Improving Car Model Classification Through Vehicle Keypoint Localization

Alessandro Simoni, Andrea D'Eusanio, Stefano Pini, Guido Borghi, Roberto Vezzani

{alessandro.simoni, andrea.deusanio, s.pini, roberto.vezzani}@unimore.it, guido.borghi@unibo.it

University of Modena and Reggio Emilia, Italy

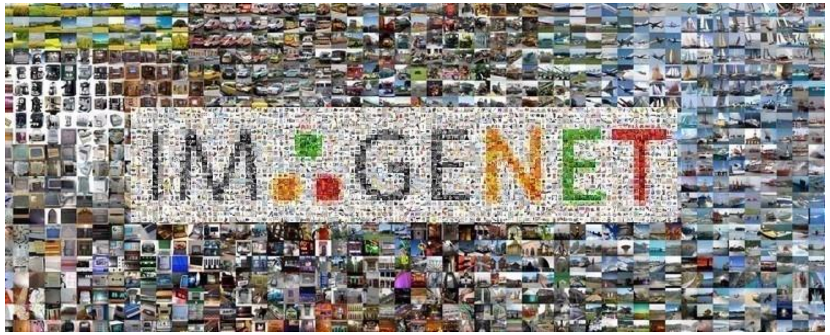
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

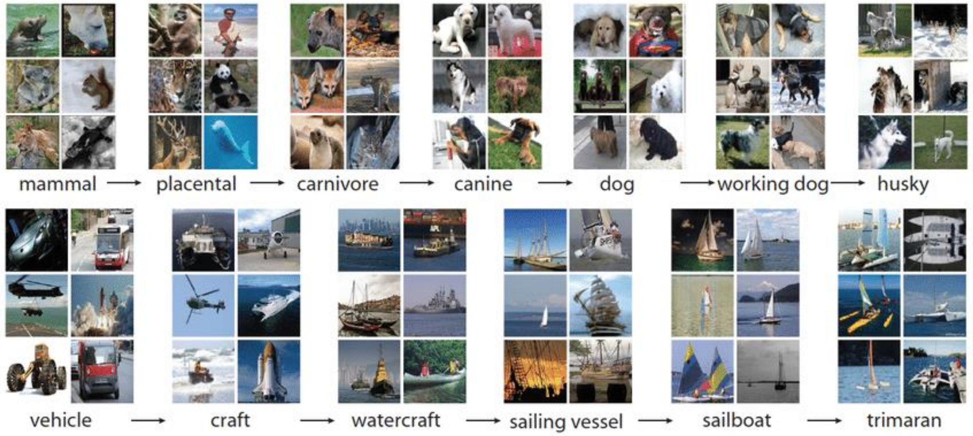
AImage Lab

softech-ict
Centro Interdipartimentale di Ricerca
Softech: ICT per le Imprese

Problem definition

- **Object classification** is a well-established technique in computer vision
- Deep learning architectures[1,2,3,4] reached impressive results on **macro-classes classification** tasks, exploiting large datasets like ImageNet
- Unfortunately, they still struggle on datasets with:
  - limited number of samples
  - classes with high similarity
  - heavy class imbalance
  - appearance differences from different viewpoints



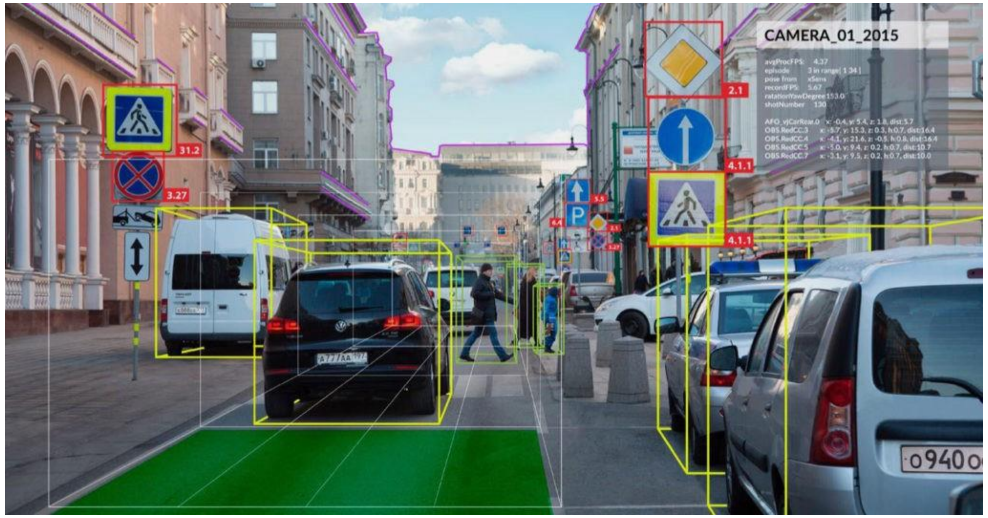


1. Simonyan, Karen et al. *"Very deep convolutional networks for large-scale image recognition". In arXiv preprint arXiv:1409.1556. 2014.*
2. He, Kaiming et al. *"Deep residual learning for image recognition". In CVPR. 2016.*
3. Huang, Gao, et al. *"Densely connected convolutional networks". In CVPR. 2017.*
4. Xie, Saining et al. *"Aggregated residual transformations for deep neural networks". In CVPR. 2017.*

AImage Lab

Why the automotive scenario?

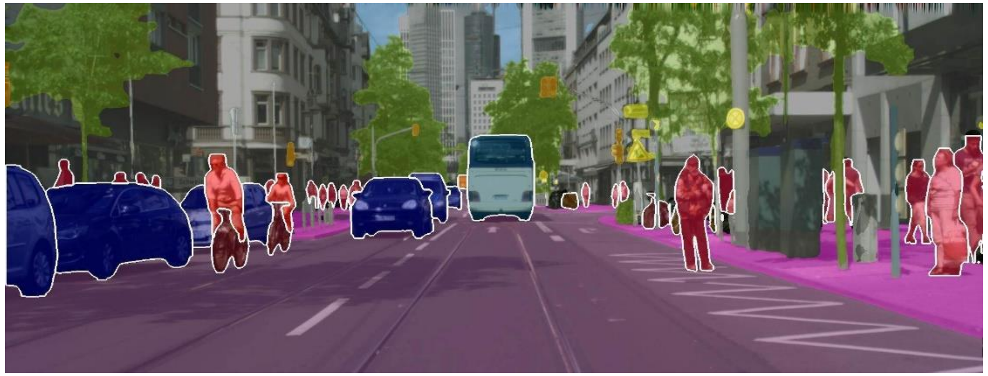
- **Automotive scenario** poses many open challenges:
    - Detection & Tracking
    - Re-identification
    - 3D object detection
    - 3D reconstruction
    - Trajectory prediction



- All tasks have a common requirement → **SAFETY**

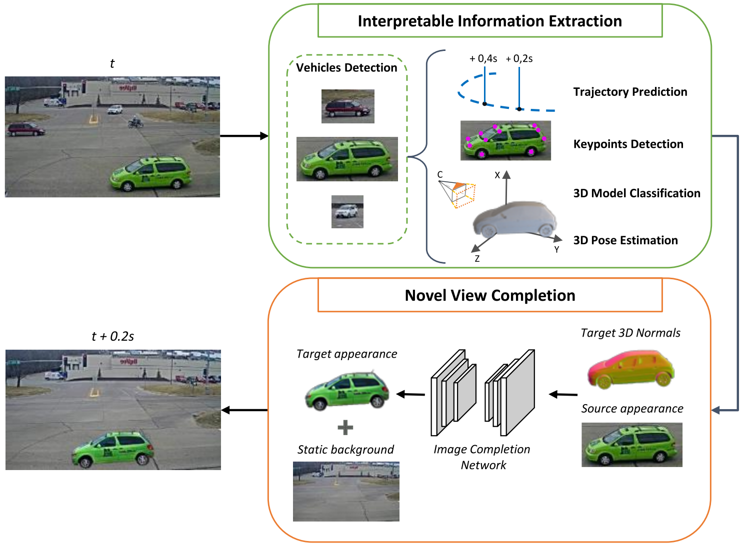
- Recognition between different agents helps having a correct perception of the scene:

➡ **object classification as an enabling solution**



Our goal

- **Previous work** [1] presented at ICPR2020:
  - Synthetic urban scene generation through vehicle synthesis
  - Exploiting interpretable information from RGB images (2D trajectory, **2D keypoints**, **vehicle class**)



<u>GOAL</u> ➡ improving the **classification module** obtaining **higher accuracy** on specific vehicle model classes

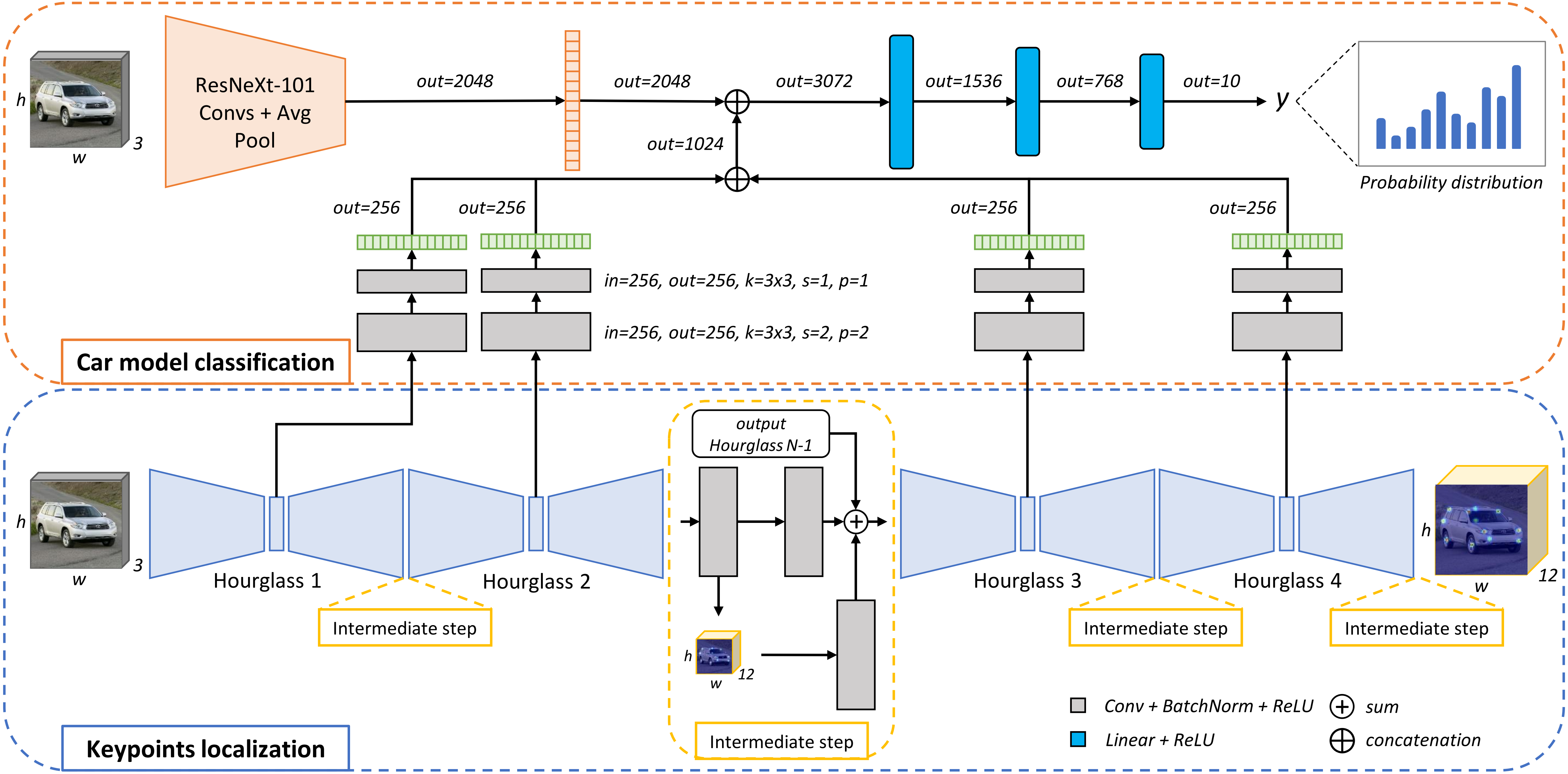
- Some works[2,3] already assess **classification together with pose estimation**, but only for predicting different macro-classes (aeroplane, bus, train, car, …)

<u>CHALLENGE</u> ➡ categorizing different **specific vehicle models**

<u>INTUITION</u> ➡ leveraging **2D keypoints localization as an additional information** to the classification method

1. *Simoni, Alessandro et al. "Future Urban Scenes Generation Through Vehicles Synthesis". In ICPR. 2020.*
2. *Grabner, Alexander et al. "3d pose estimation and 3d model retrieval for objects in the wild". In CVPR. 2018.*
3. *Afifi, Ahmed et al. "Simultaneous Object Classification and Viewpoint Estimation using Deep Multi-task Convolutional Neural Network". In VISAPP. 2018.*

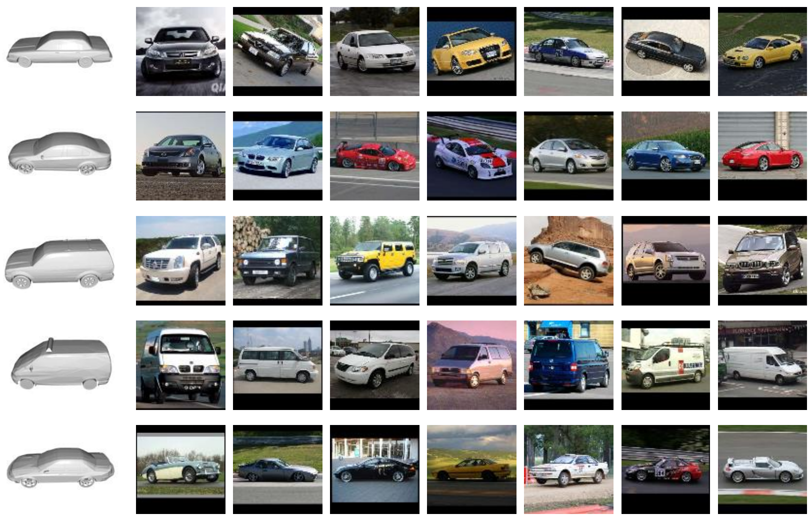
Our proposal

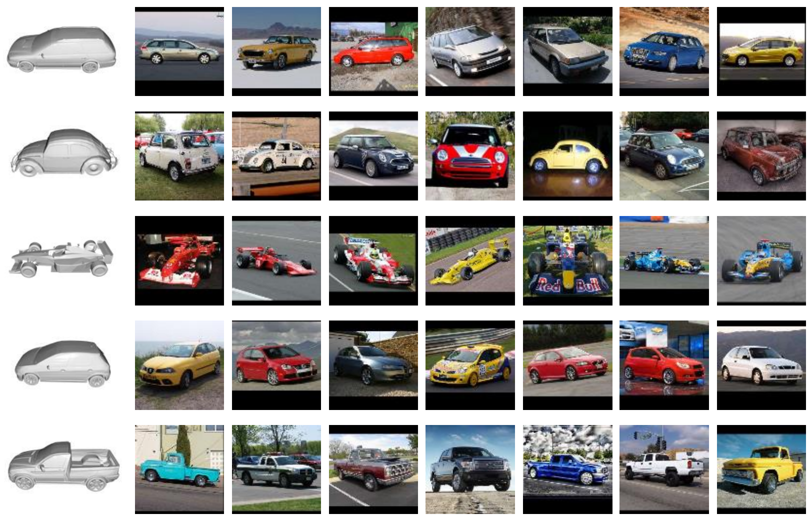


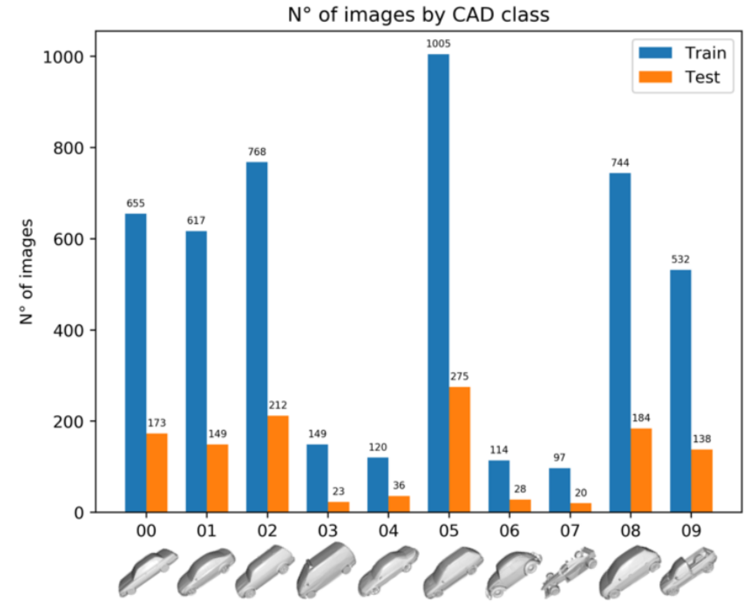
Dataset

Pascal3D+ [1]

- Collection of images from **12 different object classes**
- We take into consideration only the **"car" class** subdivided into 10 possible 3D car model sub-classes
- Annotations of **2D keypoints**, **3D model class** and **3D pose**
- 4k+ training images and 1k+ testing images





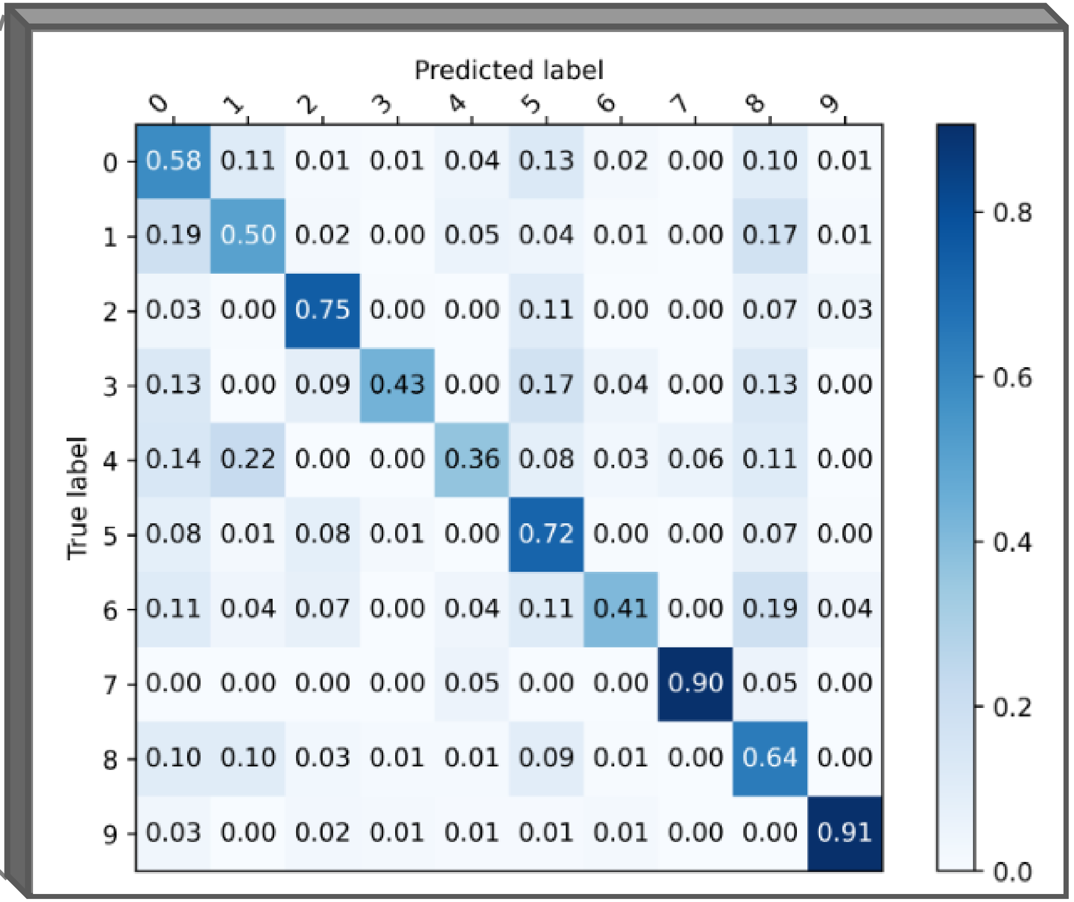


1. Y. Xiang et al., Beyond pascal: A benchmark for 3d object detection in the wild. In WACV, 2014.

Preliminary study – classification

- Searching for the **best solution among classification architectures** available in the literature
- Finetuning on models pretrained on ImageNet[2]
  (150 epochs with learning rate of 1e^{-4})

Network	Layers	Accuracy
VGG16 (Simonyan and Zisserman, 2014)	*last fc*	65.18%
VGG16 (Simonyan and Zisserman, 2014)	*all fc*	65.10%
ResNet-18 (He et al., 2016)	*last fc*	59.01%
ResNet-18 (He et al., 2016)	*all*	58.20%
DenseNet-161 (Huang et al., 2017)	*last fc*	65.02%
ResNeXt-101 (Xie et al., 2017) [1]	***last fc***	**66.96%**



1. Xie, Saining, et al. *"Aggregated residual transformations for deep neural networks". In CVPR. 2017.*
2. Deng, Jia et al. *"Imagenet: A large-scale hierarchical image database". In CVPR. 2009.*

Preliminary study – keypoints localization

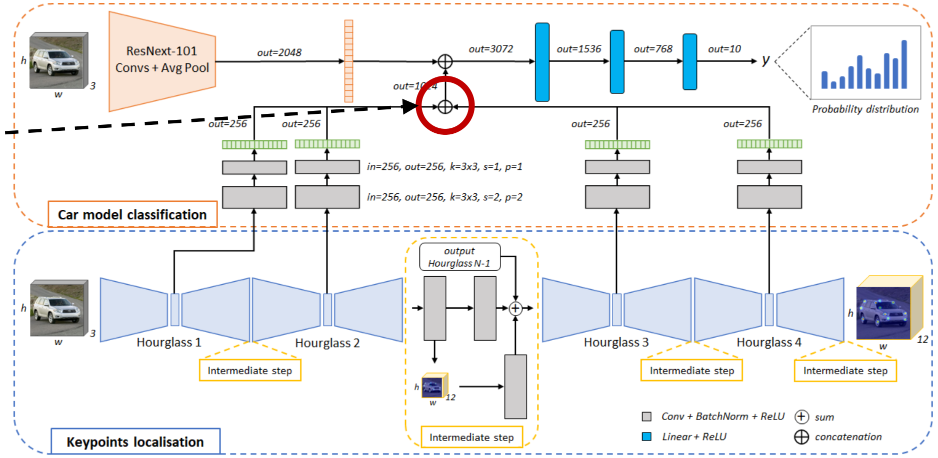
- Searching for the **best solution among keypoints localization architectures** available in the literature
- Training from scratch on Pascal3D+
  (100 epochs with initial learning rate of 1e^{-3} and decay every 40 epochs by a factor of 10)

Keypoint (∗)	HG-2	Net-W32	HRNet-W48
lb trunk	93.27	1.72	94.45
lb wheel	92.27	0.26	91.78
lf light	92.85	0.87	91.27
lf wheel	94.41	1.48	89.17
rb trunk	92.59	1.94	92.25
rb wheel	91.50	2.00	91.61
rf light	93.01	9.59	91.54
rf wheel	91.73	9.12	91.16
ul rearwindow	94.67	1.08	93.63
ul windshield	96.00	4.47	95.62
ur rearwindow	93.27	2.39	92.82
ur windshield	95.47	4.59	94.91

Model	PCKh@0.5
(Long et al., 2014)	55.7%
(Tulsiani and Malik, 2015)	81.3%
OpenPose-ResNet152 (Cao et al., 2017)	84.87%
OpenPose-DenseNet161 (Cao et al., 2017)	86.68%
(Zhou et al., 2018)	90.00%
HRNet-W32 (Wang et al., 2020)	91.63%
HRNet-W48 (Wang et al., 2020)	92.52%
(Pavlakos et al., 2017)	93.40%
Stacked-HG-2 (Newell et al., 2016)	93.41%
Stacked-HG-4 (Newell et al., 2016) [1]	**94.20%**
Stacked-HG-8 (Newell et al., 2016)	93.92%

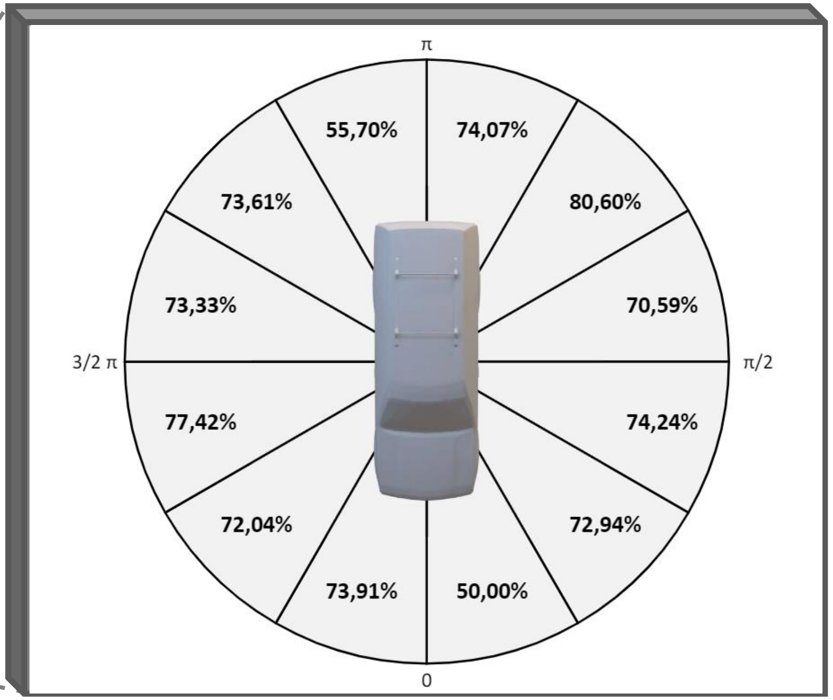
1. *Newell, Alejandro et al. "Stacked hourglass networks for human pose estimation". In ECCV. 2016.*

Quantitative results



- **Two approaches** of our framework:
  - **sum**
  - **concatenation**
- **Freeze** both ResNeXt-101 pretrained on ImageNet and Stacked-HG-4 pretrained on Pascal3D+
- Training **added fc layers** from scratch on Pascal3D+ *(100 epochs with learning rate of 1e-4)*

Method	Fusion	Accuracy
(Simoni et al., 2020)	-	65.91%
ResNeXt-101	-	66.96%
Stacked-HG-4 + (Simoni et al., 2020)	*sum*	67.61%
Stacked-HG-4 + (Simoni et al., 2020)	*concat*	69.07%
Ours	*sum*	68.26%
Ours	***concat***	**70.54%**



Performance

- Tested on a workstation with *Inter Core i7-7700K* and *Nvidia GeForce GTX 1080Ti*
- Large number of parameters
- **Multi-task framework** (classification + keypoints localization)
- **Realtime** speed with **low memory consumption**

Model	**Parameters** (M)	**Inference** (ms)	**VRAM** (GB)
VGG19	139.6	6.843	1.239
ResNet-18	11.2	3.947	0.669
DenseNet-161	26.5	36.382	0.995
ResNeXt-101	86.8	33.924	1.223
Stacked-HG-4	13.0	41.323	0.941
OpenPose	29.0	19.909	0.771
HRNet	63.6	60.893	1.103
Ours	106.8	68.555	1.389

Conclusions

- Show how **visual and pose features** can be **merged** to improve car model classification task
- **ResNeXt-101** for visual features extraction and **Stacked-Hourglass** for keypoints localization
- Combined architecture with **features concatenation** and **fc layers**

Achievements

✓ **+3.6%** improvement in classification accuracy
✓ **multitask** architecture
✓ **realtime** performance

Future work:

- Further analysis and experiments on **misclassification** due to **class imbalanced dataset**

VISAPP 2021

16th International Conference on Computer Vision Theory and Applications

Online Streaming 8 - 10 February, 2021

VISIGRAPP

Thank you for your attention

Improving Car Model Classification Through Vehicle Keypoint Localization

Alessandro Simoni, Andrea D'Eusanio, Stefano Pini, Guido Borghi, Roberto Vezzani

{alessandro.simoni, andrea.deusanio, s.pini, roberto.vezzani}@unimore.it, guido.borghi@unibo.it

University of Modena and Reggio Emilia, Italy

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

Centro Interdipartimentale di Ricerca Softech: ICT per le Imprese