Domain Adaptation with Morphologic Segmentation

Anonymous authors
Paper under double-blind review

Abstract

We present a novel domain adaptation framework that uses morphologic segmentation to translate images from arbitrary input domains (real and synthetic) into a uniform output domain. Our framework is based on an established image-to-image translation pipeline that allows us to first transform the input image into a generalized representation that encodes morphology and semantics – the edge-plus-segmentation map (EPS) – which is then transformed into an output domain. Images transformed into the output domain are photo-realistic and free of artifacts that are commonly present across different real (e.g. lens flare, motion blur, etc.) and synthetic (e.g. unrealistic textures, simplified geometry, etc.) data sets. Our goal is to establish a preprocessing step that unifies data from multiple sources into a common representation that facilitates training downstream tasks in computer vision. This way, neural networks for existing tasks can be trained on a larger variety of training data, while they are also less affected by overfitting to specific data sets. We showcase the effectiveness of our approach by qualitatively and quantitatively evaluating our method on four data sets of simulated and real data of urban scenes.

1 Introduction

Deep neural networks have shown unparalleled success on a variety of tasks in computer vision and computer graphics, among many others. To enable this paramount performance a majority of approaches relies on large and diverse data sets that are difficult to establish. Especially, for image-to-image tasks, where an image is transformed into a different representation (e.g. semantic segmentation), labels are extremely difficult to obtain. Furthermore, even large and well established data sets such as KITTI [Geiger et al. (2013)] and Cityscapes [Cordts et al. (2016a)] are biased and do not contain enough variance to allow for training models that are robust and generalize well. Synthetic data generated with modern rendering and simulation algorithms may provide a solution to this problem. However, the disparity between real and synthetic data often limits using generated data to train models so as to operate on real data. A number of approaches aim to explicitly overcome this limitation by modeling the sim-to-real transfer based on Generative Adversarial Networks (GANs) that transform synthetically generated images into images with more photo-realistic visual traits, such as shadows, highlights, or higher frequency textures [Wang et al. (2018); Bousmalis et al. (2018); James et al. (2019); Zhu et al. (2017); Yi et al. (2017)]. However, while some of the existing approaches show impressive results, it remains challenging to faithfully reconstruct the full spectrum of details in real images.

In this paper, we introduce a novel framework that translates images from various data sources to a unified representation. Synthetically generated images are enhanced, while the details of real photographs are reduced. Our goal is to establish a preprocessing step that translates data from different sources and of different modalities (real, synthetic) into a unified representation that can then be used to train networks for computer vision downstream tasks, such as classification, object detection, or semantic segmentation. This has the advantage that the downstream network can be trained on less complex data and – in turn – usually with a less complex architecture. Our image-to-image translation approach combines a state-of-the-art semantic segmentation algorithm [Soria et al. (2019a)] with an edge detection algorithm [Chen et al. (2018a)] to generate edge-plus-segmentation (EPS) maps from arbitrary input images. We refer to this intermediate representation as *morphologic segmentation*. Instead of directly working on RGB images our GAN only uses the EPS maps to generate output images. Hence, our method is agnostic to the style and details of the input images.

Figure 1: An input image is transformed into the output image using our framework: we first generate an EPS map as intermediate representation to then generate the output image in the target domain with an image-to-image translation GAN. Object classes of the EPS map are visualized based on the color scheme of Cordts et al. [Cordts et al. (2016a)]. The input image shown here for illustration is taken from the Cityscapes data set.

Our generator network builds upon the state-of-the-art *pix2pixHD* algorithm [Wang et al. (2018)]. Instead of using contour maps, we employ edge maps that also show internal object edges to provide further guidance for the image generation process. Moreover, we use a progressive learning scheme to achieve high-resolution output images and reduced training time.

To show that the automatic generation of training data is feasible and that EPS maps serve as a meaningful intermediate representation, for generating photo-realistic images with plausible amounts of detail, we train a generator on image data collected from YouTube videos. Once trained, the generator can then be used for the translation of images from various sources, including CARLA [Dosovitskiy et al. (2017); Hendrycks et al. (2019)], Cityscapes [Cordts et al. (2016a)], FCAV [Johnson-Roberson et al. (2017)], and KITTI [Geiger et al. (2013)] as well as across different modalities. To validate the impact of EPS maps, we run ablation studies and measure common metrics for semantic segmentation.

In summary, our contributions are (1) we introduce a novel framework for translating images from different sources into a unified representation; (2) we introduce *edge-plus-segmentation* (EPS) maps as a novel representation for training image-to-image translation networks so as to operate on inputs of different sources and modalities; (3) we evaluate our approach on a range of different data sources and perform ablation studies using established metrics.

2 RELATED WORK

Generative Adversarial Networks. GANs [Goodfellow et al. (2014)] provide a powerful means for image generation by training a generator network so that it generates images that are indistinguishable from a target distribution. Furthermore, it has been recognized that GANs can be a applied to a wide range of image synthesis and analysis problems, including image-to-image translation [Isola et al. (2017); Mirza & Osindero (2014); Ronneberger et al. (2015); Long et al. (2015a)] or – more generally – image generation [Arjovsky et al. (2017); Reed et al. (2016); Wang & Gupta (2016); Zhao et al. (2019)], image manipulation [Zhu et al. (2016)], object detection [Li et al. (2017)], and video generation [Mathieu et al. (2015); Vondrick et al. (2016)]. As training on large image resolutions easily becomes infeasible, many approaches aim to improve the output resolution, often based on stacked architectures [Wang et al. (2018); Huang et al. (2017); Zhang et al. (2017); Karras et al. (2018); Ledig et al. (2017)]. Similar to many of the existing approaches, we rely on a GAN-based pipeline to convert images to a common target domain.

Image-to-Image Translation. A number of methods use GANs for image-to-image translation. To this end, Isola et al. [Isola et al. (2017)] present *pix2pix*, the first approach that takes inspiration from established networks [Ronneberger et al. (2015); Long et al. (2015a)] to transfer images between domains. Although pix2pix can be applied to a wide variety of domains, it is limited by the supported output resolution and it requires a large amount of training data. Various approaches address this and related problems. Zhu et al. [Zhu et al. (2017a)] introduce cycle-consistency loss: images are first translated to the target domain and then back to the source domain; loss is measured over the generated and the input image. Cycle consistency allows generating impressive results across a large variety of image domains [Zhu et al. (2017b)]. Liu et al. [Liu et al. (2019)] present a method that can generate images from previously unseen classes from only a few examples. Similarly, *Disco GANs* learn cross-domain relations and can transform instances of one class to similar looking instances of another class without requiring extra labels [Kim et al. (2017)]. Huang et al. [Huang et al. (2018)] transfer an input image into a shared content space and a class specific style space, which allows recreating images of different classes in an unsupervised manner. Chen et al. [Chen & Koltun

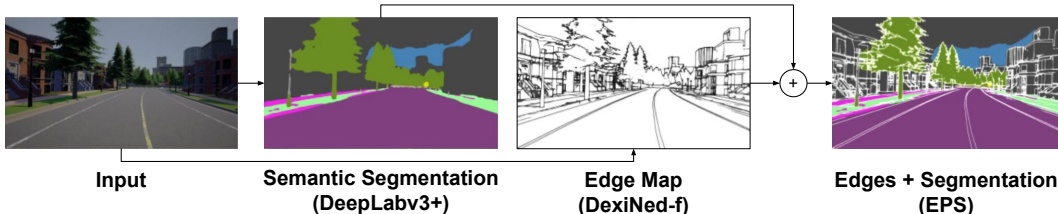| Input | Semantic Segmentation (DeepLabv3+) | Edge Map (DexiNed-f) | Edges + Segmentation (EPS) |

Figure 2: EPS maps are generated by combining a semantic segmentation map and an edge map generated by two state-of-the-art networks.

(2017)] propose cascaded refinement networks, which drastically increases the output resolution and feature quality. On a slightly different trajectory, Gatys et al. [Gatys et al. (2015)] introduce a neural style-transfer network, which is able to transfer among various artistic styles of images; for a comprehensive overview we refer to Jing et al. [Jing et al. (2019)]. While style transfer networks primarily focuses on the visual appearance of images, they do not capture any semantic properties, which limits using these algorithms for many types of image-to-image transfer.

Domain Adaptation. Methods for domain adaptation use neural networks to more generally model the domain shift of source and target domains by either converting one domain into the other or by lifting source and target domain into a shared domain. To this end, Tzeng et al. [Tzeng et al. (2015)] use a CNN-based architecture to model domain invariance for domain transfer, while Long et al. [Long et al. (2015b)] propose Deep Adaptation Network (DAN) to match domain distributions in a Hilbert space. Murez et al. [Murez et al. (2018)] propose a method for the domain adaptation by more explicitly constraining the extracted features of the encoder network, which enables them to model the domain shift of unpaired images. Moreover, it has been recognized that adversarial losses can be leveraged for modeling the domain shift. This includes GANs that model for unequally labeled data of source and target domains [Luo et al. (2017)], for improving the realism of simulated images [Shrivastava et al. (2017)], for unsupervised training setups [Sankaranarayanan et al. (2017)], and more general approaches that combine discriminative modeling and weight sharing with an adversarial loss [Tzeng et al. (2017)]. Similar to the existing approaches our goal is to model the domain shift of different source domains so as to lift samples into a common target domain. However, unlike them we use EPS maps to translate images of a variety of source domains into a common target domain.

3 METHOD

Our method consists of two main steps: first, we introduce morphologic segmentation as the combination of semantic segmentation and edge detection. Specifically, we generate an edge-plus-segmentation (EPS) map as a representation for input images. In a second step, we generate an output image by only synthesizing it from an EPS map. In the following we describe how we generate EPS maps and how we use them to synthesize photo-realistic images with a GAN.

EPS Map Generation. We introduce EPS maps as an abstract intermediate representation of an image that encodes semantic information of objects as well as their morphology defined as edges. We argue that a common semantic segmentation map is insufficient as it does not provide intra-object details. Therefore, we extend semantic segmentation maps with edges, which encode more intricate geometric details. We aim to devise a representation for images that abstracts away defining details such as textures, lighting conditions, or the style of an image, that may be unique to a specific data set. Our goal is to train a generator network so that it can synthesize an image with uniform, yet complex visual traits, by only using the EPS map.

Edge detection as well as semantic segmentation have been active areas of research for many years. We use state-of-the-art approaches for semantic segmentation (DeepLabv3) [Chen et al. (2017)] and edge detection (DexiNed-f) [Soria et al. (2019b)]. Please note that both networks are in principal interchangeable without affecting the rest of the pipeline. DeepLabv3 is the state-of-the-art extension of DeepLabv1 [Chen et al. (2015)] and DeepLabv2 [Chen et al. (2018)]. In the previous versions, the input image is processed by the network using atrous convolution increasing the receptive field while maintaining the spatial dimension of the feature maps followed by bilinear interpolation and fully connected conditional random field (CRF) postprocessing taking context into account. DeepLabv2 employs ResNet [He et al. (2016)] and VGGNet [Simonyan & Zisserman (2015)] while DeepLabv1 only uses VGGNet. DeepLabv2 further employs atrous spatial pyramid

pooling (ASPP) enabling the network to encode multi-scale contextual information. In DeepLabv3, ASPP is further augmented with features on the image-level encoding global context information and further boosting performance. It outperforms its predecessors even without the CRF postprocessing. Chen et al. [Chen et al. (2018b)] then further extended DeepLabv3 to DeepLabv3+ mainly by adding a decoder module refining the segmentation results especially along object boundaries.

DexiNed-f is an edge detector devised by Soria et al. [Soria et al. (2019b)] providing fused edge maps. It is based on holistically-nested edge detection (HED) [Xie & Tu (2015)] enabling holistic image training and prediction while allowing for multi-scale and multi-level feature learning. Its general architecture is inspired from the Xception network [Chollet (2016)].

Figure 2 shows how we compute EPS maps. The input image is passed to DexiNed-f to produce a grayscale image of the edges, while we use DeepLabv3+ to generate the segmentation map. Both maps are combined by pixel-wise addition to create the EPS map. We argue that EPS maps reduce the number of input dimensions, while providing enough meaningful details to reconstruct photo-realistic images.

Image Synthesis. Our image generator builds upon *pix2pixHD* [Wang et al. (2018)] as an extension of pix2pix [Isola et al. (2017)] overcoming the lack of fine details and unrealistic textures. In this regard, the authors introduced a novel adversarial learning objective increasing robustness together with optimized multi-scale architectures of the generator and the discriminator networks. As in pix2pix, loss functions are specifically tailored to the input domain.

The original pix2pixHD implementation supports two different encodings for the segmentation maps: a one-hot encoding for each class, or color-coded RGB maps with different colors representing different classes. Due to the encoding of the EPS maps, the latter one is used. In the original implementation an object instance map can be provided along the segmentation map. Since the edge information from the EPS maps already encode similar information (however, edge maps have not necessarily closed object boundaries and also contain inner edges), we do not make use of this additional input channel. While pix2pixHD is used without modifying its architecture, the type of input information and its encoding is notable different from the original proposed.

4 IMPLEMENTATION AND TRAINING

For generating EPS maps we use pretrained instances of DeepLabv3+ and DexiNed-f. Depending on the supported classes and used training data, different versions of DeepLabv3+ are available. We use the Xception71 variant, which is trained on the Cityscapes [Cordts et al. (2016b)] data set, which outputs segmentation classes commonly found in the Cityscapes images. While DexiNed-f is very versatile and performs well on a large set of input domains, semantic segmentation is more problem tailored. Thus, when executing our framework on different image types, the trained DeepLabv3+ network would need to be exchanged. Since DeepLabv3+, DexiNed-f, and pix2pixHD are implemented in Python, using either TensorFlow or PyTorch, our EPS pipeline is completely written in Python.

Training Data. The manual labeling of data is a time consuming and often tedious process [Johnson-Roberson et al. (2017)]. Therefore, to emphasize the independence from manually labeled training data, we abstain from using a classic data set. Instead, we use a large set of YouTube videos (according to fair use guidelines) showing urban scenes from the driver's perspective. This allows us to cover a wider variety of scenarios compared to what is contained in any of the currently available data sets, that are shown in Figure 3. In total, 98 different videos corresponding to about 45 hours of video material were used. The videos have been selected using appropriate hashtags ensuring that most scenes are showing European roads at daytime summer days, providing good weather conditions. From these videos, around 101 000 frames were randomly extracted and used for training. Many videos contain unsuitable scenes. Therefore, we manually removed video segments prior to randomly selecting frames for training.

Our data set is further augmented using a combination of three simple steps: first, each image is rotated around a randomly selected center by $\pm 7°$. A random cropping of up to half of the image is then applied in each dimension. Finally, we scale the image by $\pm 20\%$, while preserving the aspect ratio. After applying these augmentation steps, the number of training artifacts is greatly reduced.

Progressive Training. We train generator and discriminator by progressively increasing the image resolution in four steps (each doubling the resolution); starting with a resolution of 128×72 up

to a final resolution of 1024×576. Increasing the resolution results in finer details as the training progresses. Accordingly, the learning rate of $\eta = 0.0001$ used in the first step is decreased with increasing resolution to $\eta = 0.00001$ in the second and third steps. However, in the final step, the initial learning rate $\eta = 0.0001$ is applied again in order to avoid overfitting, which results in improved stability [Karras et al. (2017)].

5 EVALUATION

Examples of synthesized images from various input sources are shown in Figure 3. While the input images across the different data sets show a number of heterogeneous visual features, such as varying lighting condition or diverse environments, the translated images only show consistent visual traits. For example, while many of the input images taken from the FCAV data set show a desert environment, its corresponding output images show features of more common environments. Furthermore, it can be observed that synthetic input images are enhanced with more photo-realistic details. In contrast, images of real data sets are showing less artifacts commonly present in photos (e.g. exceedingly strong contrasts, hard shadows, gray cast, etc.). The disparity between real and synthetic data is significantly reduced.

We follow an established evaluation protocol for image-to-image translation [Wang et al. (2018); Isola et al. (2017); Zhu et al. (2017a)]. The quality of our results is evaluated by computing semantic segmentation maps of the output images and by comparing how well the resulting segmentation maps match the corresponding segmentation of the input images. This can easily be quantified by computing the mean intersection-over-union, $\mathsf{IoU} \in [0, 1] \subset \mathbb{R}$, score. Figure 3 lists the IoU score for the different cases presented therein. The scores range from $\mathsf{IoU} = 0.380$ to $\mathsf{IoU} = 0.557$ and can be considered as average values. A representative cross section of our results is shown here. Moreover, the distributions of the IoU scores per class are shown in Figure 4 (left). For this evaluation, 150 images from CARLA, 500 images from Cityscapes, 365 images from FCAV, and 285 images from KITTI data set were analysed. DeepLabv3+ for the semantic segmentation was employed as in our whole image-to-image translation pipeline. This likely causes the effect that significantly higher scores are usually obtained for the Cityscapes inputs since DeepLabv3+ is trained on the Cityscapes data set [Chen et al. (2018b)].

We carry out three ablation studies providing further evaluations of our approach. In particular, we systematically analyze the impact of quality of the input data, the availability of accurate edge details, and the quality of the segmentation. In this regard, we analyzed a reduced data set comprising 16 images of each input source CARLA, Cityscapes, FCAV, and KITTI since each image produced a series of new images as explained in the following. Figure 7 illustrates the effect caused by smoothing the input image. Specifically, a Kuwahara filter [Kuwahara et al. (1976)] is applied using different radii r to control the smoothing intensity. We quantify the effect of smoothing by measuring IoU scores between segmentations of input and output images for different smoothing intensities. This is illustrated for the different classes in Figure 4 (right). It can be observed that in the case of real data (Cityscapes and KITTI), the IoU scores are initially above the ones of synthetic data (CARLA and FCAV), while the scores are getting more similar with increasing smoothing radius. As the general quality of images decreases, the domain disparity between real and synthetic data is reduced. This study also shows significant robustness of our approach with respect to image quality as illustrated in Figure 7 (middle row) demonstrating the effect of a relatively strong Kuwahara smoothing with $r = 4$. The smoothed input image is completely blurred, while the corresponding output image contains fine traits such as structural and texture details of the vegetation located between the street and the sidewalk on the right. Thus, our approach can potentially enable the use of additional data sources for training downstream tasks.

Moreover, we investigate the importance of the accuracy of edges and segmentation with respect to the output quality. Edges are systematically carved as illustrated in Figure 5. Since we observed that most edges already show relatively small gradients, we threshold pixels at 64% intensity in each step followed by a smoothing procedure with a 1px kernel. This can be repeated to allow for different edge carving levels. The IoU scores between segmentations of input and output images for different levels are measured further underlying our previous observations. Our image-to-image transformation is robust towards a decrease of image quality while domain disparity between real and synthetic data is reduced. We carry out a similar ablation study by systematically reducing the quality of the semantic segmentation instead of edge quality. This is illustrated in Figure 6 and consistent with the previous observations. In this regard, segmentation maps are systematically warped to reduce their

Figure 3: Examples of synthesized images from various input sources CARLA, Cityscapes, FCAV, and KITTI comprising real and rendered data. The EPS maps are extracted from the input images and then used as the only input to synthesize the output images.
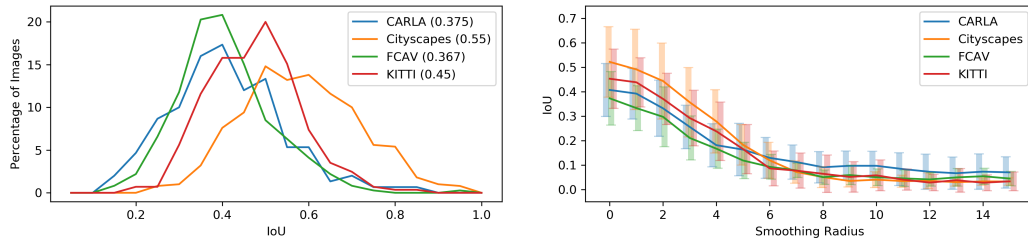

Figure 4: Illustration of the IoU scores per class considering CARLA, Cityscapes, FCAV, and KITTI (left), and the dependence of the IoU score on different smoothing radii (right).
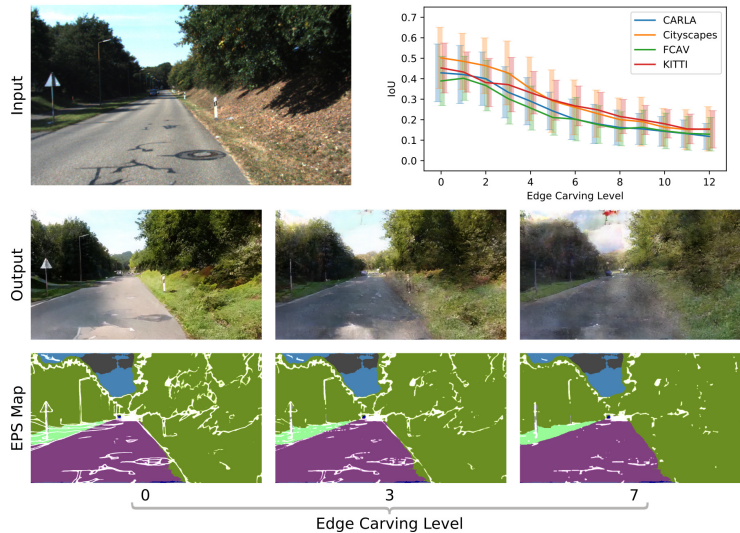

Figure 5: Illustration of the effect caused by different edge carving levels and the corresponding IoU score. The input image shown here for illustration is taken from the KITTI data set.
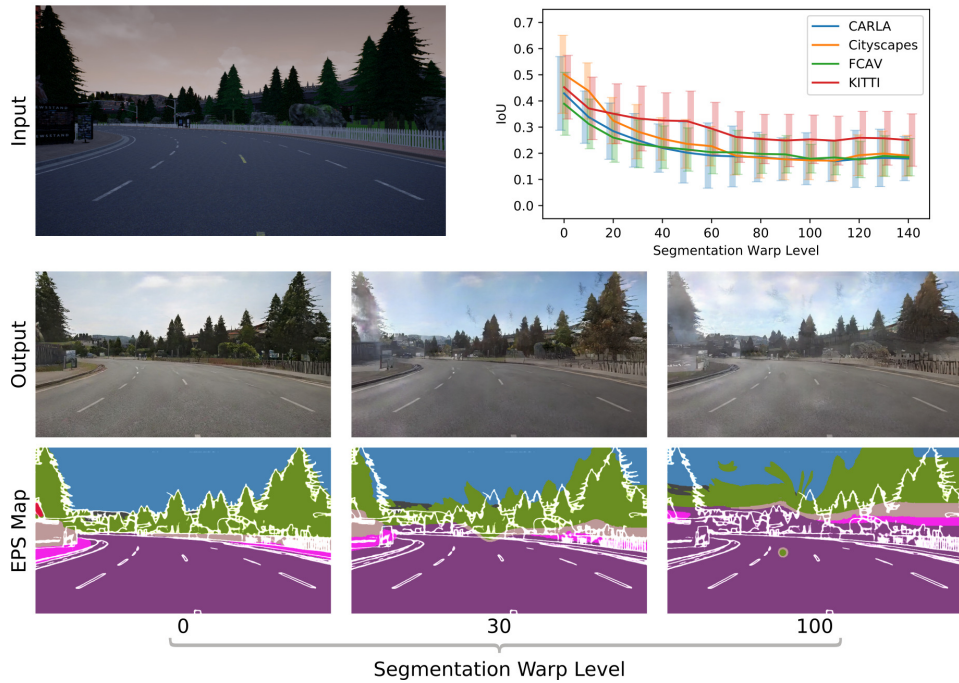
Figure 6: Illustration of the effect caused by warping semantic segmentations using different intensities, and the corresponding IoU score. The input image shown here for illustration is taken from the CARLA data set.

Figure 7: Illustration of the effect caused by smoothing the input image. A Kuwahara filter is applied with different radii. The input image is taken from the FCAV data set.

7

quality by manipulating control points. Fine details are first warped by shifting control points by distances sampled from a normal distribution with a standard deviation corresponding to the warp level. The result is then warped by a single point along a distance proportional to the warp level concluding a distorting process affecting multiple scales.[1] The analysis shows that our approach is more robust with respect to a loss of segmentation quality compared to edge quality. In other words, even if the segmentation quality is rather low, given appropriate edges usually available, proper results can still be achieved.

Since semantic segmentation is a less stable task than edge detection (for instance in Figure 7 the segmentation degrades faster) and it is unsurprising that our generator tends to prioritize edge information. In the CARLA example in Figure 3, the whole sky is classified as *building* in the EPS map, but the generator learned to ignore such obvious mistakes, as they also occur for the training data. As another example, the street sign in Figure 5 is falsely segmented as *tree*, but correctly recreated as long as the edge information is intact. However, this behaviour makes it harder to guide the image generation process by changing the segmentation map. In Figure 6 it can be observed that when the tree line is moved in the segmentation map, trees are still generated according to the edge map (although artifacts appear in the sky). We expect our method to profit most from improved semantic segmentation networks in the future.

To validate how well our method is able to overcome the domain disparity present across different data sets we visualize image embeddings generated with a convolutional autoencoder [Vincent et al. (2008)] (Figure 8). We first train the encoder so as to reconstruct the original input images of each data set (left). Second, we use the same images, convert them to EPS maps, to then generate photorealistic images with our pipeline (right). Encoding the original data generates obvious clusters, which indicates that the autoencoder latches on to the unique visual features of each data set (e.g. desert next to the road for FCAV data, low frequency textures in CARLA, etc.). Images generated with our pipeline are devoid of these artifacts and only show similar visual traits, which results in a more uniform distribution of embeddings. This shows that our method allows us to remove domain shifts that are present in different data sets and for different data modalities.
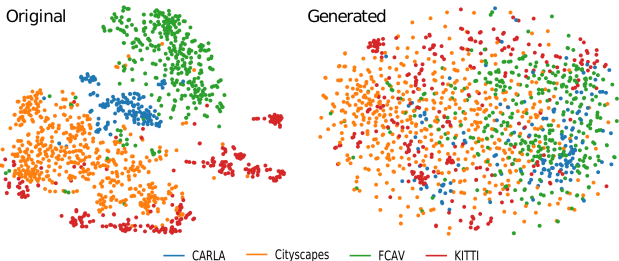

Figure 8: t-SNE plots of 1.3k image embeddings of the original images (left) and the generated ones (right).

6 CONCLUSION

We have presented a novel framework that allows us to translate images of various input sources into a unified target domain. Our approach uses EPS maps as an intermediate representation to generate images with photo-realistic, but simplified visual traits. Our goal is to remove common image artifacts, while we maintain enough plausible details to faithfully represent an image. This way, our pipeline can be used to generate training data that facilitates training downstream computer vision tasks, such as classification, object detection, or semantic segmentation. By generating images from EPS maps our method is able to remove a wide range of image artifacts of real data, including seasonal and daytime shifts, camera artifacts such as lens flares, as well as to compensate for common artifacts of synthetic data, such as simplified geometry or low frequency textures. We have shown that our pipeline is able to generate images with similar visual properties from four different data sources. Furthermore, we have evaluated our approach through a number of ablation studies that show its robustness against common artifacts present across the source domains. However, since the quality of edges is of superior importance compared to segmentations, it is hard to ignore edges impeding the generation of more abstract recreations of the input images. As avenues for future work, it would be interesting to explore the usefulness of our method for an even large number of data sources and for the training of computer vision downstream tasks.

---

[1]The resolution of our segmentation maps is given by 1024×576. A 7×5 grid is used in which the inner points are shifted. The result is further transformed by shifting the center point in a random direction by a distance of twice the warp level.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor Sampedro, Kurt Konolige, Sergey Levine, and Vincent Vanhoucke. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. 2018. URL https://arxiv.org/abs/1709.07857.

L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In Yoshua Bengio and Yann LeCun (eds.), *ICLR*, 2015. URL http://arxiv.org/abs/1412.7062.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. URL http://arxiv.org/abs/1706.05587.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *ECCV*, 2018a. URL https://github.com/rishizek/tensorflow-deeplab-v3-plus.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *The European Conference on Computer Vision (ECCV)*, September 2018b.

Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. *IEEE International Conference on Computer Vision (ICCV)*, 2017.

François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016. URL http://arxiv.org/abs/1610.02357.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016a. URL https://www.cityscapes-dataset.com/.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016b.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16, 2017.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint*, 2015.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems (NIPS)*, 2014.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. A benchmark for anomaly segmentation. *arXiv preprint arXiv:1911.11132*, 2019.

X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie. Stacked generative adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1866–1875, 2017.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *European Conference on Computer Vision (ECCV)*, 2018.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://phillipi.github.io/pix2pix/.

S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12619–12629, 2019.

Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2019. doi: 10.1109/tvcg.2019.2921336.

Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *IEEE International Conference on Robotics and Automation (ICRA)*, 2017. URL https://fcav.engin.umich.edu/projects/driving-in-the-matrix.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. URL http://arxiv.org/abs/1710.10196.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations (ICLR)*, 2018.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *International Conference on Machine Learning (ICML)*, March 2017.

M. Kuwahara, K. Hachimura, S. Eiho, and M. Kinoshita. *Processing of RI-Angiocardiographic Images*, pp. 187–202. Springer US, Boston, MA, 1976. ISBN 978-1-4684-0769-3. doi: 10.1007/978-1-4684-0769-3_13. URL https://doi.org/10.1007/978-1-4684-0769-3_13.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan. Perceptual generative adversarial networks for small object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1951–1959, 2017.

Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. *IEEE International Conference on Computer Vision (ICCV)*, 2019.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015a.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, ICML'15, pp. 97–105. JMLR.org, 2015b.

Zelun Luo, Yuliang Zou, Judy Hoffman, and Li Fei-Fei. Label efficient learning of transferable representations across domains and tasks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 164–176, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *CoRR*, abs/1511.05440, 2015.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint*, 2014.

Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim. Image to image translation for domain adaptation. In *CVPR*, pp. 4500–4509, 2018.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *International Conference on Machine Learning (ICML)*, 2016. URL http://proceedings.mlr.press/v48/reed16.html.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. URL https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/.

Swami Sankaranarayanan, Yogesh Balaji, Carlos D. Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. *CVPR*, pp. 8503–8512, 2017.

A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2242–2251, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun (eds.), *ICLR*, 2015. URL http://arxiv.org/abs/1409.1556.

Xavier Soria, Edgar Riba, and Angel Sappa. Dense extreme inception network - towards a robust cnn model for edge detection. *arXiv preprint*, 2019a.

Xavier Soria, Edgar Riba, and Angel D. Sappa. Dense extreme inception network: Towards a robust cnn model for edge detection, 2019b.

E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4068–4076, 2015.

E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pp. 2962–2971, 2017.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pp. 1096–1103, 2008. ISBN 9781605582054. doi: 10.1145/1390156.1390294. URL https://doi.org/10.1145/1390156.1390294.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 613–621, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL https://github.com/NVIDIA/pix2pixHD.

Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. *European Conference on Computer Vision (ECCV)*, 2016.

Saining Xie and Zhuowen Tu. Holistically-nested edge detection. *CoRR*, abs/1504.06375, 2015. URL `http://arxiv.org/abs/1504.06375`.

Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2868–2876, 2017.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial networks. January 2019. 5th International Conference on Learning Representations, ICLR 2017 ; Conference date: 24-04-2017 Through 26-04-2017.

J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *IEEE International Conference on Computer Vision (ICCV)*, 2017a.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in Neural Information Processing Systems (NIPS)*, 2017b. URL `https://junyanz.github.io/BicycleGAN`.