CLIPTraVeLGAN for Semantically Robust Unpaired Image Translation

Yevgeniy Bodyanskiy, Nataliya Ryabova and Roman Lavrynenko

Kharkiv National University of Radio Electronics, Nauky av., 14, Kharkiv, 61166, Ukraine

Abstract

In this paper a novel approach for semantically robust unpaired image translation is presented. CLIPTraVeLGAN replaces the Siamese network in TraVeLGAN with a contrastively pretrained language-image model (CLIP) with frozen weights. This approach significantly simplifies the model selection and training process of TraVeLGAN, making it more robust and easier to use.

Keywords

Image-to-image translation, GAN, CLIP, Transfer knowledge

1. Introduction

Currently generative deep learning has become the most promising direction in the development of IT technologies, in which modern generative neural network models are being developed. There is a lot of research going on in this rapidly growing field of machine learning, and a large proportion of it is focused on generative adversarial networks [1, 2]. The main property of generative adversarial networks (GAN) is unsupervised learning, thanks to which GANs successfully demonstrate a wide range of creative potential and in particular, the ability to generate images. Interesting tasks in this direction belong to image-to-image translation (I2I) which includes image translation from one subject domain to another while maintaining the main content. In recent years, many different GAN models have been developed that solve this type of problem with different variations of architectures. Their classification, advantages and disadvantages are considered concerning analysis methods and applications for image-to-image translation problems in [3].

Semantic robustness is an essential aspect of image translation. In the context of unpaired image translation, semantic robustness is particularly important to ensure that the translations are accurate and meaningful. Image translation from one domain to another is a challenging task that requires the generated image to belong to the target domain while also retaining the individuality of the input image. TraVeLGAN [4] was designed to address this task by using a Siamese network to encode the high-level semantics that characterizes the domains. However, the Siamese network selection and cooperative training process are complex. In this paper, we propose a novel approach to simplify the training process of TraVeLGAN. Our approach is based on the use of a contrastively pretrained language-image model (CLIP [5]) with frozen weights instead of the Siamese network. We call this new model CLIPTraVeLGAN.

Our approach aims to solve the problems associated with cooperative learning in TraVeLGAN. In TraVeLGAN, the generator and the Siamese network have the same goal, which leads to additional difficulties in determining the effectiveness of each solution. Our approach eliminates the need for choosing and training a Siamese network and thus avoids these difficulties. At the same time, the generator still receives the TraVeL loss as proposed in [4]. This makes our approach simpler and

COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine

EMAIL: yevgeniy.bodyanskiy@nure.ua (Y. Bodyanskiy); nataliya.ryabova@nure.ua (N. Ryabova); roman.lavrynenko.cpe@nure.ua (R.Lavrynenko);
ORCID: 0000-0001-5418-2143 (Y. Bodyanskiy); 0000-0002-3608-6163 (N. Ryabova); 0000-0003-1969-1107 (R. Lavrynenko)

more straightforward, while still ensuring the high-level semantics are captured in the generated image.

The transfer of knowledge from CLIP to CLIPTraVeLGAN enables the generator to understand the relationships between words and images without any additional training. In this paper, we present the results of our experiments and compare CLIPTraVeLGAN with the original TraVeLGAN. Our results show that CLIPTraVeLGAN outperforms TraVeLGAN in terms of both stability and quality of the generated images. This paper is a contribution to the field of image translation and provides a promising new direction for further research.

The effectiveness of our proposed method, CLIPTraVeLGAN, is evaluated on a benchmark dataset for unpaired image translation. The results are compared with other methods to demonstrate the performance of our method.

Our code is available on www.kaggle.com/code/unfriendlyai/cliptravelgan-gta-cityscapes

2. Related Works

2.1. TraVeLGAN

The field of image translation has seen significant advancements in recent years with the introduction of Generative Adversarial Networks. One such approach, CycleGAN [6], introduced the idea of using a cycle consistency loss to preserve the content and structure of the input image during translation. Despite its success, CycleGAN has some limitations such as the assumption of cyclic consistency which can result in blurry translations and mode collapse. To address these limitations, TraVeLGAN [4] was introduced as a competitor to CycleGAN for one-sided image translation.

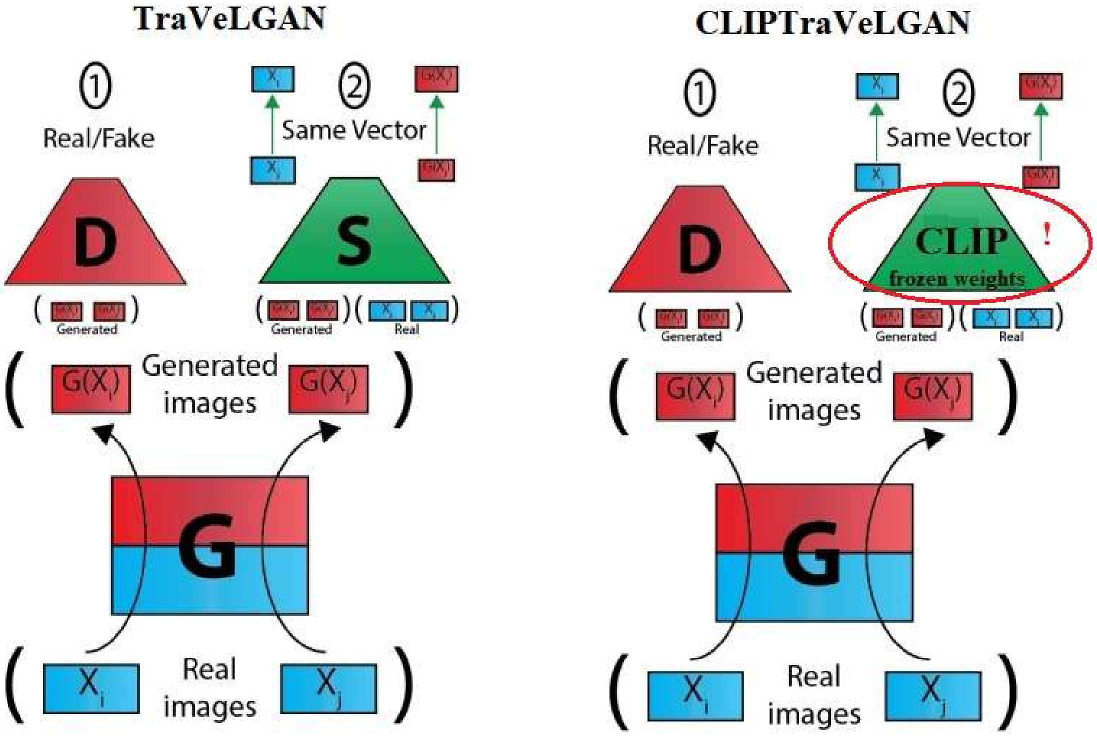


Figure 1: The only difference between our solution and the TrAVeLGAN [4] is the pretrained CLIP instead of the Siamese network S

Problem setting. Let X and Y be two subject regions of images. The training dataset consists of a finite number of elements $\{x_i\} \in X$ and $\{y_i\} \in Y$, which do not have pairwise connections among themselves. That is, no image y_i is a translation of the image x_i, which means unsupervised learning. The task of unpaired translation of images from one domain to another is to train the generator G_{XY} to perform the transformation X → Y. At the same time, the generator G_{XY} should not just generate images belonging to the domain Y, but between the input image from the domain X and the output the image of the generator G_{XY} should be a significant and understandable connection, that is, the individual features and semantic content of the input image should be reflected in the generated image.

Thus, this task of unpaired image translation consists of two components: the generated image must be a member of the target domain and have the individuality of the input image.

Membership in the target domain. The generator must ensure that G_{XY} (X) ∈ Y. To do this, a standard GAN architecture is used, in which a discriminator D_Y tries to distinguish generated images from real samples from Y. The goal of optimizing the generator parameters is to maximize D_Y (G_{XY} (X)), and the goal of optimizing the discriminator parameters is, conversely, to minimize D_Y (G_{XY} (X)) and maximize D_Y (Y). The networks compete with each other.

Individuality. If x_i, $x_j \in X$, $i \neq j$, then there must be a relationship between x_i and G_{XY} (x_i), which explains why G_{XY} (x_i) is a representation in the domain Y of the image x_i, and not x_j. In [4] for this purpose the Siamese network S is trained, which will transform the images of both subject areas into a vector in some hidden space. For training S, pairwise differences between images are determined for each batch of input images: Vij = S (x_i) – S (x_j) and Zij = S (Gxy (x_i)) – S (Gxy(x_j)). The goal of optimizing the parameters of the Siamese network S and the generator Gxy is to maximize the cosine similarity between Vij and Zij for all cases when $i \neq j$. The network S is a proof that there is an explanation according to which the generated image G_{XY} (x_i) differs from any G_{XY} (x_j) as much as the samples x_i and x_j are different from each other. Since networks S and Gxy have the same goal, they do not compete, but help each other - they cooperate.

The authors [4] introduce the concept of a transformation vector between two points. In natural language processing tasks, words are represented by points in a space in which if a certain vector would transform the word "man" into the word "woman", then the word "king" into the word "queen" would be transformed by a very similar vector. A Siamese network S represents an image by points in a certain space. In the image translation task, instead of changing the gender of the word, the transformation vector can change the background colour, size or shape of the image. But the main idea is that the vector of transformation of a point obtained from the Siamese network from one image S(x_i) ("man") to a point of another original image S(x_j) ("woman") will also transform the point of the generated image S (G_{XY} (x_i)) "king" to the point of the generated S(G_{XY}(x_j)) "queen".

TraVeLGAN used an additional Siamese network to encode high-level semantics between the source and target domains. This idea seemed like a breakthrough as the Siamese network was believed to outperform CycleGAN in terms of translation quality. However, TraVeLGAN has not received much development due to the difficulties in choosing the architecture of the Siamese network and the parameters of its training. This results in a large set of possible solutions and makes it difficult to determine the effectiveness of each of them.

2.2. Contrastive language-image pretraining

CLIP [5] was introduced as a language-image model for the transfer of knowledge without any further training. After pretraining the model, it can be used for any purpose with any images without any tuning. Trained on a dataset of billions of image-caption pairs from the Web (WIT), the model can successfully classify images with text class labels for a wide range of tasks, even quite far from its training set: geolocation, car brands. CLIP trained on WIT shows better accuracy on ImageNet than ResNet50 trained on ImageNet. The worst performance of knowledge transfer without additional training is shown on very specialized data sets, such as classification of satellite images, medical images, and object counting in synthetic images.

The authors discovered an unexpected feature that, on many data sets, knowledge transfer without any post-training performs better than adding logistic regression on the top of the frozen network and post-training in 4 epochs on a new data set. Even worse indicators were obtained when trying not to freeze CLIP, but to fit all layers on a new data set.

The internal representation of CLIP. One of the side effects of CLIP is that encoders learn the internal representation of images in a shared space with the internal representation of natural language texts. Although there is no consensus in the scientific community on what is a "perfect" representation, one common option for testing the quality of a representation is to train a linear classifier attached to a frozen model and determine the performance of that model on different datasets. According to the results of experiments, all CLIP models regardless of encoder and size outperform any other known models in this test.

Natural language encodes semantic content and hierarchical relations between concepts with words. Contrastive learning of a visual model using natural language texts as a learning cue led to the learning and generalization of such special knowledge about image elements as expressed in image-relevant texts in human language. The extent to which the visual model learns the hierarchy of concepts that exists in a human language requires separate research. Currently, CLIP reflects the meaning of the image in hidden representation most effectively among other well-known models.

The vector into which CLIP transforms images is the best choice for finding similar images. Other options for using CLIP in the search task are finding images that are most relevant to the content of some text and finding the text that most relevantly describes the image.

Overall, CLIP's powerful internal representation of images and text make it a valuable tool for a wide range of applications, with potential future uses that have yet to be imagined.

In our work, CLIPTraVeLGAN, we use CLIP as a means of preserving high-level semantics between the source and target domains in unpaired image-to-image translation.

2.3. Semantic robustness

In [7], the concept of semantic stability of unpaired translation of images was introduced and the reasons for the conflict between compliance with the subject area and accuracy of the translation, and the reasons for hallucinating objects that are absent in the input image were highlighted. SRUNIT model is proposed to provide translation semantically robust, which is simultaneously trained with a generator and a discriminator similar to TraVeLGAN's Siamese network. CLIP is not used. In [8], the use of Vector Symbolic Architectures was proposed to improve the semantic robustness of unpaired image translation, which showed even better indicators of semantic translation accuracy than SRUNIT. CLIP is not used also.

The robustness of CLIP under natural skew of data distribution was tested in [5]. If the model is trained on one set of data, and then the efficiency is determined on the updated (sometimes synthetically corrected), then the efficiency of the models is significantly reduced. CLIP is more reliable in distribution bias problems compared to models pretrained on ImageNet. This property is especially important for the case of ensuring semantic robustness when translating images between domains.

The use of CLIP in our work adds the ability to preserve the high-level semantics between the source and target domains, making the translations semantically robust.

In conclusion, our work, CLIPTraVeLGAN, builds upon the idea of TraVeLGAN and adds the advantages of CLIP to improve the quality of unpaired image translation while preserving the high-level semantics between the source and target domains.

3. Method

The core of the CLIPTraVeLGAN approach is the use of a pre-trained language-image model (CLIP) as a Siamese network in TraVeLGAN setup. The proposed CLIPTraVeLGAN model is composed of a generator, a discriminator and pretrained CLIP model. The generator takes an image from one domain and generates an image that belongs to the other domain. The discriminator is responsible for distinguishing between real and fake images. In CLIPTraVeLGAN, we replace the Siamese network in TraVeLGAN with the pre-trained language-image model CLIP. The CLIP model is used to encode the high-level semantics of the input and target domains.

We train the CLIPTraVeLGAN model using the adversarial loss and the TraVeL loss. The adversarial loss is used to ensure that the generated image belongs to the target domain, while the TraVeL loss encourages the generator to preserve the high-level semantics of the input image. Thus, the final objective terms of the generator are:

$$Lg = Ladv + \lambda Ltravel, \tag{1}$$

where λ controls the relative importance of TraVeL loss.

TraVeL loss is the same as in TraVeLGAN:

$$Ltravel = \Sigma\Sigma i \neq j \; Dist[\, S(Xi) - S(Xj), \qquad S\big(G(Xi)\big) - S\big(G(Xj)\big)] \,, \tag{2}$$

where Dist is a distance metric, such as cosine similarity.

One advantage of our approach is that it eliminates the need for choosing and training a Siamese network, which can be complex and time-consuming. Instead, the transfer of knowledge from CLIP to CLIPTraVeLGAN enables the generator to understand the relationships between images without any additional training. This makes our approach simpler and more straightforward, while still ensuring the high-level semantics are captured in the generated image.

In this context, the use of CLIP in CLIPTraVeLGAN adds the ability to preserve high-level semantics between the source and target domains, making the translations semantically robust. Therefore, our work builds upon the idea of TraVeLGAN and leverages the advantages of CLIP to improve the quality of unpaired image translation while maintaining semantic robustness.

4. Experiment

For experiments we use Kaggle environment with TensorFlow and TPU support. Pretrained CLIP weights were loaded from https://huggingface.co transformers package openai/clip-vit-large-patch14. The dataset contains images from two different domains that are not aligned with each other. We preprocess the dataset by resizing all the images to a fixed size of 256x256 and normalising the pixel values to lie in the range [-1, 1]. CLIP model receives a central crop with the size of 224x224 of images preprocessed according to its configuration.

To test the effectiveness of the proposed approach, we used the models studied in the Kaggle competition "I'm Something of a Painter Myself" [9] as a basis for the experiments. CycleGAN showed the best results in the competition, while TraVeLGAN was the most promising one-side image translation model. We replaced the Siamese network in TraVeLGAN with a contrastively pre-trained language-image model (CLIP) to create the CLIPTraVeLGAN model. Batch size 128 was chosen for the experiments, following the example of other CLIP applications with large batch-size values.

The generators and discriminators of all models were identical. The purpose of the experiments was to establish the effect of using the proposed improvement, that is replacing the Siamese network with pretrained CLIP. We use the Adam optimizer to train both the generator and discriminator.

To eliminate the irrelevant effects associated with struggling with overfitting on a small number of Monet paintings in competition, as well as the difficulty of visually evaluating Monet paintings, models must generate realistic photos of the landscape from Monet paintings. Some of Monet's paintings did not participate in the training. It is designed to test the quality of image translation and determine the FID metric.

We evaluate CLIPTraVeLGAN on GTA (Grand Theft Auto) [10] to Cityscapes dataset [11] which is a benchmark dataset for unpaired image translation because it involves translating images from one domain to another, where the source and target domains are vastly different.

The GTA and Cityscapes datasets represent two different domains of real-world urban environments. The GTA dataset consists of images of urban scenes generated from a video game, while the Cityscapes dataset comprises real-world urban scenes captured by a camera mounted on a car. The images in these datasets differ in terms of lighting conditions, weather, time of day, and many other factors. The main problem is that GTA images have more sky than Cityscapes. The discriminator can easily distinguish fake image by that criterion. Cityscape images have more vegetation instead. Thus, models may hallucinate vegetation in open sky regions which is semantic mistake.

We made an ablation study using CLIP model without pretrained weights to understand if semantic CLIP knowledge is necessary for the correct translation of an image. TraVeL loss turned into cosine similarity of noise vectors instead of semantic vectors.

5. Results

The results of the translation of Monet's paintings into photographs were the following. The basic CycleGAN model showed on the test dataset FID = 6.9. Basic TraVeLGAN had 7.7. The investigated CLIPTraVeLGAN showed a result between the two base models, FID = 7.3. An example of resulted images is shown in Figure 2.

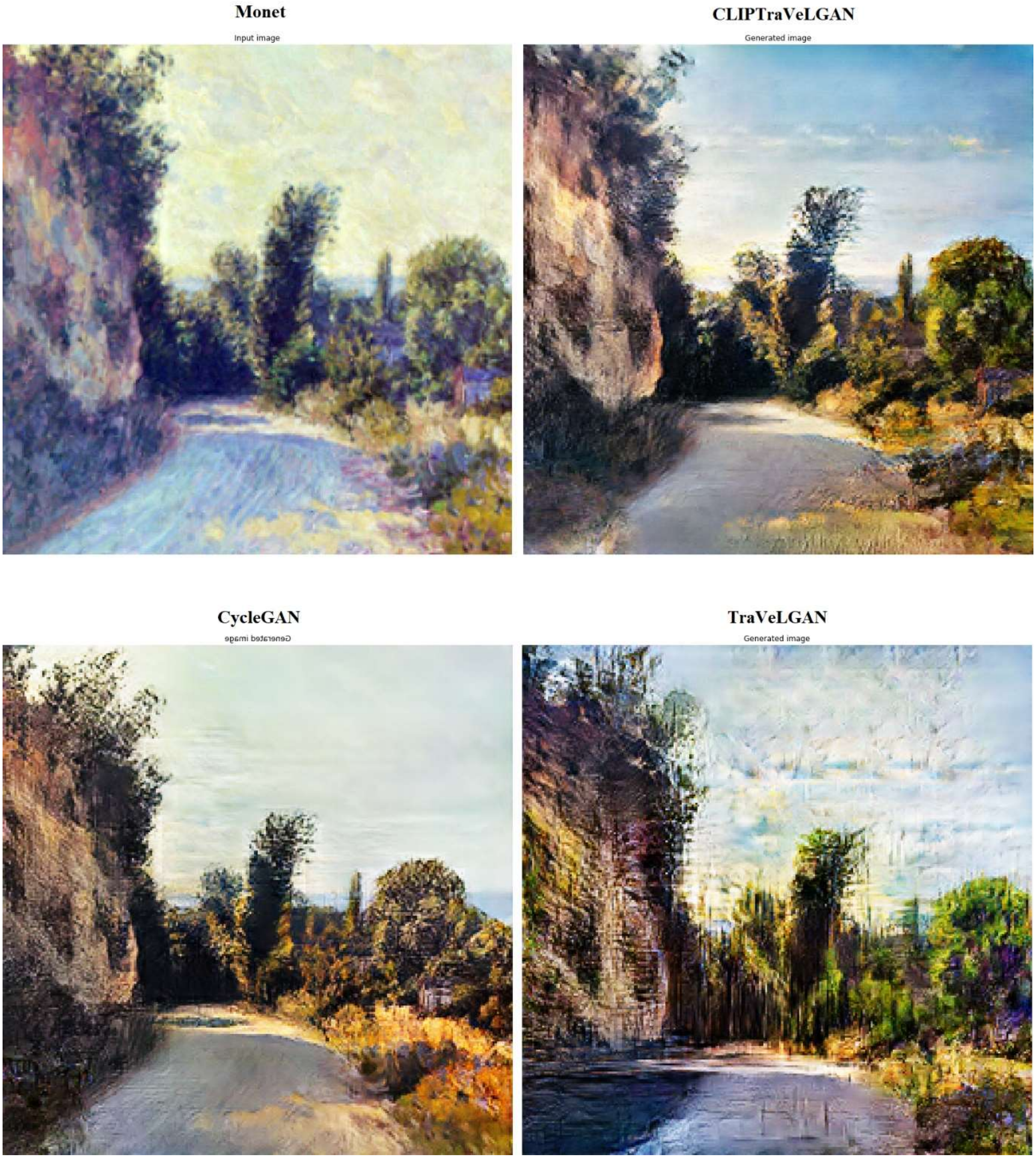


Figure 2: CLIPTraVeLGAN, TraVeLGAN and CycleGAN in Monet-Photo task

We compare the results of CLIPTraVeLGAN with those of TraVeLGAN to evaluate the effect from using pretrained CLIP instead of the Siamese network using GTA – Cityscapes benchmark (Figure 3).

Yellow lane lines from GTA should be translated into white lane lines. All real Cityscapes images have Mercedes hood ornaments. GTA has more sky than Cityscapes. An example of semantic flipping is hallucinations of trees instead of the sky.

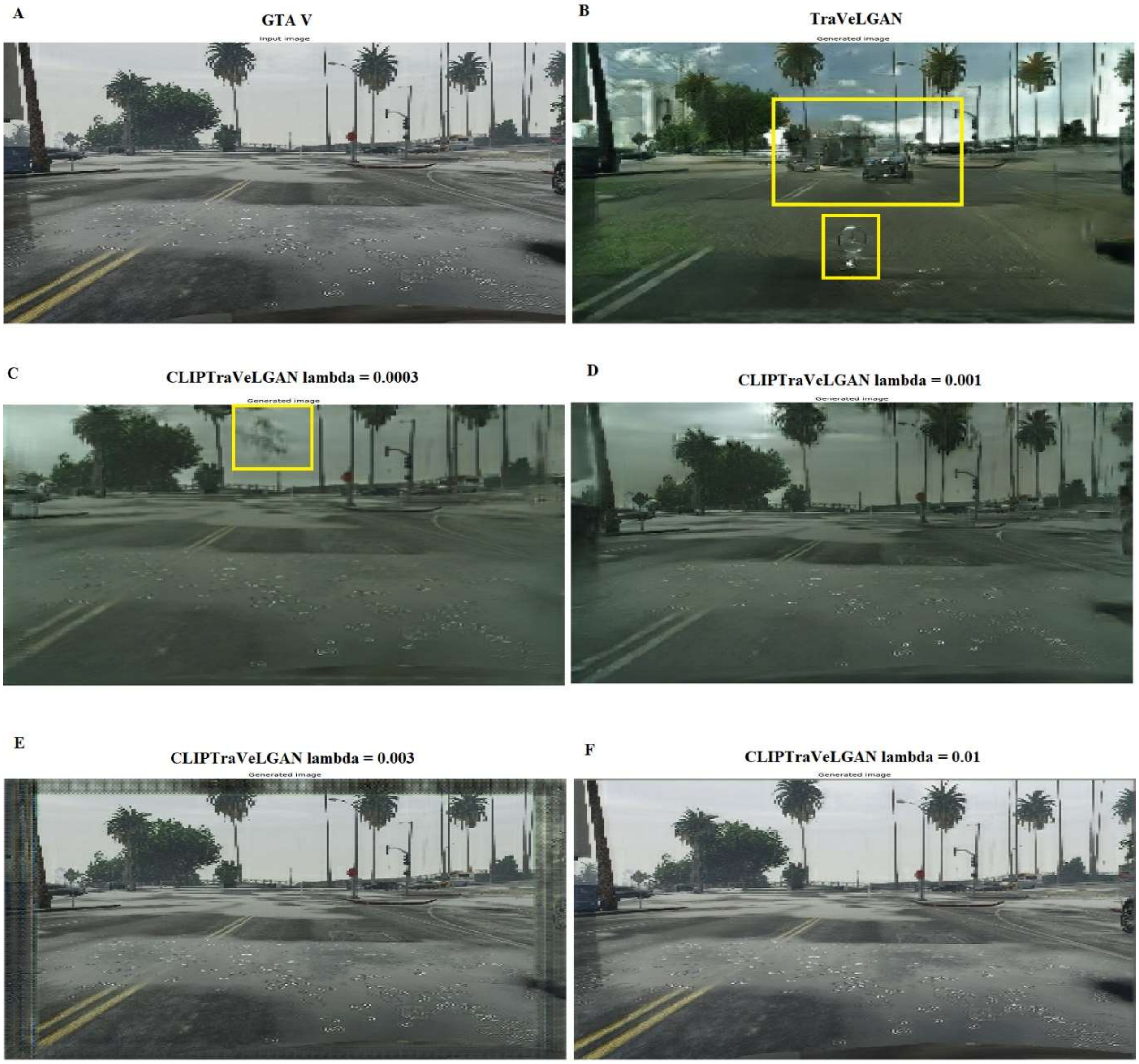

A
GTA V
Input image
B
TraVeLGAN
Generated image
C
CLIPTraVeLGAN lambda = 0.0003
Generated image
D
CLIPTraVeLGAN lambda = 0.001
Generated image
E
CLIPTraVeLGAN lambda = 0.003
Generated image
F
CLIPTraVeLGAN lambda = 0.01
Generated image

Figure 3: TraVeLGAN and CLIPTraVeLGAN (B, C and D) translate yellow lane lines into white lane lines that are commonly found in Cityscapes images. All models dislike open sky regions of the image causing discriminator easily distinguish fake image by that criterion. GTA has more sky than Cityscapes. TraVeLGAN (B) hallucinate car, CLIPTraVeLGAN with lambda=0.0003 (C) paints something unnatural in the sky. When lambda grows to 0.001 (D) translation is almost correct. With lambda>=0.003 (E, F) output images are equal to input (no translation at all, yellow lane lines).

We compared results with different values of λ that controls the relative importance of TraVeL loss and a different number of updates. FID metric conflicts with semantic robustness. When the coefficient of the importance of TraVeL loss λ is small it still prevents the model from collapse mode but allows semantic flipping. The FID metric is the best and the output image fits target domain. Increasing the value of λ results in worse FID values, but helps control semantic flipping. When the λ value is too high, the model renders the input images unchanged. This is the easiest way to ensure TraVeL consistency between input and output pairs.

The results of the GTA-Cityscapes experiment are shown in Table 1 and in Figures 3,4. There is a trade-off between image quality and semantic robustness. In our experiments, we searched for the optimal value of λ to achieve satisfactory results according to both criteria.

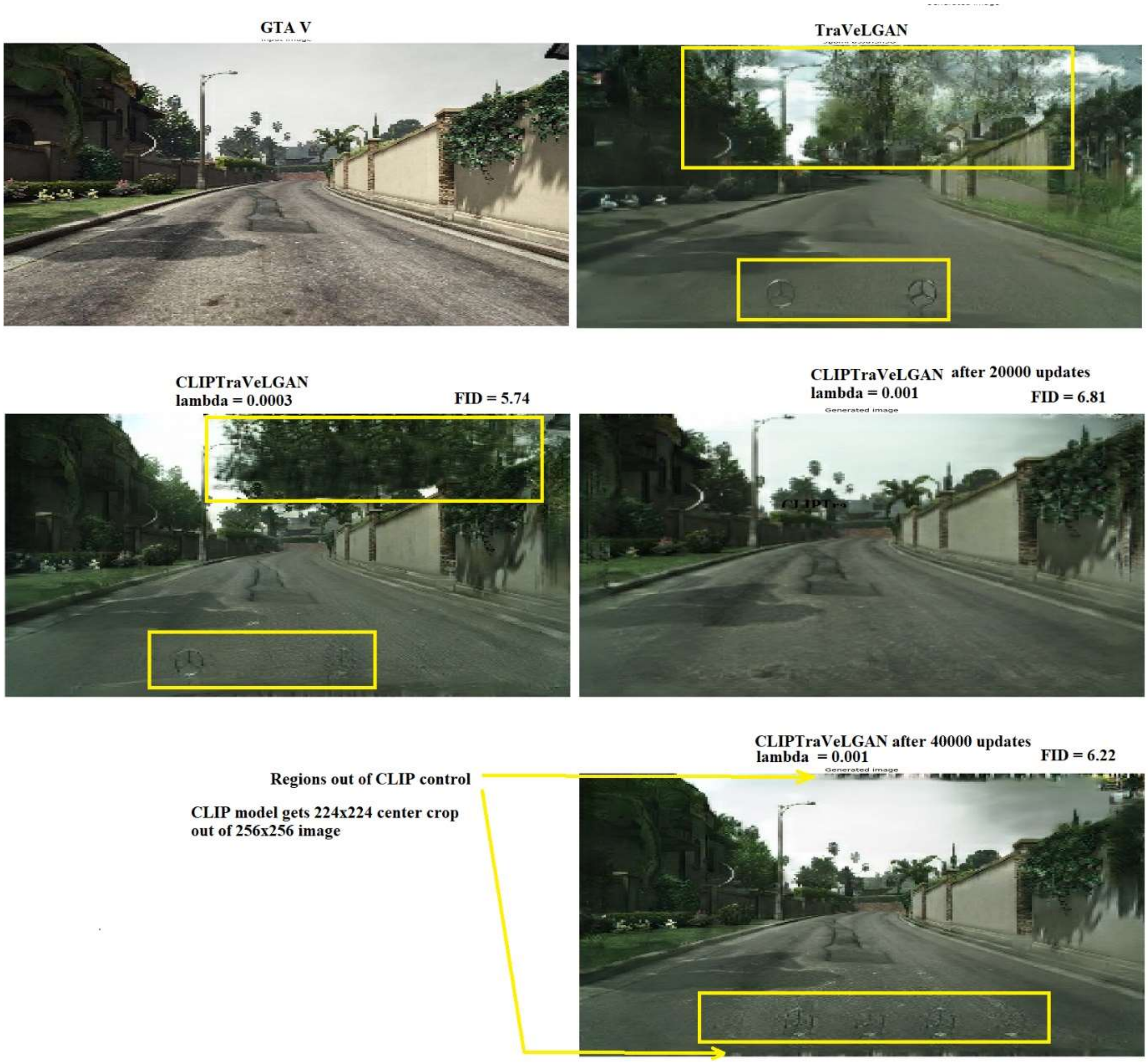


Figure 4: Results for different lambda and number of updates

Table 1
CLIPTraVeLGAN results

λ	Number of updates	FID	Translation results
0.0003	40 000	**5.74**	The model generates trees and artefacts in the sky. Yellow lane lines are translated correctly into white lane lines
0.001	40 000	6.22	Model hallucinates multiple Mercedes hood ornaments. There are artefacts in regions out of CLIP control. The rest of the sky is translated correctly
0.001	20 000	6.81	**There is no semantic flipping. Yellow lane lines are translated correctly into white lane lines**
0.003	17 000	9.09	High values of λ prevent translation. Yellow lane lines remain the same. Output images are almost equal to the input except some artefacts in regions out of CLIP control
0.01	17 000	9.08	
0.01	40 000	8.90	

In Figure 5 we show the results of translation of the same images that were used to compare the state-of-the-art VSAIT model against other models in [8]. We demonstrate that our method does not exhibit semantic flipping. CLIPTraVeLGAN results are very close to VSAIT.

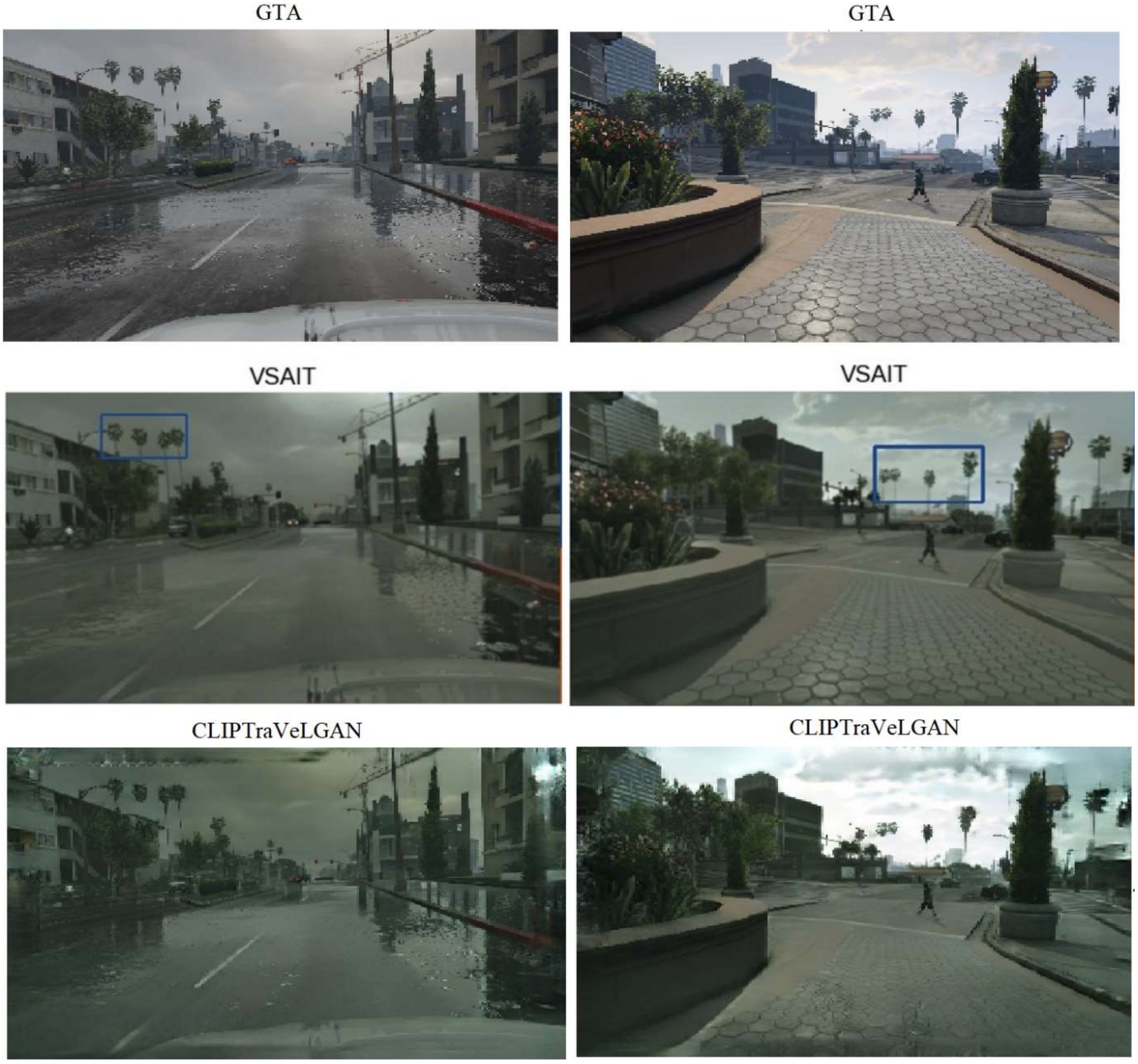


Figure 5: Examples of semantic flipping for VSAIT (images from [8]) and CLIPTraVeLGAN in GTA to Cityscapes experiment.

In the Ablation study (Figure 6) we use CLIP model without pretrained weights. Semantic CLIP knowledge is necessary for the correct translation of an image.

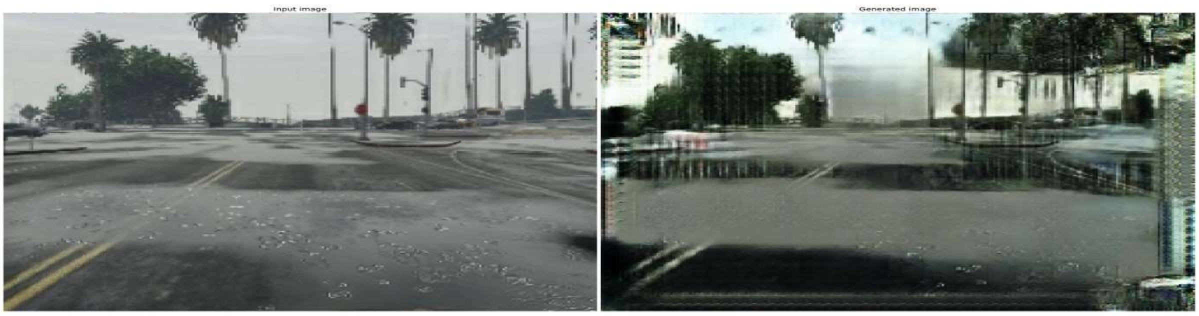
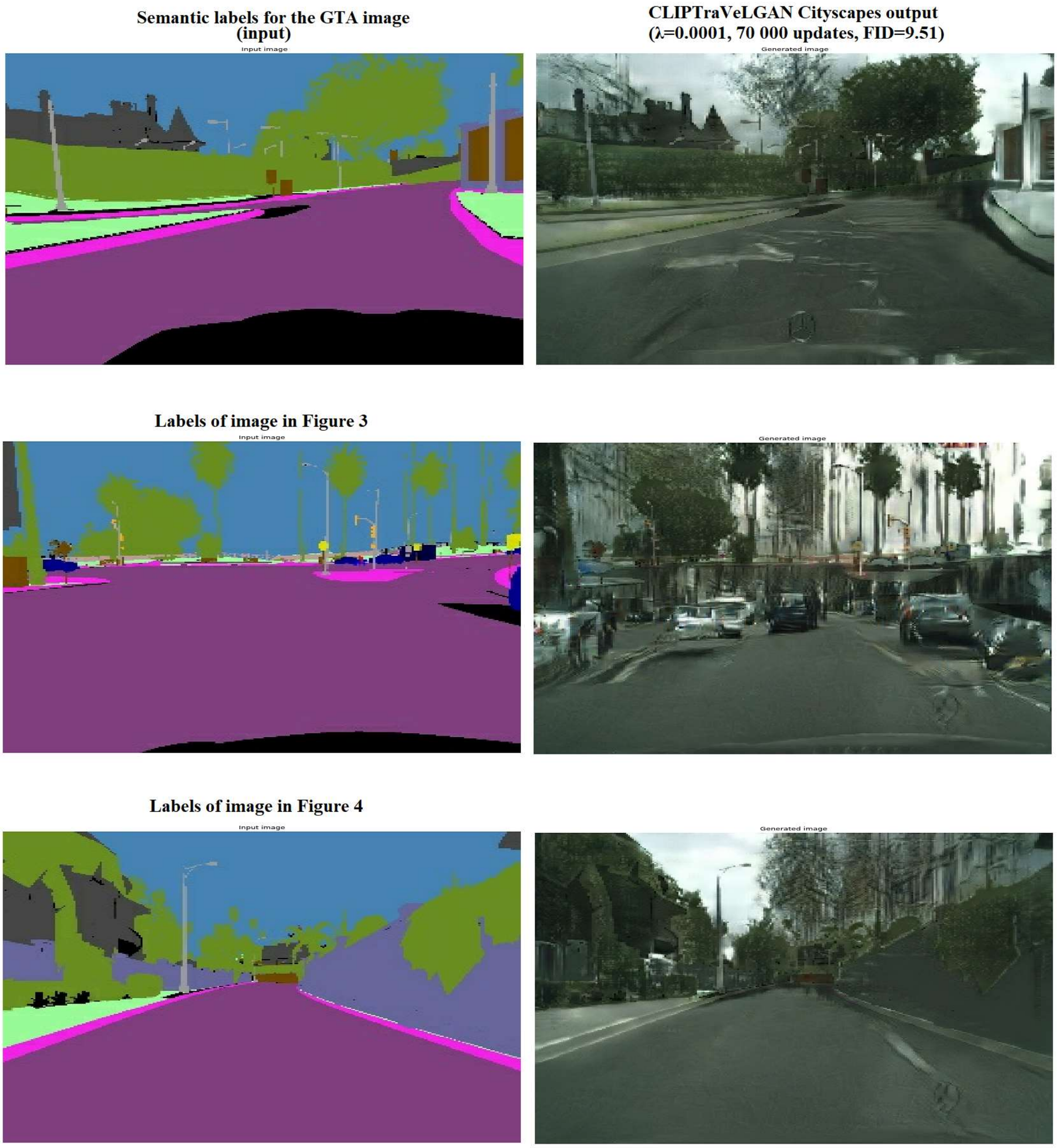


Figure 6: Ablation study results. Using CLIP model without pretrained weights


Semantic labels for the GTA image (input)

CLIPTraVeLGAN Cityscapes output (λ=0.0001, 70 000 updates, FID=9.51)

Labels of image in Figure 3

Labels of image in Figure 4

Figure 7: Visual results of the GTA Label to Cityscapes Image

We trained CLIPTraVeLGAN to translate GTA Label to Cityscapes image. Our method exhibits semantic flipping when λ is small. Increasing the value of λ leads to the disappearance of the gradient from the discriminator. This seems to be a limitation of our method when domain images (like semantic labels) were not found in the CLIP training dataset.

6. Discussions

Our biggest impression from the experiments is that the CLIPTraVeLGAN is very easy to train compared to many image-to-image translation models we tried before. With this model, we did not meet the collapse mode. The only TraVeL loss from pretrained CLIP is enough to prevent it.

We evaluated our proposed method on a benchmark dataset for unpaired image translation and compared it with other methods. The results demonstrated the effectiveness of our approach and the potential for further research in this area. However, it should be noted that our approach has some limitations and there is still room for improvement.

The main problem is that perfect semantically robust translation and perfect membership in the target domain are incompatible. The FID metric itself is only an additional indicator of the quality of the generated images but has nothing to do with the accuracy of the translation. Even more, the accuracy of the translation conflicts with the FID metric - the more the image matches the target domain, the more semantic changes were made during the translation. Thus, an approximate correspondence of the FID metric different models is enough to consider the model as a candidate for further research. Trying to demonstrate SOTA by metric FID on any dataset for an image translation task is not proof of its superiority. The evaluation on the GTA to Cityscapes dataset showed that the proposed CLIPTraVeLGAN approach outperformed TraVeLGAN for unpaired image translation in terms of both stability and quality of the generated images. The FID score of CLIPTraVeLGAN is lower than that of TraVeLGAN when λ is small and translation is not perfect. But CLIPTraVeLGAN can provide such semantic stability of translation that is not available for TraVeLGAN. The results of the experiment demonstrate that CLIPTraVeLGAN is a promising approach to image translation and provides a new direction for further research.

Additionally, we evaluated our method on Kaggle competition dataset. Since the task of the Kaggle competition is not to translate images from one domain to another, but to generate images, the FID metric displays only quality generation and diversity of images of the target domain. Semantic robustness of the translation is not evaluated in any way.

One of the key features of CLIP is its internal representation of images and natural language texts in a shared space learned through contrastive learning. This shared space allows for finding similar images and relevant text descriptions of images, among other potential applications. Additionally, CLIP's encoders learn to represent images in a way that reflects the meaning of the image most effectively among other well-known models.

Specifically, we tried using CLIP to evaluate the similarity of the semantic content between the input image and the generated image. During training, the generator network tries to produce an output image that is not only visually similar to the input image but also semantically similar in the sense that it is classified similarly by CLIP.

To achieve this, we used a pre-trained CLIP model as a feature extractor and obtain the CLIP embedding of both the input and generated images. We then use the cosine similarity between these embeddings to evaluate the semantic similarity between the two images. Such straightforward method is not effective even for GTA to Cityscapes translation. It prevents translating yellow lane lines into white lane lines that are commonly found in Cityscapes images. Translating “men into women” or “horse to zebra” conflicts with such semantic identity too. To avoid such effect, we developed an idea using CLIP in the TraVeLGAN setup. By using CLIP in this way, we aim to improve the semantic robustness of our image-to-image translation model, allowing it to translate details that must be translated. There is one more limitation of our method. When domain images are very specific and were not found in the CLIP training dataset, our method is almost useless. In the example, we are trying to translate GTA labels into Cityscapes photos. We could not prevent semantic flipping in this task. Finally, we suggest exploring the use of multiple CLIP models with different pre-training data and architecture to improve the robustness and accuracy of the generated images. This could lead to even better performance and expand the range of applications for image translation. Possible use of two or three various CLIP models (for example based on ResNET and on visual transformers) to increase the reliability of Siamese network.

7. Conclusions

In this paper, we proposed a novel approach for semantically robust unpaired image translation, CLIPTraVeLGAN. CLIPTraVeLGAN simplifies the Siamese network selection and training process of TraVeLGAN by using a contrastively pretrained language-image model (CLIP) with frozen weights. To our knowledge, we were the first to utilize CLIP to enforce that the individual features and semantic content of the input image are reflected in the generated image during image-to-image translation. The proposed model CLIPTraVeLGAN proved to be much easier to train than the original

TraVeLGAN. The training is quite stable, the results are quite comparable with other models built on the same generators and discriminators. Our results show that CLIPTraVeLGAN outperforms both CycleGAN and TraVeLGAN in terms of semantic robustness while being easier to train and producing comparable results in terms of quality.

The methodology for achieving semantic robustness in image translation typically involves training a model to effectively preserve the high-level semantics between the source and target domains. This can be done using a variety of techniques, including the use of specialized loss functions and Vector Symbolic Architectures [8]. Our results are close to state-of-the-art Vector Symbolic Architectures but our approach is simpler and more straightforward.

There is a trade-off between image quality and semantic robustness. In our experiments, we manually searched for the optimal value of λ to achieve satisfactory results according to both criteria. In future works, the automated estimation of the optimal value of λ during training has the potential to improve the results and stability of training. The proposed model showed promising results in terms of semantic robustness and ease of training and can be used as a starting point for future research on efficient image translation models. Possible future work includes exploring the use of multiple CLIP models and new solutions of generators and discriminators.

8. Acknowledgements

The work was performed as part of the state budget research project “Development of methods and algorithms for combined learning of deep neuro-neo-fuzzy systems under short training set conditions” (state registration number 0122U001701) of Artificial Intelligence Department of Kharkiv National University of Radio Electronics.

9. References

[1] D. Foster, Generative Deep Learning. Teaching Machines to Paint, Write, Compose and Play, O’Reilly Media, Inc., 2019.

[2] J. Langr, V. Bok, GANs in Action. Deep Learning with Generative Adversarial Networks, Manning Publications Co, 2019.

[3] Y. Pang, J. Lin, T. Qin and Z. Chen, "Image-to-Image Translation: Methods and Applications." IEEE Transactions on Multimedia, 24 (2022): 3859-3881. doi: 10.1109/TMM.2021.3109419.

[4] M. Amodio, S. Krishnaswamy, TraVeLGAN: image-to-image translation by transformation vector learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 8983–8992. doi 10.1109/CVPR.2019.00919.

[5] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, Learning transferable visual models from natural language supervision, in: Proceedings of the International Conference on Machine Learning, 2021.

[6] 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, ICCV, Venice, Italy, 2017, pp. 2242-2251. doi: 10.1109/ICCV.2017.244.

[7] Jia Zhiwei et al., Semantically robust unpaired image translation for data with unmatched semantics statistics, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, Montreal, QC, Canada, 2021, pp. 14253-14263. doi: 10.1109/ICCV48922.2021.01401.

[8] Justin D. Theiss et al., Unpaired image translation via vector symbolic architectures, in: Proceedings of the European Conference on Computer Vision, 2022.

[9] Amy Jang, Ana Sofia Uzsoy, Phil Culliton, I’m something of a painter myself, Kaggle, 2020. URL: https://kaggle.com/competitions/gan-getting-started.

[10] Stephan R Richter et al., Playing for data: Ground truth from computer games, in: Proceedings of the European Conference on Computer Vision. Springer. 2016, pp. 102–118.

[11] M. Cordts et al., The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Las Vegas, NV, USA, 2016, pp. 3213-3223, doi: 10.1109/CVPR.2016.350.