.07	39.66	83.39	69.68	86.67	87.10	86.92	84.39	65.69	86.66	57.39	75.28	75.94	76.58
	CRF-5	**94.60**	84.28	**41.83**	88.00	68.81	76.56	87.69	87.90	**93.79**	40.35	84.92	**70.26**	**88.84**	**89.22**	87.39	85.73	67.45	87.95	58.80	75.31	77.38	77.96
	RMI	94.57	**84.77**	41.67	**89.99**	**69.11**	**77.86**	**90.02**	**90.17**	93.14	**42.97**	**85.70**	64.74	87.45	86.63	**88.25**	**87.04**	**68.78**	**90.42**	**59.13**	**79.67**	**78.05**	**78.58**
DeepLabv3+	CE	94.37	90.03	42.40	82.07	70.46	75.77	93.36	88.07	90.70	36.50	86.50	67.17	86.04	90.18	87.23	85.02	68.36	88.46	57.34	84.13	78.62	78.23
	CRF-1	94.57	**92.13**	42.48	83.25	71.07	**76.61**	93.47	87.96	91.45	36.82	87.04	67.21	87.28	90.87	**87.63**	85.86	69.22	89.23	**58.04**	**84.43**	**79.46**	78.86
	RMI	**94.97**	91.57	**42.93**	**93.72**	**74.84**	76.23	**93.68**	**89.09**	**93.59**	**41.99**	**87.63**	**68.79**	**88.23**	**91.33**	87.12	**88.62**	**70.24**	**92.00**	57.77	82.53	76.60	**80.16**

2. The qualitative results. They are shown in Fig. 1. The lack of the these results in the main paper is due to the limit of paper length. It is clear that the predictions of DeepLabv3+&RMI have more accurate boundaries and richer details.

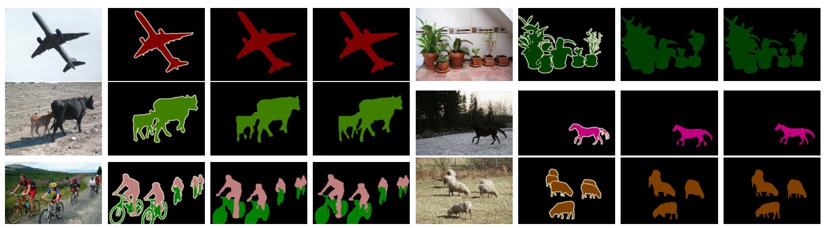

Figure 1: The qualitative results on PASCAL VOC 2012 *val* set. **Best view in color with 400% zoom**.

3. The significance of RMI. PASCAL VOC dataset is well-studied and DeepLabv3+ is still the best model on it (`http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?challengeid=11&compid=6`). The improvement gained by RMI with a well-developed model can definitely demonstrate its effectiveness. ResNet101-DeepLabv3 have a 77.21% mIoU on *val* set [6, Tab. 5], and ResNet101-DeepLabv3+ obtain a 78.85% mIoU (add a decoder based on DeepLabv3) [7, Tab. 3]. In contrast, with the same settings, the best mIoU of ResNet101-DeepLabv3&RMI is 79.09% (Tab. 3a in the main paper). Furthermore, RMI can get consistent improvements with DeepLabv3+. It worth noticed that the top performance of DeepLabv3+ also comes from other paralleled aspects, *e.g.*, small output stride, more powerful backbone network (Xception), COCO&JFT pretraining, and multi-scale&flipping evaluation. They cost more computational resources.

Besides above, RMI provides a practical method to estimate the mutual information through statistics of the data when their corresponding distributions are unknown. Moreover, the idea of RMI is not restricted in semantic segmentation. It can be applied in many other structured output tasks (like some image-to-image tasks), and this is our future work.

4. Some other questions. Q (Reviewer #1): The influence of λ in Eq.(16). **A:** Following the SSIM index [37], the importance of pixel similarity and structure similarity is equal, so we simply set $\lambda = 0.5$. We further study the influence of λ on VOC *val* set with DeepLabv3: {0.1, 0.3, 0.5, 0.7, 0.9} – {77.49, 77.88, **78.71**, 78.50, 77.40}(%, mIoU).

Q (Reviewer #2): Assumption (11), training set up, implementation, other tricks, and other base models. **A: (1)**. Given assumption (11), we can calculate an approximate value of $\boldsymbol{I}_l(\boldsymbol{Y};\boldsymbol{P})$, and the difference between estimated $\boldsymbol{I}_l(\boldsymbol{Y};\boldsymbol{P})$ and real value of $\boldsymbol{I}_l(\boldsymbol{Y};\boldsymbol{P})$ is restricted in certain range by Theorem 3.1. This assumption is discussed in a more general way in [14]. We are standing on giants' shoulders, check [14] for more details. **(2)**. Training set up is clear in Sec 5.1. We keep the set up same as [6, 7] for fair comparisons, and all models in our paper are training end-to-end with a certain loss following the same routine. No fine-tune and pretrained DeepLab models. **(3)**. It is discussed in Line.149-162. Code will also be public. **(4)**. Atrous convolution and various up-sampling modules are already used in DeepLabv3 and DeepLabv3+ (ASPP and Encoder-Decoder design). Check [5, 6, 7] for more details. **(5)**. According to [6, 7], DeepLabv3 and DeepLabv3+ follow two different design principles – the former is plain and the latter is Encoder-Decoder based. Nevertheless, we provide results of PSPNet on VOC *val* set: CE (77.58%), RMI (**78.63%**).

Q (Reviewer #3): mixture of RMI/CRF/Affinity, "pyramid" loss. **A: (1)**. CRF shows minor improvments when base model is powerful enough (Tab. 1b) and negative effect on CamVid (Tab. 4). Affinity loss shows negative effect on PASCAL VOC. In contrast, the improvements of RMI are consistent, so we think it is unnecessary to mix them up. **(2)**. Common downsampling methods may produce the same interpolated result form different regions, so "pyramid" loss is not locality-aware. We examed this idea with DeepLabv3 on VOC *val* set: CE (**77.14%**) , "pyramid" loss (76.11%, simply averaged over scales [0.25, 0.5, 0.75, 1.0]). The "pyramid" loss shows negative effect.