InfraParis: A multi-modal and multi-task autonomous driving dataset
—— Supplementary material ——

Gianni Franchi
U2IS, ENSTA Paris, IP Paris
gianni.franchi@ensta-paris.fr

Marwane Hariat
U2IS, ENSTA Paris, IP Paris
marwane.hariat@ensta-paris.fr

Xuanlong Yu
SATIE, Paris-Saclay University; U2IS, ENSTA Paris, IP Paris
xuanlong.yu@ensta-paris.fr

Nacim Belkhir
Safrantech, Safran Group
nacim.belkhir@safrangroup.com

Antoine Manzanera
U2IS, ENSTA Paris, IP Paris
antoine.manzanera@ensta-paris.fr

David Filliat
U2IS, ENSTA Paris, IP Paris
david.filliat@ensta-paris.fr

InfraParis: A multi-modal and multi-task autonomous driving dataset
—— Supplementary Material ——

A. Implementation details

A.1. Semantic segmentation

Table A1 furnishes comprehensive insights into the hyperparameters and elaborate implementation details pertaining to the semantic segmentation task. In the case of RGB images, our approach incorporates classical data augmentation techniques such as random crop, random color jitters, and random horizontal flip, alongside normalization transformations that involve mean subtraction and division by the standard variation. However, for infrared images, a similar augmentation process is applied, albeit without the inclusion of random color jitters.

A.2. Supervised monocular depth estimation

As we mentioned in the main paper §4.2, the hyperparameters we used are the same as the official ones applied on the KITTI dataset, except that we use 4 instead of 8 as the batch size when we train the NeWCRFs model on the InfraParis RGB images. We find that batch size 4 works slightly better than 8 during evaluation.

Additionally, we found that sometimes the radar produced some depth information on parts of the front of the car. Thus, during training on InfraParis, we also crop the training images from the top down to a position of $1.8 * 352$.

A.3. Object detection

Table A2 furnishes comprehensive insights into the hyperparameters and elaborate implementation details regarding the object detection task. In the case of RGB images, our approach incorporates classical data augmentation techniques such as random crop, random color jitters, and random horizontal flip, alongside normalization transformations that involve mean subtraction and division by the standard variation. Please refer to the detectron2 library for more specific details on the architecture name mentioned in Table A2.

A.4. Unsupervised monocular depth estimation

We here provide a benchmark for unsupervised monocular depth estimation. We started from Monodepth2 pretrained on KITTI and then fine-tuned it on the InfraParis training and validation set. Then we evaluated it on the InfraParis test set. Results are shown in Table A3. We followed the same evaluation protocol as Monodepth2 to solve the scale ambiguity. We applied a mask adapted to the InfraParis dataset which gets rid of pixels corresponding to both the front of the car and to the sky.

B. Resolving ambiguous images

Throughout the annotation process, we encountered various instances of ambiguity. For a visual representation of these complexities, please consult Figures A1. To ensure the integrity of our dataset, particularly in instances where uncertainty prevailed, we collaborated with the annotation company to categorize dubious instances as ”unlabeled.” This approach was adopted to prevent any adverse impact

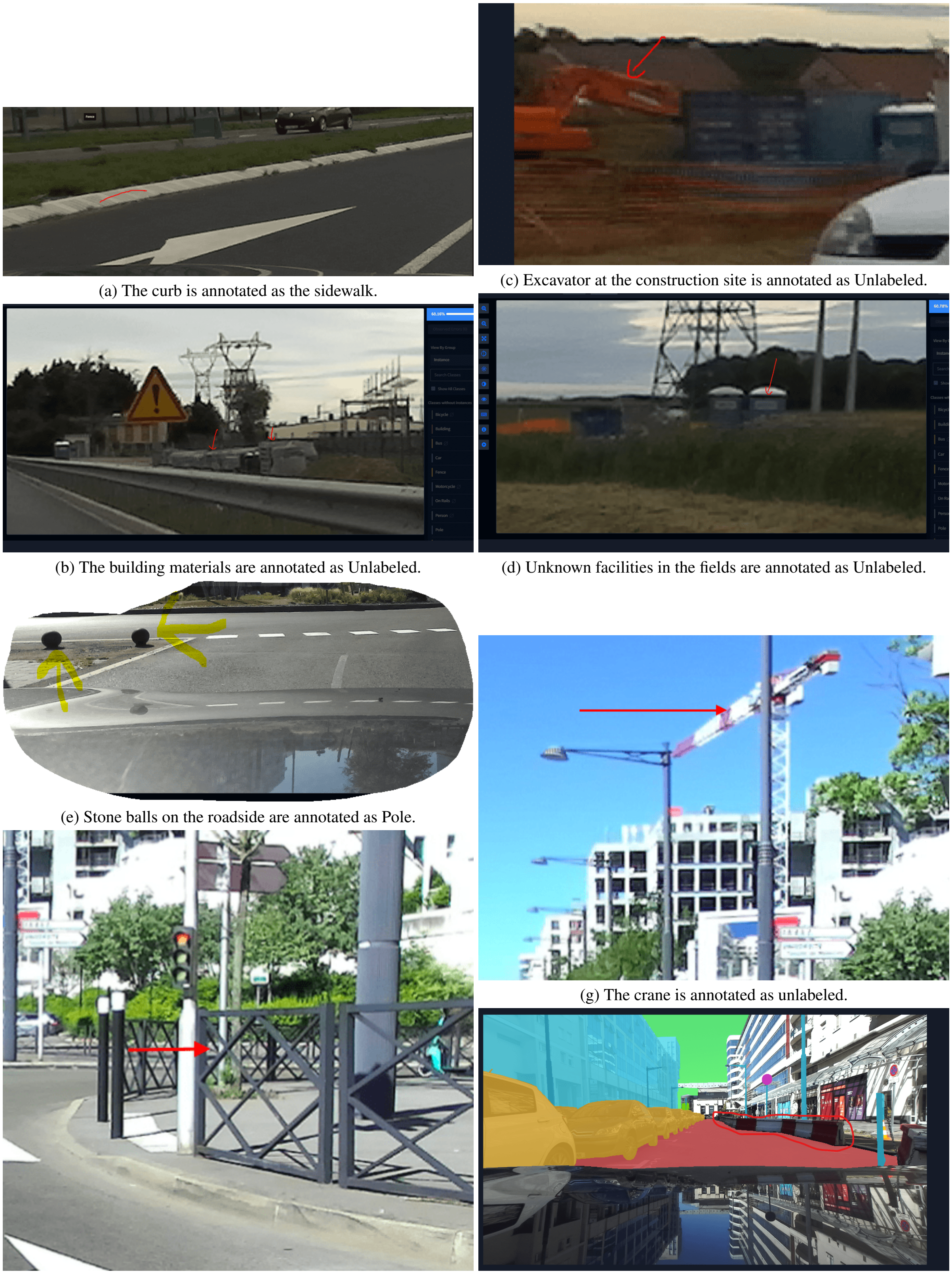
Architecture	Deeplab v3+	Segformer B0,B1,B2	Segformer B3,B4,B5
backbone	ResNet101 and mobilenet	NA	NA
output stride	8	NA	NA
learning rate	0.1	0.1	0.0001
batch size	16	8	3
number of train iterations	100	60	60
weight decay	0.0001	0.01	0.01
Optimizer	SGD	AdamW	AdamW
random crop of training images	768	768	1024

Table A1. **Hyper-parameter configuration used in the semantic segmentation experiments.**

Architecture	faster_rcnn_R_50_C4_1x	mask_rcnn_R_50_FPN_3x
backbone	ResNet50	ResnNt50
learning rate	0.001	0.001
roi.heads batch size	256	256
batch size	8	8
number of train iterations	50	50
weight decay	0.0001	0.0001
Optimizer	SGD	SGD

Table A2. **Hyper-parameter configuration used in the object detection experiments.**

on the performance of DNNs stemming from incorrect annotations.



(a) The curb is annotated as the sidewalk.

(c) Excavator at the construction site is annotated as Unlabeled.

(b) The building materials are annotated as Unlabeled.

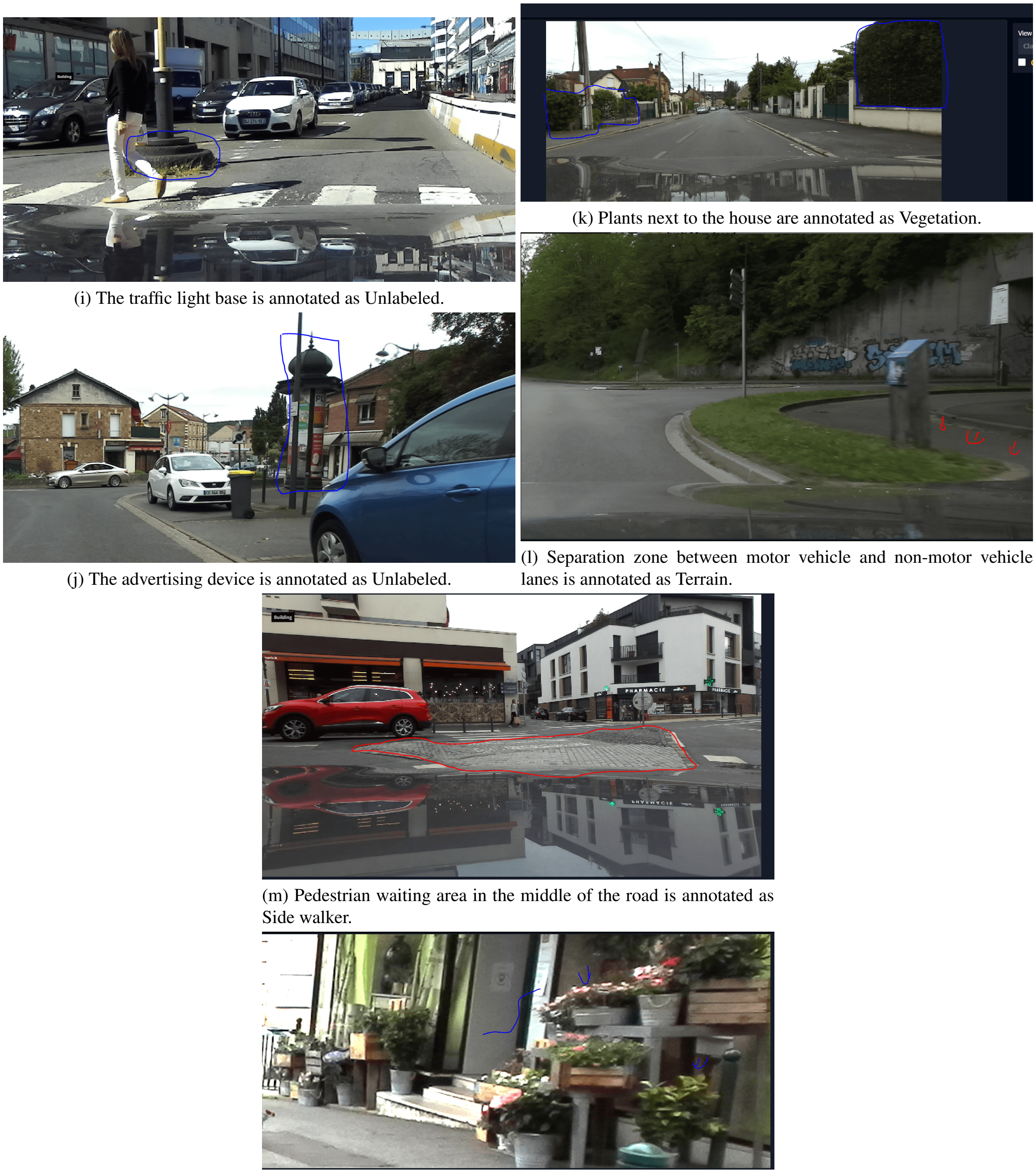
(d) Unknown facilities in the fields are annotated as Unlabeled.

(e) Stone balls on the roadside are annotated as Pole.

(g) The crane is annotated as unlabeled.

(f) Featured fences are annotated as Fence.

(h) Blocking device at the construction site is annotated as Fence.



(i) The traffic light base is annotated as Unlabeled.

(k) Plants next to the house are annotated as Vegetation.

(j) The advertising device is annotated as Unlabeled.

(l) Separation zone between motor vehicle and non-motor vehicle lanes is annotated as Terrain.

(m) Pedestrian waiting area in the middle of the road is annotated as Side walker.

(n) Flowers and pots at the flower shop are annotated as Unlabeled.

Figure A1. **Examples of the ambiguous objects encountered in the annotation process.**

Training set	Eval set	Abs Rel ↓	Sqr Rel ↓	RMSE ↓	RMSElog ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
KITTI	InfraParis	0.236	0.984	3.870	0.341	0.573	0.832	0.933

Table A3. **Comparative results for unsupervised monocular depth estimation.** The evaluation depth range is 0-40 meters. The model chosen is Monodepth2.