# Supplementary Material for NoVA: Learning to See in Novel Viewpoints and Domains

Benjamin Coors[1,3] Alexandru Paul Condurache[2,3] Andreas Geiger[1]
[1]Autonomous Vision Group, MPI for Intelligent Systems and University of Tübingen
[2]Institute for Signal Processing, University of Lübeck [3]Robert Bosch GmbH

## Abstract

*In this supplementary document, we present more details on the datasets which were used in our experiments. In addition, we provide further qualitative image translation and semantic segmentation results for NoVA as well as the baselines.*

## 1. Datasets

In this section, we provide details on the semantic segmentation classes and camera viewpoint transformations which were used in the datasets of our Sim2Sim and Sim2Real experiments.

### 1.1. Semantic Segmentation Classes

In total, we use 9 semantic segmentation classes in our experiments. The mapping from dataset IDs and classes to training labels for CARLA [3] and CityScapes [1] is defined in Table 1, where -1 denotes the classes which are ignored during training. For CARLA, the pedestrian class is removed as there are no pedestrians in the generated dataset. For CityScapes, classes which are semantically equivalent to CARLA classes are mapped to the respective CARLA training labels. For example, as CARLA features both cars and trucks in its vehicles class, the corresponding CityScapes classes are mapped to the same training label as CARLA's vehicles class.

### 1.2. Viewpoint Transformation

NoVA utilizes the transformation $V_{S\to T}$ between the source and target domain viewpoints in order to warp the source domain images and labels to the target domain viewpoint. The viewpoint transformation $V_{S\to T}$ is defined wrt. a coordinate system in which the $x$-axis points in the driving direction, the $y$-axis points left and the $z$-axis points up.

#### 1.2.1 Sim2Sim

For Sim2Sim, the transformation between the source car viewpoint and the target truck viewpoint in CARLA is given by $V_{Sim2Sim} = (K_{CARLA}, K_{CARLA}, R_{Sim2Sim}, t_{Sim2Sim})$ where $K_{CARLA}$ for both cameras is:

$$K_{CARLA} = \begin{bmatrix} 1024 & 0 & 1024 \\ 0 & 1024 & 512 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

The rotation and translation between the two viewpoints is derived from the extrinsic parameters of the two cameras. While the car camera points straight forward (no rotation along any axis) and is translated by $t_{car} = \begin{bmatrix} 0.30 & -0.11 & 1.30 \end{bmatrix}^T$, the truck camera is rotated around the pitch axis by $\theta_y = -22.5°$ and translated by $2\,\mathrm{m}$ along the $z$-axis wrt. the car camera $t_{truck} = \begin{bmatrix} 0.30 & -0.11 & 3.30 \end{bmatrix}^T$. Thus, the extrinsic transformation between the two viewpoints for the Sim2Sim task is given by $R_{Sim2Sim} = R_y(22.5°)$ where $R_y$ is the rotation matrix around the $y$-axis and $t_{Sim2Sim} = \begin{bmatrix} 0 & 0 & 2 \end{bmatrix}^T$.

(a) CARLA Mapping

| ID | Class | Label |
|---|---|---|
| 0 | *None* | -1 |
| 1 | Buildings | 0 |
| 2 | Fences | 1 |
| 3 | *Other* | -1 |
| 4 | *Pedestrians* | -1 |
| 5 | Poles | 2 |
| 6 | RoadLines | 3 |
| 7 | Roads | 3 |
| 8 | Sidewalks | 4 |
| 9 | Vegetation | 5 |
| 10 | Vehicles | 6 |
| 11 | Walls | 7 |
| 12 | TrafficSigns | 8 |

(b) CityScapes Mapping

| ID | Class | Label |
|---|---|---|
| -1 | *License Plate* | -1 |
| 0 | *Unlabeled* | -1 |
| 1 | *Ego Vehicle* | -1 |
| 2 | *Rectification Border* | -1 |
| 3 | *Out of RoI* | -1 |
| 4 | *Static* | -1 |
| 5 | *Dynamic* | -1 |
| 6 | *Ground* | -1 |
| 7 | Road | 3 |
| 8 | Sidewalk | 4 |
| 9 | *Parking* | -1 |
| 10 | *Rail Track* | -1 |
| 11 | Building | 0 |
| 12 | Wall | 7 |
| 13 | Fence | 1 |
| 14 | *Guard Rail* | -1 |
| 15 | *Bridge* | -1 |
| 16 | *Tunnel* | -1 |
| 17 | Pole | 2 |
| 18 | *Polegroup* | -1 |
| 19 | Traffic Light | 8 |
| 20 | Traffic Sign | 8 |
| 21 | Vegetation | 5 |
| 22 | *Terrain* | -1 |
| 23 | *Sky* | -1 |
| 24 | *Person* | -1 |
| 25 | *Rider* | -1 |
| 26 | Car | 6 |
| 27 | Truck | 6 |
| 28 | *Bus* | -1 |
| 29 | *Caravan* | -1 |
| 30 | *Trailer* | -1 |
| 31 | *Train* | -1 |
| 32 | *Motorcycle* | -1 |
| 33 | *Bicycle* | -1 |

Table 1: Mapping of dataset IDs and classes to training labels in CARLA and CityScapes.

#### 1.2.2 Sim2Real

For Sim2Real, the transformation between the truck viewpoint in CARLA and the car viewpoint in CityScapes is given by $V_{Sim2Real} = (K_{CARLA}, K_{CityScapes}, R_{Sim2Real}, t_{Sim2Real})$ where $K_{CARLA}$ is defined as in Eq. (1) and $K_{CityScapes}$ is:

$$K_{CityScapes} = \begin{bmatrix} 2262.52 & 0 & 1096.98 \\ 0 & 2265.30 & 513.14 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

The extrinsic parameters of the truck camera in CARLA remain unchanged. In CityScapes, the camera is rotated around the pitch axis by $\theta_y = 2.18°$ and around the yaw axis by $\theta_z = 1.12°$. It is translated by $t_{CityScapes} = \begin{bmatrix} 1.70 & 0.10 & 1.22 \end{bmatrix}^T$. Thus, the extrinsic transformation between the truck viewpoint in CARLA and the car viewpoint in CityScapes is given by $R_{Sim2Real} = R_z(1.12°)R_y(-20.32°)$, $t_{Sim2Real} = \begin{bmatrix} 1.40 & 0.21 & -2.08 \end{bmatrix}^T$.

## 2. Additional Qualitative Results

In this section, we present additional qualitative image translation and semantic segmentation results for NoVA and the baselines on the Sim2Sim and Sim2Real tasks.

### 2.1. Sim2Sim

Figures 1 to 4 visualize the output of the individual steps of NoVA's image and label translation pipeline for the different NoVA variants. For variants of NoVA which use depth estimation models instead of ground truth depth maps, we find that a bilinear upsampling of the predicted depth maps from a resolution of $512 \times 256$ to a resolution of $2048 \times 1024$ causes some artifacts in the forward warped images and labels. However, the refinement model is able to remove some of these artifacts and outputs images that are visually close to real target view images.

In contrast, the CyCADA [4] and SPLAT [6] domain adaptation baselines struggle with a semantically consistent translation of the source view images to the target domain viewpoint (see Fig. 5). One reason for this is that, unlike NoVA, CyCADA and SPLAT do not utilize an explicit depth representation of the scene but instead need to learn the perspective transformation to the novel viewpoint end-to-end. Furthermore, they only consider a translation of the source view images to the target domain but not a translation of the source view labels, which leads to a mismatch when using the translated source view images in combination with the original source view labels for training the task segmentation network.

As a result, NoVA's semantic segmentation performance compares favorably to the CyCADA and SPLAT baselines (see Fig. 7 for a VGG16-FCN8s [5], which is utilized in the experiments of the main paper, and Fig. 6 for a DRN-26 [7] model). NoVA also performs better than SceneAdapt [2] and the source segmentation model, which is trained with the original source view data, and comes close to matching the output of a target oracle model which is trained with labeled target view data.

### 2.2. Sim2Real

Figures 8 to 11 show qualitative examples of NoVA's image and label translation on the Sim2Real task. We find that as the visual styles of CARLA and CityScapes differ significantly, the refinement network now does not only inpaint the forward warped images but also adapts their overall style to the style of the target domain. Importantly, NoVA's refinement retains the overall semantic structure of the scene so that the refined images are still consistent with the warped source labels.

In comparison to Sim2Sim, CyCADA and SPLAT now display an improved image translation performance (see Fig. 12) as they successfully adapt the source domain images to the target domain style. However, they do not correctly warp semantic objects in the scene (e.g. cars) to the target domain viewpoint, which, in the case of SPLAT, can be explained by its semantic consistency loss that encourages semantic objects to reappear in the same image locations as in the source segmentation map. This suggests that CyCADA and SPLAT are better suited for the regular domain adaptation task, for which they are originally proposed and in which the focus is on an adaptation to a novel style, rather than for the adaptation to a novel viewpoint.

Semantic segmentation results for NoVA and the baselines are visualized in Fig. 14 for a VGG16-FCN8s [5] and in Fig. 13 for a DRN-26 [7] model. Due to their improved image translation performance, CyCADA and SPLAT now perform better than the source segmentation model, while the SceneAdapt baseline does not appear to be well suited for adapting to a novel viewpoint in a novel domain. As before, all NoVA variants compare favorably to the baselines.

## References


[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] D. Di Mauro, A. Furnari, G. Patanè, S. Battiato, and G. M. Farinella. Scene adaptation for semantic segmentation using adversarial learning. In *Proc. of International Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, 2018.

[3] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *Proc. Conf. on Robot Learning (CoRL)*, 2017.

[4] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In *Proc. of the International Conf. on Machine learning (ICML)*, 2018.

[5] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[6] E. Tzeng, K. Burns, K. Saenko, and T. Darrell. SPLAT: semantic pixel-level adaptation transforms for detection. *arXiv.org*, 1812.00929, 2018.

[7] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

(a) Source Images (b) Forward Warped Images (c) Refined Warped Images (d) Forward Warped Labels

Figure 1: **NoVA**$_{mono-self}$ **Translation on Sim2Sim.**

(a) Source Images (b) Forward Warped Images (c) Refined Warped Images (d) Forward Warped Labels

Figure 2: **NoVA**$_{mono-sup}$ **Translation on Sim2Sim.**

(a) Input Images (b) Forward Warped Images (c) Refined Warped Images (d) Forward Warped Labels

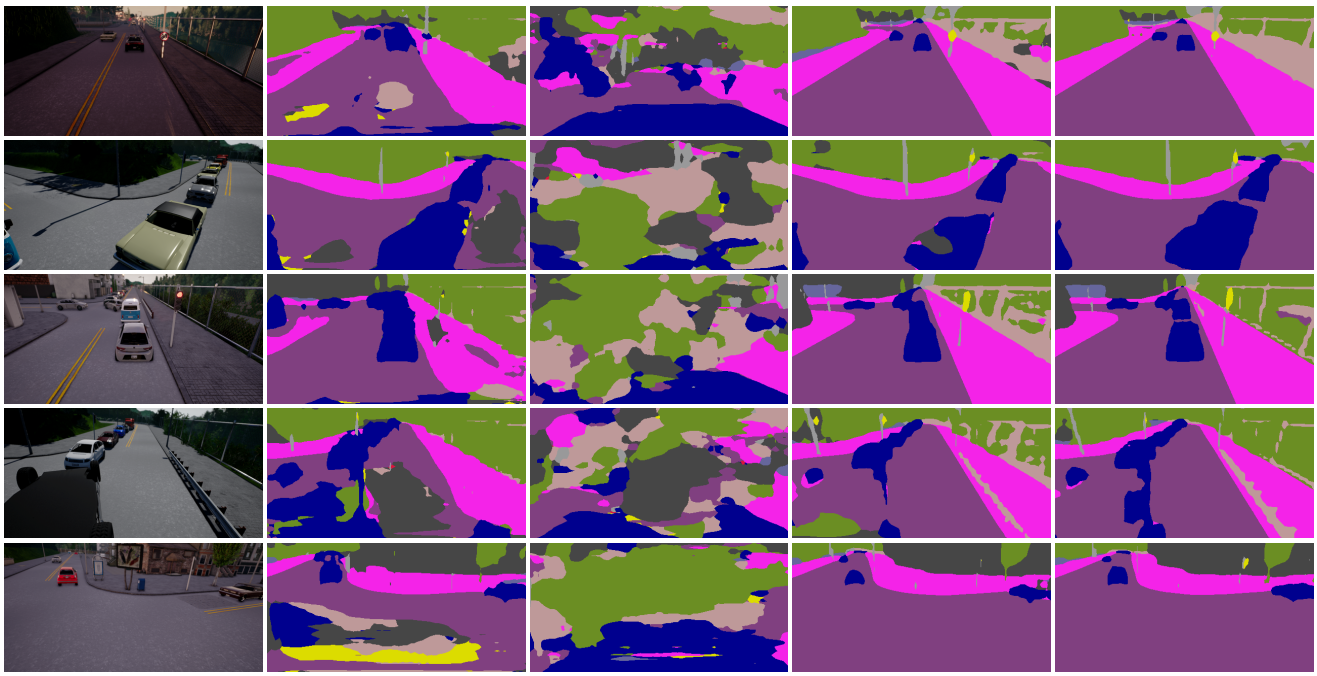
Figure 3: **NoVA**$_{stereo-sup}$ **Translation on Sim2Sim.**

(a) Source Images (b) Forward Warped Images (c) Refined Warped Images (d) Forward Warped Labels

Figure 4: **NoVA**$_{GT}$ **Translation on Sim2Sim.**

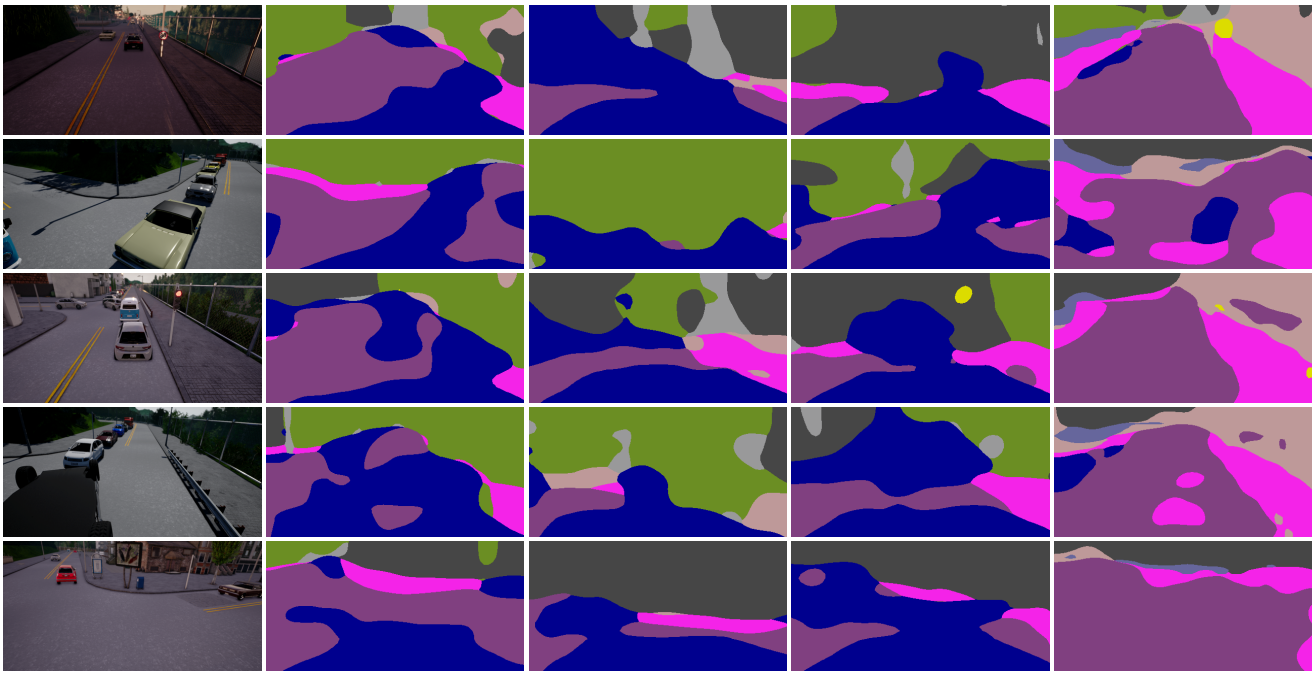
(a) Source Images (b) CyCADA [4] (c) SPLAT [6] (d) Source Labels

Figure 5: **CyCADA [4] and SPLAT [6] Translation on Sim2Sim.**

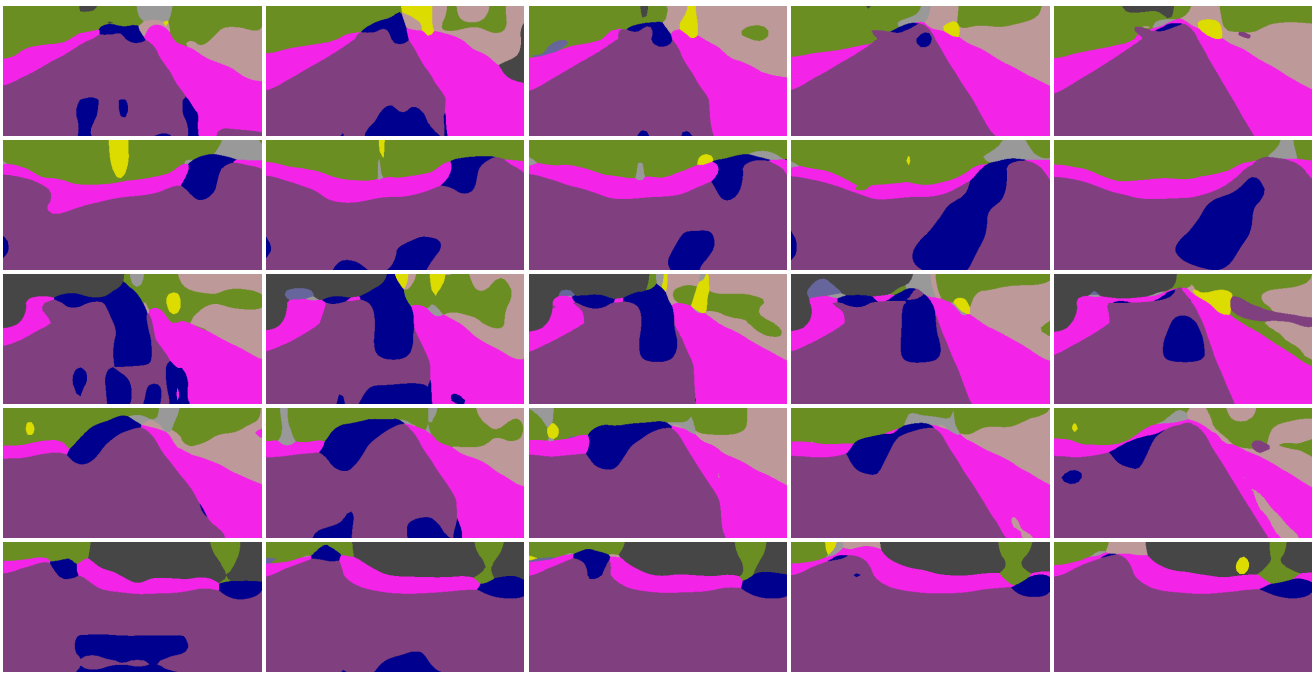


(a) Target Images (b) Source Model (c) CyCADA [4] (d) NoVA$_{GT}$ (e) Target Oracle

Figure 6: **Comparison of Semantic Segmentation Performance on Sim2Sim for a DRN-26 Model [7].**



(a) Target Images (b) Source Model (c) CyCADA [4] (d) SPLAT [6] (e) SceneAdapt [2]



(f) NoVA$_{mono-self}$ (g) NoVA$_{mono-sup}$ (h) NoVA$_{stereo-sup}$ (i) NoVA$_{GT}$ (j) Target Oracle

Figure 7: **Comparison of Semantic Segmentation Performance on Sim2Sim for a VGG16-FCN8s Model [5].**

(a) Target Images (b) Forward Warped Images (c) Refined Warped Images (d) Forward Warped Labels

Figure 8: **NoVA**$_{mono-self}$ **Translation on Sim2Real.**

(a) Source Images (b) Forward Warped Images (c) Refined Warped Images (d) Forward Warped Labels

Figure 9: **NoVA**$_{mono-sup}$ **Translation on Sim2Real.**

(a) Source Images (b) Forward Warped Images (c) Refined Warped Images (d) Forward Warped Labels

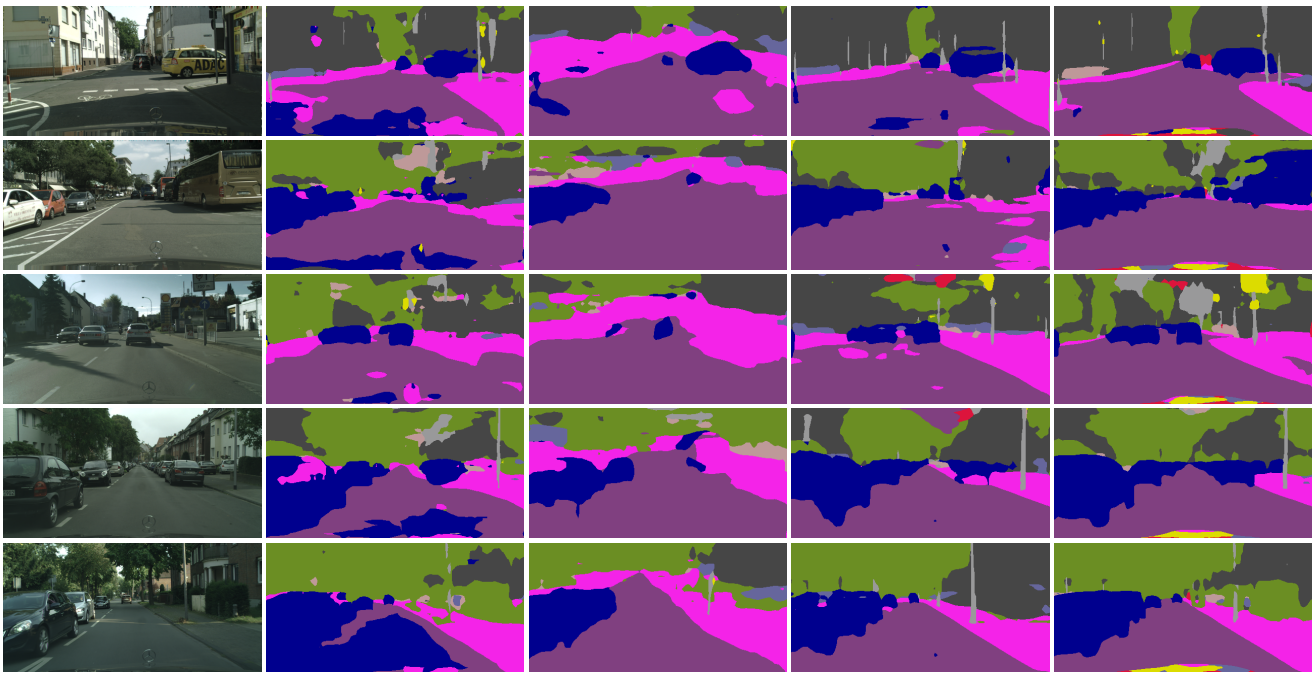
Figure 10: **NoVA**$_{stereo-sup}$ **Translation on Sim2Real.**

(a) Source Images (b) Forward Warped Images (c) Refined Warped Images (d) Forward Warped Labels

Figure 11: **NoVA**$_{GT}$ **Translation on Sim2Real.**

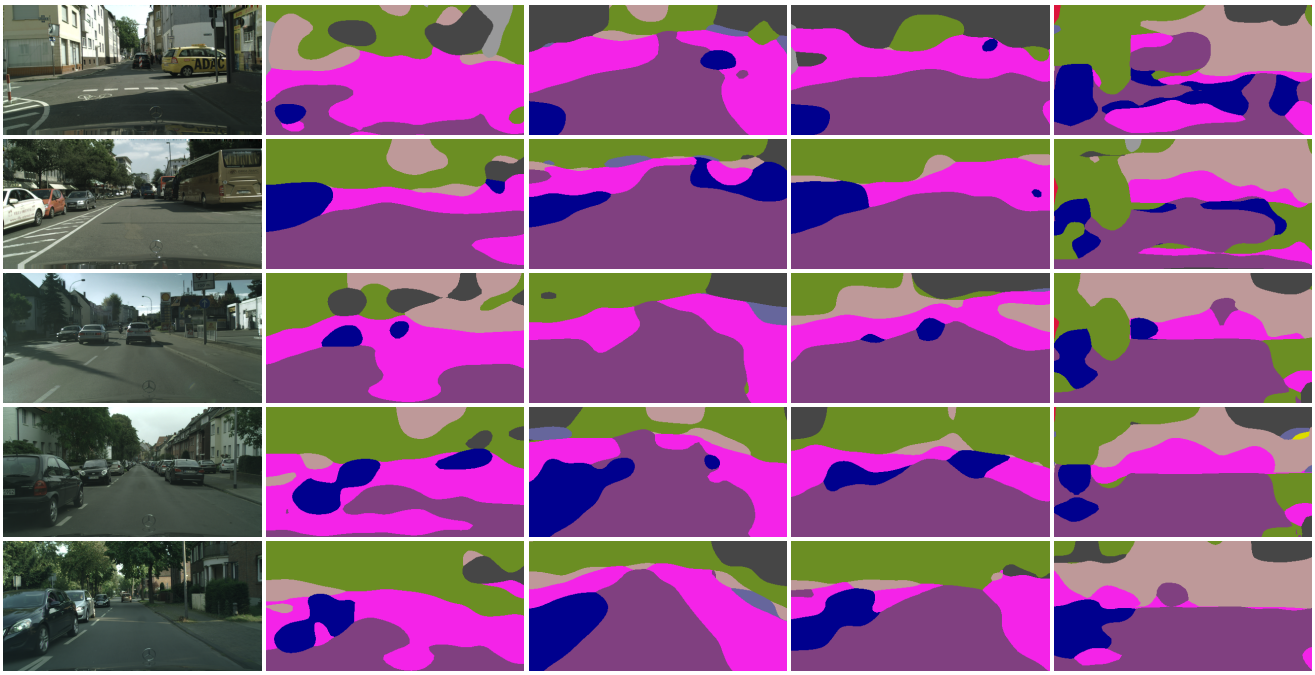
(a) Source Images (b) CyCADA [4] (c) SPLAT [6] (d) Source Labels

Figure 12: **CyCADA [4] and SPLAT [6] Translation on Sim2Real.**

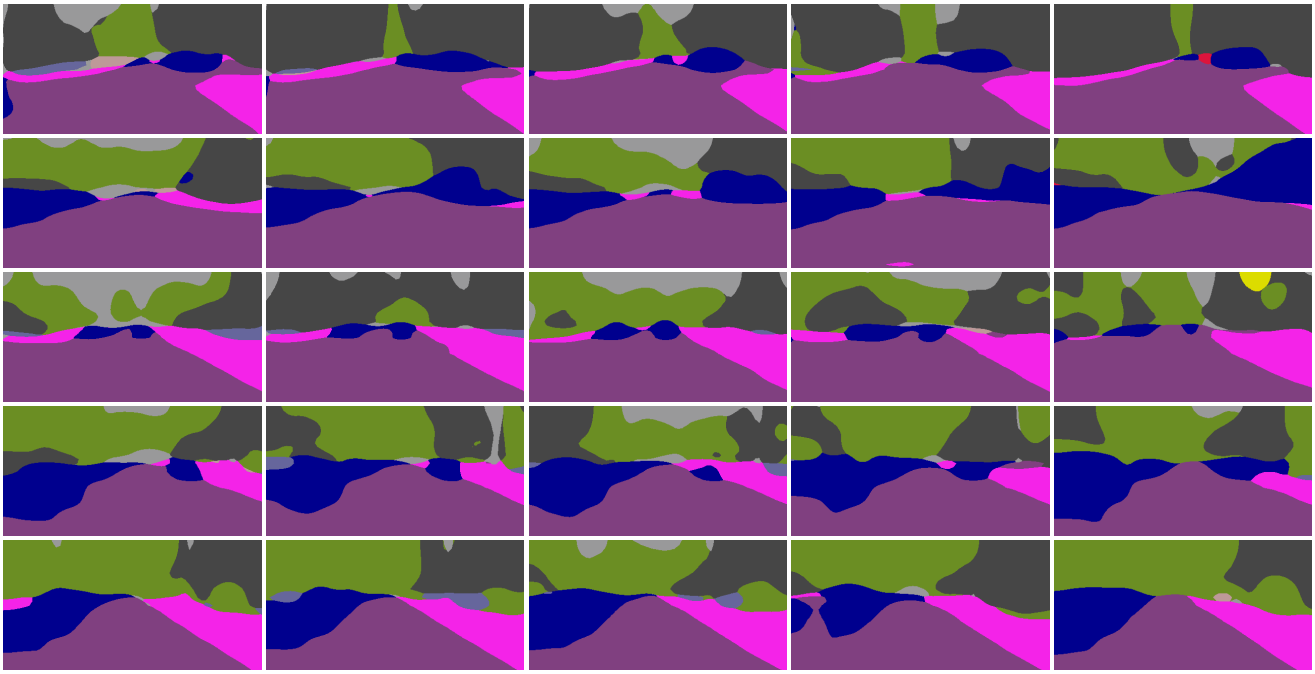


(a) Target Images (b) Source Model (c) CyCADA [4] (d) NoVA$_{GT}$ (e) Target Oracle

Figure 13: **Comparison of Semantic Segmentation Performance on Sim2Real for a DRN-26 Model [7].**



(a) Target Images (b) Source Model (c) CyCADA [4] (d) SPLAT [6] (e) SceneAdapt [2]



(f) NoVA$_{mono-self}$ (g) NoVA$_{mono-sup}$ (h) NoVA$_{stereo-sup}$ (i) NoVA$_{GT}$ (j) Target Oracle

Figure 14: **Comparison of Semantic Segmentation Performance on Sim2Real for a VGG16-FCN8s Model [5].**