Your Image is My Video: Reshaping the Receptive Field via Image-To-Video Differentiable AutoAugmentation and Fusion

Supplementary Material

A. Overview

This supplementary material is organized as follows. We first introduce the proof for Eq. (5) and Eq. (6), which motivate our choice in the final transformations selection, based on a perturbation approach (Sec. B). We then present the implementation details for our method, giving an overview of the hyper-parameters utilized (Sec. C), and provide the details for the best transformations found by DAS for the Image-to-Video setup (Sec. D). We conduct an additional ablation study to further prove the effectiveness of our approach by testing the results under a re-shuffle operation, which motivates the need for the GSF as temporal shift mechanism (Sec. E). For completeness, we report the result of DAS in the "standard" setup of Image-to-Image, where auto-augmentation methods are usually employed and compare with two SOTA approaches, *i.e.* AutoAugment (AA) and RandAugment (RA), conducting ablations studies to highlight the difference with respect to DAS (Sec. F). We conclude with more qualitative results for the semantic segmentation datasets (Sec. G) and with an additional ablation on Cifar100 testing performance robustness with reduced data (Sec. H).

B. Proof of Equations 5, 6

Let $\theta_I = Softmax(\tau_I)$ and $\theta_T = Softmax(\tau_T)$. Then the mixed operation in Eq. 4 can be re-written as $\overline{m}(x) = \theta_T x_T + \theta_I x_I$. The objective can be formally formulated as:

$$\min_{\theta_I, \theta_T} = \text{var}(\overline{m}(x) - m^*) \qquad s.t. \qquad \theta_I + \theta_T = 1 \quad (10)$$

This constraint optimization problem can be solved with Lagrangian multiplies:

$$\begin{aligned}
L(\theta_I, \theta_T, \lambda) &= \text{var}(\overline{m}(x) - m^*) - \lambda(\theta_I + \theta_T - 1) && (11)\\
&= \text{var}(\theta_T T(x) + \theta_I x - m^*) && (12)\\
&\quad - \lambda(\theta_I + \theta_T - 1)\\
&= \text{var}(\theta_T T(x) + \theta_I x - (\theta_I + \theta_T) m^*) && (13)\\
&\quad - \lambda(\theta_I + \theta_T - 1)\\
&= \text{var}[\theta_T (T(x) - m^*) + \theta_I (x - m^*)] && (14)\\
&\quad - \lambda(\theta_I + \theta_T - 1)\\
&= \text{var}(\theta_T (T(x) - m^*) = \text{var}(x - m^*) && (15)\\
&\quad + 2\text{cov}[\theta_T (T(x) - m^*), \theta_I (x - m^*)]\\
&\quad - \lambda(\theta_I + \theta_T - 1)\\
&= \theta_T^2 \text{var}(T(x) - m^*) + \theta_I^2 \text{var}(x - m^*) && (16)\\
&\quad + 2\theta_T \theta_I \text{cov}[T(x) - m^*, x - m^*]\\
&\quad - \lambda(\theta_I + \theta_T - 1)
\end{aligned}$$

Setting partial derivatives to 0

$$\begin{aligned}
\frac{\partial L}{\partial \lambda} &= \theta_I + \theta_T - 1 = 0 && (17)\\
\frac{\partial L}{\partial \theta_T} &= 2\theta_T \text{var}(T(x) - m^*) && (18)\\
&\quad + 2\theta_I \text{cov}[T(x) - m^*, x - m^*] - \lambda = 0\\
\frac{\partial L}{\partial \theta_I} &= 2\theta_I \text{var}(x - m^*) && (19)\\
&\quad + 2\theta_T \text{cov}[T(x) - m^*, x - m^*] - \lambda = 0
\end{aligned}$$

we obtain equations whose solution are

$$\begin{aligned}
&\theta_T \text{var}(T(x) - m^*) + \theta_I \text{cov}[T(x) - m^*, x - m^*] && (20)\\
&= \theta_I \text{var}(x - m^*) + \theta_T \text{cov}[T(x) - m^*, x - m^*]
\end{aligned}$$

Substituting θ_I with $(1 - \theta_T)$ we get:

$$\theta_T^* = \frac{\text{var}(x - m^*) - \text{cov}[T(x) - m^*, x - m^*]}{z} \quad (21)$$

where $z = \text{var}(T(x) - m^*) + \text{var}(x - m^*) - 2\text{cov}[T(x) - m^*, x - m^*]$. Similarly we obtain

$$\theta_I^* = \frac{\text{var}(T(x) - m^*) - \text{cov}[(T(x) - m^*, x - m^*]}{z} \quad (22)$$

Given that $\theta_i = \frac{e^{\tau_i}}{e^{\tau_i} + e^{\tau_j}}$, with $i = I, T$, we obtain:

$$\tau_T^* = \log[\text{var}(x - m^*) - \text{cov}(T(x) - m^*, x - m^*)] + C \quad (23)$$

$$\tau_I^* = \log[\text{var}(T(x) - m^*) - \text{cov}(T(x) - m^*, x - m^*)] + C \quad (24)$$

where the only difference between τ_I and τ_T is the first term inside the logarithm. Therefore, if we choose the operation associated to the largest τ, assuming it is related to the strength of the transformation, we will always end up choosing identity operations. This proof applies also for search spaces with more than two operations, as the transformation T previously defined as a translation can be seen as the composition of multiple transformations.

C. Implementation details

Our hyper-parameters are summarized in Tab 9. We kept the same hyper-parameters during the search phase and the training from scratch, with the only difference in the additional optimizer needed for the Architect neural network. Such a network, responsible for the topology optimization, was trained with Adam optimizer, with $3e-4$ as learning rate and $1e-3$ as weight decay-rate. For all image datasets we applied standard augmentation techniques, such as random horizontal flip, random crop,and cutout, on the inputs to the DAS cell. Every image, after being augmented, undergoes the temporal expansion, achieved through an image replication. Transformations are then applied inside the DAS cell to each frame so that smoothness and continuity are kept during the video generation. The image replication module acts as a "stem" module to allow multiple cells with multiple nodes. Inference cost is not affected, as DAS is involved only during training.

D. Example of cells found by DAS

We provide the graph visualization for the cells found by DAS for Cifar100 (Fig. 7a), ImageNet (Fig. 7b), Pascal-VOC (Fig. 7c) and Cityscapes (Fig. 7d). We do not report the results for Cifar10 and Tiny ImageNet as the found cell is the same as Cifar100 and ImageNet, respectively. This justifies the results previously introduced in Tab. 8 for Cifar-10.

E. Additional ablations on DAS for Image-to-Video

Tab. 10 compares the results we previously showed in Tab. 5 in the main paper, with an additional experiment to prove the need for the GSF component. To this aim, the frames of the video input (obtained with the best transformations found by DAS) are randomly shuffled with the goal of loosing the temporal continuity. This experiment aims at showing that both components, DAS and GSF, are needed, but does not imply a limitation of DAS in the search space definition. As the optimization of the DAS cell to find the optimal transformations occurs during the training of the network, even given a huge search space with non continuous transformations, DAS will optimize to find the best transformations that lead to the highest validation accuracy for *that* architecture. As a result, as we show with further experiments in Sec. F the approach stays robust even under noisy transformations. The experiments are run with a PSP-Net with ResNet-50 backbone for Pascal-VOC dataset and with ResNet-18 for Cifar10 dataset. For each dataset, we show the accuracy (first row) the # of parameters (second row) and the number of flops (third row) with an input size 32×32 and 400×400 for Cifar10 and Pascal-VOC, respectively. Finally, Fig. 8 gives an example of our RF (left) and standard 2d CNN (right) for an ImageNet sample.

The little difference in the "re-shuffle" experiment performed for Pascal-VOC and Cifar-10 datasets with respect to the baseline and DAS Aug *S* is probably due to perturbations. The temporal shift mechanism, *i.e.* GSF, is designed to learn to shift features among adjacent frames. However, if those features are not consistent across the time dimension, GSF correctly learns not to route gated features. As a result, the experiment reconducts to processing data augemnted as in DAS Aug *S* with a 2D backbone integrated with a temporal shift mechanism that learns not to shift.

F. Experiments on DAS for Image-to-Image

F.1. Comparison with SOTAs

Tab. 11 compares our Differentiable Augmentation Search with other SOTA auto-augmentation techniques, *i.e.* AA [1] and RA [2] for the task of image-to-image. This means that no temporal expansion is performed, and a comparable search-space usually deployed for finding standard data-augmentation is defined. Similar to AA and RA, we define in our search space the following set of transformations: Shear X/Y, Translate X/Y, Rotate, AutoContrast, Invert, Equalize, Solarize, Posterize, Color, Brightness, Sharpness, Cutout, and Identity that corresponds to applying no transformation. We run experiments on Cifar-10, Cifar-100, SVHN, and ImageNet, for this set of experiments, we did not fix a budget time for the required search time. Following RA setup, for comparison purposes, we employed a Wide-ResNet-28-2 for the first three datasets, and a ResNet-50 model for ImageNET.

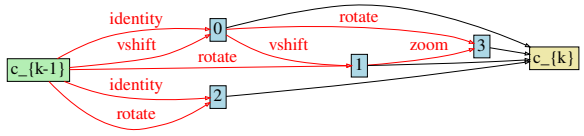
DAS out-performs previous auto-augmentation methods in all datasets but SVHN, where it equals RA performance.

F.2. Advantages of DAS

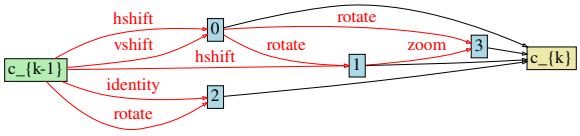
We ablate now on the importance of introducing our differentiable algorithm highlighting the two main drawbacks of the cited competitors. On the one hand, AA is extremely competitive in terms of obtained accuracy, surpassing RA in Cifar-10, Cifar-100, and having equal performance on Imagenet. However, AA is extremely slow, requiring 15000 GPU hours to look for the optimal policy on a *reduced* ImageNet. On the other hand RA is extremely efficient, as it reduces the search space to 10^2 different choices, but we argue it is not robust when introducing not relevant transformations. The authors of [2] indeed show that when introducing color transformations in the Cifar-10 experiments, they experience a degradation of validation accuracy on average. This implies that one needs to carefully design the search space, and cannot include transformations that potentially may harm the performance on the dataset. A justification for such a behaviour is due to their search space definition,

	Cifar10	Cifar100	Tiny	ImageNet	Pascal-VOC	CityScapes
Optimization						
Image size	(32,32)	(32,32)	(64,64)	(224,224)	(380, 380)	(1024,1024)
Optimizer	SGD	SGD	SGD	SGD	SGD	SGD
Batch size	96	96	64	32	32	16
Learning rate scheduler	step decay	step decay	step decay	step decay	poly	poly
Base Learning rate	0.1	0.1	0.1	0.1	0.03	0.03
Weight decay	1e-4	1e-4	5e-4	1e-4	1e-4	1e-4
Epochs	90	90	90	100	80	130
Number of segments	8	8	8	8	8	8

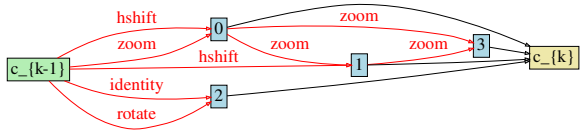
Table 9. Hyperparameters employed for our experiments.



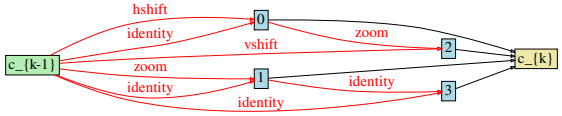
(a) Cifar100.



(b) ImageNet.



(c) Pascal.



(d) Cityscapes.

Figure 7. Best transformations found by DAS.

Figure 8. Receptive Field shape difference between our method (left) and standard 2D CNNs (right).

where a transformation is selected with uniform probability $1/K$. This implies that as the number of K transformations in the search space increases, the probability is reduced, and the time required to find the best transformations increases. On the other hand, under a fixed searching-time budget, this results in a higher variability when the procedure is run multiple times. Evidence supporting this is displayed in Fig. 9, where we fix for Cifar-10 a searching budget time of 24 hours, and exacerbate this behaviour by progressively adding a noise transformation.

	Baseline	DAS Aug (*S*)	Re-shuffle	Ours
Pascal	85.40	85.51	85.44	86.10
	51.32 M 16.55 Gflops		51.43 M 16.67 Gflops	
Cifar10	94.12	94.23	94.15	95.12
	11.18 M 37.12 Mflops		11.20 M 37.12 Mflops	

Table 10. Additional ablation experiments. Baseline was obtained with the 2D backbone with standard augmentation techniques. "DAS Aug (*S*)" stands for the inclusion of additional DAS augmentations in Space *S*, meaning that the data is processed by a 2D backbone. Re-shuffle processes the input in the same way as DAS Aug (*S*) but stacks the transformations in the temporal dimension to create a video, and subsequently re-shuffles the frames of the video. "Ours" processes the input obtained with DAS with temporal continuity preserved. The backbone for the last two experiments is 2D+temporal shift.

G. Segmentation results

We provide more segmentation results on Pascal (Fig. 11) and CityScapes (Fig. 12) datasets. In our the figures we provide the original image (first column), the ground truth (second column), results from DeepLabv3 (column 3) and results with our methods (column 4). We highlight with a square the details where attention should be put to appreciate the difference in the results. We observe in our method,

	search space	Cifar-10 WRN	Cifar-100 WRN	SVHN WRN	ImageNet ResNet
Baseline	0	94.90	75.40	96.70	76.30
AA	10^{32}	95.90	78.50	98.00	77.60
RA	10^{2}	95.80	78.30	**98.30**	77.60
DAS	10^{13}	**96.10**	**78.90**	**98.30**	**77.90**

Table 11. Comparison among different auto-augmentation methods. WRN stands for Wide-ResNet-28-2, while ResNet is the ResNet-50 model. Best results are bolded.

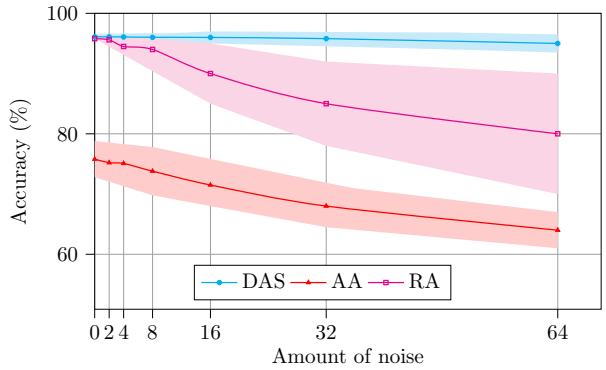


Figure 9. Results on Cifar-10 for each auto-augmentation technique. Experiments are run 5 times, with the shaded area representing the variance.

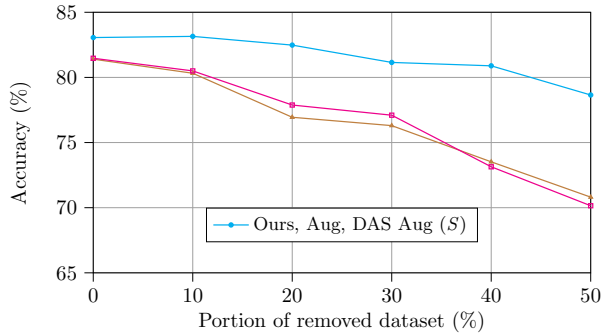


Figure 10. Top-1 test accuracy on Cifar-100 dataset given different portion of removed training dataset.

as general behaviour, a stronger capability in reconstructing details, *e.g*. the back part of the airplane, the details in the motorcycle, plants in Pascal-VOC, street lamps in Cityscapes. We also see that, with respect to the baseline, fewer classes are misclassified, as it can be seen for the portion of the table in the sixth row of Pascal-VOC results, in traffic lights in the third row of Cityscapes results, and in the sidewalk of the sixth row of Cityscapes.

H. Generalizability with reduced training data

To strengthen our point we run a further ablation on Cifar100, shown in Fig. 10. When reducing the size of the dataset, we barely experience a performance degradation (compared to standard augmentations (Aug) and to DAS augmnentations not concatenated in time (DAS Aug *S*)), finding a very useful application in scenarios where few data are available. Compared to finding new data, the cost of representing an image as a video is largely reduced.

References

[1] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 113–123, 2019.*

[2] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

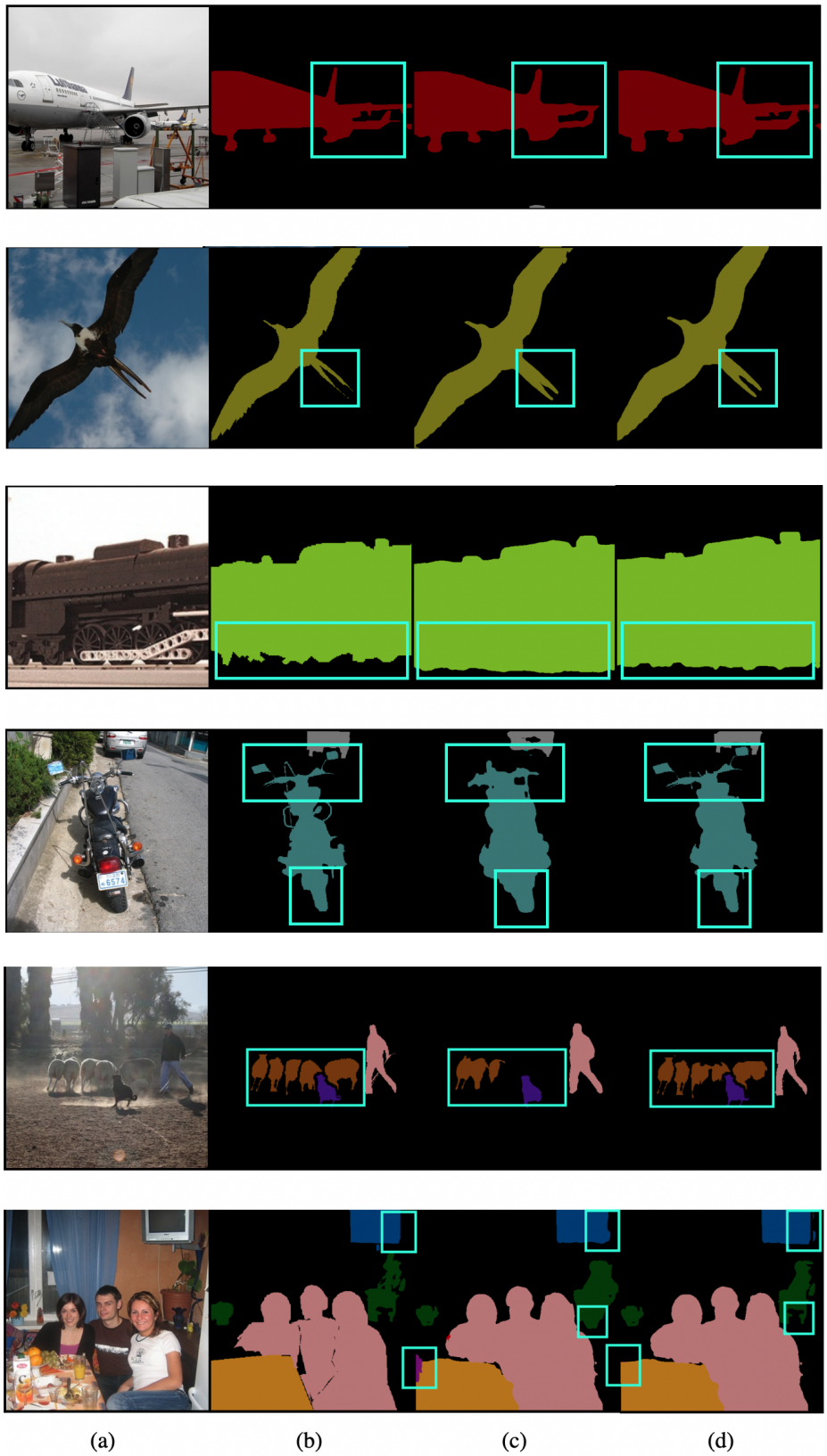


Figure 11. VOC qualitative results. Original image (a), Ground Truth (b), DeepLabv3 (c) and Ours (d) images are displayed.

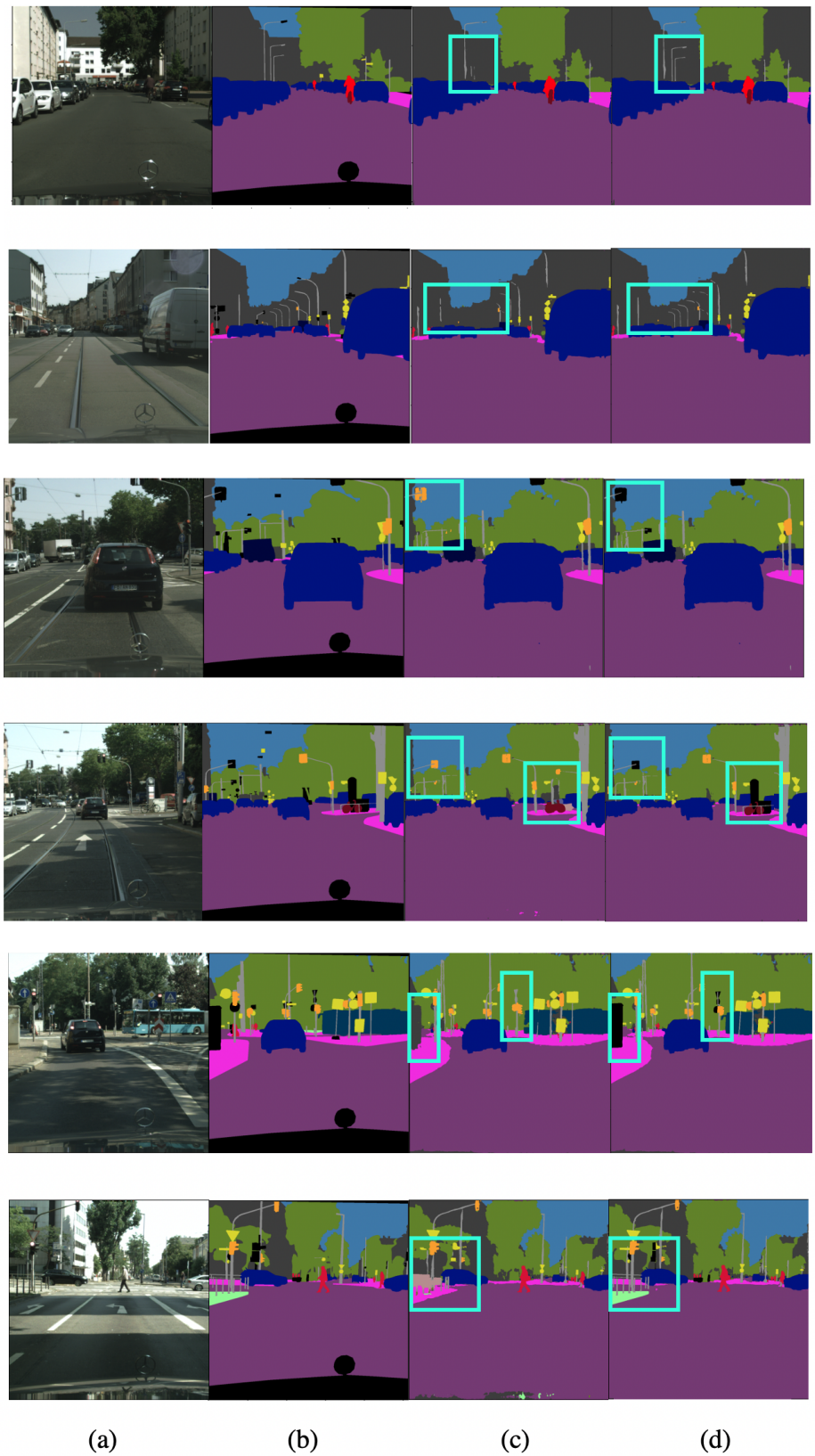


Figure 12. City qualitative results. Original image (a), Ground Truth (b), DeepLabv3 (c) and Ours (d) images are displayed.