On the Role of Receptive Field in Unsupervised Sim-to-Real Image Translation

Nikita Jaipuria[1], Shubh Gupta[2]*, Praveen Narayanan[1], Vidya N. Murali[1]
[1]Ford Motor Company, [2]Stanford University
njaipuri@ford.com, shubhgup@stanford.edu, pnaray11@ford.com, vnariyam@ford.com

Abstract

Generative Adversarial Networks (GANs) are now widely used for photo-realistic image synthesis. In applications where a simulated image needs to be translated into a realistic image (sim-to-real), GANs trained on unpaired data from the two domains are susceptible to failure in semantic content retention as the image is translated from one domain to the other. This failure mode is more pronounced in cases where the real data lacks content diversity, resulting in a content *mismatch* between the two domains - a situation often encountered in real-world deployment. In this paper, we investigate the role of the discriminator's receptive field in GANs for unsupervised image-to-image translation with mismatched data, and study its effect on semantic content retention. Experiments with the discriminator architecture of a state-of-the-art coupled Variational Auto-Encoder (VAE) - GAN model on diverse, mismatched datasets show that the discriminator receptive field is directly correlated with semantic content discrepancy of the generated image.

1 Introduction

Advanced Driver Assistance Systems (ADAS) and self-driving cars utilize Deep Neural Networks (DNNs) for a variety of perception tasks. The main bottleneck in training and validating DNNs is the collection and annotation of large quantities of diverse data. As an alternative, gaming-engine based simulations can quickly generate large quantities of diverse synthetic data with accurate ground truth. However, models trained on synthetic data often fail to generalize to the real-world because of lack of realism in synthetic data.

Prior works on GANs [2] and VAEs [7] have shown promising results in photo-realistic image synthesis, in both supervised and unsupervised settings. In the unsupervised setting, the training data consists of *unpaired* images in the simulated and real domains. Here, unpaired refers to the: *(i)* absence of one-to-one correspondence between the simulated and real images used for training; and *(ii)* absence of any form of annotation, such as semantic segmentation masks or edge masks. We are also interested in the unsupervised setting, since it is less restrictive in terms of training data requirements. Additionally, we assume lack of content diversity in the real training data, which causes a content mismatch between the two datasets. For instance, consider an object detection task on trailer images where the real dataset has images of trailer type A and the simulated dataset has images of other trailer types to add diversity in trailer types to the full dataset. While such a problem setting is more challenging, it is also more practical, intentional and often encountered in real-world deployments.

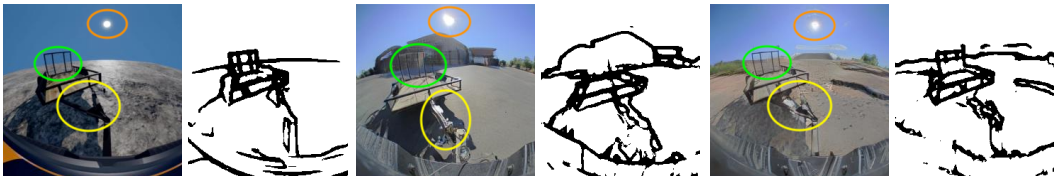
In the unsupervised setting, prior works such as Unsupervised Image-to-Image Translation (UNIT) [8], combine the VAE reconstruction error, the adversarial GAN loss, cycle consistency constraints [12]

*The author contributed to this work during his time as an intern with Ford Motor Company.

Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

and perceptual losses [5] to show promising results. However, when such unsupervised methods are applied to datasets that are heavily mismatched in terms of semantic content, these constraints are not enough to ensure retention of low-level scene semantics. Fig. 1a and Fig. 1b show sim-to-real results with UNIT trained on unpaired simulated and real trailer images. Here all simulated images are of trailers with *A-frame* couplers, while all real images are of trailers with *straight* couplers. Note that while high level content is preserved during sim-to-real translation from Fig. 1a to Fig. 1b, lower-level details, such as, trailer coupler shape (A-frame vs. straight), structure (number of vertical bars) and sun position, are not preserved. This renders ground truth labels from simulation unusable in any downstream perception task.

We argue that these discrepancies are due to the large content mismatch between the simulated and real datasets. GANs simultaneously train adversarial networks – a generator (`G`) and a discriminator (`D`) – that compete to generate realistic imagery and distinguish between synthetic and real imagery, respectively. Since all the real images that `D` sees during training have trailers with straight couplers, `G` quickly converges to a point where it starts replacing A-frame couplers with straight couplers. To address this issue, we introduce a simple yet surprisingly powerful notion - *what if we reduce the receptive field of `D` such that it can only see parts of, but never the full trailer coupler?* Fig. 1c shows sim-to-real results with a modified UNIT architecture, where the original `D` (with a receptive field of 46×46) is replaced with a modified `D` (with a receptive field of 16×16). Observe the improved retention in trailer coupler shape, number of vertical bars on trailer and sun position. This is primarily because the modified `D` penalizes structure at the scale of image patches that are much smaller than the size of the trailer coupler itself. The rest of the paper describes experimentation with the network architecture of `D`, which establish the hypothesis that a reduction in `D`'s receptive field is directly correlated to improved semantic content retention during sim-to-real translation with GANs.



(a) Input Simulated Image (b) Sim-to-real with UNIT (46×46) (c) Sim-to-real with Model 4 (16×16)

Figure 1: The effect of reducing `D`'s receptive field on unsupervised sim-to-real translation with mismatched data: **(a)** Simulated image of an A-frame trailer (left) and its Holistically-nested Edge Detections (HED) (right); **(b)** Sim-to-real translation on (a) with UNIT (left) and its HED (right). Note the shift in position of sun (orange), change in trailer coupler shape from A-frame to straight (yellow) and change in number of vertical bars on trailer (green); **(c)** Sim-to-real translation on (a) with modified UNIT, where the original `D` was replaced with one with a reduced receptive field (left) and its HED (right). Note the improved semantic content retention between (a) and (c) as compared to that between (a) and (b).

Model	Discriminator Architecture	Receptive Field
UNIT	$A \rightarrow A \rightarrow A \rightarrow A \rightarrow C$	46×46
1	$D \rightarrow D \rightarrow A \rightarrow C$	40×40
2	$A \rightarrow A \rightarrow A \rightarrow C$	22×22
3	$A \rightarrow B \rightarrow B \rightarrow B \rightarrow C$	22×22
4	$A \rightarrow B \rightarrow B \rightarrow C$	16×16
5	$A \rightarrow A \rightarrow C$	10×10
6	$A \rightarrow B \rightarrow C$	10×10
7	$B \rightarrow B \rightarrow B \rightarrow C$	10×10
8	$B \rightarrow B \rightarrow C$	7×7

Figure 2: List of UNIT and its modifications. A, B, C and D denote 2D convolutional layers with kernel sizes of $4, 4, 1, 4$ and strides of $2, 1, 1, 3$ respectively.

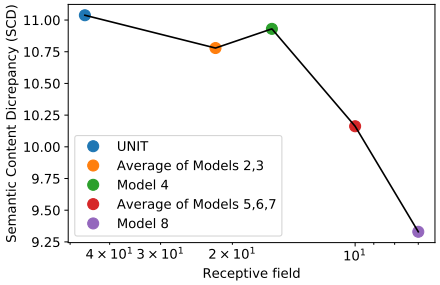


Figure 3: Plot of SCD vs. receptive field. SCD values for models with the same receptive field are averaged.

2 Preliminaries

PatchGAN: PatchGAN [4] has a convolutional `D` that penalizes structure at the scale of image patches, while simultaneously capturing local style statistics. As opposed to classifying the full image

as real or fake, a PatchGAN D classifies $N \times N$ image patches as real or fake. The final output for an image is the average of the patch-wise real/fake predictions. Isola et al. [4] showed high quality results even with patch sizes (N) that were much smaller than the image size. Given mismatched data, a D that classifies large image patches as real or fake prevents generation of semantic content (e.g. trailer shapes) not present in the real domain (e.g. A-frame). To this end, limiting D's *field-of-view* to smaller image patches, wherein the semantic entities present are only visible in part, can help preserve content during translations.

Unsupervised Image-to-Image Translation (UNIT): The UNIT architecture is a coupled VAE-GAN architecture, which makes use of a shared latent space assumption [8] to learns sim-to-real and real-to-sim models simultaneously. Thus, UNIT comprises of two Ds, one for each domain. Each D is a 5-layer PatchGAN. The first 4 convolutional layers have a kernel size of 4 and stride of 2, while the last layer has a kernel size of 1 and stride of 1. The effective receptive field (N) of each of the two Ds is 46. UNIT is used as the baseline architecture in this work.

3 Modifications to UNIT

We explored two different ways of reducing D's receptive field in UNIT: (i) reducing the number of layers; and (ii) reducing stride; and studied the effect of each on semantic content retention. Architectural variants of D with a kernel size other than the original size of 4×4 [8] caused a drastic drop in visual quality of the sim-to-real translations. Thus, kernel size was not varied in our experiments. Models 2-8 in Fig. 2 are the architectural variants of D that were used to investigate the role of D's receptive field in sim-to-real image translation.

Since reducing the number of layers in D also reduces its representation capacity, improved sim-to-real translations with Models 2 and 4-8 in Fig. 2) could also be attributed to reduced overfitting, as opposed to a reduction in D's receptive field. To investigate the role of overfitting, an additional experiment was conducted in which D's receptive field was kept similar to that in UNIT, while reducing the number of layers from 5 to 4 (refer Model 1 in Fig. 2 with a discriminator receptive field of 40×40 vs. UNIT with a discriminator receptive field of 46×46). For this purpose, no further experimentation was done as it is not possible to achieve a receptive field similar to that of 46×46 with fewer layers (i.e. < 4) for a kernel size of 4×4.

4 Experiments

Datasets: The trailer dataset used in this work has 9349 real images and 9000 simulated images (from Unreal Engine[2]) of trailers with straight couplers and A-frame couplers respectively, in daytime on various ground textures. The mismatch in this dataset arises from the difference in trailer coupler shapes between the two domains. We also created another, more conventional dataset that was intentionally mismatched between the two domains. This dataset is called the *parking-highway* dataset. It has 10438 simulated images of open parking lots in daytime (also from Unreal Engine) and 14172 real images of daytime highway scenes from BDD100K [11]. Note that the mismatch in the parking-highway dataset arises because of the different background semantics of the simulated and real world images (parking lots vs. highways).

Experiment Details: Our training procedure followed [8] and used ADAM [6] optimization for training with a learning rate of 0.0001 and momentums of 0.5 and 0.999. Each mini-batch comprised of an image from both the domains. All images were resized and then cropped to 256×256 for training. The VGG-16 [10] based perceptual loss was included only in the training of UNIT as-is and not in the training of any of the modified UNIT architectures, as listed in Fig. 2.

Qualitative Results: Fig. 4 shows sim-to-real results for all variants of D listed in Fig. 2. Since the datasets used for training are mismatched in terms of semantic content, as argued in Section 1, reducing D's field-of-view (as we move from left to right in Fig. 4), helps generate realistic images while also preserving semantic content. For instance, in the first row, note the gradual decrease in trailer scale and change in shape of the trailer coupler (from straight to A-frame) with a decrease in D's receptive field (as we move from Column (b) to (f)), to perfectly match the semantic content of the input image in Column (a). In the second row, the input trailer (Column (a)) is masked by

[2] https://en.wikipedia.org/wiki/Unreal_Engine

(a) Input (b) 46 × 46 [8] (c) 22 × 22 (d) 16 × 16 (e) 10 × 10 (f) 7 × 7

Figure 4: Effect of reducing D's receptive field on semantic content retention in sim-to-real translations on unpaired and mismatched data: Column (a) shows the input simulated image. Column (b) shows the output of UNIT as-is (with a receptive field of 46 × 46). Columns (c)-(f) show the sim-to-real outputs of the variants of UNIT with Ds with reduced receptive fields. In scenarios where there are multiple models listed with the same receptive field in Fig. 2, the results shown here are from the model with the most visually appealing results. The top two rows show sim-to-real results on the trailer dataset, while the bottom two rows show results on the parking-highway dataset. Note the gradual improvement in shape, scale and structure of objects present in the scene with a decrease in D's receptive field, as we move from left to right. Columns (e)-(f) show the best sim-to-real translation results on images in Column (a) in terms of realism and structural coherence.

shadows. Since most of the real images used for training had a trailer in the image center, the baseline UNIT translates the input image to a realistic looking image with a trailer in the center (Column (b)). Again, as we move from Column (b) to (f), there is an anticlockwise rotation in the orientation of the generated trailer to match the orientation in the input image in Column (a). We observed a similar trend in sim-to-real translations on the parking-highway dataset. In both the third and fourth rows in Fig. 4, the baseline UNIT (Column (b)) masks out the semantic content of the input image (Column (a)), and instead generates a highway road-type scene, as all the real images used for training were from dashcams of vehicles driving on highways. Gradually, as we move from Column (b) to (f), realistic looking parking lot images start getting generated, with semantic content matching that of images in Column (a).

Recall that the baseline UNIT trains sim-to-real and real-to-sim models simultaneously. Fig. 5 shows that a reduction in D's receptive field leads to improved semantic content retention in real-to-sim translations as well. Similar to the results shown in Fig. 4, reducing D's field-of-view (as we move from left to right in Fig. 5), helps generate images that look synthetic while also preserving semantic content. For instance, in the first row, note the gradual change in shape of the trailer coupler (from A-frame to straight) with a decrease in D's receptive field (as we move from Column (b) to (f)), to perfectly match the trailer couple shape (straight) in the input image in Column (a). In the second row, the baseline UNIT completely masks out the trailer and replaces it with wall shadows (Column (b)). Again, as we move from Column (b) to (f), we can notice the gradual appearance of a trailer in the same position as that in the input image in Column (a). A similar trend is observed in real-to-sim translations on the parking-highway dataset. In both the third and fourth rows in Fig. 5, the baseline UNIT completely ignores the semantic content of the input images in Column (a) to generate parking lot images, similar to those in the simulated parking data. Again, as we move from Column (b) to (f),

(a) Input (b) 46×46 [8] (c) 22×22 (d) 16×16 (e) 10×10 (f) 7×7

Figure 5: Effect of reducing D's receptive field on semantic content retention in real-to-sim translations on unpaired and mismatched data: Column (a) shows the input real image. Column (b) shows the output of UNIT as-is (with a receptive field of 46×46). Columns (c)-(f) show the real-to-sim outputs of the variants of UNIT with Ds with reduced receptive fields. In scenarios where there are multiple models listed with the same receptive field in Fig. 2, the results shown here are from the model with the most visually appealing results. The top two rows show real-to-sim results on the trailer dataset, while the bottom two rows show results on the parking-highway dataset. Again, similar to Fig. 4, note the gradual improvement in shape, scale and structure of objects present in the scene with a decrease in D's receptive field, as we move from left to right. Columns (e)-(f) show the best real-to-sim translation results on images in Column (a) in terms of visual quality and structural coherence.

highway images that look synthetic start getting generated, with semantic content matching that of images in Column (a).

Quantitative Results: Prior works use metrics such as the IOU over segmentation masks of the input and translated images to quantify content retention [9]. However, none of the open source datasets used for training popular deep segmentation networks (such as Mask R-CNN [3]) have a trailer class. Thus, we were unable to use these metrics in this work. The same is true for classification-based metrics, such as the FCN score [4]. Therefore, we define a new metric called 'Semantic Content Discrepancy' (SCD), which is measured as the Modified Hausdorff Distance (MHD) [1] between the Holistically-nested Edge Detections (HED) of the input simulated image and the sim-to-real translated image (see Fig. 1 for example HED outputs). While this metric does have some limitations (refer Appendix for details), in general, a lower SCD indicates that edges, shapes and scales of objects present in the input image are preserved during the sim-to-real translations. This metric was used for quantitative analysis of results from sim-to-real translation of 500 images from the trailer dataset (see Fig. 3). Note the consistent drop in SCD with a decrease in D's receptive field (plotted on a log scale).

Role of Overfitting: Fig. 6 shows that reducing the number of parameters/layers in D (which in effect also reduces its representation power and consequently reduces the generator's chances of overfitting) while keeping its receptive field the same as the baseline UNIT architecture, does not improve structural coherence in sim-to-real translations (see Columns (a)-(f)). A quantitative analysis of the sim-to-real translation results on 500 simulated trailer images gives an average SCD of 12.20 with Model 1 (4 layers), as compared to that of 11.03 with UNIT (5 layers). Thus, simply reducing the number of parameters in the discriminator does not result in improved semantic content retention for a receptive field of 46×46. The same trend is observed with discriminator models with a receptive

field of 22×22 in all scenarios except for the second example from the parking-highway dataset (see Columns (g)-(l)). Again, a quantitative analysis of the sim-to-real translation results on 500 simulated trailer images gives an average SCD of 11.21 with Model 2 (4 layers), as compared to that of 10.35 with Model 3 (5 layers). However, as shown in Columns (m)-(r), reducing the representation power of the discriminator in the case of a receptive field of 10×10 improves semantic content retention during translation as we see better results with Model 5 (3 layers) as compared to those with Model 7 (4 layers). Quantitatively also Model 5 gives an average SCD of 9.51 as compared to that of 11.06 with Model 7. All these results combined lead us to conclude that reducing the number of parameters in the discriminator architecture while keeping its receptive field the same does not always lead to better results. However, for discriminator architectures with a small enough receptive field (10×10 in this case), designing an architecture with fewer parameters could lead to a lower SCD, while maintaining the desired level of photorealism (again, see Columns (m)-(r)).

(a) Simulated trailer (b) UNIT (46×46) (c) Model 1 (40×40) (d) Simulated parking (e) UNIT (46×46) (f) Model 1 (40×40)

(g) Simulated trailer (h) Model 3 (22×22) (i) Model 2 (22×22) (j) Simulated parking (k) Model 3 (22×22) (l) Model 2 (22×22)

(m) Simulated trailer (n) Model 7 (10×10) (o) Model 5 (10×10) (p) Simulated parking (q) Model 7 (10×10) (r) Model 5 (10×10)

Figure 6: *Role of overfitting:* Effect of reducing the number of parameters in D while simultaneously maintaining its receptive field on sim-to-real translation with mismatched data. Columns (a), (b), (c); (g), (h), (i) and (m), (n), (o) show results on the trailer dataset, while Columns (d), (e) and (f); (j), (k), (l) and (p), (q), (r) show results on the parking-highway dataset. Columns (a), (d), (g), (j), (m) and (p) show input simulated images. The rest of the columns show sim-to-real translations results, for the simulated images to their left, with the models listed in the column captions.

5 Discussion

Realism vs. Semantic Content Retention: Fig. 4 and Fig. 5 establish the fact that reducing D's field of view, i.e. receptive field, which is equivalent to weakening the realism signal during GAN training, helps reduce SCD in unsupervised image translation. However, we would like to note that this SCD improvement comes at the cost of photorealism, specifically in the case of the trailer dataset. For instance, in the first row in Fig. 4, sim-to-real translation using the baseline UNIT architecture (with a receptive field of 46×46) gives a more photorealistic image than that using Model 8 (with a receptive field of 7×7). Thus, there exists a two-way tradeoff between photorealism and semantic content retention during translation. Extensive experimentation with D's architecture is required to find the sweet spot of receptive field that simultaneously achieves desired levels of photorealism and semantic content retention during translation.

Limitations of the SCD metric: Recall that SCD is defined as the MHD between the Holistically-nested Edge Detections (HED) of the input simulated image and the sim-to-real translated image. Thus, in effect, SCD quantifies the difference between edge maps of the input and translated images. This is beneficial with respect to object entities, such as trailers and cars, for which simulation provides ground truth labels in the form of bounding boxes. However, a difference between edge masks also penalizes visual artifacts generated during translation, such as trees and building in the background (see Fig. 1b). The generation of such artifcats might infact be desirable for certain perception tasks, such as trailer detection, as such artifacts add diversity to the training images. Thus, whether or not SCD is the right metric to quantify which sim-to-real model does best in terms of both realism and semantic content retention is very much dependent on the downstream perception task.

6 Conclusion

Experiments with D's architecture in UNIT [8] on two mismatched datasets show that D's receptive field is directly correlated with semantic content discrepancy of the generated image in sim-to-real image translation. Thus, reducing D's receptive field reduces semantic content discrepancy during translation, as shown both qualitatively and quantitatively in Section 4. However, to effectively learn sim-to-real translation, the two datasets cannot be completely mismatched. For e.g., one cannot expect to learn what real trees look like if the real images are of cars only. Furthermore, diversity in the training datasets from the two domains, regardless of the content mismatch between them, is still essential as GANs very quickly 'memorize' and thus, are highly susceptible to overfitting.

Acknowledgments

The authors would like to sincerely thank Rohan Bhasin, Gautham Sholingar and Xianling Zhang for generating the synthetic data used in this work. We would also like to thank Gaurav Pandey for insightful discussions; Jinesh Jain, Marcos P. Gerardo-Castro and Sandhya Sridhar for reviewing the submission draft; and Raju Nallapa and Ken Washington for their continuous support and encouragement.

References

[1] M.-P. Dubuisson and A. K. Jain. A modified hausdorff distance for object matching. In *Proceedings of 12th international conference on pattern recognition*, volume 1, pages 566–568. IEEE, 1994.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[5] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[7] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[8] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[9] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[11] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.