CoMFormer: Continual Learning in Semantic and Panoptic Segmentation

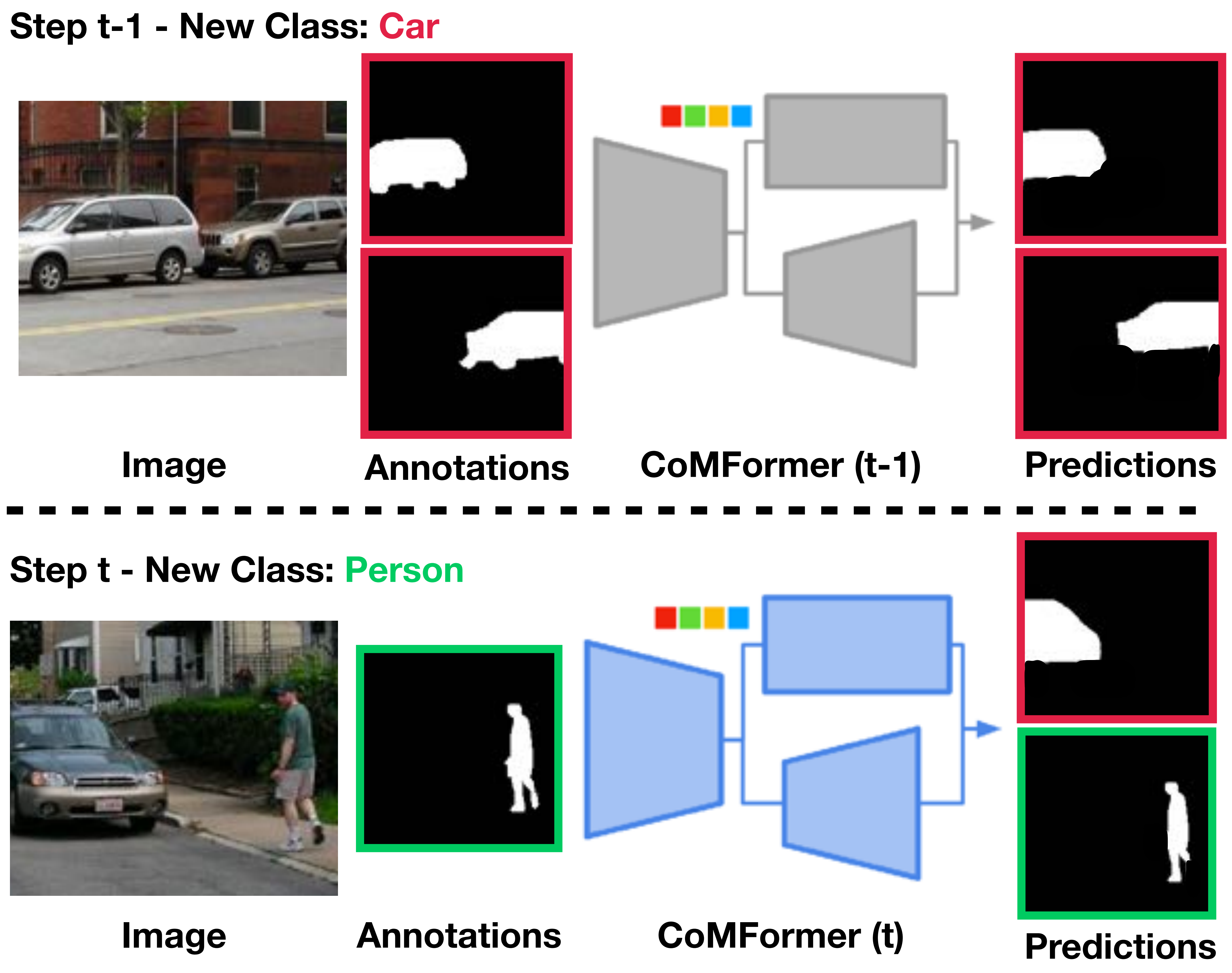
Fabio Cermelli, Matthieu Cord, Arthur Douillard

POSTER: TUE-AM-286

Contributions

CoMFormer

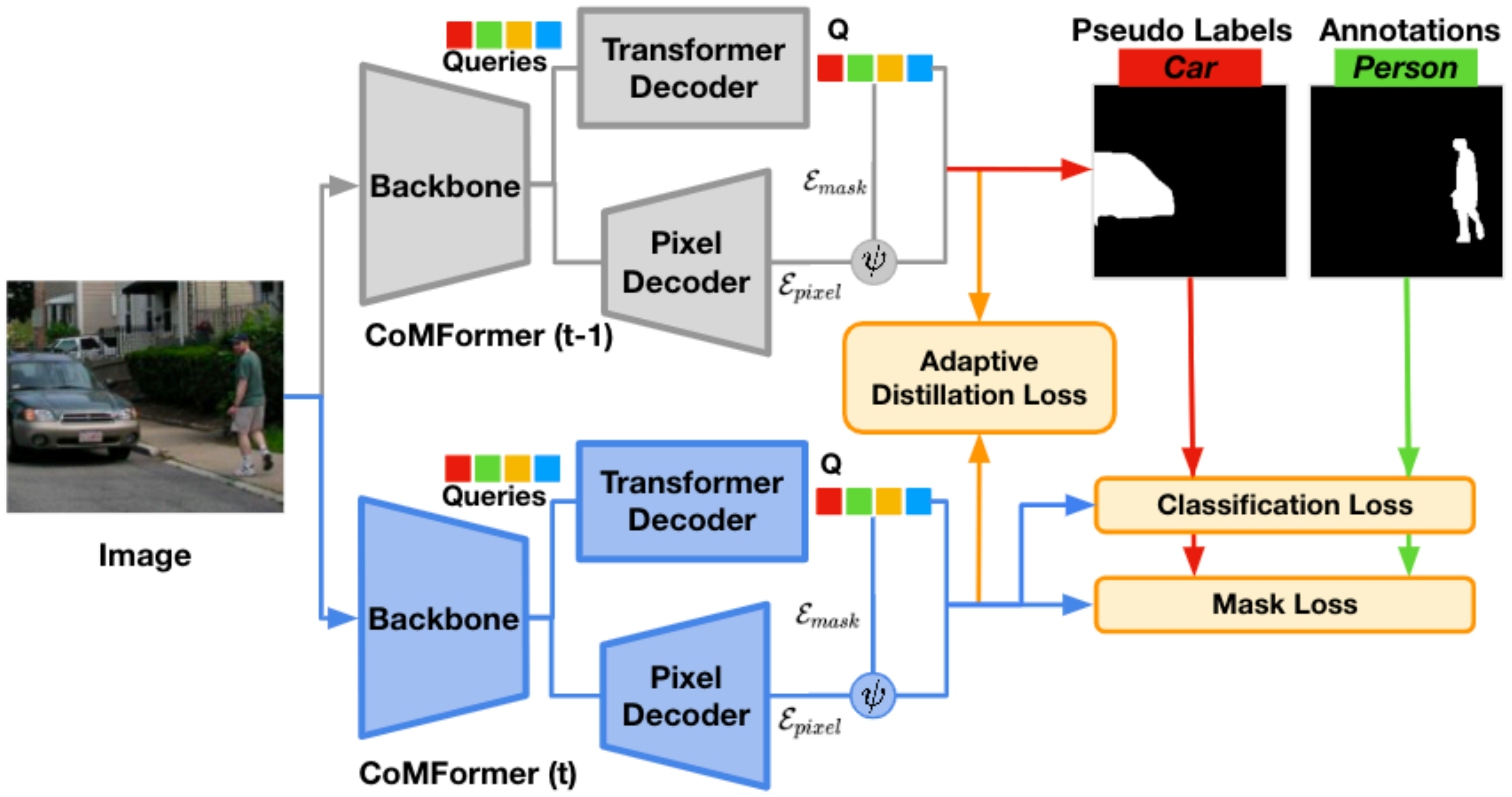
We present a **continual segmentation setting**, including semantic and panoptic segmentation.



CoMFormer

We present a **continual segmentation setting**, including semantic and panoptic segmentation.

We introduce **CoMFormer**, a **novel method** based on the MaskFormer architecture. It avoids forgetting using a Adaptive Distillation Loss and a Mask-based Pseudo-labeling strategy.



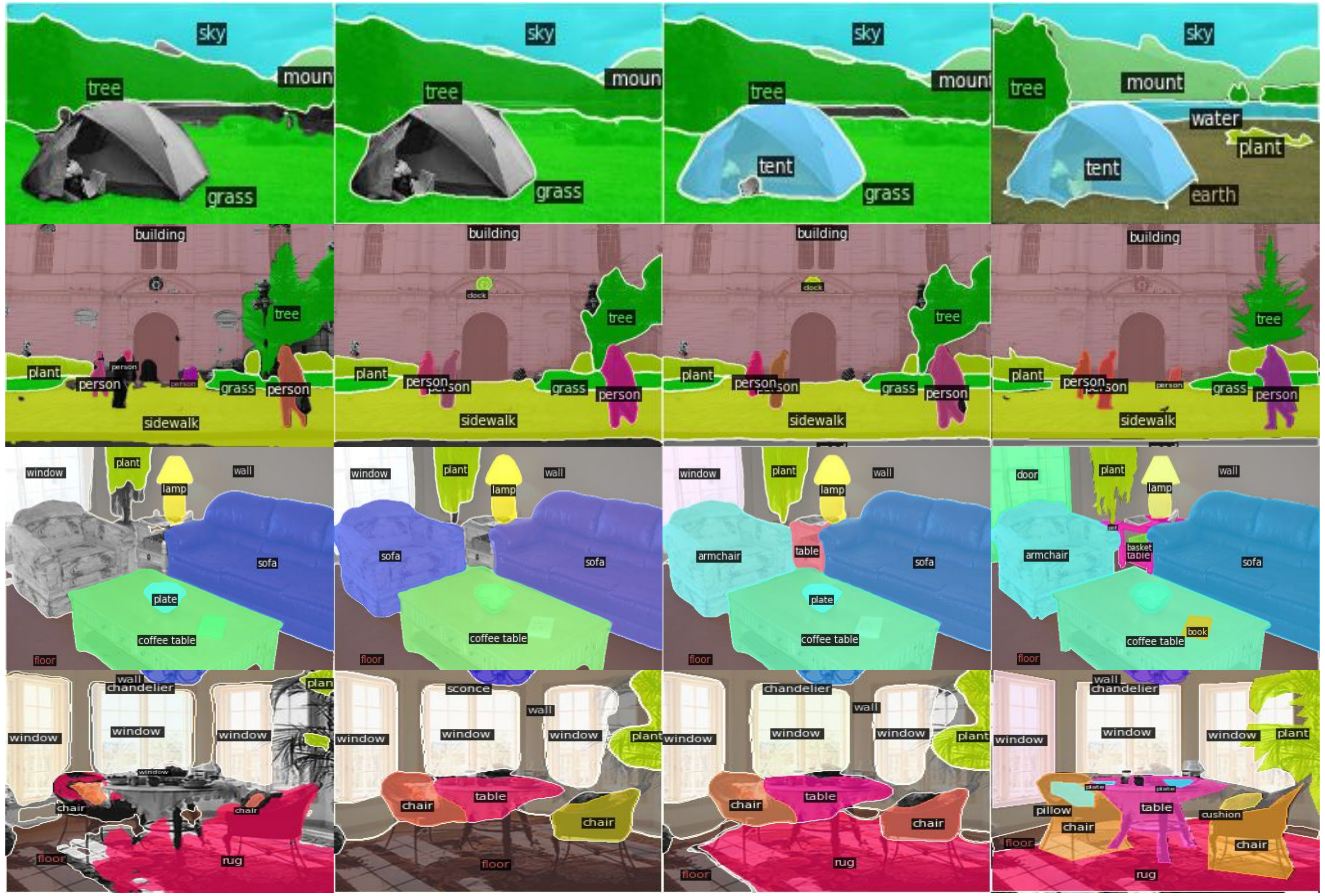
CoMFormer

CONTRIBUTIONS

We present a **continual segmentation setting**, including semantic and panoptic segmentation.

We introduce **CoMFormer**, a **novel method** based on the MaskFormer architecture. It avoids forgetting using a Adaptive Distillation Loss and a Mask-based Pseudo-labeling strategy.

We propose a **novel benchmark** on both semantic and panoptic segmentation, where CoMFormer outperforms previous baselines.



MiB PLOP **CoMFormer** Ground-truth

Motivation

PROBLEM

a) Semantic Segmentation

b) Instance Segmentation

c) Panoptic Segmentation

Motivation

PROBLEM

Segmentation tasks require to **cluster pixels** given their **semantic** category, separating or not instances of the same class.


a) Semantic Segmentation

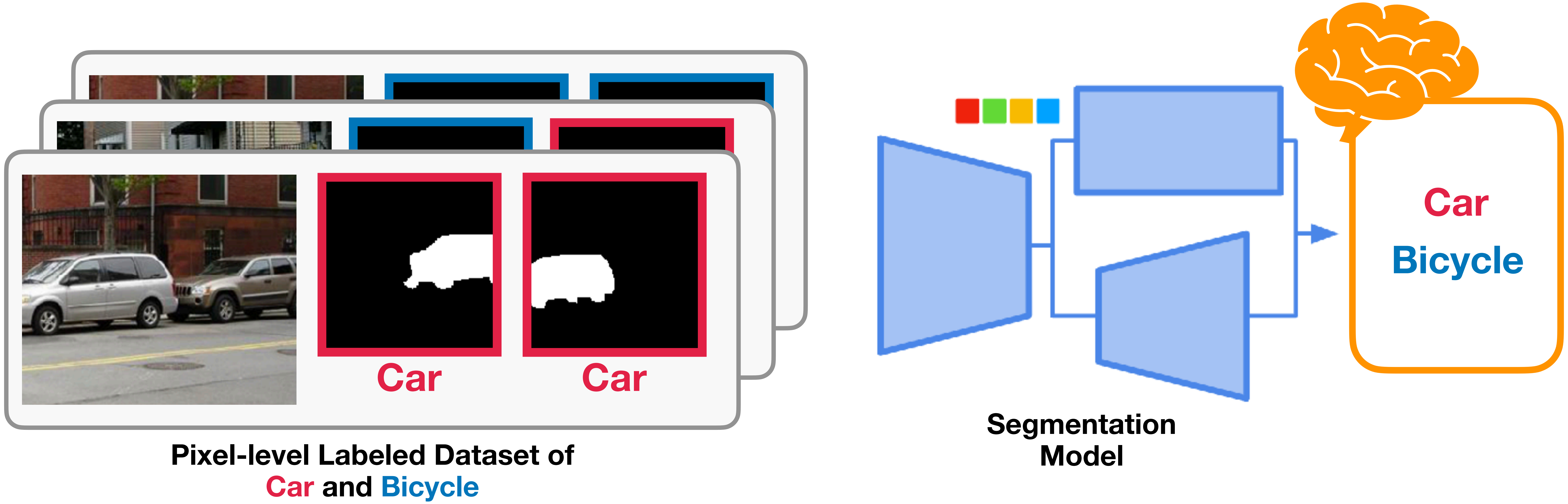

b) Instance Segmentation


c) Panoptic Segmentation

Motivation

Segmentation tasks require to **cluster pixels** given their **semantic** category, separating or not instances of the same class.

Current segmentation models are able to predict only the set of classes provided in the dataset.



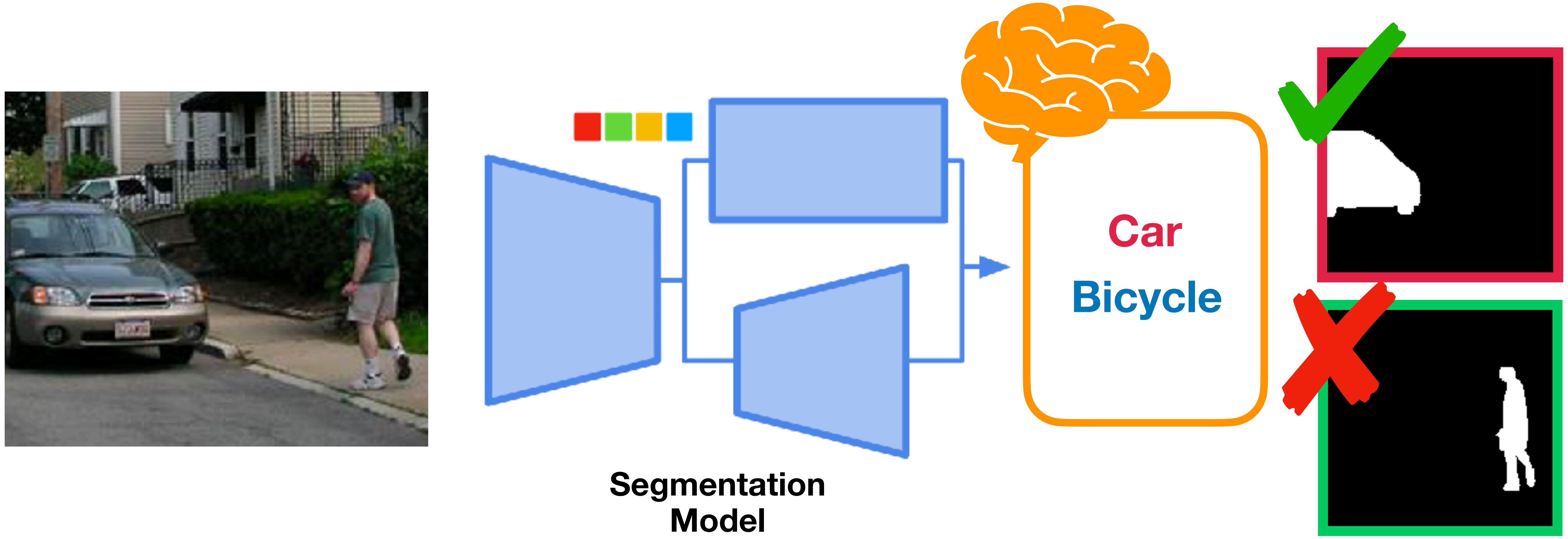
Motivation

Problem

Segmentation tasks require to **cluster pixels** given their **semantic** category, separating or not instances of the same class.

Current segmentation models are able to predict only the set of classes provided in the dataset.

Moreover, they **cannot be updated** as novel classes are discovered, requiring to restart training.



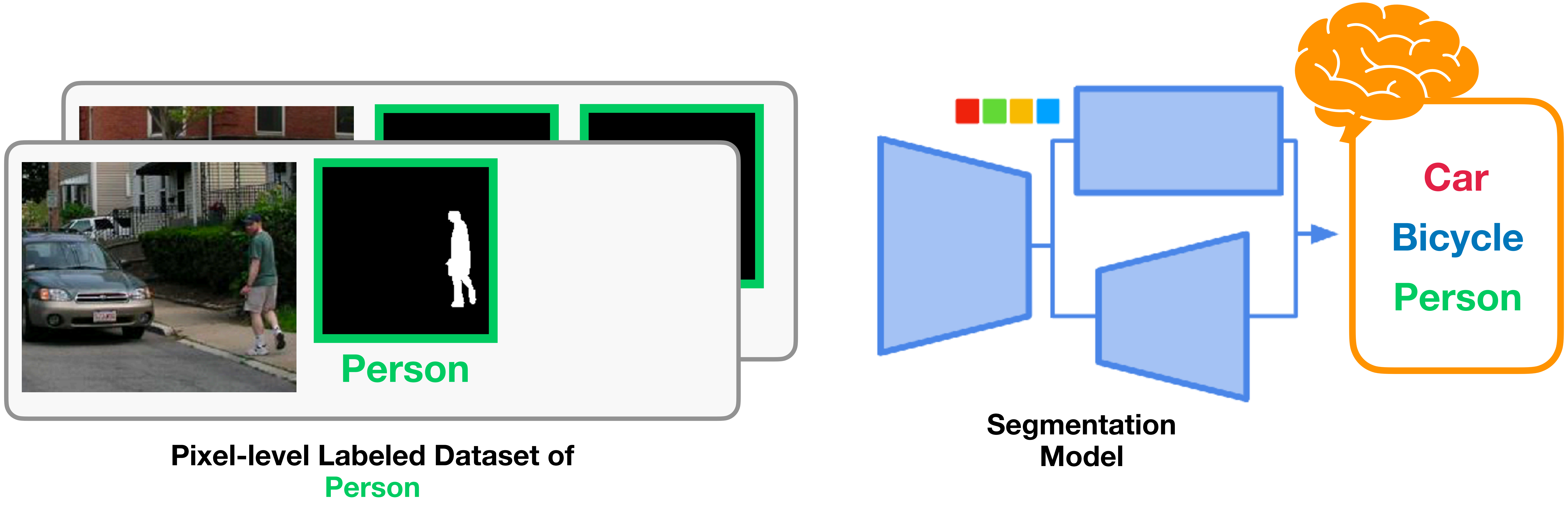
Motivation

Segmentation tasks require to **cluster pixels** given their **semantic** category, separating or not instances of the same class.

Current segmentation models are able to predict only the set of classes provided in the dataset.

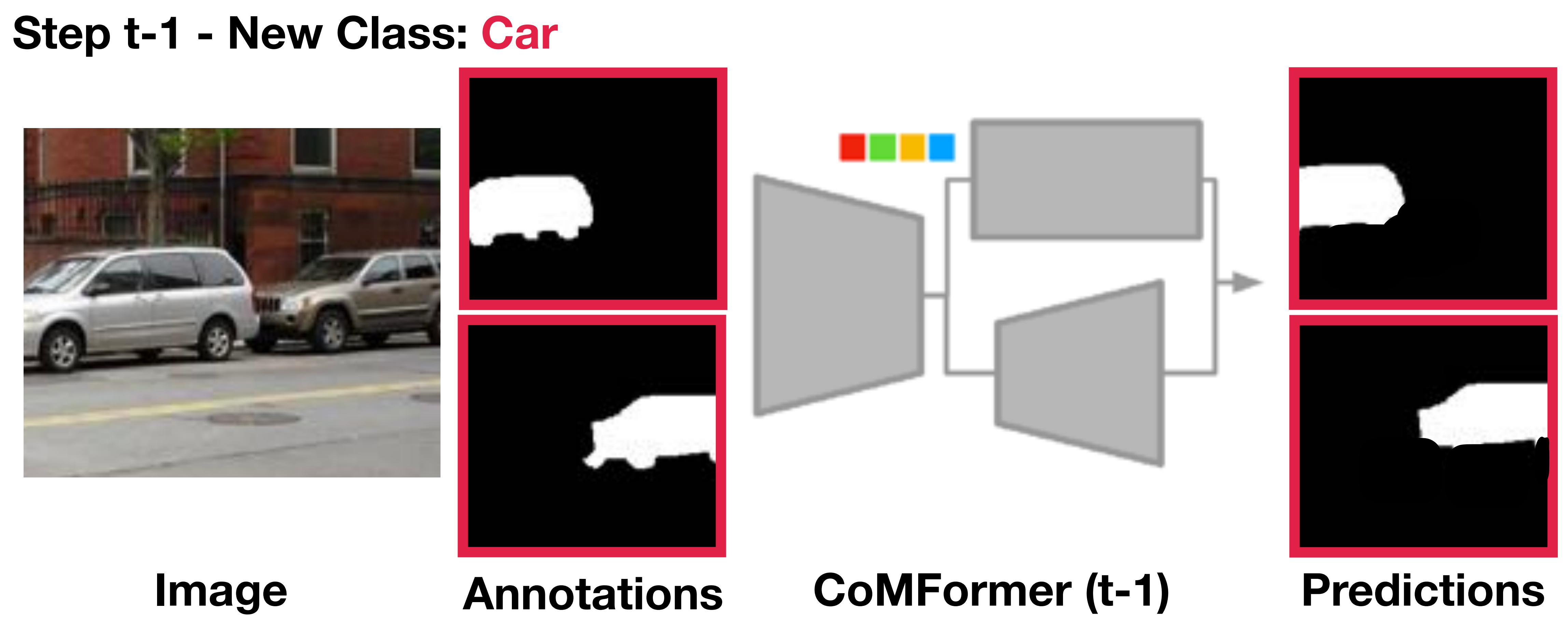
Moreover, they **cannot be updated** as novel classes are discovered, requiring to restart training.

We aims to extend the models' capabilities, enabling to **learn novel classes without forgetting**.



Continual Segmentation

We propose a **continual learning setting** unifying semantic and panoptic segmentation. The training is done in **multiple learning steps** *t=1…T*, each introducing a new set of classes.

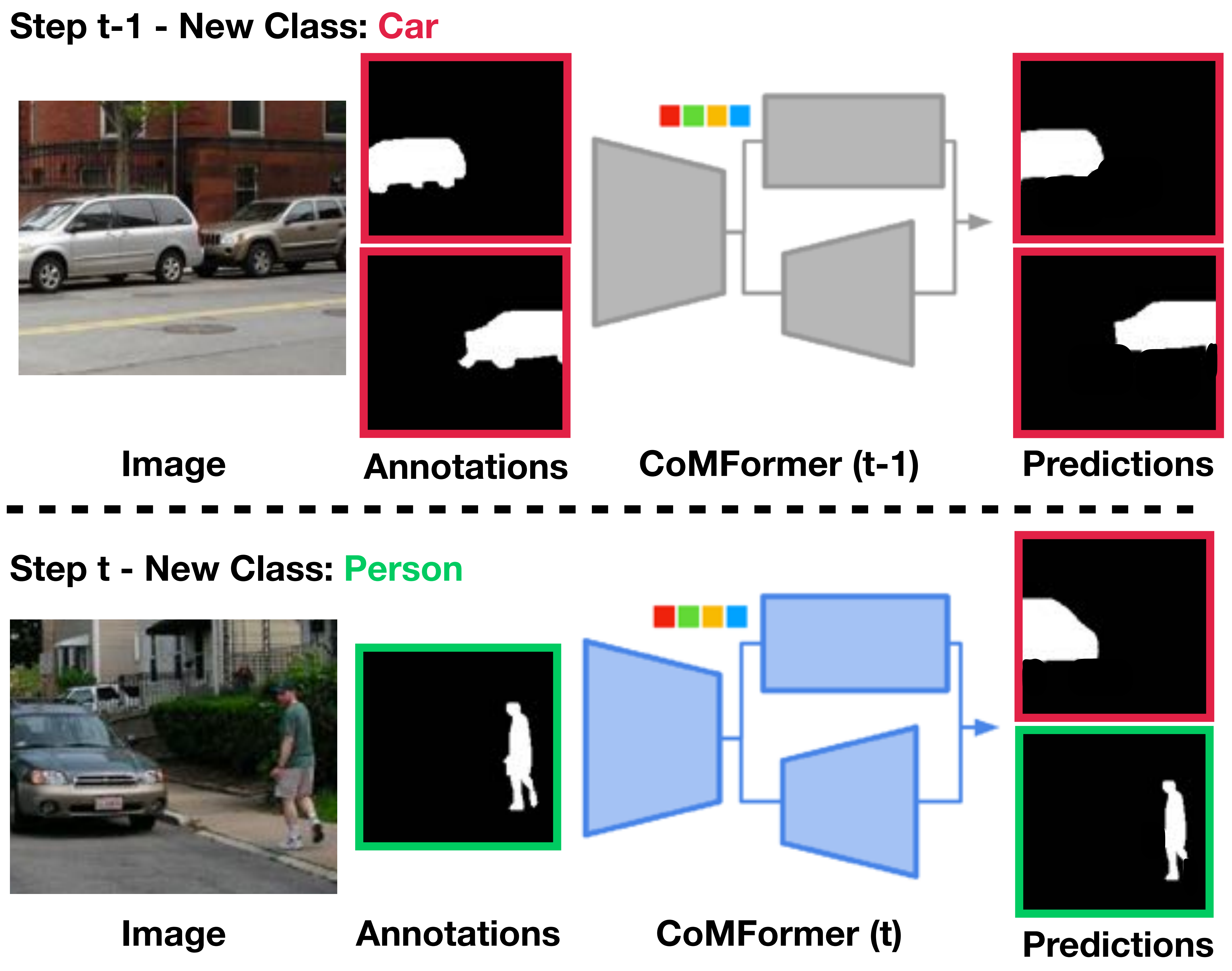


Continual Segmentation

A Unified Setting

We propose a **continual learning setting** unifying semantic and panoptic segmentation. The training is done in **multiple learning steps** *t=1…T*, each introducing a new set of classes.

At training step *t,* the **annotation** is provided only for the novel classes, while for the old ones is not present.



Continual Segmentation

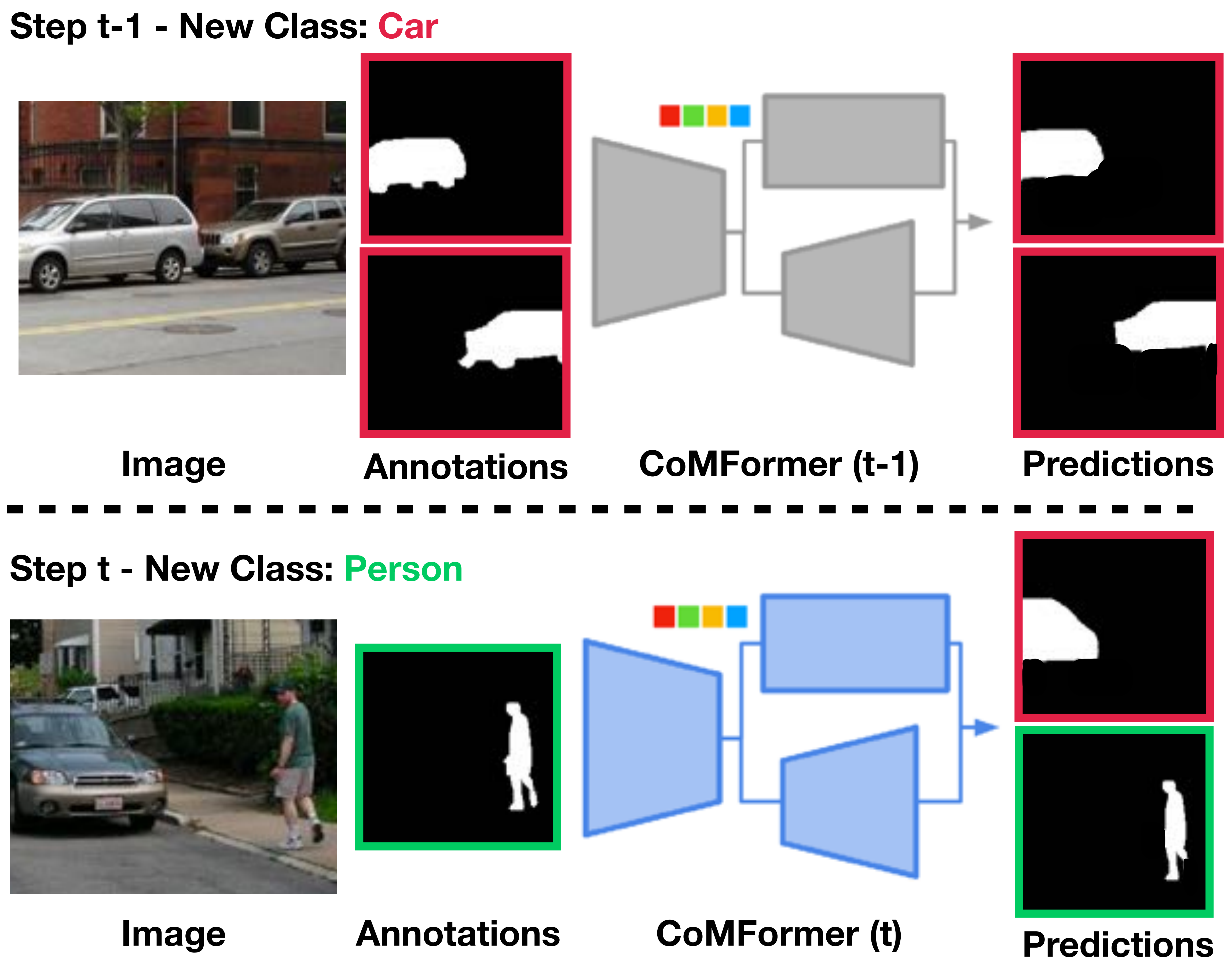
A Unified Setting

We propose a **continual learning setting** unifying semantic and panoptic segmentation. The training is done in **multiple learning steps** *t=1…T*, each introducing a new set of classes.

At training step *t,* the **annotation** is provided only for the novel classes, while for the old ones is not present.

The label is composed by a set of pairs made by the **ground-truth class and the binary mask**, indicating where the object appears.

The **goal** is to obtain a model able to predict all the seen classes, without forgetting.



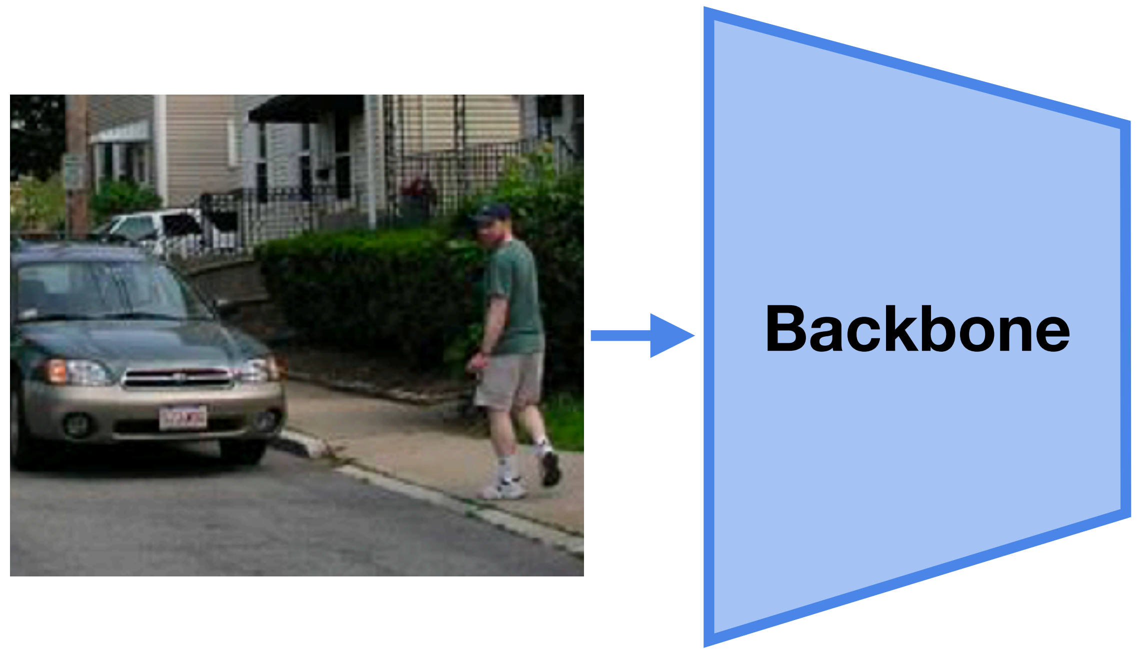
CoMFormer

ARCHITECTURE

Mask2Former: Masked-attention mask transformer for universal image segmentation. B. Cheng, I. Misra, A. G Schwing, A. Kirillov, R. Girdhar in CVPR 22

CoMFormer

The **architecture** is based on **Mask2Former**. A **backbone** extract image features. The **transformer decoder** takes image features and N learnable queries and outputs N per-segment embeddings Q. The **pixel decoder** takes the image features and extract per-pixel embeddings $\mathscr{E}_{pixel}$.

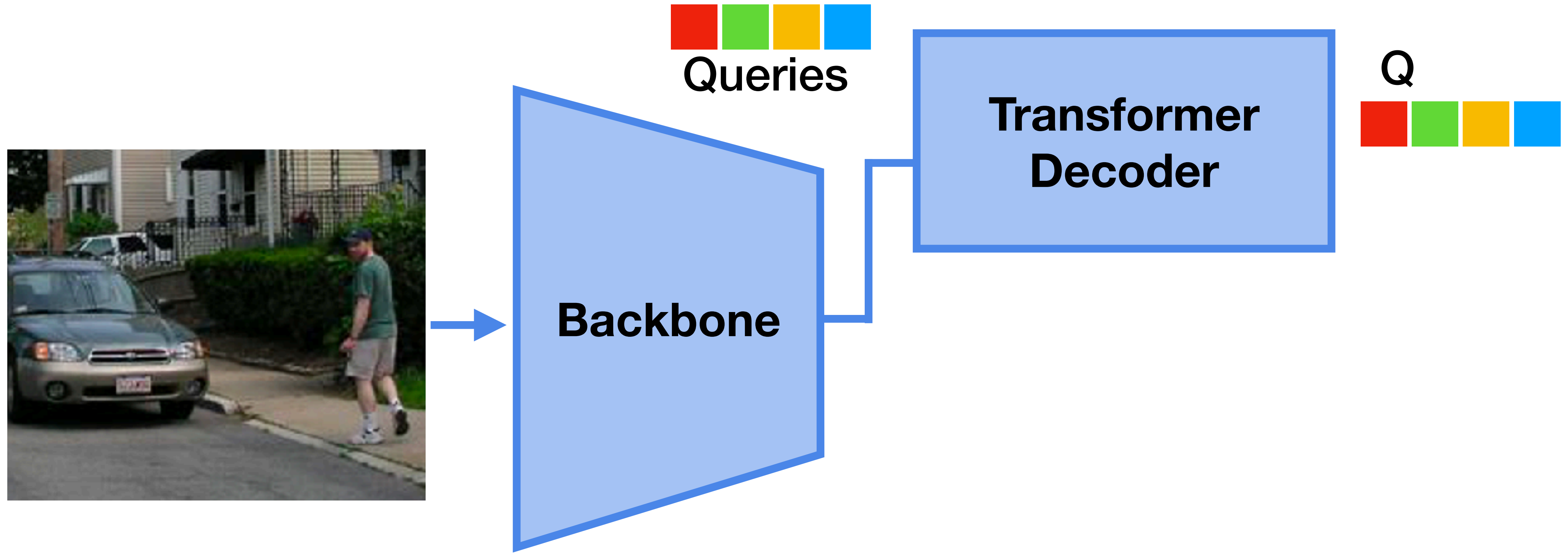


Mask2Former: Masked-attention mask transformer for universal image segmentation. B. Cheng, I. Misra, A. G Schwing, A. Kirillov, R. Girdhar in CVPR 22

CoMFormer

Architecture

The **architecture** is based on **Mask2Former**. A **backbone** extract image features. The **transformer decoder** takes image features and N learnable queries and outputs N per-segment embeddings Q. The **pixel decoder** takes the image features and extract per-pixel embeddings $\mathscr{E}_{pixel}$.

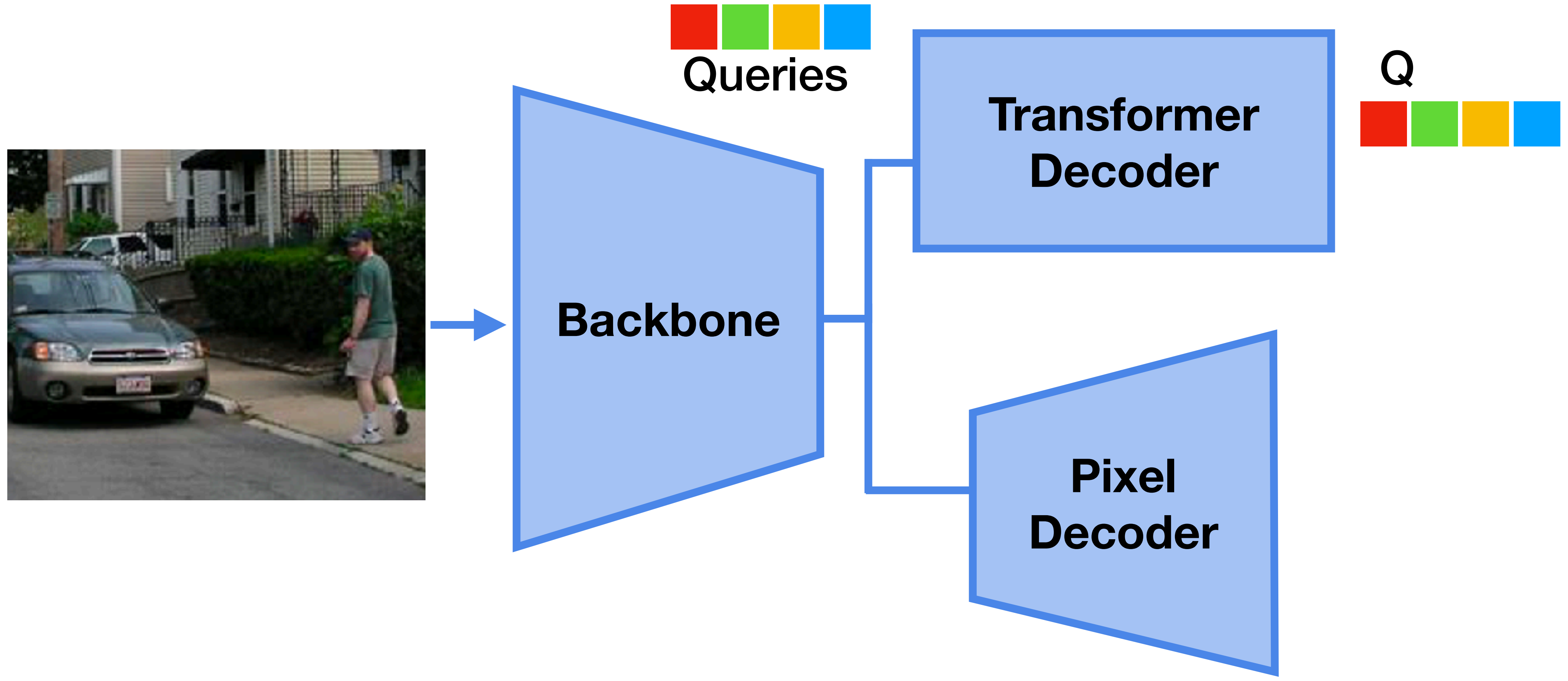


Mask2Former: Masked-attention mask transformer for universal image segmentation. B. Cheng, I. Misra, A. G Schwing, A. Kirillov, R. Girdhar in CVPR 22

CoMFormer

Architecture

The **architecture** is based on **Mask2Former**. A **backbone** extract image features. The **transformer decoder** takes image features and N learnable queries and outputs N per-segment embeddings Q. The **pixel decoder** takes the image features and extract per-pixel embeddings $\mathscr{E}_{pixel}$.



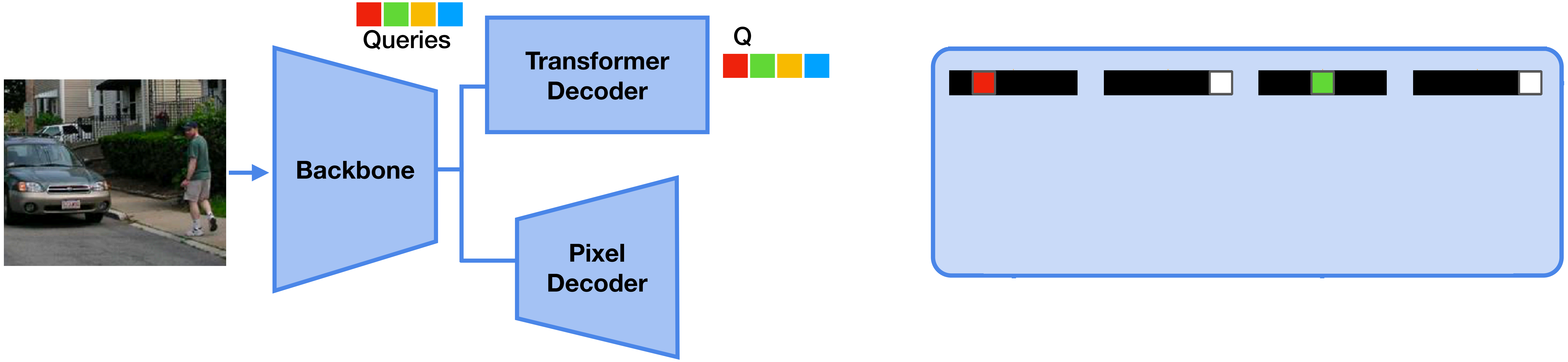
Mask2Former: Masked-attention mask transformer for universal image segmentation. B. Cheng, I. Misra, A. G Schwing, A. Kirillov, R. Girdhar in CVPR 22

CoMFormer

ARCHITECTURE

The **architecture** is based on **Mask2Former**. A **backbone** extract image features. The **transformer decoder** takes image features and N learnable queries and outputs N per-segment embeddings Q. The **pixel decoder** takes the image features and extract per-pixel embeddings $\mathscr{E}_{pixel}$.

A classifier is then applied to Q, obtaining **class probabilities** for each segment.



Mask2Former: Masked-attention mask transformer for universal image segmentation. B. Cheng, I. Misra, A. G Schwing, A. Kirillov, R. Girdhar in CVPR 22

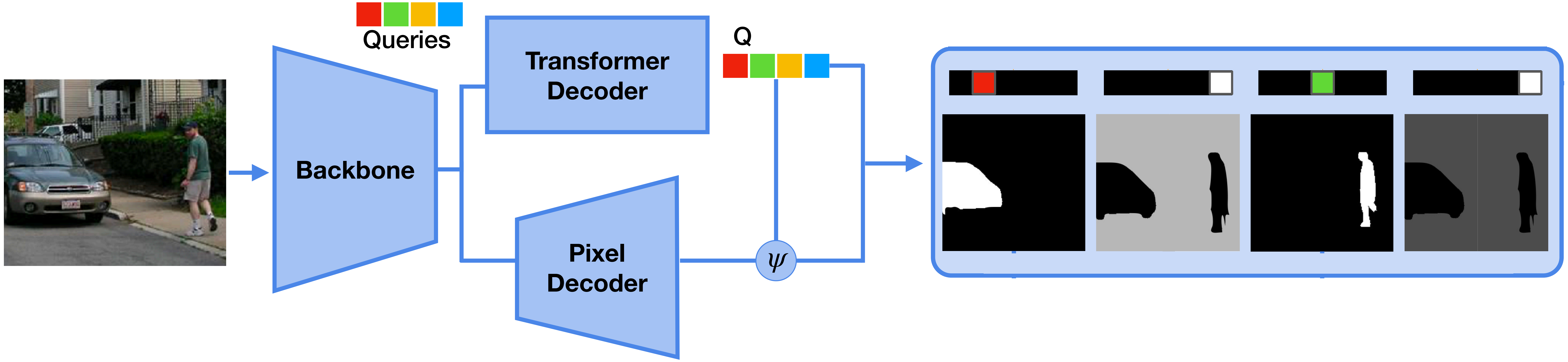
CoMFormer

ARCHITECTURE

The **architecture** is based on **Mask2Former**. A **backbone** extract image features. The **transformer decoder** takes image features and N learnable queries and outputs N per-segment embeddings Q. The **pixel decoder** takes the image features and extract per-pixel embeddings $\mathscr{E}_{pixel}$.

A classifier is then applied to Q, obtaining **class probabilities** for each segment.

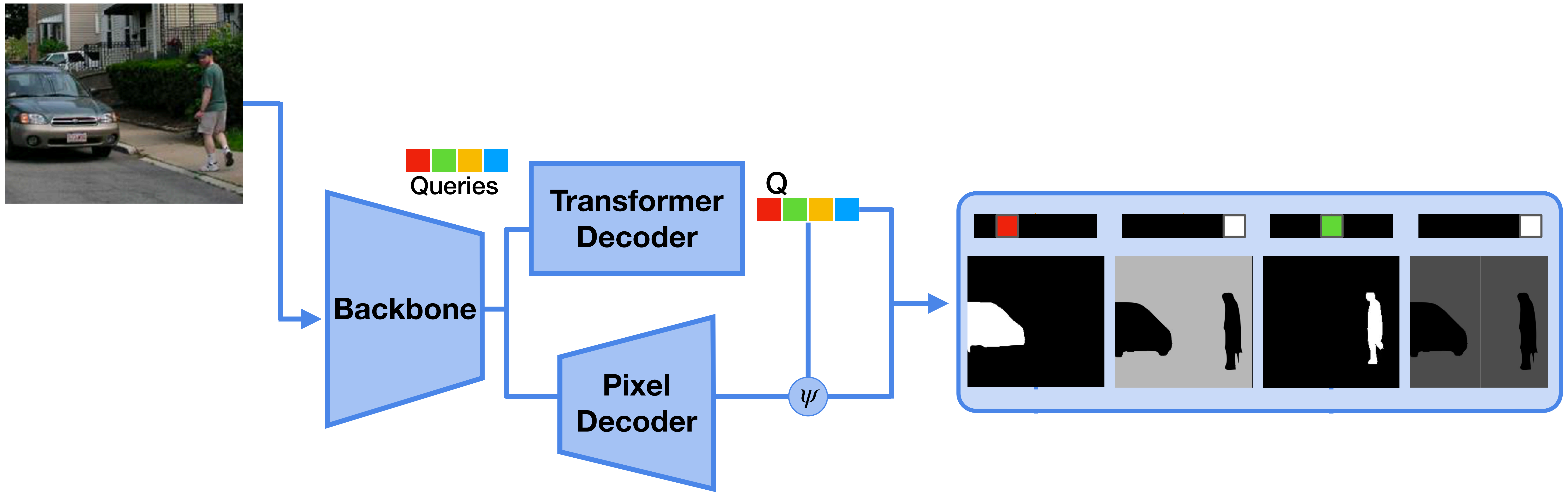
To obtain the **binary masks**, Q and $\mathscr{E}_{pixel}$ are multiplied and binarized. To operate in continual learning, we replace the *sigmoid in Mask2Former with* the ***softmax* for binarizing** to introduce inter-segment competition.



Mask2Former: Masked-attention mask transformer for universal image segmentation. B. Cheng, I. Misra, A. G Schwing, A. Kirillov, R. Girdhar in CVPR 22

CoMFormer

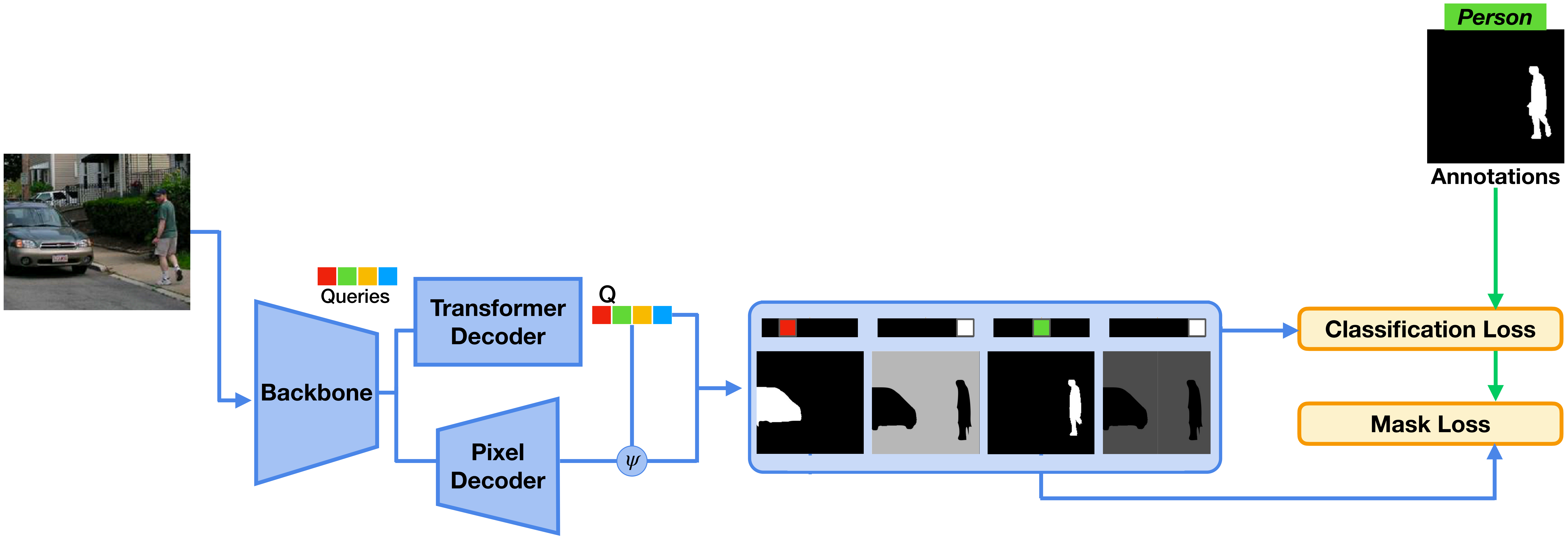
Learning without Forgetting



CoMFormer

Learning without Forgetting

To **learn the novel classes**, we use two **losses** exploring the provided annotations.

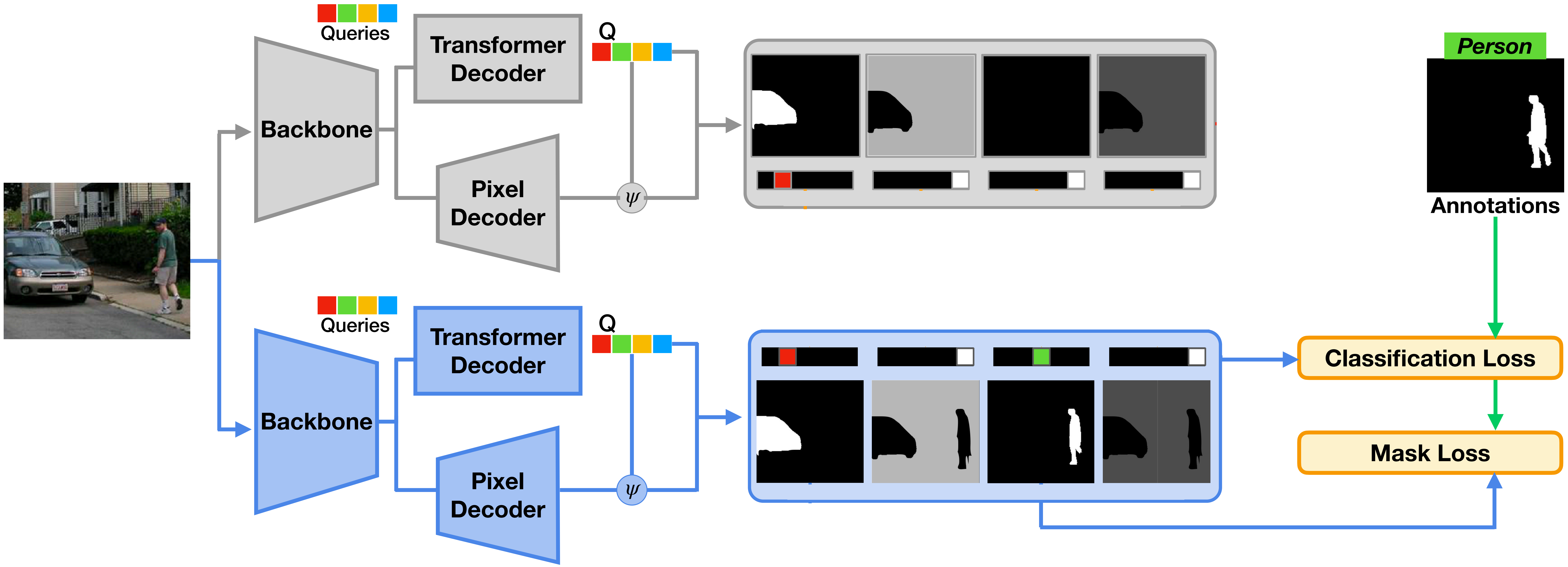


CoMFormer

Learning without Forgetting

To **learn the novel classes**, we use two **losses** exploring the provided annotations.

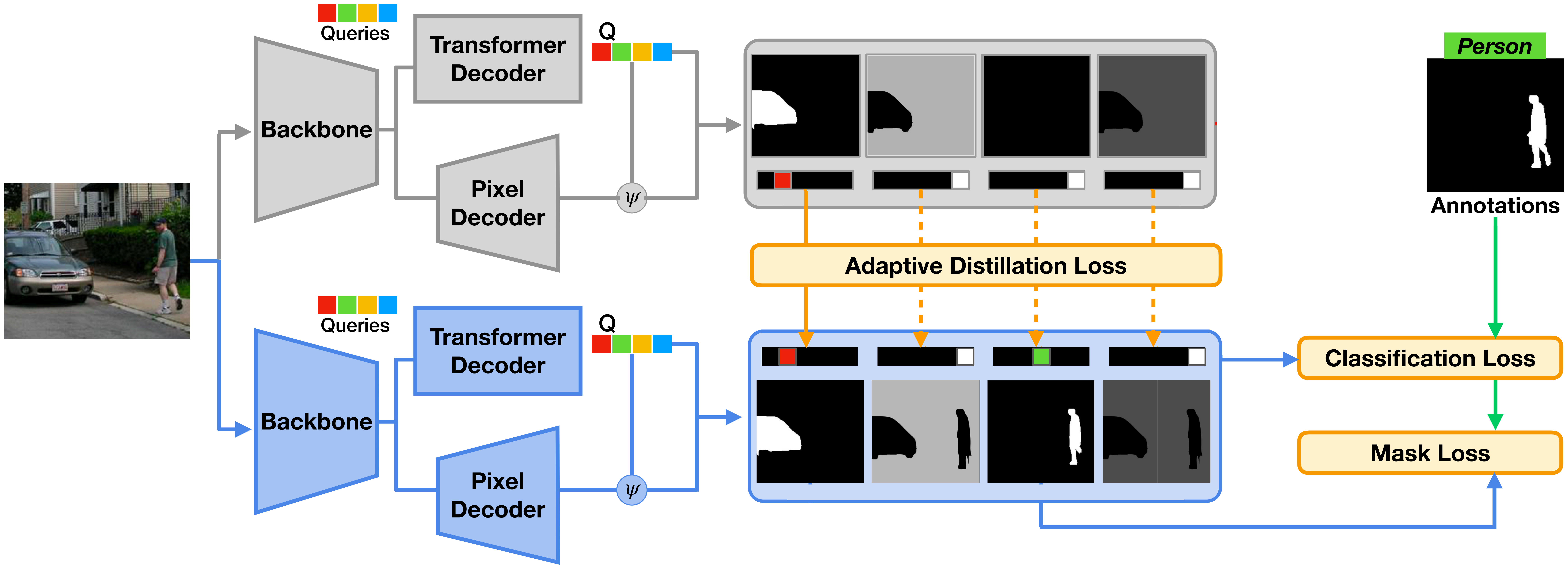
To **avoid forgetting**, we design a **knowledge distillation framework** bases on two components:



CoMFormer

LEARNING WITHOUT FORGETTING

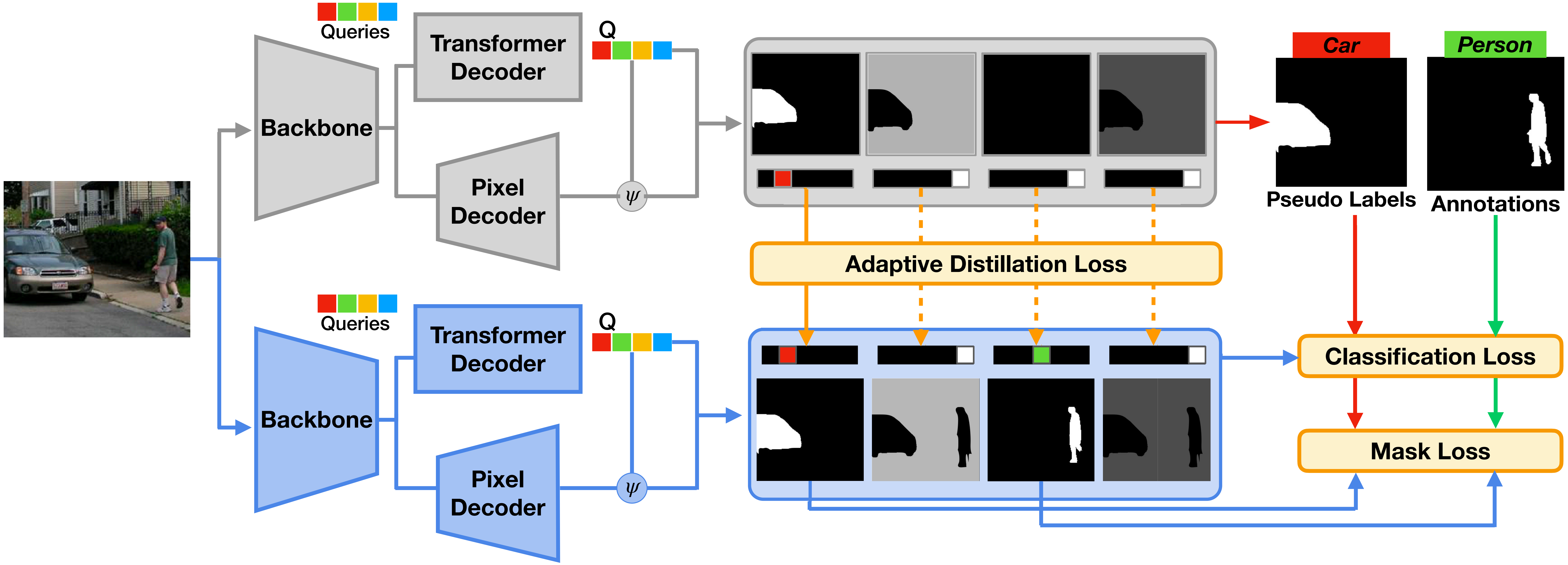
To **learn the novel classes**, we use two **losses** exploring the provided annotations.
To **avoid forgetting**, we design a **knowledge distillation framework** bases on two components:
To regularize the classifier, we use an **adaptive distillation loss**, weighting each mask contribution.



CoMFormer

LEARNING WITHOUT FORGETTING

To **learn the novel classes**, we use two **losses** exploring the provided annotations.
To **avoid forgetting**, we design a **knowledge distillation framework** bases on two components:
To regularize the classifier, we use an **adaptive distillation loss**, weighting each mask contribution.
Finally, we employ a **mask-based pseudo-labeling** to annotate old classes appearing in the image.



Results

Results on the ADE20K [3] dataset starting from 100 classes and performing a single step of 50 (100-50), or five steps of 10 (100-10), or ten steps of 5 (100-5) classes.

Results are reported in Panoptic Quality (PQ) after performing all the training steps.

Panoptic Segmentation
Joint
FT
MiB [1]
PLOP [2]
CoMFormer

CoMFormer Ground-truth

[1] Modeling the background for incremental learning in semantic segmentation. F. Cermelli, M. Mancini, S. Rota Bulò, E. Ricci, and B. Caputo in CVPR 20.
[2] Plop: Learning without forgetting for continual semantic segmentation. A. Douillard, Y. Chen, A. Dapogny, and M. Cord in CVPR 21.
[3] Scene parsing through ade20k dataset. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba in CVPR 17

Results

Results on the ADE20K [3] dataset starting from 100 classes and performing a single step of 50 (100-50), or five steps of 10 (100-10), or ten steps of 5 (100-5) classes.

Results are reported in Panoptic Quality (PQ) after performing all the training steps.

Panoptic Segmentation	100-50 (PQ)		
	1-100	101-150	All
Joint	43.2	32.1	39.5
FT	0.0	25.8	8.6
MiB [1]	35.1	19.3	29.8
PLOP [2]	41.0	26.6	36.2
CoMFormer	**41.1**	**27.7**	**36.7**

CoMFormer Ground-truth

[1] Modeling the background for incremental learning in semantic segmentation. F. Cermelli, M. Mancini, S. Rota Bulò, E. Ricci, and B. Caputo in CVPR 20.
[2] Plop: Learning without forgetting for continual semantic segmentation. A. Douillard, Y. Chen, A. Dapogny, and M. Cord in CVPR 21.
[3] Scene parsing through ade20k dataset. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba in CVPR 17

Results

Results on the ADE20K [3] dataset starting from 100 classes and performing a single step of 50 (100-50), or five steps of 10 (100-10), or ten steps of 5 (100-5) classes.

Results are reported in Panoptic Quality (PQ) after performing all the training steps.

Panoptic Segmentation	100-50 (PQ)			100-10 (PQ)		
	1-100	101-150	All	1-100	101-150	All
Joint	43.2	32.1	39.5	43.2	32.1	39.5
FT	0.0	25.8	8.6	0.0	2.9	1.0
MiB [1]	35.1	19.3	29.8	27.1	10.0	21.4
PLOP [2]	41.0	26.6	36.2	30.5	**17.5**	26.1
CoMFormer	**41.1**	**27.7**	**36.7**	**36.0**	17.1	**29.7**

CoMFormer Ground-truth

[1] Modeling the background for incremental learning in semantic segmentation. F. Cermelli, M. Mancini, S. Rota Bulò, E. Ricci, and B. Caputo in CVPR 20.
[2] Plop: Learning without forgetting for continual semantic segmentation. A. Douillard, Y. Chen, A. Dapogny, and M. Cord in CVPR 21.
[3] Scene parsing through ade20k dataset. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba in CVPR 17

Results

CONTINUAL PANOPTIC SEGMENTATION

Results on the ADE20K [3] dataset starting from 100 classes and performing a single step of 50 (100-50), or five steps of 10 (100-10), or ten steps of 5 (100-5) classes.

Results are reported in Panoptic Quality (PQ) after performing all the training steps.

Panoptic Segmentation	100-50 (PQ)			100-10 (PQ)			100-5 (PQ)		
	1-100	101-150	All	1-100	101-150	All	1-100	101-150	All
Joint	43.2	32.1	39.5	43.2	32.1	39.5	43.2	32.1	39.5
FT	0.0	25.8	8.6	0.0	2.9	1.0	0.0	1.3	0.4
MiB [1]	35.1	19.3	29.8	27.1	10.0	21.4	24.0	6.5	175.7
PLOP [2]	41.0	26.6	36.2	30.5	**17.5**	26.1	28.1	15.7	24.0
CoMFormer	**41.1**	**27.7**	**36.7**	**36.0**	17.1	**29.7**	**34.4**	**15.9**	**28.2**

CoMFormer Ground-truth

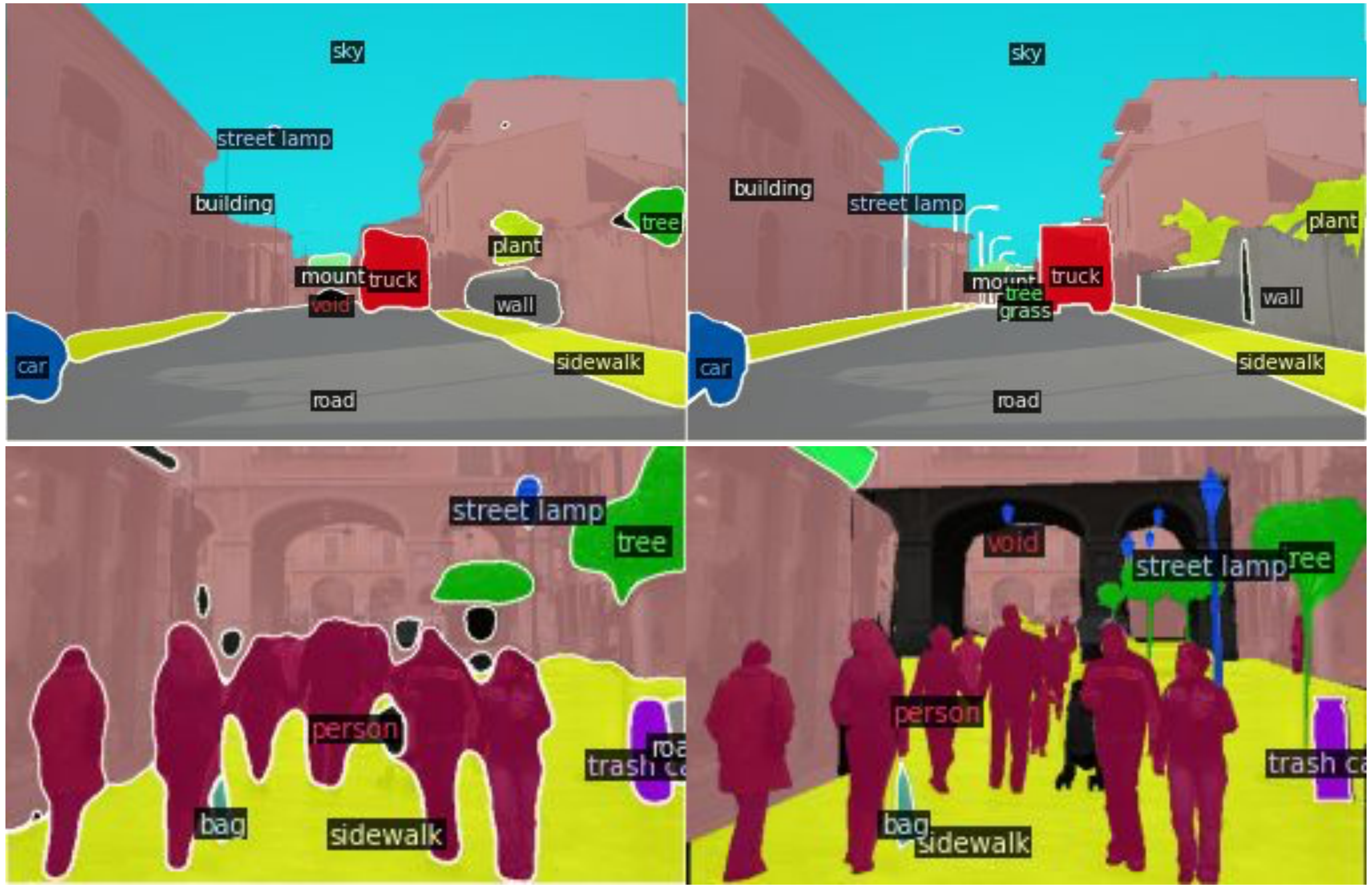
[1] Modeling the background for incremental learning in semantic segmentation. F. Cermelli, M. Mancini, S. Rota Bulò, E. Ricci, and B. Caputo in CVPR 20.
[2] Plop: Learning without forgetting for continual semantic segmentation. A. Douillard, Y. Chen, A. Dapogny, and M. Cord in CVPR 21.
[3] Scene parsing through ade20k dataset. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba in CVPR 17

Results

Results on the ADE20K [3] dataset starting from 100 classes and performing a single step of 50 (100-50), or five steps of 10 (100-10), or ten steps of 5 (100-5) classes.

Results are reported in mean Intersection over Union (mIoU) after performing all the training steps.

Semantic Segmentation
Joint
FT
MiB [1]
PLOP [2]
CoMFormer



CoMFormer Ground-truth

[1] Modeling the background for incremental learning in semantic segmentation. F. Cermelli, M. Mancini, S. Rota Bulò, E. Ricci, and B. Caputo in CVPR 20.
[2] Plop: Learning without forgetting for continual semantic segmentation. A. Douillard, Y. Chen, A. Dapogny, and M. Cord in CVPR 21.
[3] Scene parsing through ade20k dataset. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba in CVPR 17

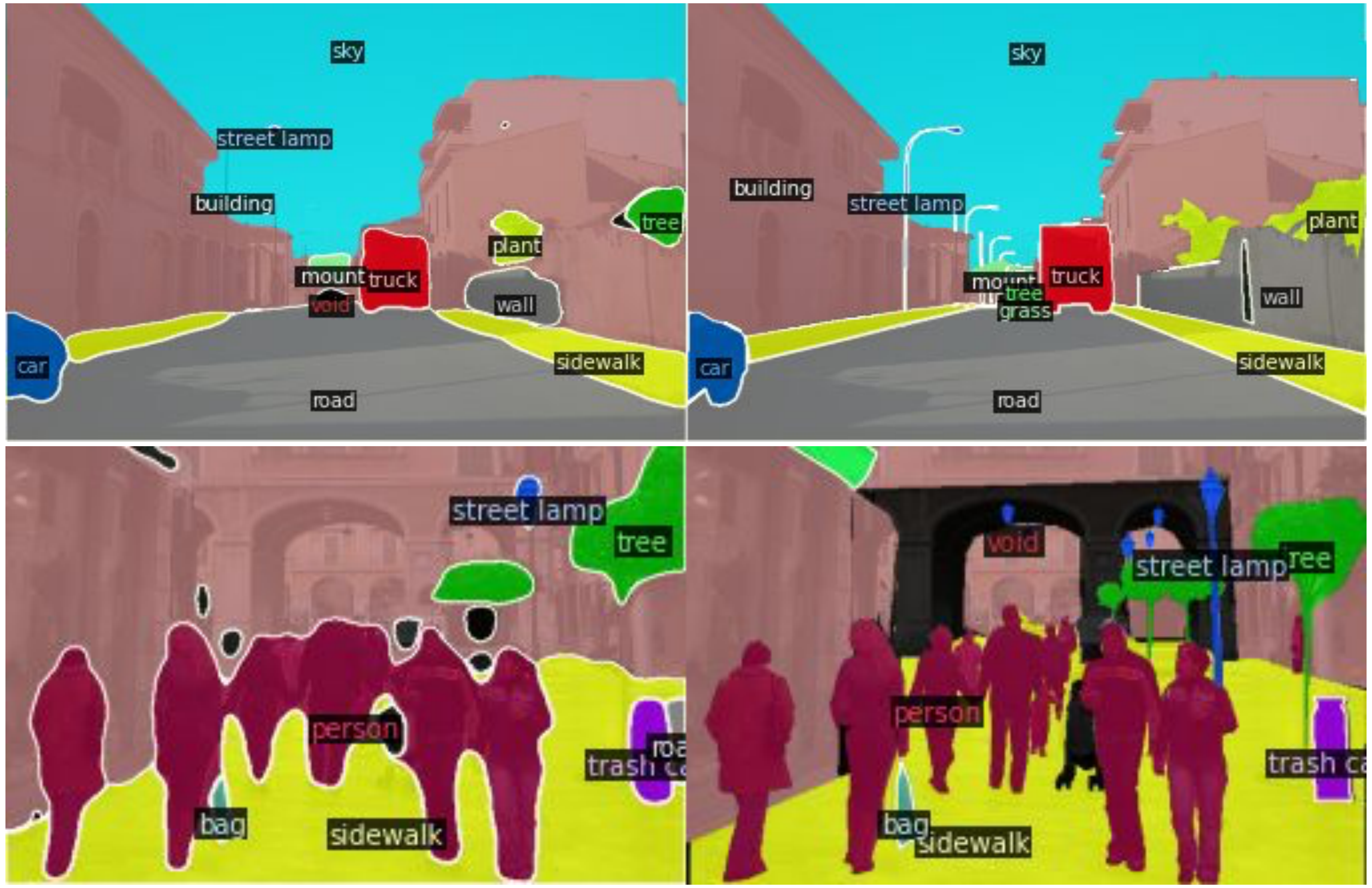
Results

Continual Semantic Segmentation

Results on the ADE20K [3] dataset starting from 100 classes and performing a single step of 50 (100-50), or five steps of 10 (100-10), or ten steps of 5 (100-5) classes.

Results are reported in mean Intersection over Union (mIoU) after performing all the training steps.

Semantic Segmentation	100-50 (mIoU)		
	1-100	101-150	All
Joint	46.9	35.6	43.1
FT	0.0	**26.7**	8.9
MiB [1]	37.0	24.1	32.6
PLOP [2]	44.2	26.2	38.2
CoMFormer	**44.7**	26.2	**38.4**



CoMFormer Ground-truth

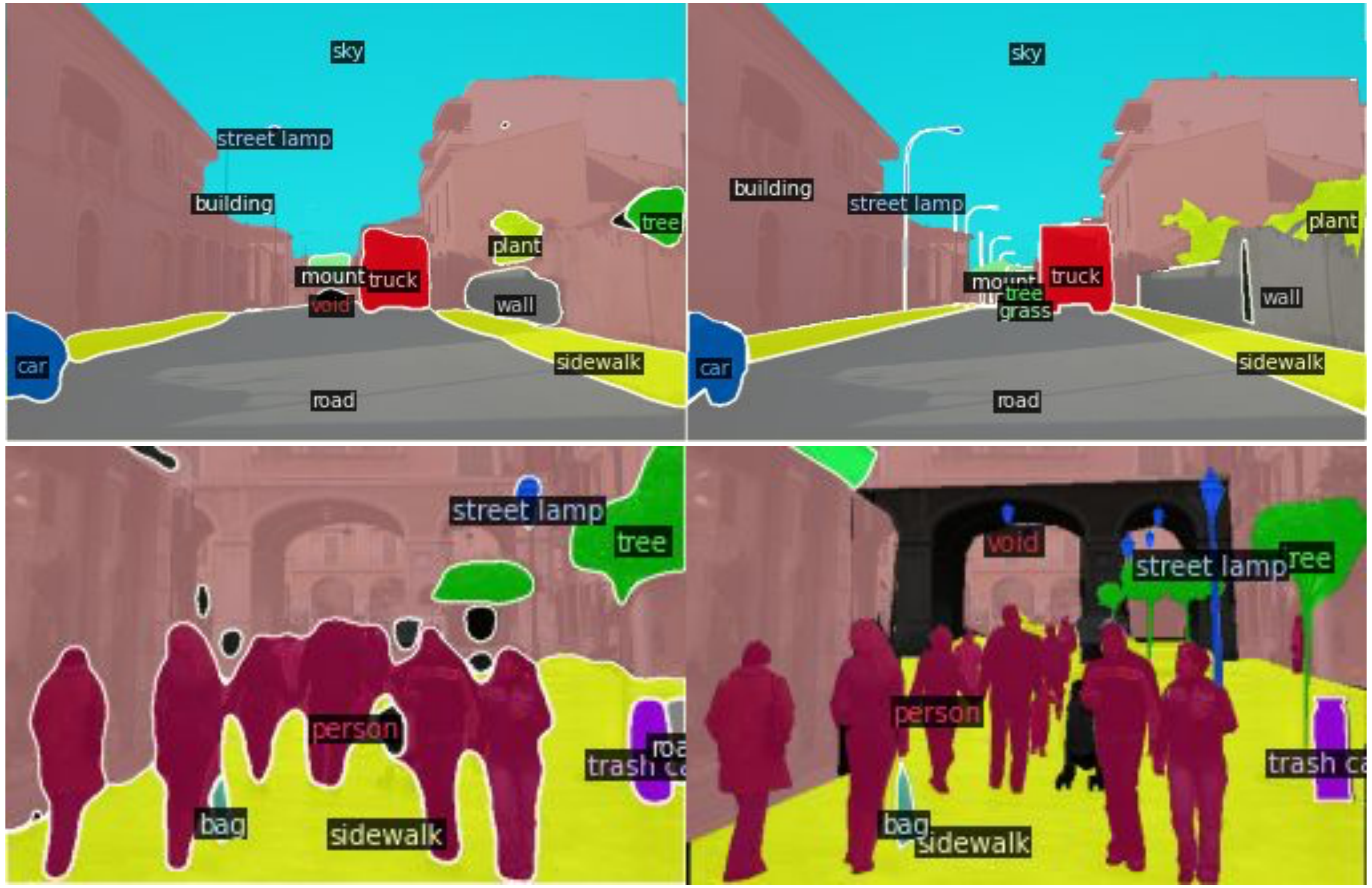
[1] Modeling the background for incremental learning in semantic segmentation. F. Cermelli, M. Mancini, S. Rota Bulò, E. Ricci, and B. Caputo in CVPR 20.
[2] Plop: Learning without forgetting for continual semantic segmentation. A. Douillard, Y. Chen, A. Dapogny, and M. Cord in CVPR 21.
[3] Scene parsing through ade20k dataset. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba in CVPR 17

Results

Results on the ADE20K [3] dataset starting from 100 classes and performing a single step of 50 (100-50), or five steps of 10 (100-10), or ten steps of 5 (100-5) classes.

Results are reported in mean Intersection over Union (mIoU) after performing all the training steps.

Semantic Segmentation	100-50 (mIoU)			100-10 (mIoU)		
	1-100	101-150	All	1-100	101-150	All
Joint	46.9	35.6	43.1	46.9	35.6	43.1
FT	0.0	**26.7**	8.9	0	2.3	0.8
MiB [1]	37.0	24.1	32.6	23.5	10.6	26.6
PLOP [2]	44.2	26.2	38.2	34.8	**15.9**	28.5
CoMFormer	**44.7**	26.2	**38.4**	**40.3**	15.6	**32.3**



CoMFormer Ground-truth

[1] Modeling the background for incremental learning in semantic segmentation. F. Cermelli, M. Mancini, S. Rota Bulò, E. Ricci, and B. Caputo in CVPR 20.
[2] Plop: Learning without forgetting for continual semantic segmentation. A. Douillard, Y. Chen, A. Dapogny, and M. Cord in CVPR 21.
[3] Scene parsing through ade20k dataset. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba in CVPR 17

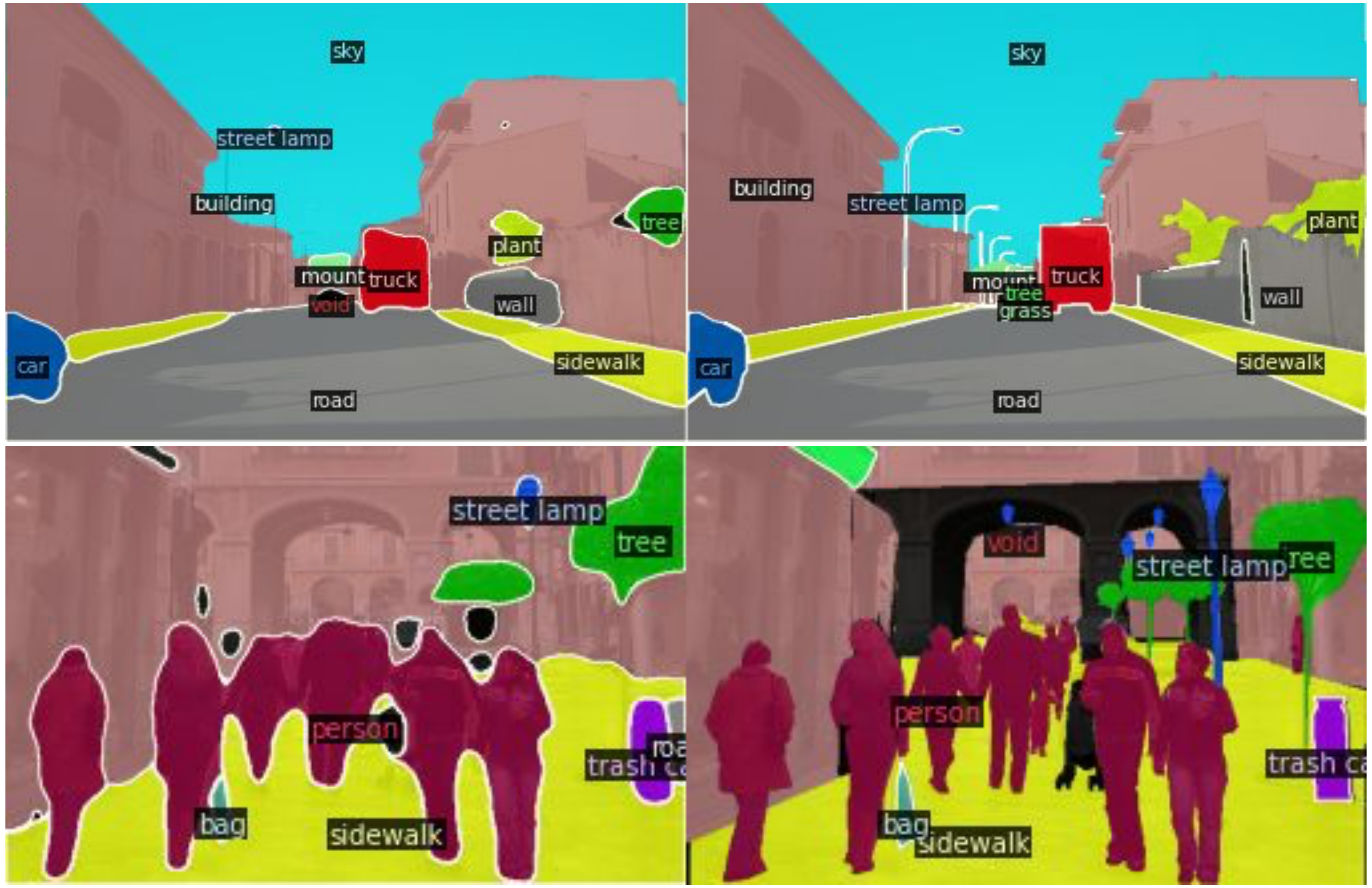
Results

CONTINUAL SEMANTIC SEGMENTATION

Results on the ADE20K [3] dataset starting from 100 classes and performing a single step of 50 (100-50), or five steps of 10 (100-10), or ten steps of 5 (100-5) classes.

Results are reported in mean Intersection over Union (mIoU) after performing all the training steps.

Semantic Segmentation	100-50 (mIoU)			100-10 (mIoU)			100-5 (mIoU)		
	1-100	101-150	All	1-100	101-150	All	1-100	101-150	All
Joint	46.9	35.6	43.1	46.9	35.6	43.1	46.9	35.6	43.1
FT	0.0	**26.7**	8.9	0	2.3	0.8	0.0	1.1	0.3
MiB [1]	37.0	24.1	32.6	23.5	10.6	26.6	21.0	6.1	16.1
PLOP [2]	44.2	26.2	38.2	34.8	**15.9**	28.5	33.6	**14.1**	27.1
CoMFormer	**44.7**	26.2	**38.4**	**40.3**	15.6	**32.3**	**39.5**	13.6	**30.9**



CoMFormer   Ground-truth

[1] Modeling the background for incremental learning in semantic segmentation. F. Cermelli, M. Mancini, S. Rota Bulò, E. Ricci, and B. Caputo in CVPR 20.
[2] Plop: Learning without forgetting for continual semantic segmentation. A. Douillard, Y. Chen, A. Dapogny, and M. Cord in CVPR 21.
[3] Scene parsing through ade20k dataset. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba in CVPR 17

Thank You!
Visit our Poster TUE-AM-286!

Scan for the code!

CoMFormer: Continual Learning in Semantic and Panoptic Segmentation

Fabio Cermelli, Matthieu Cord, Arthur Douillard