TRAINING AN EMBEDDED OBJECT DETECTOR FOR INDUSTRIAL SETTINGS WITHOUT REAL IMAGES

Julia Cohen, Carlos Crispim-Junior, Jean-Marc Chiappa, Laure Tougne

{julia.cohen, carlos.crispim-junior, laure.tougne}[at]liris.cnrs.fr
jean-marc.chiappa[at]groupe-dems.fr

Paper ID: 2829

Laboratoire d'InfoRmatique en Image et Systèmes d'information
LIRIS UMR 5205 CNRS / INSA de Lyon / Université Claude Bernard Lyon 1 / Université Lumière Lyon 2 / Ecole Centrale de Lyon

1 Introduction – Industrial object detection

Challenges

- Specific objects, absent from the common massive datasets
  - ➢ **Real images** of the objects of interest **unavailable**
- Real-time inference on embedded devices
  - ➢ **Low-memory** footprint, **real-time inference** required

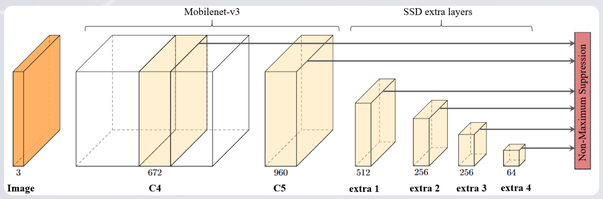
Our approach

- **Single Shot Detector** (SSD) [Liu *et al.*, 2016] with **MobileNet-V3** backbone [Howard *et al.*, 2019]
- **Synthetic images** as training set, generated from 3D models
- Curated **data augmentation** to bridge the synthetic-real domain gap

2 Our method

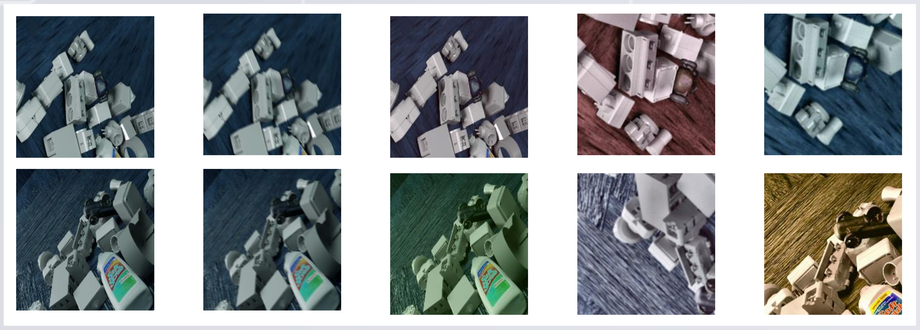
Architecture: MobileNet backbone + SSD

- MobileNet-V3 Large
- MobileNet-V3 Small
- MobileNet-V2



Data augmentation

- Strong distortions (aug2) VS. common distortions (aug1)
  - ➢ *Color jitter (brightness, contrast, saturation, RGB shift)*
  - ➢ *Histogram Equalization (CLAHE)*
  - ➢ *Blur (mean, Gaussian, median, motion)*
  - ➢ *Noise (Gaussian, multiplicative, ISO)*
  - ➢ *Vertical flip*
  - ➢ *Blur*
  - ➢ *Sharpness*
  - ➢ *Contrast*
  - ➢ *Color*



Examples on T-LESS training images. From left to right: no transform, motion blur, RGB shift, full pipeline, full pipeline.

3 Dataset

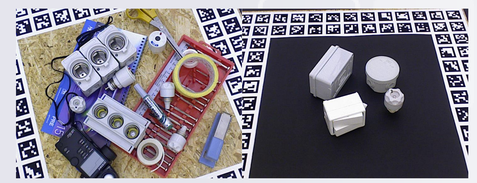
Training with 50,000 synthetic images

- Use of the objects 3D models
- Physically-based rendering using **BlenderProc** [Denninger *et al.*, 2019]

Testing on real images

- ➢ Example of the T-LESS dataset [Hodan *et al.*, 2017]

Synthetic T-LESS images

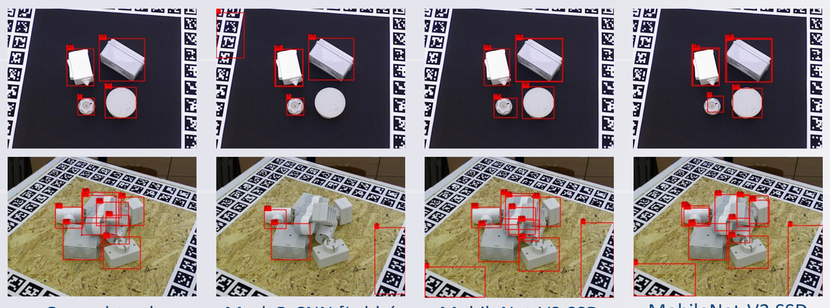


Real T-LESS images from a PrimeSense camera

4 Results

- MobileNet-V2 and MobileNet-V3 + SSD perform better than a Mask R-CNN model [Labbé *et al.*, 2020]

	mean Average Precision (mAP, %)	
Augmentation method	Low augmentation (aug1)	Strong augmentation (aug2)
Mask R-CNN	32.8	-
MobileNet-V3 Small + SSD	18.6	23.5
MobileNet-V3 + SSD	36.3	46.1
MobileNet-V2 + SSD	**38.3**	**47.7**



Ground-truth | Mask R-CNN [Labbé *et al.*, 2020] | MobileNet-V3 SSD (ours) | MobileNet-V2 SSD (ours)

5 Conclusion

- MobileNet-V2 + SSD object detector outperforms the other models, with a better generalization to real images and real-time inference: less than 30ms per image on a GPU.
- Industrial objects can be detected using 3D models only, using synthetic images and strong data augmentation.

This work is supported by grant CIFRE n.2018/0872 from ANRT.