Master's Thesis

SyntStereo2Real: Edge aware GAN for remote sensing image translation while maintaining stereo constraints.

Vasudha Venkatesan

Examiner: Prof. Dr. Thomas Brox
Prof Dr. Peter Reinartz

Advisers: Dr. Ksenia Bittner
Mario Fuentes Reyes
Daniel Pananigan

Albert-Ludwigs-University Freiburg
Faculty of Engineering
Department of Computer Science
Chair for Thesis Templates

December 28th, 2023

Writing period

01. 11. 2023 – 25. 12. 2023

Examiner

Prof. Dr.Thomas Brox

Prof. Dr.Peter Reinartz

Advisers

Dr. Ksenia Bittner, Mario Fuentes Reyes, Daniel Pananigan

Declaration

I hereby declare, that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work.
I hereby also declare, that my Thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

Place, Date	Signature

Abstract

In the field of remote sensing, the scarcity of stereo-matched data often hinders the training of deep neural networks. The use of synthetically generated images as an alternative alleviates this difficulty but suffers from the problem of domain generalization. Unifying the capabilities of image-to-image translation and stereo-matching presents an effective solution to address the problem of domain generalization. Current methods involve combining two networks—an unpaired image-to-image translation network and a stereo-matching network—while jointly optimizing them. This work proposes a single edge-aware GAN-based network that effectively tackles both tasks simultaneously. We obtain edge maps of input images from the sobel operator and use it as an additional input to the encoder in the generator to enforce geometric consistency during translation. Additionally, we include a warping loss from translated images to maintain the stereo consistency. This work performs qualitatively and quantitatively better than existing models, and its applicability extends to diverse domains, including autonomous driving.

Zusammenfassung

Zusammenfassung Im Bereich der Fernerkundung wird das Training von tiefen neuronalen Netzen oft durch die Knappheit von Stereodaten behindert. Die Verwendung von synthetisch generierten Bildern als Alternative mildert diese Schwierigkeit, leidet aber unter dem Problem der Domänengeneralisierung. Die Vereinheitlichung der Fähigkeiten der Image-To-Image-Übersetzung und des Stereo-Matchings stellt eine effektive Lösung dar, um das Problem der Domänengeneralisierung zu lösen. Bei den derzeitigen Methoden werden zwei Netzwerke kombiniert - ein ungepaartes Image-To-Image-Übersetzungsnetzwerk und ein stereo matching netzwerk - und gemeinsam optimiert. Wir schlagen vor ein edge-aware-GAN-basiertes Netzwerk vor, das beide Aufgaben effektiv und gleichzeitig bewältigt. Wir erhalten Kantenkarten der Eingabebilder vom Sobel-Operator und verwenden sie als zusätzliche Eingabe für den Encoder im Generator, um geometrische Konsistenz während der Übersetzung zu erzwingen. Wir beziehen zusätzlich einen Warping-Verlust aus übersetzten Bildern ein, um die Stereokonsistenz zu erhalten. Wir zeigen, dass unser Modell qualitativ und quantitativ bessere Ergebnisse liefert als bestehende Modelle und dass seine Anwendbarkeit sich auf verschiedene Bereiche erstreckt, einschließlich des autonomen Fahrens.

Contents

List of Figures

List of Tables

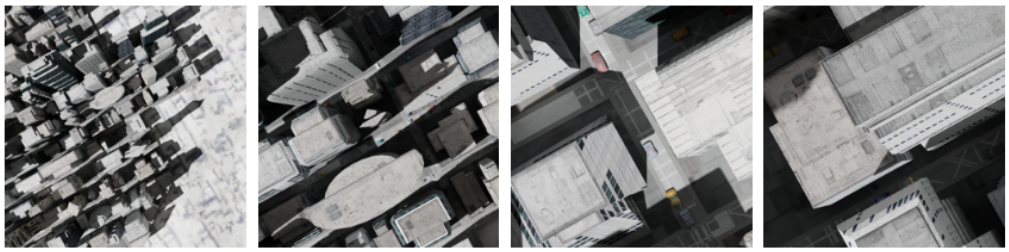
1 Introduction

> Translation is that which transforms
> everything so that nothing changes.
>
> *Günter Grass*

The challenges in obtaining ground truth images in the remote sensing domain stem from the difficulty in capturing matching images due to temporal changes, sparse measurements and a significantly large baseline. Correspondence tasks like disparity estimation or stereo reconstruction for these images, can be both cumbersome and expensive. An additional layer of complexity in obtaining stereo-paired images for aerial applications arises from the use of single cameras along an acquisition line. In this scenario, the challenge extends beyond the initial capture, as the stereo matching process demands meticulous consideration of the acquisition geometry. The singular camera approach, while efficient in terms of data acquisition, introduces the intricacy of orthorectification due to variations in terrain and elevation along the acquisition line. The need for precise orthorectification in such cases is important to maintain the accuracy of the derived 3D information from the stereo-paired images.

The concept of using synthetic data for training deep neural networks arises from the persistent problems posed by data scarcity, privacy concerns, and the overall difficulty in acquiring authentic data. Synthetic data generation allows for the creation of simulated datasets that provide essential ground truth information, including accurate labels and stereo disparity maps. The quality of synthetic data generation techniques and their ability to mimic real-world characteristics are critical factors in determining the success of vision model training using synthetic datasets.

While the synthetic data is obtained from a simulation of real-world scenario, it may not perfectly represent the complexities and variations in real-world data in remote sensing domain due to the subtleties and variations of buildings, bridges and vegetation. This can result in domain shift, where the model struggles to generalize to real world data. Unpaired image-to-image translation algorithms have been used to address the problem of domain shift. They provide promising results to reduce the domain gap between the domains. However they can alter the structural information as



(a) Original

(b) Translated

Figure 1: Aerial images translated using CUT[1]. The model tends to hallucinate when translating images with diverse scenes, where the target distribution is more likely to be unbalanced.

shown in Figure 1. This can pose as a serious challenge when training on downstream tasks such as stereo matching or instance segmentation because the translated images no longer align with their corresponding labels. Our approach focuses on the specific task of translating synthetic images to realistic domain while maintaining the stereo constraints. Some of the existing methods such as StereoGAN [8] have addressed this task for autonomous driving datasets with joint optimization of image translation and disparity estimation networks. Moreover images from remote sensing domain are rich with diverse content. Existing methods suffer from the problem of an increased likelihood of hallucinations and discrepancies to preserve epipolar geometry and fail to perform good quality image translations.

We address this problem using a *light-weight single edge-based GAN network*, that performs unpaired image-to-image translation while maintaining the stereo constraints. At first, the edge maps of input images are obtained from Sobel operator. They are provided as an additional input along with image pairs from both domains to the generator. The encoder of the generator computes the content and edge code separately from the input image and its edge map. The content code is added together with the edge code as *content edge code*. The content edge code is provided to the

decoder along with a random style to generate images of different domain as shown in Figure 9. The style code is generated randomly from a normal distribution for each domain and is maintained as a constant throughout the training. The use of edge maps ensures that the structure of the image is retained and not lost in translation and thus helps in matching to calculate the disparity map. Additionally, we use warping loss, where we warp the left translated image with its respective disparity map and compare it to the right translated image to enforce stereo constraints. Extensive experiments across multiple datasets demonstrate our method outperforms the existing methods quantitatively and qualitatively. Moreover, we use a single light-weight network to perform optimization on two tasks without the use of any pre-trained networks.

To sum up, our main contributions are:

- Developing a framework for *image-to-image translation of stereo pairs* considering a consistent translation of left and right images that preserves the matching.
- Employing *edge maps* in the generator to retain essential geometric content and enhance the preservation of sharp boundaries within the translated images.
- *Incorporating warping loss* to enforce stereo consistency without the need for an additional stereo matching network.

Figure 2: Examples of aerial scene translated by SyntStereo2Real. Our model can produce semantically consistent realistic translations.

2 Related Work

2.1 Unpaired image-to-image translation

The concept of image-to-image translation was first introduced by Hertzmann et al[9] called Image analogies, which used non parametric texture model to generate a new image that is analogous to a given image based on the analogy with another pair of images. More recent approaches use a parametric translation function approach using CNN such as pix2pix[10], where the mapping function is learned for the input-output image pair. But these approaches required image pairs in the input and output domain. On the other hand, unpaired image data is more abundant and accessible than paired data. In many real-world scenarios, obtaining a large and diverse set of paired images with corresponding translations or annotations is impractical or expensive. CycleGAN [2] has been a pioneer in solving this task by identifying the key mappings in unpaired data from two different domains. The authors introduced cycle consistency loss to constrain the one-to-one mapping space by reconstructing the original image back from the translated image. This loss, in conjunction with adversarial loss and identity loss, plays a pivotal role in image-to-image translation, leading to remarkable visual results. The CUT [1] model extends this concept for one-sided image translation with a contrastive loss. It is calculated using negative samples obtained from the same input, thus enabling faster training. However, the addition of contrastive loss function does not translate well to images from certain domains such as remote sensing as shown previously in Figure 1. Satellite images are typically high-resolution, which leads to a large number of features learned by the model, thereby inducing the problem of hallucination. UNIT [11] carries out unsupervised image-to-image translation under the assumption that images from both domains consist of a shared latent space. The model uses weight sharing between the layers of generators and discriminators to learn the joint distribution of data. MUNIT [12] extends this architecture to handle multiple styles using the disentanglement principle to obtain content and style code separately. The content code from the image is combined with a random style code from cross-domains to

obtain diverse styled images.

2.2 Stereo matching

Semi-global matching (SGM) [13], is a classical stereo method that uses pixel-wise matching cost for computing the disparities between two images. It produces an approximate global optimal solution and is still one of the best performing classical techniques for disparity estimation in certain domains with the advantage of an efficient implementation. MC-CNN [14] introduced disparity estimation techniques based on convolutional neural networks where pairs of small image patches are compared to initialise the matching cost. A number of post processing steps are applied to the matching cost which includes cross-based cost aggregation and semiglobal matching refinement, followed by a left-right consistency check to eliminate errors in the occluded regions. DispNet [6] is one of the pioneering networks that involves direct estimation of disparity maps. It includes a 1D correlation layer which is used to estimate a cost volume and then is refined using subsequent convolutional networks for accurate disparity estimation. PSMNet [15] introduces spatial pyramid pooling (SPP) to estimate cost volume at different scales of the image and a stacked hourglass 3D CNN to process the cost volume.

2.3 Domain adaptation

The task of translating synthetic images to realistic has been an active research topic with multiple applications such as semantic segmentation, stereo matching and pseudo label learning. StereoGAN [8] is specifically designed for the task of translating synthetic images to realistic domain while maintaining the stereo constraints. It utilizes a CycleGAN for image translation and a DispNet [6] for disparity estimation. Secogan [16] utilized content disentanglement architecture from MUNIT for translating synthetic images of autonomous driving datasets to realistic domain. SDA [17] utilizes the spatial feature transform to fuse features of edge maps with source images. The authors use CycleGAN for unapaired image translation along with warping loss to enforce the stereo matching.

The task of translating images to a realistic style while maintaining the content structure for stereo matching is a dual optimization task. Although the existing networks address this problem, they suffer when applied to remote sensing images due to large disparity values, seasonal effects and temporary objects. The models

developed are predominantly applied in the field of autonomous driving and struggle in achieving domain generalisation. Another challenge is the training of existing models tends to become computationally expensive, as it is a combination of two deep learning networks, one for image translation and the latter for stereo matching. The number of parameters required for training is high and can slow the training process. We address both of the above concerns in our work by employing a single edge-based image translation GAN model trained additionally with warping loss to enforce the stereo constraints.

3 Background

3.1 GANs for Image to Image translation

Some of the prominent methods used for domain adaptation are using GAN-based models. The initial works on image-to-image translation using GANs were carried out by Shrivastava et al [18] for generating highly realistic images for gaze and hand pose estimation. They used the SimGAN architecture where synthetic images from the simulator were enhanced using a refiner network (similar to a generator) and a discriminator to distinguish real from generated images.

GANs consists of two networks: the generator and the discriminator. The job of the generator is to create new examples, while the discriminator aims to distinguish between real and generated examples. They go back and forth in a competitive process. The generator gets better at creating realistic examples, and the discriminator gets better at telling real from fake. The training continues until the generator can produce realistic examples which that cannot be labelled as fake by the discriminator.

Much recent work on domain adaptation is based on pix2pix [10] architecture for paired image translation and CycleGAN [2] architecture for unpaired image translation. The challenge in obtaining accurate remote sensing data for paired translation of synthetic images makes pix2pix network infeasible to apply image translation.

3.1.1 CycleGAN

CycleGAN consists of a pair of generator and discriminator for each domain as shown in Figure 3. In order to learn the translation from one domain to other domain images, the generator creates images from the first domain that match the output distribution of the other domain. The task of the discriminator is similar to other GANs where it distinguishes the translated image from the original image. To prevent mode collapse during adversarial training, where all images map to the same output image, a cycle consistency loss is added along with adversarial loss during training. This loss ensures that the translated image retains the original structure of the image

and modifies only the style.

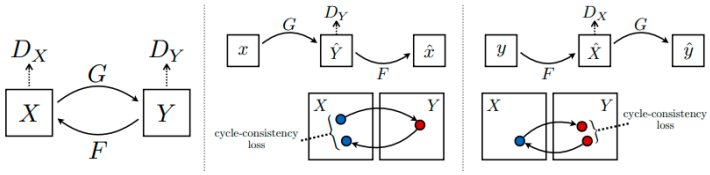


Figure 3: (a) CycleGAN model architecture[2]. The model consists of two mapping functions G and F, $G : X \rightarrow Y$, F: $Y \rightarrow X$ and its associated discriminators to distinguish real images from translated images. Cycle consistency ensures that the same image is be obtained on translating one domain to other and the generated image is translated back to first domain.

3.1.2 PatchGAN based discriminator

PatchGAN based discriminator is a type of discriminator that has been used in the image generation tasks extensively. It was introduced by Isola et al[19], for image to image translation tasks. The PatchGAN based discriminator learns to classify whether each $N \times N$ patch in an image is real or fake. This is done by passing the discriminator convolutionally across the image and averaging all responses from patches to provide the final output of D. By only looking at local image patches, PatchGAN can capture textures or styles of an image, rather than just its overall structure. Through its emphasis on local patches instead of the broader structure, PatchGAN has the capability to generate images with better detailed textures and styles. PatchGAN exhibits increased robustness to image distortions and inconsistencies by concentrating on small, localized regions rather than the complete image. It is also efficient to use on large datasets as it assumes independence among pixels and thus minimizing the computational workload.

3.2 Autoencoder

Autoencoder is a type of neural network which is used to obtain efficient encoding in a latent space. The autoencoder processes input data through the encoder function, denoted as f, to obtain a compressed representation in a lower-dimensional latent space. The compressed image is then reconstructed using the decoder function, represented as g, with the aim of faithfully reproducing the original input.

$$h = f(X), \tilde{X} = g(h) \tag{1}$$

Theoretically, the weight W_f is a pseudo inverse of the weight W_g . It is important to note that the functions f and g are typically nonlinear, allowing for more expressive mappings.

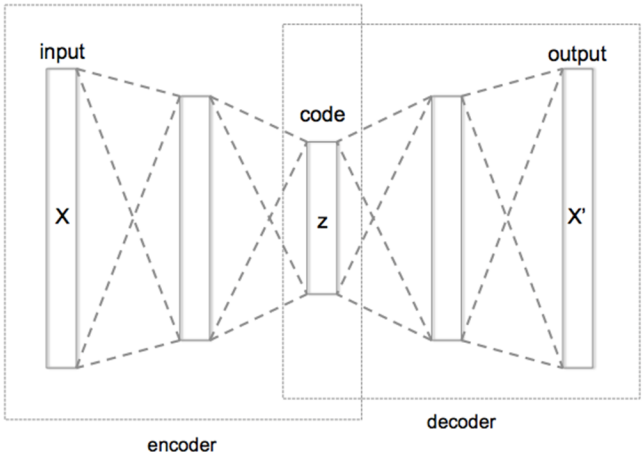


Figure 4: Architecture of autoencoder. The input X is compressed using a function f, encoder to project it to a lower dimension and is then reconstructed to the original data using function g, decoder. [3].

3.3 Adaptive instance normalisation

Adaptive Instance Normalization (AdaIN)[20] is a technique used in neural networks for image style transfer and other tasks where adjusting the visual appearance of an image is important. It builds upon the concept of Instance Normalization but introduces a more flexible and adaptable approach.

In Instance Normalization, you normalize the input data to a specific style using parameters that adjust the mean and variance of each channel. AdaIN takes it a step further by allowing you to specify both content and style inputs. AdaIN aligns the mean and variance of the content input to match those of the style input. This is done separately for each channel, allowing AdaIN to capture detailed style characteristics.

In equation 2, z refers to the activation of the output from encoder, γ and β are the style code parameters of the target domain.

$$AdaIN(z, \gamma, \beta) = \gamma \left(\frac{z - \mu(z)}{\sigma(z)} \right) + \beta \tag{2}$$

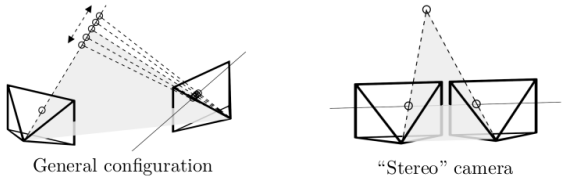


Figure 5: In general configuration of two pinhole cameras, varying the 3D position of a point along its projection ray in one camera leads to pixel motion along a different line in the other camera. In stereo camera, a point projects to same scan-line in both cameras. The gray triangle represents the epipolar plane in which the 3D point, its projections and the camera centers are co-planar [4].

3.4 Stereo Matching

The task of stereo matching is to find corresponding points or features in two images taken from different viewpoints. To find the corresponding points, we need to enforce the epipolar geometry constraint that guarantees that a point seen on one camera's view projects onto a line in another camera. Since the images are taken in a calibrated setup, the epipolar line is known. The epipolar constraint restricts the disparity estimation into a 1D search problem as shown in Figure 5. Disparity is the signed distance between images of the same 3D point in two views. Disparity estimation can be considered as a special case of optical flow, since it is a scalar quantity rather than vector field.

The disparity value is directly related to the depth value in the pixel. It is given by the equation 3. Focal length is an attribute of camera and the baseline is the distance between the two camera projection centers. Since the images are captured after camera calibration both the values are known, and hence the depth values can be estimated. This can help in creating 3D models using stereo reconstruction from disparity values.

$$depth = \frac{focallength \cdot baseline}{disparity} \tag{3}$$

Stereo matching for high-resolution satellite images or remote sensing data has been an active research topic in the field of photogrammetry and remote sensing. Generally, stereo matching consists of four steps : matching cost computation, cost aggregation, disparity computation and optimization, and disparities refinement [21]. The methods to solve stereo matching problem can be classified into three categories : variational methods, combinatorial optimization and deep learning based approaches.

Variational methods formulate the disparity estimation as an energy minimization

problem. The goal is to find a set of flow vectors that minimize the energy function. Variational methods typically assume that disparities vary smoothly within a local neighborhood. This assumption is reasonable for small disparities, where neighboring pixels are likely to have similar disparities. However, for larger disparities, this assumption can break down, leading to inaccurate results. Larger disparities are preferred for disparity estimation as they can lead to accurate estimation of height of buildings for 3D model. Due to this reason, combinatorial approaches are preferred for the task of disparity estimation. The most common combinatorial technique is the Semi-global matching technique (SGM) by Hirschmüller [13].

3.4.1 Semi-global Matching

Semi-global matching uses pixel-wise matching cost for computing the disparities between two images. Thanks to the stereo rectification, the corresponding matching pixel for each pixel needs to be looked upon only on the same row in the second image. The matching cost is defined by a unary cost function $\theta(x_p)$, where each pixel p is matched with the pixel that looks closest to it in the other image. A pairwise cost $\theta_{pq}(x_p, x_q)$ is added as a smoothness constraint, where p is the matching pixel and q are the neighboring pixels in the same row as shown in Equation 4. This leads to approximations only along the x-axis and no smoothness in the y-direction as shown in Figure 6. Thus in SGM, the author proposes to consider not only horizontal direction but lines in multiple directions. Applying this idea leads to quite an accurate estimation of disparity values for each pixel. This method does not provide a globally optimal solution (it is an NP-Hard problem), but a good approximate solution for disparity estimation. SGM is still one of the best-performing classical stereo method and is efficient for real-time implementations.

$$E(x) = \sum_{p} \theta(x_p) + \sum_{p,q \in \mathcal{N}} \theta(x_p, x_q) \tag{4}$$

Figure 6: (a) Disparity estimated with dynamic programming along x-axis (b) Disparity estimated using SGM (c) Ground truth disparity [5].

3.4.2 Deep learning based techniques

Deep learning based approaches have also been implemented to solve the problem of disparity estimation. Siamese network is one of the initial methods to perform patch based disparity estimation. The network takes input of two image patches from the left and right images, and use convolutional layers to compute feature representation. The network then consists of fully connected layers on top to calculate the similarity scores. The network is trained to compute the similarity between left and right patch. This method is slow and the matching score is not great for all resolutions, because score is not calculated for different disparity ranges.

DispNet

DispNet is the first neural network to process stereo camera images as a complete unit, and to predict a dense disparity map in a single inference step. The model follows a similar architecture to FlowNet[22] with the addition of only two structural changes: a correlation layer and additional convolutional layers in between upsampling layers. FlowNet consists a contractive part and an expanding part with long-range links between them. The contracting path captures fine-grained details through convolutional and pooling layers, while the expanding path recovers spatial information and synthesizes a holistic understanding of the image. This dual pathway enables the network to seamlessly integrate local features and context, enhancing its ability to comprehend and interpret global structures within the image. The correlation layer in DispNetC (adapted from the FlowNet-C network) computes the scalar product between a feature vector from the first image and a 2D region of feature vectors in the second image, spread around the location of the first image's reference feature. With the extra convolutions, the network can produce outputs with significantly reduced visible artifacts. Predicted disparity maps are subpixel accurate.

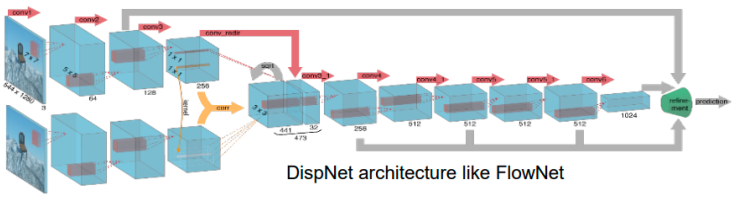


Figure 7: DispNet architecture. [6].

3.4.3 AANet

In this work, we use AANet to evaluate the performance of stereo matching of our network with other existing works[7]. In Figure 8, the Adaptive Aggregation Network (AANet) architecture is presented. The process begins with a stereo pair, from which a downsampled feature pyramid is extracted at resolutions of 1/3, 1/6, and 1/12 using a shared feature extractor. Following this, multi-scale cost volumes are generated by correlating left and right features at corresponding scales. The raw cost volumes undergo aggregation through six stacked Adaptive Aggregation Modules (AAModules). Each AAModule is composed of three Intra-Scale Aggregation (ISA, as detailed in Sec. 3.1) modules and a Cross-Scale Aggregation (CSA, as explained in Sec. 3.2) module, catering to three pyramid levels. Subsequently, multi-scale disparity predictions are regressed. It is noteworthy that dashed arrows serve a specific role during training and can be omitted during inference. Finally, the disparity prediction at 1/3 resolution undergoes hierarchical upsampling and refinement to achieve the original resolution.

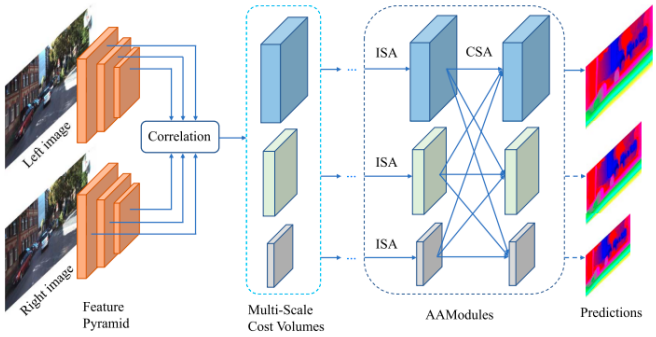


Figure 8: AANet architecture. [7].

4 Approach

For correspondence tasks like disparity estimation or stereo reconstruction in the field of remote sensing, obtaining aerial or satellite images of cities using a stereo-rectified camera can be cumbersome and expensive. Some stereo matching for aerial is done with single cameras but along an acquisition line, which adds the difficulty of orthorectification. The idea of using synthetic data for training deep neural networks is motivated by data scarcity, privacy concerns, avoiding manual annotation costs, and difficulty in obtaining data. While the synthetic data consists of the simulation of real-world scenario, it may not perfectly represent the complexities and variations in real-world data. This can result in domain shift, where the model struggles to generalize from synthetic data to real-world data. Image-to-image translation algorithms have been used to address the problem of domain shift. Our approach focuses on this problem, if we can translate a photo to an artistic-styled Van Gogh painting using image-to-image translation algorithms, why not use it for translating the synthetically generated images to real-world data for models to learn from? Some of the existing methods have addressed this task for autonomous driving datasets (Driving[6] and SYNTHIA[23] for synthetic datasets, KITTI2012[24] and KITTI2015[25] for real-world datasets) using GAN based approaches[8][17]. Although these methods produce promising results, the translations of images from remote sensing domain are poor due to large disparity values and occlusions as shown previously in Figure 1. We carry out the translation of synthetic to realistic domain images under the assumption that both domains share universal features that describe the elements in the scene (such as buildings, roads, vegetation), as well as distinctive features specific to the particular domain, focusing on visual attributes like appearance or style.

Given a synthetic *left-right-disparity* tuple $(xc_l, xc_r, x_d)_a \in \mathcal{X}_a$ denoting the stereo pair of left and right image with its corresponding disparity for source domain, a real image $xc_b \in \mathcal{X}_b$ representing the target domain, and two randomly sampled style codes s_a, s_b for each domain, our model synthesizes a realistic stereo matched pair of the synthetic image.

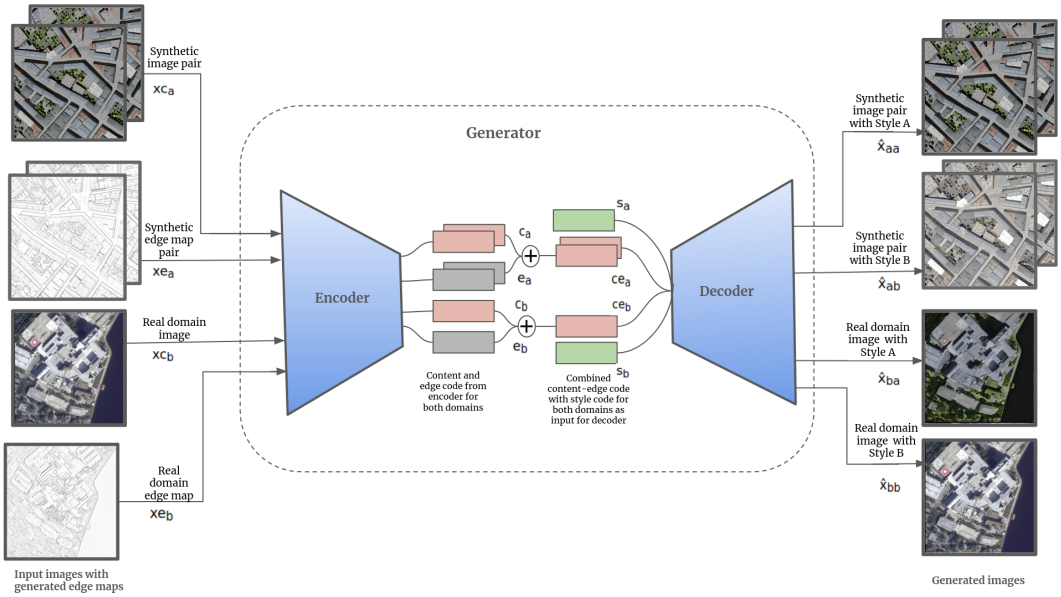


Figure 9: Illustration of the generator architecture in an autoencoder with edge map integration. The image along with its corresponding edge map is encoded and added together as content edge code before applying it as an input to the decoder. The decoder merges the content-edge code with style code from every domain to generate content that is contextually fitting. xc_a, xc_b represents the input images from both domains (content), xe_a, xe_b represents the corresponding edge maps. c_a, c_b, e_a, e_b represents the content and edge code from encoder for both domains. s_a, s_b are the randomly initialized style code before the training. x_{aa}, x_{ab}, x_{ba}, x_{bb} represents the respective output images from the decoder.

Our work draws inspiration from MUNIT [12] and Secogan [16] to learn disentangled representations from two domains without supervision. Similar to [16], our translation model consists of an autoencoder (encoder E and decoder G) as a generator for both domains. The encoder factorizes each input into latent content code $c_i(i = a, b)$, where $c_i = E(xc_i)$. Style code is initialized before the training using normal distribution as $s_i = (\gamma_i, \beta_i)$ for each domain and remains constant during the training. Edge maps of the corresponding input images are obtained from the Sobel operator $xe_i = SO(xc_i)$ and are given as additional input to preserve structural information. The encoder generates the latent edge code $e_i = E(xe_i)$ from the edge maps. The edge code is added to the content code as content-edge code $ce_i = c_i + e_i$ and is provided as an input to the decoder as shown in Figure 9. The decoder generates the output image by swapping the content and style codes. The discriminator distinguishes the original image to the generated image by adversarial training. Since we have a real

and synthetic domain, we have two discriminators D_A and D_B.

Multiple losses help in constraining and generating images in a meaningful manner in GAN based networks. Figure 10 shows an overview of the losses used in the training of the model. A reconstruction loss

$$\mathcal{L}_{rec}^{aa}(E, G) = E_{x_a \sim X_a} \| G(E(x_a), s_a) - x_a \|_1 \tag{5}$$

ensures that the model generates accurate reconstruction of images after content disentanglement.

In image-to-image translation, it is essential that the generated images in the target domain are not only realistic but also faithfully represent the original content. Cycle consistency loss [2],

$$\mathcal{L}_{cycle}^{aba}(E, G) = E_{x_a \sim X_a} \| G(E(x_{ab}), s_a) - x_a \|_1 \tag{6}$$

enforces this constraint by calculating the loss between original image and the transformation of original image to another domain (x_{ab}), and transform it back again to original domain (x_{aba}).

Since we use a GAN based approach to train the model, we use an adversarial loss

$$\mathcal{L}_{adv}^{a}(E, G, D_a) = E_{x_a \sim X_a} log D_a(p(x_a)) + E_{x_b \sim X_b} log(1 - D_a(p(x_{ba})) \tag{7}$$

to match the data distribution of translated images to the distribution of target domain. The adversarial loss is employed by both the discriminator and generator, whereas the other mentioned loss exclusively guides the training of the generator. Since we use a patch based discriminator, the p in Equation 7 refers to random patches of image.

Considering the images from one domain are synthetically generated, we assume to have access to additional information like ground truth labels, disparity maps, and segmentation masks. Warping loss as an additional constraint can be a useful addition, especially in tasks where the images are later used for training disparity estimation models. We compute the warping loss

$$\begin{aligned} \mathcal{L}_{\text{warp}} = \lambda_4 \cdot \mathcal{L}_1 (& G(E(x_{r_a}), s_b) - W(G(E(x_{l_a}), s_b), x_d) \\ & + \lambda_5 \cdot (1 - \text{SSIM}(G(E(x_{r_a}), s_b) - W(G(E(x_{l_a}), s_b), x_d))) \end{aligned} \tag{8}$$

by comparing the warped left image $W(x_{l_{ab}}, x_d)$ (which has undergone translation)

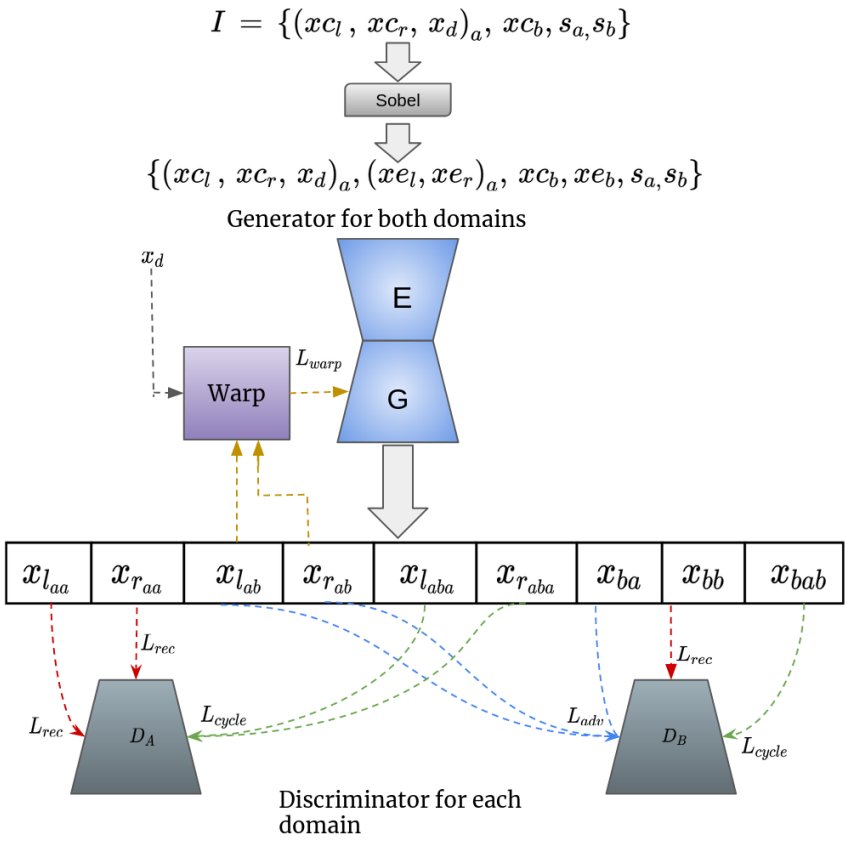


Figure 10: Illustration of the GAN-based model architecture featuring multiple loss functions. The design incorporates a combination of adversarial, reconstruction, cycle and warping losses. Adversarial loss promotes realistic image generation, while reconstruction loss ensures faithful reproduction of input data, cycle loss enforces the correct mapping between domains and warping loss enforces geometrical stereo constraints.

and the right image after translation $x_{r_{ab}}$. We use a combination of $\mathcal{L}_1$ loss and SSIM loss for calculating the warping loss.

The corresponding losses from other domain $\mathcal{L}_{rec}^{bb}$, $\mathcal{L}_{cycle}^{bab}$ and $\mathcal{L}_{adv}^{b}$ are calculated in a similar manner. Therefore, the overall loss function for the generator is given by

$$\min_{E,G} \max_{D_a,D_b} \mathcal{L}(E, G, D_a, D_b) = \lambda_1 \cdot (\mathcal{L}_{rec}^{aa} + \mathcal{L}_{rec}^{bb}) + \lambda_2 \cdot (\mathcal{L}_{cyc}^{aba} + \mathcal{L}_{cyc}^{bab}) + \lambda_3 \cdot (\mathcal{L}_{adv}^{a} + \mathcal{L}_{adv}^{b}) + \mathcal{L}_{warp} \tag{9}$$

5 Experiments

5.1 Network architecture

We adopt the architecture of secogan with one generator and two discriminators. The autoencoder, with a pair of encoder and decoder for generator is based on MUNIT architecture[12]. The discriminators are implemented using PatchGAN architecture. The input to the network consists of images from two domains and their corresponding edge map and the output consists of translated images with the style from other domain.

5.2 Datasets

We use two sets of datasets from different application areas to study the generalisability of our model architecture. For remote sensing data, we use SyntCities dataset for synthetic data and Urban semantic 3D dataset for real domain data[26][27].

For autonomous driving data, we use the Driving dataset from Sceneflow for synthetic domain and KITTI(2015) dataset in real domain[6][25].

5.2.1 Syntcities

Syntcitites is a large dataset set consisting of synthetically generated images of remote sensing imagery. It is specially developed to train deep learning networks for disparity estimation. It consists of 8100 pairs of images resembling three cities : New York, Paris and Venice. Large city models were generated using CityEngine, a software to build cities in 3D environment. The models were then refined using Blender software based on different illumination conditions, camera properties and reflection properties. RGB images were then rendered from these models along with their corresponding depth and segmentation maps. Disparity maps can be obtained from depth maps using the equation 3. New York dataset consists of images with tall buildings that help us study various lighting conditions such as shadows and reflection, where as Paris and Venice consists of more street views and city maps which help us understand

traffic and urban data. The size of the images is of 1024×1024. We use 4000 sets of images taken evenly from all the three cities for training.

5.2.2 Urban 3D semantic dataset

Urban Semantic 3D Dataset is a large-scale public data set hosted on IEEE DataPort[1]. It consists of more than 320GB of data for training and evaluation of urban areas based on some of the states in the USA. It offers multiview satellite images, airborne lidar data for estimating digital surface models (DSMs) and semantic labels for important features in the urban data. Since we primarily focus on urban areas for image translation, we filter images which consists of atleast 15% of area as buildings and not completely vegetation based on the label map. We obtain 1683 images each of 1024×1024 size for training after applying this filter.

5.2.3 Sceneflow

We use the Driving dataset from Sceneflow as synthetic data for training. Driving consists of 4,400 images that decribe a virtual environment simulating car driving scenarios. It encompasses fast and slow sequences of images, comprising scenes of both forward and backward driving directions with accurate disparity maps. Each image is of 540×960 size and we use the complete dataset for training the network.

5.2.4 KITTI 2015

KITTI 2015 dataset is a subset of the KITTI Vision Benchmark Suite, specifically designed for evaluating stereo vision and optical flow algorithms. KITTI is a widely used benchmark in computer vision, particularly for tasks related to autonomous driving. The KITTI 2015 dataset focuses on stereo and optical flow challenges and provides ground truth annotations for evaluation purposes. The dataset includes image pairs captured by stereo cameras, consisting of left and right images. For stereo vision tasks, the dataset includes pixel-level annotations for disparities, representing the perceived depth in each pixel.

We use 1000 tuples of images taken evenly from Syntcities dataset from all the three cities for training. In US3D dataset, a significant portion of the images primarily consists of vegetation with limited urban content. To address this, we filtered images

[1]The authors would like to thank the Johns Hopkins University Applied Physics Laboratory and IARPA for providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for organizing the Data Fusion Contest.

based on label data, retaining only those images that contain a minimum of 15% building-related content. We randomly selected 1000 samples each of size 1024×1024 for training.

We use the complete dataset from Driving consisting of 4400 images of size 540×960 and the 160 training images each of size 385×1242 provided by KITTI2015 benchmark. We resize the images of 512×512 for remote sensing dataset and 256×512 for autonomous driving dataset during training due to memory and time constraints.

5.3 Loss functions

The adversarial loss L_{adv} aims to train the discriminator to distinguish between real and fake samples. Using L2 loss in this context helps ensure that the discriminator assigns higher scores to real samples and lower scores to fake samples. The squared error L2 loss is a common choice for this task as it produces smoother gradients which can lead to more stable training and avoid vanishing gradients.

In case of cycle loss L_{cycle}, we measure the difference between the original image and the image that has been created through the generator. It is the same case with recreation loss L_{rec}. Using L1 loss here encourages the generator to produce images that are close in pixel-wise similarity to the original. It performs well on sparse data, and can help preserve fine details in the generated images, which is crucial for tasks like image-to-image translation. Using L2 loss for cycle-consistency might result in overly smooth images, as it tends to blur details due to its sensitivity to outliers.

For warping loss L_{warp}, we use a combination of L1 loss and SSIM as given in equation 8[28]. Holes or gaps, can occur as a result of warping when a pixel in the transformed image does not have a corresponding pixel in the original image. We avoid calculating the loss in those regions, by masking those areas. This helps in reducing error while calculating the loss. SSIM computes the perceptual distance between the translated image and its ground truth. It assesses the similarity between two images based on their luminance, contrast, and structure. This can be helpful to measure the quality of warping.

5.4 Training

The network model is implemented using Pytorch [29] and the training is carried out for 100 epochs with a batch size of 4. The hyperparameter values for λ_1, λ_2, λ_3, λ_4 and λ_5 in 8 and 9 are set to 10,10,1,1 and 0.8 respectively. We use stochastic mini

batch gradient descent with Adam optimizer [30]. Beta coefficients of Adam are set to 0.5 and 0.999 respectively.

5.5 Evaluation metrics

We compare the two models based on three criteria : performance of stereo matching, performance of unpaired image-to-image translation, and the number of learnable parameters required to train the model.

We acquire translated images and assess their performance on disparity estimation by training them on a disparity network. Specifically, we employ AANet [7] for the training and evaluation of estimation. For the case of SyntCities to US3D we trained for 400 epochs and for Driving to KITTI 2015 for 120 epochs, as this is a larger dataset. In both cases we used a batch size of 20 and the maximum disparity was set to 192. To evaluate the predicted disparity maps, we removed the areas where the ground truth is not defined. 60 samples from US3D are used for testing and 40 for KITTI 2015 (these samples were not included in the GAN training). The cases where the original data (before translation) is taken as input is named as Inference.

Given the scarcity of models specializing in synthetic-to-real domain adaptation with stereo constraints, we conduct a comparative analysis of our model against StereoGAN. We use MAD (Median Absolute Deviation) [31], 3px accuracy percentage and 1px accuracy percentage for evaluation of stereo matching. MAD is a robust statistic, being resilient to outliers in a dataset compared to standard deviation because it is calculated by obtaining the median of the absolute difference of pixels and not the squared mean as in standard deviation. 3px accuracy represents the percentage of pixels in the disparity map for which the estimated disparity is within a range of ± 3 pixels from the ground truth disparity and 1px refers to the same metric but for a 1 pixel range. The results are given in Table 1. To evaluate the performance of image translation, we compute the FID [32] score to calculate the similarity between distributions of feature vectors for two separate datasets of image as given in Table 2.

5.6 Quantitative Results

As indicated in Table 1, our approach demonstrates enhancements, showcasing a notable improvement with respect to StereoGAN of +3.14% in 3px accuracy and +2.30% in 1px accuracy for remote sensing images. Additionally, the model exhibits improvements of +1.727% in 3px accuracy and +1.621% in 1px accuracy for

Datasets	Metrics	Inference	StereoGAN	SyntStereo2Real(ours)
SyntCities to US3D	MAD ↓	1.801	1.520	**1.319**
	3px-acc% ↑	63.097	66.765	**69.906**
	1px-acc% ↑	30.790	33.619	**35.928**
Driving to KITTI	MAD ↓	0.721	0.626	**0.575**
	3px-acc% ↑	88.871	89.646	**91.373**
	1px-acc% ↑	61.832	64.271	**65.892**

Table 1: Comparison of metrics for SyntCities to US3D and Driving to KITTI. The table illustrates the performance across datasets, showcasing results for the original synthetic dataset (Inference), StereoGAN, and SyntStereo2Real(ours). Bold values highlight superior performance in MAD reduction and accuracy enhancement.

autonomous driving datasets. Please note that the ground truth in the KITTI dataset is sparse and can not be evaluated for all the pixels. Despite that, we can visually compare the reconstruction capabilities for not labelled pixels. The disparity maps illustrated in Figure 12 and Figure 13 highlight a more complete prediction without empty regions. FID in Table 2 shows the significant difference in the quality of image translation compared to StereoGAN as the value is significantly lower indicating the similarity of the nature of data. Comparing the number of parameters in Table 3, our model has a significantly smaller number of learnable parameters for training, making it ideal for applications with limited storage and processing capabilities.

Datasets	Models	FID ↓
SyntCities to US3D	StereoGAN	188.913
	SyntStereo2Real(ours)	**152.863**
Driving to KITTI	StereoGAN	188.112
	SyntStereo2Real(ours)	**154.055**

Table 2: Frechet Inception Distance (FID) Comparison between StereoGAN and SyntStereo2Real (ours).

Model	n_{params}
StereoGAN	54M
SyntStereo2Real(ours)	11M

Table 3: Comparison of the number of learnable parameters to train model between StereoGAN and SyntStereo2real models.

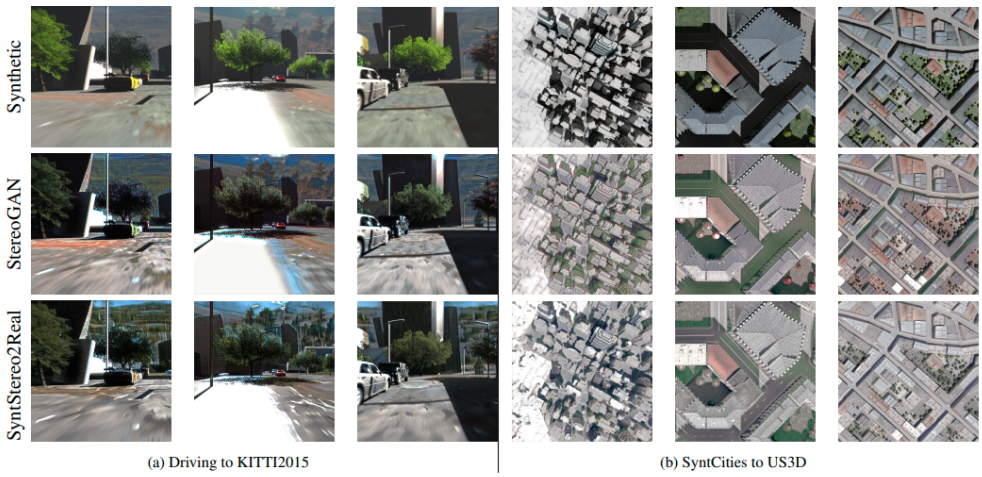


Figure 11: Comparison of image translations: The first row showcases original synthetic images, the second row presents images translated using StereoGAN, and the third row exhibits images translated using our SyntStereo2Real.

5.7 Qualitative Results

Figure 11 shows the results of translation of synthetic images using StereoGAN and our network SyntStereo2Real. The main challenge in translating in remote sensing images is maintaining the structural information for all resolution of images. Our model effectively captures and reproduces the content such as architectural details of building rooftops, bridges and roads. StereoGAN, while proficient in certain aspects of disparity estimation, fails in the translation of shadows by hallucinating green patches instead of building shadows. We can also notice StereoGAN generates small colorful artifacts on the generated images as shown in Figure 11. Our method shows consistent prediction of disparity maps for complete objects without empty gaps or unclear boundaries.

5.8 Ablation Studies

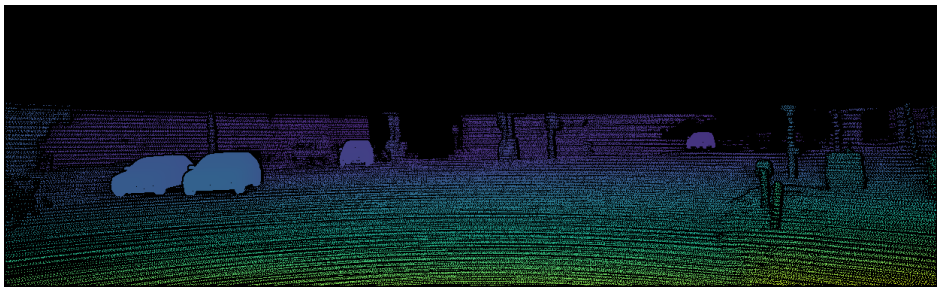
In the Table 4, various configurations of the model are evaluated based on the presence or absence of edge information and warping loss for disparity. Firstly, the inclusion of edge information results in a decrease in the Mean Absolute Deviation (MAD), indicating improved results in predicting deviations from the ground truth. This decrease, coupled with a corresponding increase in both 3px accuracy and 1px accuracy indicates the importance of addition of edge maps. Similarly, addition of

warping loss helps in improving the accuracy and MAD of the model significantly. Thus the the ablation study demonstrates that incorporating both edge information and disparity significantly improves the model's performance across all evaluated metrics for the used datasets.

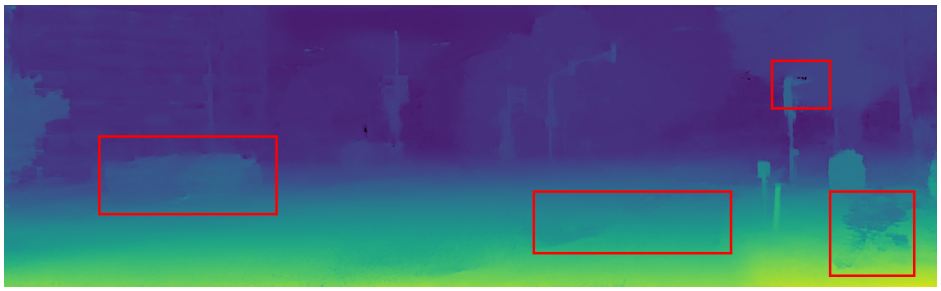
Metrics	**No Edge and No Disp**	**With Edge**	**With Disp**	**With Edge and Disp**
MAD ↓	1.779	1.755	1.646	**1.319**
3px-acc% ↑	62.670	62.887	63.847	**69.906**
1px-acc% ↑	31.503	31.853	32.600	**35.929**

Table 4: Ablation studies. Here the *Edge* refers to the addition of edge information along with input image and *Disp* refers to the additional use of warping loss to enforce disparity constraints.

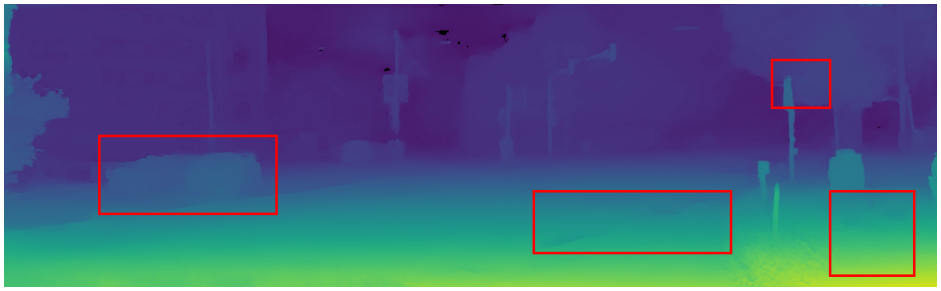
(a) Reference



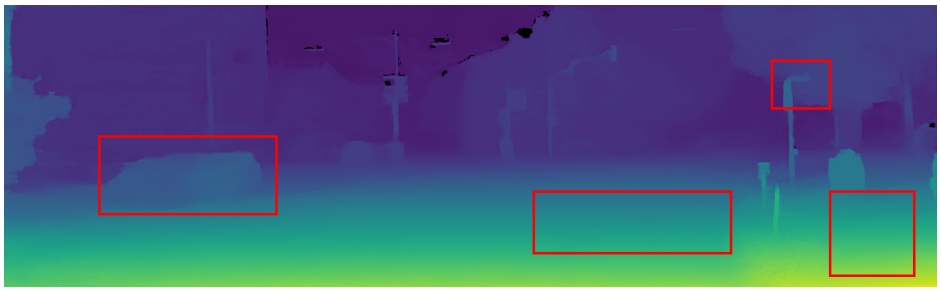
(b) Ground Truth



(c) Inference

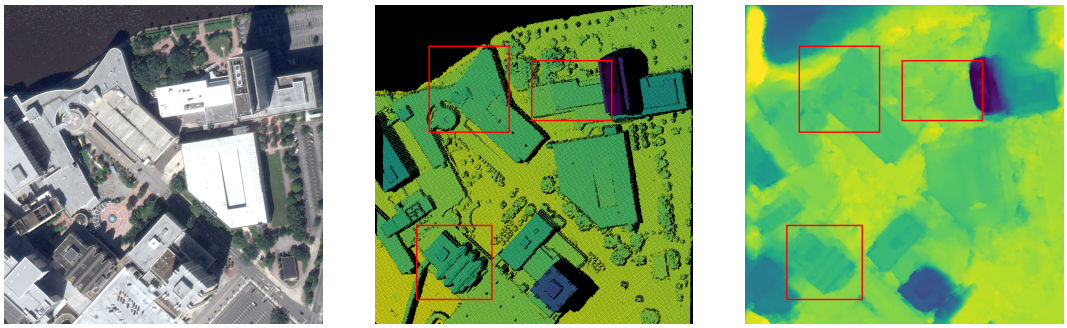


(d) StereoGAN



(e) SyntStereo2Real(ours)

Figure 12: Results of disparity estimation from the AANet for the KITTI 2015 dataset. Three models are computed for the image shown in (a) RGB reference image, (b) Ground truth, (c) Model trained on Driving (Inference), (d) Model trained on Driving translated using StereoGAN (e) Model trained on Driving translated using SyntStereo2Real(ours).



(a) Reference **(b)** Ground Truth **(c)** Inference

(d) StereoGAN **(e)** SyntStereo2Real(ours)

Figure 13: Results of disparity estimation from the AANet for the US3D dataset. Three models are computed for the image shown in (a) RGB reference image, (b) Ground truth, (c) Model trained on SyntCities (Inference), (d) Model trained on SyntCities translated using StereoGAN (e) Model trained on SyntCities translated using SyntStereo2Real(ours).

6 Conclusion

This work introduces a novel, lightweight Generative Adversarial Network (GAN) model tailored for unpaired image-to-image translation from synthetic to real data, while emphasizing adherence to stereo constraints. Unlike traditional approaches, the proposed model incorporates the significance of edge maps and integrates a warping loss into the translation process. By leveraging edge maps, the model preserves structural information during translation, aiding in reducing hallucination and enhancing the shadowing effects. The inclusion of a warping loss ensures accuracy in the estimation of disparities, crucial for maintaining the integrity of translated images in the stereo context. The experimental results showcase the model's state-of-the-art performance in single synthetic-to-real image translation networks, demonstrating its potential to contribute to 3D reconstruction tasks with reduced domain gap dependence.

Moreover, here we highlight the broader applicability of the proposed model by emphasizing its flexibility across different domains. The method not only improves the completeness of disparity predictions but also showcases advantages in terms of reduced memory resource requirements when compared to existing StereoGAN techniques. The flexibility of the model for diverse domains underscores its adaptability to various synthetic-to-real translation scenarios, making it a promising solution for real-world applications where image fidelity, disparity accuracy, and resource efficiency are essential considerations.

7 Acknowledgments

I extend my heartfelt gratitude to my supportive family and friends for their unwavering encouragement. Special thanks to my dedicated supervisors and esteemed professors for their invaluable guidance and mentorship throughout my academic journey.

Bibliography

[1] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Computer Vision – ECCV 2020* (A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds.), (Cham), pp. 319–345, Springer International Publishing, 2020.

[2] 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*, pp. 2223–2232, 2017.

[3] O. Vechtomova, "Autoencoders." University Lecture slides, 2019.

[4] N. Mayer, *Synthetic Training Data for Deep Neural Networks on Visual Correspondence Tasks*. PhD dissertation, University of Freiburg, 2020.

[5] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, pp. 7–42, 2002.

[6] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.

[7] H. Xu and J. Zhang, "AANet: Adaptive aggregation network for efficient stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1959–1968, 2020.

[8] R. Liu, C. Yang, W. Sun, X. Wang, and H. Li, "Stereogan: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12757–12766, 2020.

[9] C. Jacobs, D. Salesin, N. Oliver, A. Hertzmann, and A. Curless, "Image analogies," in *Proceedings of Siggraph*, pp. 327–340, 2001.

[10] 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*, pp. 1125–1134, 2017.

[11] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–189, 2018.

[13] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[14] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research*, vol. 17, pp. 1–32, 2016.

[15] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418, 2018.

[16] M. Keser, A. Savkin, and F. Tombari, "Content disentanglement for semantically consistent synthetic-to-real domain adaptation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3844–3849, IEEE, 2021.

[17] X. Li, Y. Fan, Z. Rao, G. Lv, and S. Liu, "Synthetic-to-real domain adaptation joint spatial feature transform for stereo matching," *IEEE Signal Processing Letters*, vol. 29, pp. 60–64, 2021.

[18] 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.

[19] 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 (CVPR)*, July 2017.

[20] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 1510–1519, IEEE Computer Society, 2017.

[21] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, pp. 7–42, 2002.

[22] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[23] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.

[24] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[25] M. Menze, C. Heipke, and A. Geiger, "Joint 3d estimation of vehicles and scene flow," in *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015.

[26] M. F. Reyes, P. D'Angelo, and F. Fraundorfer, "Syntcities: A large synthetic remote sensing dataset for disparity estimation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 10087–10098, 2022.

[27] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown, "Semantic stereo for incidental satellite images," in *2019 IEEE Winter Conference on Applications of Computer Vision*, pp. 1524–1532, 2019.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[31] J. Höhle and M. Höhle, "Accuracy assessment of digital elevation models by means of robust statistical methods," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 64, no. 4, pp. 398–406, 2009.

[32] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.