Appendix

Resolution	Backbone	Model	Speed(ms)
Varied Clothing Dataset			
800× 1333	R50	Mask R-CNN	92
		Mask R-CNN-IMP	94
		Semantic-P2	76
		Semantic-FPN	103
		Panoptic-P2	109
		Panoptic-FPN	110
		Panoptic-P2-IMP	110
		Panoptic-FPN-IMP	111
ModaNet			
600×1000	R50	Panoptic-FPN-IMP	72
	R101	Panoptic-FPN-IMP	87
Cityscapes			
1024×2048	R50	Mask R-CNN	151
		Panoptic-FPN	194
		Panoptic-FPN-IMP	195
	R101	Panoptic-FPN-IMP	243
	X101	Panoptic-FPN-IMP	401

Table 6: Speed performance analysis. In this table, we show the speed performance for each model. For simplicity, we use the following abbreviations:**R50**:ResNet-50. **R101**:ResNet-101. **X101**:ResNeXt-101

Varied Clothing DatasetClasses

Class	Super Class	# Train	# Val	Area(x^2)
Hair	Body	7,260	635	192
Skin	Body	34,795	3,074	119
Top/T-shirt	G-Top	4,364	424	221
Sweater/Cardigan	G-Top	1,906	148	266
Jacket/Blazer	G-Top	2,360	183	261
Coat	G-Top	1,597	161	279
Shirt/Blouse	G-Top	2,650	244	229
Vest	G-Top	266	20	220
Pants/Jeans	G-Bottom	2,763	217	261
Tights/Leggings	G-Bottom	930	116	214
Shorts	G-Bottom	532	60	203
Socks	G-Bottom	803	80	174
Skirt	G-Bottom	1,281	114	262
Dress	G-Whole	2,728	241	340
Jumpsuit	G-Whole	273	31	370
Shoes	Footwear	6,619	591	118
Boots	Footwear	1,801	109	142
Hat/Headband	Accessories	983	111	192
Scarf/Tie	Accessories	909	88	274
Watch/Bracelet	Accessories	2,627	206	86
Bag	Accessories	3,284	263	186
Gloves	Accessories	431	41	210
Necklace	Accessories	1,711	134	131
Glasses	Accessories	1,329	129	89
Belt	Accessories	1,035	95	110

Table 7: Varied Clothing Dataset Class Definition and statistics.

Table 7 shows the class definition and statistics of the Varied Clothing Dataset. Because we convert each segment(connected component) of semantic segmentation into an instance annotation, the number of training instance is much more than usual. The details can be found in Sec. 4.1 in the main submission. Another is the diverse classes. In contrast to ModaNet [43], in Varied Clothing Dataset, the confusing classes are not grouped. For example, Jacket/Blazer to Coat. This makes it more challenging for semantic segmentation approaches to generate clean results.

In Figure 6, we show more qualitative examples besides Figure 2. We use ResNet-50-FPN as the backbone model and train the model on the Varied Clothing Dataset. Figure 6 contains more diverse photos, such as vintage photos, layflat photos and images with full or half-bodies visible. Although Mask R-CNN-IMP can generate cleaner results than Panoptic-FPN, Mask R-CNN-IMP also incurs poor performance on boundaries of large objects which was caused by the low resolution output of Mask R-CNN[3]. Our final model Panoptic-FPN-IMP can generate sharp semantic segmentation results but also makes labeling of pixels from the same objects consistent.

class	Difference		#Instances	Total area
		DA		
Person	0.7	1.1	17,395	64,901,113
Rider	4.4	5.1	1,660	7,169,330
Car	0.2	0.4	26,180	380,112,819
Truck	2.5	9.4	466	14,657,648
Bus	1.2	6.9	350	12,684,337
Train	9.8	5.5	158	11,643,940
Motorcycle	6.5	4.3	705	5,037,718
Bicycle	0.6	0.8	3,433	14,646,908
Average	3.2	4.2		

Table 8: Analysis of Semantic Segmentation classes which are also Instance Segmentation. There is a correlation if the class has fewer instances and area, it gets more improvement from Instance Mask Projection. **DA**: with Data Augmentation.

6.1. Ablation Study on Cityscapes datasets.

For Cityscapes, we focus evaluations on the FPN-Panoptic network (ablation study in Table 9) and shows the effectiveness of each component. *Color Jitter* shows the marginally improvement in Table 9a. For *Hard Boostraping*, we see consistent improvements when setting the lower ratio in Table 9b. *Multi-scale Training* definitely helps a lot and also reduce overfitting on BBox/Mask prediction in Table 9c. *Instance Mask Projection* provides around 1.35/1.5 improvement without any data augmentation and with all data augmentations.

[3] 28×28

CJ	BBox	Mask	mIOU
	36.9	32.7	72.74
Y	36.8	**32.8**	**73.12**

(a) **Color Jitter**: Adding Color Jitter improves the performance marginally.

BS	Box	Mask	mIOU
0.50	37.8	34.0	73.81
0.25	38.4	34.1	73.93
0.10	**38.7**	**34.7**	**74.94**

(b) **Hard Boostraping**: Lower Bootstrapping provides the better accuracy. Color Jitter is used.

MS	Box	Mask	mIOU
	38.7	34.7	74.94
Y	**40.7**	**36.5**	**76.11**

(c) **Multi-scale training**: consistently improves three different measures. Color Jitter is used and Bootstrapping is set as 0.10

IMP	Box	Mask	mIOU
Without all the Data Augmentation			
	36.9	**32.7**	72.74
Y	**36.9**	32.5	**74.09**
With all the Data Augmentation			
	40.7	**36.5**	76.11
Y	39.8	35.8	**77.49**

(d) **IMP**:improves the two scenarios with and without data augmentation. See Table 4 for more details.

Table 9: Performance Analysis of each module used on Cityscapes `val` set. For simplicity, we use the following abbreviation: **MS**:*multi-scale training*, **CJ**:*Color Jitter*, **BS**:*Hard Boostraping*, **IMP**:*Instance Mask Projection*,

More discussions on Cityscapes dataset.

Table 8 shows the mIOU difference of Thing classes of Cityscapes with and without the data augmentation. This Table is part of Table 4 but adds number of instances and area information. We found out the improvement is also similar to the clothing datasets. First, the classes with less examples are improved more. See Train(#158), Bus(#350), Truck(#466), and Motorcycle(#705). Another is the improvement among the confusing classes. Although Rider contains enough examples, its similarity to Person, makes its mIOU lower. Our model is useful to distinguish these cases and increases the mIOU of Rider significantly.

Figure 7 shows the visualization examples of results of our models. We found that the qualitative results are also similar to the clothing datasets. Our final model, Panoptic-FPN-IMP, provides leaner results. See the better results of segments of Bus and Truck in Figure 7a and 7b. Another interesting case is Rider which means the person on the motorcycle or bicycle. The top part of Rider of Panoptic-FPN in Figure 7c and 7d are misclassified as Person. But with Instance Mask Projection, our final model shows correct labeling of all pixels of Rider.

Preliminary results on Pascal VOC dataset

In order to demonstrate the generalization of the proposed method in the general object dataset and properly utilize the instance segmentation results, here we add new results on the dataset from PASCAL in Detail Challenge at CVPR'17. [4], This version of PASCAL VOC contains 4,996(train), and 5,104(val) images which include both semantic segmentation and instance segmentation labeling. As the evaluation server is not available, we train on the training set and report preliminary results on the validation set. Table 10 shows the respective performance improvement from multitask training and IMP operator. As we can see, the improvement is not trivial. The IMP operator improves 2.74% absolutely improvement for mean IOU. The improvement due to IMP is similar to the other datasets.

Model	Instance	IMP	Semantic (mIOU)
Semantic-FPN			55.85
Panoptic-FPN	Y		63.34
Panoptic-FPN-IMP	Y	Y	**66.06**

Table 10: Ablation study of semantic segmentation accuracy on the PASCAL in Detail Challenge dataset from CVPR'17. We use the same models which were proposed in Section 3 . The backbone network is ResNet-50.

[4] https://sites.google.com/view/pasd/dataset?authuser=0

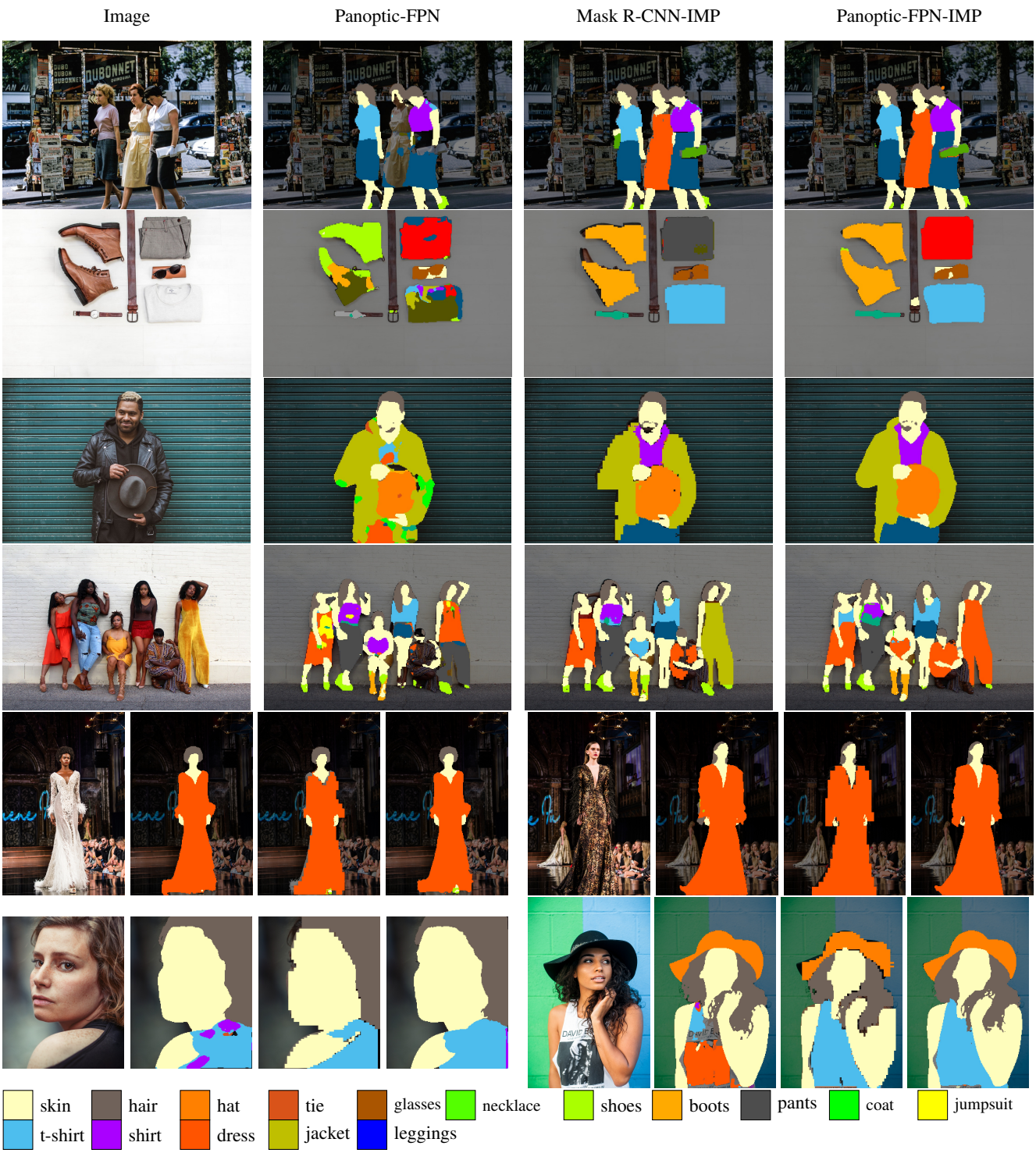
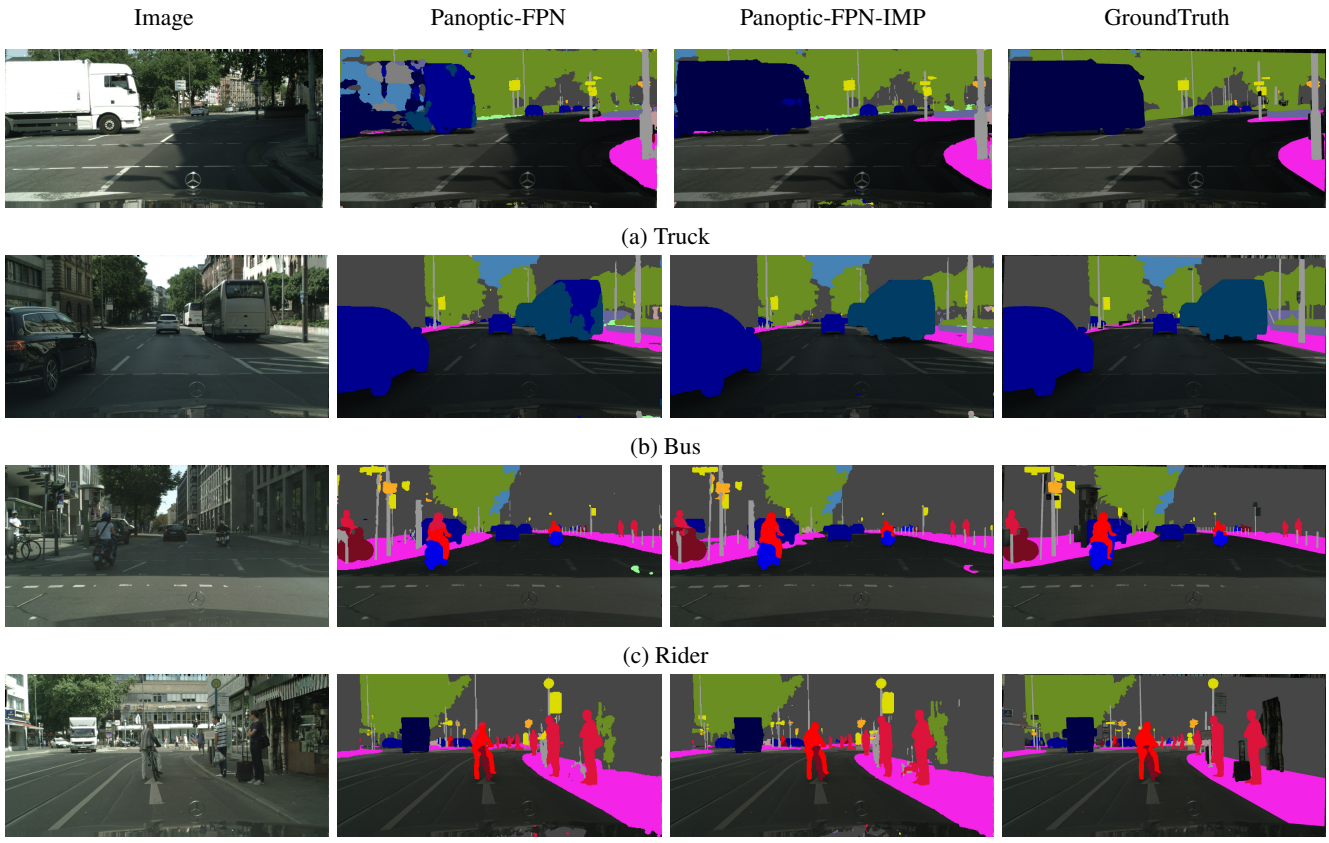


Figure 6: This Figure is an extension of Figure 2. From left to right, images, results of Panoptic-FPN, results of Mask R-CNN-IMP, results of our final model, Panoptic-FPN-IMP. The proposed method, IMP, works well on different types of clothing parsing examples, from vintage images, layflat images, street-fashion examples, fashion-runway photos, and photos with full or partial-bodies visible.



(a) Truck

(b) Bus

(c) Rider

(d) Rider

Figure 7: From left to right, images, results of Panoptic-FPN, Panoptic-FPN-IMP and GroundTruth. With the Instance Mask Projection, our final model, shows cleaner results on Truck(a), Bus(b), and Rider(c,d) classes.