Reliability in Semantic Segmentation: Can We Use Synthetic Data? — *Supplementary Material* —

Thibaut Loiseau[2]†, Tuan-Hung Vu[1], Mickael Chen[1], Patrick Pérez[3], and Matthieu Cord[1,4]

[1] valeo.ai, Paris, France
[2] LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-Vallee, France
[3] kyutai, Paris, France
[4] Sorbonne Université, Paris, France

In this document, we provide technical details for covariate shifts generation and OOD object inpainting (Appendix A), additional calibration details and results (Appendix B), analysis on direct real-*vs*.-synthetic data correlation (Appendix C) and class-wise PCC scores (Appendix D). Appendix E discusses the limitations and Appendix F showcases more qualitative examples.

A Technical Details

A.1 Covariate Shifts Training

We train a ControlNet on top of a frozen Stable Diffusion 1.5 for 2100 steps. The ControlNet used here is a trainable copy of the Stable Diffusion encoder only, as in the original paper [9]. We use a batch size of 8 with 32 gradient accumulation steps, which makes an effective batch size of 256, and a learning rate of 10^{-5}. We use the training set of Cityscapes, and do a random horizontal crop of the images to get square images, and then resize them to 512×512, convenient of Stable Diffusion 1.5. All other training hyperparameters are the per default settings on the official ControlNet repository. The objective is to reconstruct the original images of Cityscapes using its semantic masks as input to the ControlNet, and the captions extracted with CLIP-interrogator as input to Stable Diffusion.

Similarly for SDXL, we train a ControlNet for 27500 steps and use a batch size of 8 with 4 gradient accumulation steps, which makes an effective batch size of 32. The learning rate is 10^{-5}. The images are square cropped to get 1024×1024 images, convenient for SDXL, and there is no need to resize them.

A.2 Covariate Shifts Generation

To generate images with new styles with SD 1.5, we take a semantic mask from the validation set of Cityscapes, crop and resize it as explained in the previous part. We use nearest neighbor interpolation to keep good values for

† Work done during an internship at valeo.ai

specific classes. We only use the part of the caption extracted with CLIP-interrogator that corresponds to a BLIP caption. To this new caption, we add `[, in <domain>]` depending on the domain we want to generalize to. Starting from pure noise, we use 25 DDIM steps with a guidance scale of 8. On a RTX 2080, one new image is generated in about 4 seconds. All other sampling hyper-parameters are the per default settings on the official ControlNet repository.

For SDXL, we tune the conditioning strength and the prompt strength to get better images. To get both mask and prompt aligned images, we set the ControlNet conditioning strength to 0.65 and the prompt guidance to 10. The prompt tuning is the same as used for SD 1.5; we also use 25 denoising steps.

A.3 OOD Objects Generation

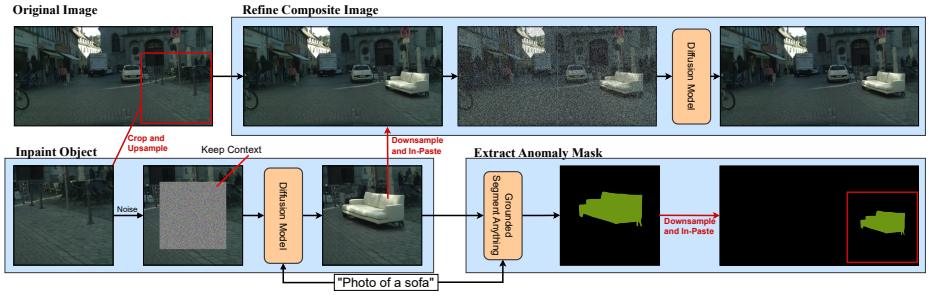


Fig. 12: OOD object data generation pipeline. We use pretrained Stable Diffusion for inpainting and refining steps, and pretrained Grounded Segment Anything for mask extraction.

The OOD object generation pipeline described in Sec. 4.1 is further illustrated in Fig. 12. It is composed of three steps. Given a text prompt containing an object, the inpainting step generates a zoomed-in version of the object with the appropriate close-range background given as context. The mask extraction step infers the anomaly mask from the zoomed-in generated image and the name of the object. Both are in-pasted back in the original complete image or mask. To reduce some composition artifacts, the composite image is refined with a noise/denoise step.

In details, we first randomly choose a box size for the new object, uniformly sampled between a quarter and half the minimum dimension of the original image. We also uniformly sample a location for the box in the bottom three quarters of the image. This box will contain the new object we wish to add, and we refer to it as *inpainted region*. In addition to the inpainted region, we create a larger box, 1.5× its height and width, with the inpainted region in its center. The contour outside of the inpainted region will serve as *context* for the inpainting process. We then crop and resize the inpainted region with its context

to a resolution of 512 × 512. We fully noise the inpainted region, but leave the clean context. We then denoise the inpainted zone with the prompt *"A photo of an [object]"*, with a guidance scale of 15. The full patch is then resized and pasted on the original image, at its original position. Some artifact might be still present as shown in Fig. 13. To remedy this, we refine the inpainted zone by noising and denoising it with 0.65 strength, with the default guidance scale of 7.5. The effect of this refining step is shown in Fig. 13.

We list here all 42 objects used in our experiments, which are not present in Cityscapes' classes: arcade machine, armchair, baby, bag, bathtub, bench, billboard, book, bottle, box, chair, cheetah, chimpanzee, clock, computer, desk, dolphin, elephant, flamingo, giraffe, gorilla, graffiti, hippopotamus, kangaroo, koala, lamp, lion, microwave, mirror, panda, penguin, pillow, plate, polar bear, radiator, refrigerator, sofa, table, tiger, toilet, vase, and zebra.

A.4 OOD Detection Training

For the OOD detection method in Sec. 4.2, we used the codebase of [5]. We adapt the code to be able to use our generated data with the binary masks extracted from Grounded-SAM, as explained in Appendix A.3. As in the original paper [5], we fine-tune the mask prediction MLP and classification layer after the transformer decoder. To obtain all OOD detection results reported in the main paper (Tab. 2), we used the recommended hyperparameters, and train the models for 5000 iterations.

A.5 Segmentation Models

We list here all models used in our experiments: ANN-R101, ANN-R50, APCNet-R101, APCNet-R50, BiSeNetV1-R50, BiSeNetV2-FCN, CCNet-R101, CCNet-R50, ConvNext, ConvNext-B-In1K, ConvNext-B-In21K, DLV3+ResNet101, DLV3+ResNet18, DLV3+ResNet50, GCNet-R101, GCNet-R50, ICNet-R101, ICNet-R18, ICNet-R50, MobileNetV3, PSPNet-R101, PSPNet-R18, PSPNet-R50, SETR-MLA, SETR-Naive, SETR-PUP, SegFormer-B0, SegFormer-B1, SegFormer-B2, SegFormer-B3, SegFormer-B4, SegFormer-B5, SegFormer-B5-v2, SegFormer-B5-v3, Segmenter, SemFPN-R101, SemFPN-R50, UpperNet-R101, UpperNet-R18, UpperNet-R50.

B Calibration Details and Additional Results

We elaborate on our strategy for performing per-class calibration to obtain the synthetic results (▮) presented in Fig. 6. Utilizing our synthetic data, a temperature scaling (TS) scalar is learned for each class. When calibrating models on shifted domains, we choose the corresponding TS scalar based on model predictions. In the case of calibration results with real-shift data (▮), only one scalar is learned for each model.

In Fig. 14a, we compare per-class TS *vs.* standard TS with one scalar per model, both applied on our synthetic data. Both strategies enhance calibration results, highlighting the advantage of employing synthetic data for calibration.

Per-class TS demonstrates superiority for more robust models (right part of the plots), while its performance is weaker for less robust ones (left part of the plots).

In Fig. 14b, we compare Cityscapes *vs.* our synthetic data; in this experiment we adopt the standard TS with one scalar per model. The results obtained from Cityscapes are clearly inferior to those achieved using our synthetic data, demonstrating the limitations when relying solely on in-domain data for calibration in shifted domains.

Generalization. We ask ChatGPT the generic question like *"give me different cities/weathers that would be representative of the whole world?"*. We then use all ChatGPT's answers as prompts to generate OOD data, referred to as "all-domains". For confidence calibration, using "all-domains" shows comparable results as using domain-specific data with manual prompts (see Tab. 4), e.g. 100% ECE improvement for *rain*, 78.6% for *india* (cf. Fig.6). Interestingly, using "all-domains-but-rain" (no *rain* prompts), we also obtain similar improvement for *rain*. Detailed results are reported in Tab. 3. That experiment hints at the generalization potential of our framework. We list here all the domains we used to form the "all-domains" calibration set: Beijing, Cairo, clouds, Dubai, fall, fog, hurricane, India, Istanbul, Johannesburg, lightning, London, Moscow, Mumbai, night, Paris, rain, sandstorm, snow, spring, summer, sun, Sydney, Tokyo, tornado, Toronto, wind, winter. This calibration set is comprised of 64 images per shift.

Prompt	India	Fog	Rain	Snow	Night
domain-specific	72.5	92.5	100	95	90
all domains	78.6	100	100	100	100
all domains but-rain	-	-	100	-	-

Table 3: Generalization. We experiment the effects of prompting and showcase the generalization potentials of our framework. Here ECE improvements are reported (cf. Fig.6).

C Direct Data Correlation

We provide in Tab. 4 the FID scores(↓) for direct correlation between synthetic and real data. Our zero-shot approach outperforms all. Layout differences between Cityscapes and ACDC cause high FIDs – one critical limitation of FID on structured data. As our work focuses on performance testing, we prioritize metrics like mIoU and FPR_{95}, aligned with recent works.

Reflected in the FID study, in-domain data (CS) serves as a strong baseline for measuring OOD performance [3, 4]. Hendrycks' augmentations are unrealistic and inadequate for assessing real OOD performance, as previously revealed in [8]. Using untouched CS data could be even better; for example, in terms

	OOD expertise?	OOD data?	**Night**	**Rain**	**Snow**	**Fog**
Cityscapes (no aug.)	no	no	236.5	188.3	194.8	184.8
Hendrycks' aug. [1]	required	no	-	-	210.7	191.1
GAN-based TSIT [2]	no	required	254.2	223.2	225.6	-
Physics-based Fog Sim. [6]	required	no	-	-	-	182.8
Ours w/ SD1.5	no	no	**180.2**	**177.5**	**164.5**	**163.3**

Table 4: Direct Data Correlation. Our pipeline achieves better FID scores while does not require any OOD knowledge.

of PCC score for 'fog' (cf. Fig.3 and Tab.1), Hendrycks's data obtains 0.70 as compared to 0.78 of CS data and 0.89 of our SDXL data. Our work, consistent with [8], demonstrates that simple augmentations are insufficient for OOD testing, advocating for more realistic testing data. While showing the zero-shot advantage over standard domain-specific augmentations, we acknowledge that a well crafted augmentation approach like physics-based fog augmentation [6] can obtain very good results, potentially better than ours if improved. Of note, *we do not aim to obtain best results in all cases; we advocate for a generic zero-shot testing framework as the starting reference in arbitrary domains.*

D Class-wise analysis.

We conduct a class-wise analysis and report the PCC scores per class in Tab. 5. Interestingly, we notice that 'bicycle' and 'bus' have the least correlation for *india* and *night*, respectively, which actually corresponds to the low occurrences of those classes in such conditions. High correlation indicates that either the corresponding classes are easy ('building', 'person', 'car', or 'truck') or those classes are difficult ('sign' or 'pole') and make segmenters struggle either on real or synthetic data. We conjecture that using synthetic data may provide us with hints about the inherent bias of the pretrained models.

E Limitations

We focus our research on the task of semantic segmentation while keeping open the extension possibility to other critical tasks, such as object detection. Our quantitative assessments are confined to existing publicly available datasets. However, our framework is fully zero-shot and can be applied to any domain of interest. On the generative side, our study is restricted to Stable Diffusion and ControlNet due to our resource constraints. Of note, although improvements in this area should enhance the results, similar insights are expected to be achieved, as primarily shown with SDXL results. Another limitation is our primary focus on autonomous driving (AD) data. That is because the field of

Class	Fog	Night	Rain	Snow	India
road	0.54	0.69	0.57	0.60	0.74
sidewalk	0.62	0.62	0.68	0.64	0.69
building	0.67	0.82	0.76	0.57	0.84
wall	0.03	0.76	0.60	0.71	0.58
fence	0.21	0.66	0.71	0.64	0.38
pole	0.33	0.75	0.60	0.29	0.47
light	0.36	0.53	0.63	0.70	0.30
sign	0.33	0.82	0.47	0.55	0.26
vegetation	0.57	0.71	0.77	0.52	0.00
terrain	0.31	0.55	0.62	0.76	-
sky	0.70	0.44	0.07	0.17	0.30
person	0.56	0.72	0.52	0.69	0.86
rider	0.27	0.55	0.43	0.31	0.16
car	0.73	0.71	0.72	0.78	0.61
truck	0.83	0.67	0.74	0.75	0.72
bus	0.67	0.32	0.57	0.69	0.19
train	0.35	0.65	0.35	0.41	-
motocycle	0.25	0.39	0.37	0.62	0.40
bicycle	0.49	0.64	0.73	0.72	-0.16

Table 5: Class-wise analysis. We provide the PCC scores per class for each shift. The most and least scores are colored.

AD released diverse datasets to stress-test models across various OOD scenarios. Unlike generalist datasets like MS-COCO, they clearly distinguish between domains, enabling the covariate shift studies in this work.

Robustness Assessment. In order to accommodate the in-domain GTs, we make the semantic preservation assumption, similar to the one in domain adaptation. We realize that this common assumption might raise questions when an element like snow is added; nevertheless, the issue is complex and hinges on the annotation policies. The ACDC dataset [7], for instance, uses clean images to inform the annotators of what is behind the snow, and our approach mimics this. But they are also very conservative about ambiguity, with an explicit label. Moving forward, we could take inspiration and try to handle ambiguity.

F Qualitative Examples

We show more qualitative examples for synthetic covariate shifts in Fig. 15 and synthetic OOD objects in Fig. 16.

References

1. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019)

2. Jiang, L., Zhang, C., Huang, M., Liu, C., Shi, J., Loy, C.C.: Tsit: A simple and versatile framework for image-to-image translation. In: ECCV (2020)
3. de Jorge, P., Volpi, R., Torr, P.H., Rogez, G.: Reliability in semantic segmentation: Are we on the right track? In: CVPR (2023)
4. Miller, J.P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P.W., Shankar, V., Liang, P., Carmon, Y., Schmidt, L.: Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In: ICLR (2021)
5. Nayal, N., Yavuz, M., Henriques, J.F., Güney, F.: Rba: Segmenting unknown regions rejected by all. In: ICCV (2023)
6. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. IJCV (2018)
7. Sakaridis, C., Dai, D., Van Gool, L.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: ICCV (2021)
8. Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L.: Measuring robustness to natural distribution shifts in image classification. NeurIPS (2020)
9. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)

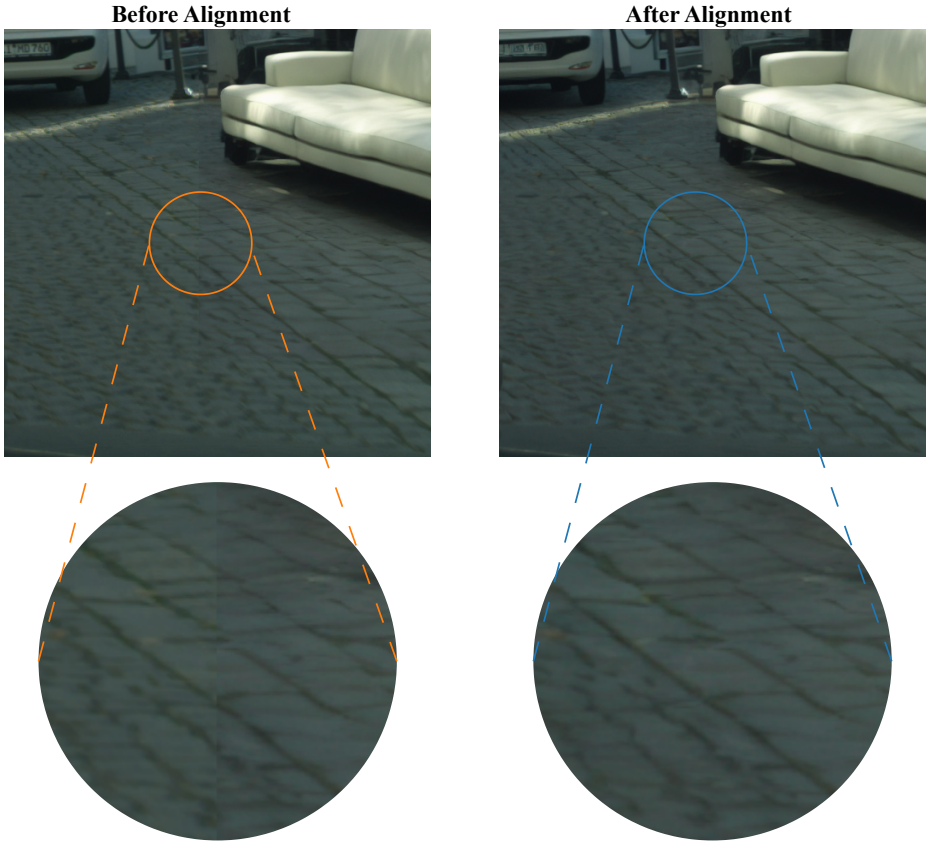
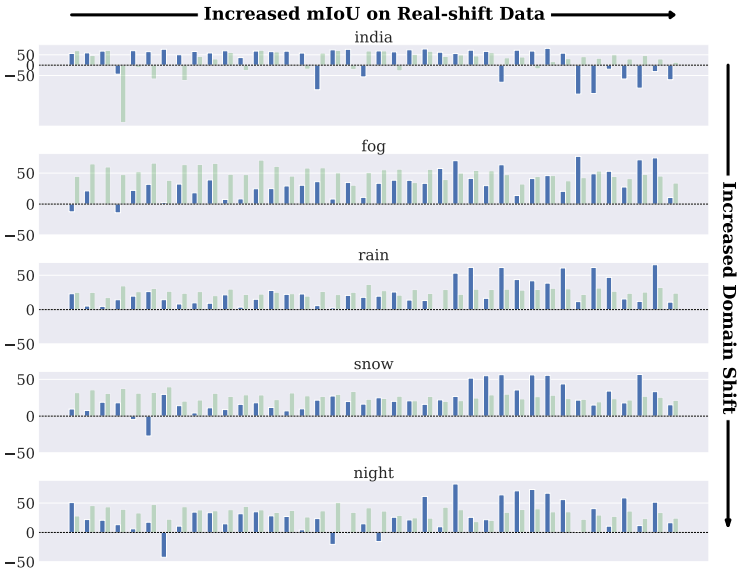
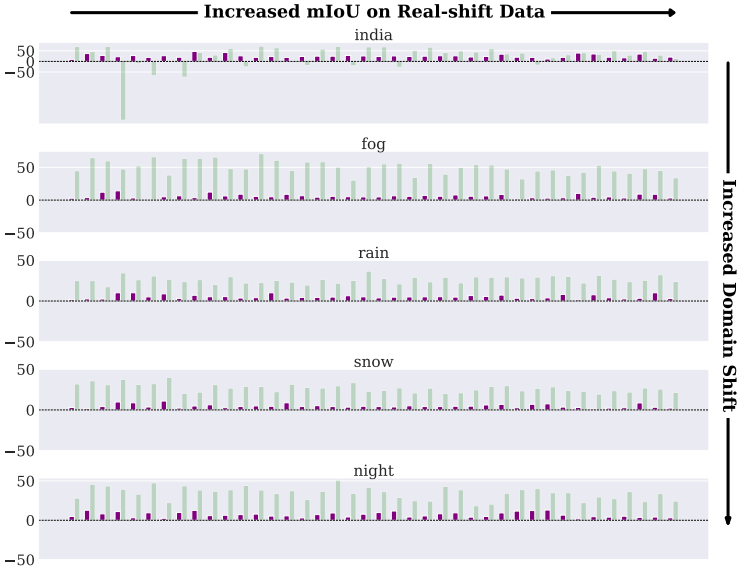


Fig. 13: Refinement. This example highlights the importance of the refinement step. The left image shows the state before refinement, whereas the right image displays the refined version. Upon zooming into the edge of the inpainting box, a clear distinction between the left and right is evident in the first image. Such difference is eliminated in the second image. Must be viewed in color.



(a) Per class Temperature Scaling (■) *vs.* One Temperature Scaling (■). The Figure has the same structure as of Fig. 6 and the bars show relative ECE improvements. We compare the two strategies for performing calibration using synthetic data; both enhance calibration in shifted domains.



(b) Cityscapes (■) *vs.* our synthetic data with one TS (■). The Figure has the same structure as of Fig. 6 and the bars show relative ECE improvements. Our synthetic data is superior to Cityscapes in improving calibration in shifted domains.

Fig. 14: Additional Calibration Results.

Fig. 15: Qualitative results. Examples of rare conditions generated for testing and predictions from different models. Results of the strong model like SegFormer-B5 is visibly better than the Semantic-FPN and MobileNetV3.

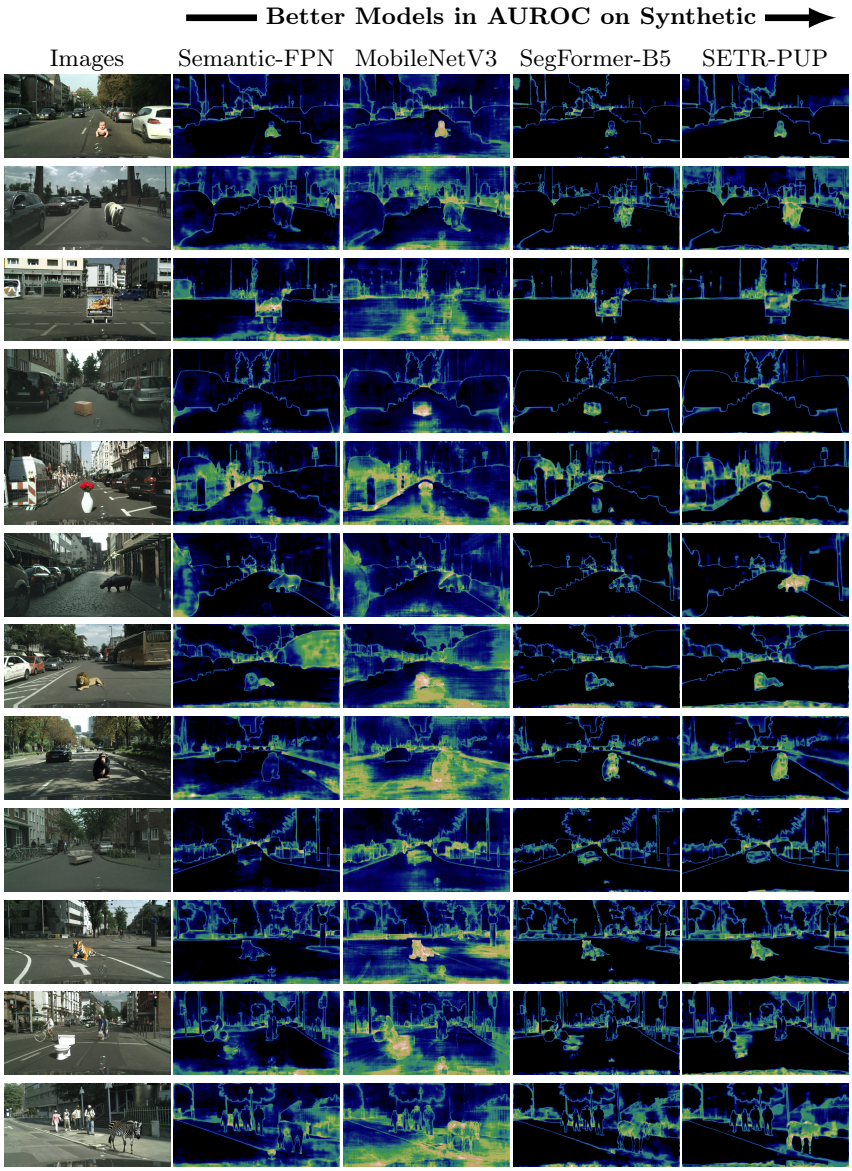


Fig. 16: Qualitative results. Confidence maps are visualized for the four exemplified models on synthetic inpainted data. Hotter colors correspond to higher OOD likelihood. Ideally, results should exhibit hot colors in OOD areas and cold colors everywhere else.