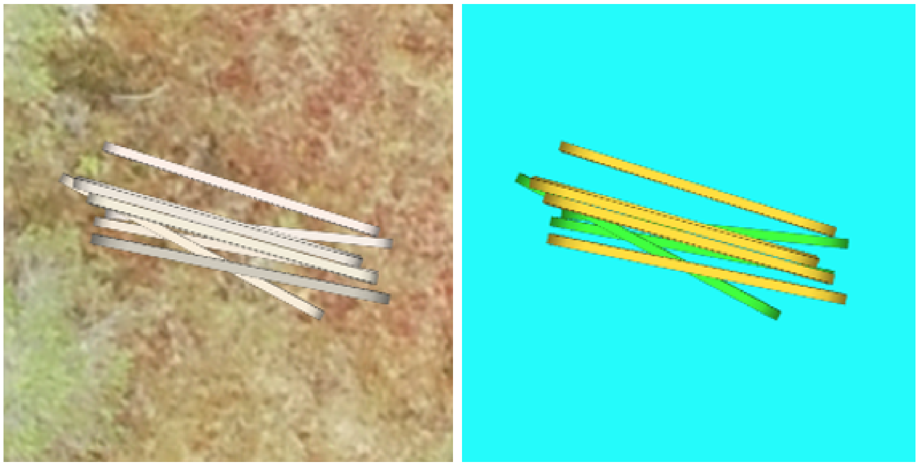
Predicting Gripability Heatmaps using conditional GANs

Jennifer Andersson and Martin Servin

2021-12-22

Technical Report



UMIT Research Lab - Digital Physics
Department of Physics, Umeå University, Umeå, Sweden

Abstract

The feasibility of using conditional GANs (Generative Adversarial Networks) to predict gripability in log piles is investigated. This is done by posing gripability heatmap prediction from RGB-D data as an image-to-image translation problem. Conditional GANs have previously achieved impressive results on several image-to-image translation tasks predicting physical properties and adding details not present in the input images. Here, piles of logs modelled as sticks or rods are generated in simulation, and ground-truth gripability maps are created using a simple algorithm streamlining the data collection process. A modified SSIM (Structural Similarity Index) is used to evaluate the quality of the gripability heatmap predictions. The results indicate promising model performance on several different datasets and heatmap designs, including using base plane textures from a real forest production site to add realistic noise in the RGB data. Including a depth channel in the input data is shown to increase performance compared to using pure RGB data. The implementation is based on the general $Pix2Pix$ network developed by Isola et al. in 2017. However, there is potential to increase performance and model generalization, and the adoption of more advanced loss functions and network architectures are suggested. Next steps include using terrains reconstructed from high-density laser scans in physics-based simulation for data generation. A more in-depth discussion regarding the level of sophistication required in the gripability heatmaps should also be carried out, along with discussions regarding other specifications that will be required for future deployment. This will enable derivation of a tailored gripability metric for ground-truth heatmap generation, and method evaluation on less ideal data.

Contents

1 Introduction

Applied artificial intelligence is rapidly transforming automation in manufacturing, logistics, and transportation. The large investemens in research and in software and hardware infrastructure are highly beneficial to automation in closely related industries. Forestry is no exception. The research program Mistra Digital Forest[1] aims to create digital solutions for a sustainable and efficient forest bioeconomy. The sub-project "AI-forwarding" investigate how forestry machine autonomy can be develeoped using simulated environments and machine learning techniques. In particular, the reinforcement learning framework is used to automate the log-grasping task of a digital forwarder, with the ultimate goal of providing models transferable to physical machines, learned using high-fidelity simulation software. The research will cover the possibility of achieving end-to-end learning as well as successfully combining the reinforcement learning approach with automatic control and human operator assistance. This report addresses the problem of predicting target locations for grasping logs and the opportunity of using synthetic sensor data from simulators.

Synthetic environments can achieve high resemblance to real forest terrains using scanned models of real forest environments. A digital forwarder can be equipped with virtual sensors allowing for real-time observations of the scene using RGB- and depth-cameras, and sensors capable of capturing the current state of the actuators, as required for reliable control.

While successfully automating manipulation tasks in the forestry context requires precise control in very complex and unstructured environments, this only poses the first of a longer sequence of technical challenges. The log-grasping task alone is likely to require a combination of high- and low-level controllers cooperating in seamless manner, where high-level controllers are responsible for high-level scene understanding and decision making such as strategic log-selection, and low-level controllers are trained to perform the requested motion and manipulation necessary to accomplish these strategies in practice.

Using high-dimensional camera data quickly complicates the learning process of a reinforcement learning agent, as the observation space quickly grows very large. This is also associated with a simulation-to-reality gap, since the camera data generated in simulation will not map perfectly to real camera data, even when high-accuracy simulation environments are available. If the machine relies on an internal 3D model of the environment, it may be feasible to process sensor data in real-time, continuously processing incoming data and updating the internal model. This can allow for perception based machine intuition and scene understanding optimizing the final behavior, avoiding risk and analysing the impact of actions on the environment, as well as mapping camera data to simulated data similar to the training data used to optimize the controllers. The goal would be to assist autonomous and semi-autonomous systems operating in the environment, including for example manipulators used for log-grasping and forwarders navigating a forest harvesting site.

[1] https://www.mistradigitalforest.se

This project aims to initiate research in this direction. Examples of research directions include:

- Perception-based segmentation of objects with dynamical properties, allowing the machine to update an internal model of its surroundings and create semantic label maps of the observed scene, for example segmenting rigid bodies and vegetation.
- Prediction of post-interaction response based on observations and an internal model of the environment.
- Intelligent analysis based on visual observations of the environment, aiding in the decision making process.

Previous research related to scene understanding based on visual sensory information has mainly focused on applications in controlled environments, focused on for example 3D reconstruction of occluded objects and prediction of position or shape response to controlled manipulation of objects of different geometrical shapes and physical properties. Another line of research concerns inverting physics engines based on visual observations. Both of these research areas require the extraction of additional knowledge, since it is not possible to infer this information from two-dimensional images. The question is if machines can develop an intuition allowing for advanced scene understanding from experience of exposure to limited observations, similar to the advanced human ability to analyse her surroundings and draw conclusions based on nothing but two-dimensional observations.

This project aims to tackle this question within the log-grasping context. Given enough experience, it should be possible for an intelligent system to predict which objects are logs and which logs are optimal to grasp based on camera data. We frame the problem as an image-to-image translation problem, and investigate the possibility of predicting gripability heatmaps based on raw RGBD-data. The data is generated in simulation to allow for automatic ground-truth heatmap generation, with the ultimate goal of learning a model capable of real-time processing of real-world RGBD-data generating reliable gripability maps that can aid forwarders in the grasping process. This can also serve as a form of representation learning to speed up the reinforcement learning training process, allowing the reinforcement learning agent to observe the environment in segmented gripability maps, removing high frequencies and redundant information present in the raw camera data.

2 Image-to-Image Translation using Machine Learning

2.1 Background

The problem of predicting an output image from the corresponding input image is found in various application areas in computer vision and graphics artificial intelligence contexts. The goal is to find a mapping between the input and output domain, allowing for automatic image translation. An RGB-image rendered from a real-world scene may not only be translated into a heatmap to visualize some measured quantity, but also into for example semantic label maps, edge maps or images in different color spaces.

Traditional machine learning approaches have required application-specific models with tailored loss functions. For example, learning a model from pixel-wise $L2$-loss often results in blurry output images. Hand-engineering of more sophisticated loss functions can be an exhausting and time-consuming task, often relying on expert knowledge specific to the desired properties of the output images and their use case. Though the CNN architecture has proved successful in many of these cases, this has been an obstacle, especially in terms of generalization to a broader collection of image translation problems.

In 2014, Goodfellow et al. proposed an image synthesizing network in which loss functions are learned automatically using an adversarial setup. These dual networks, known as *Generative Adversarial Networks* (GANs), consist of a classifier, the *discriminator*, and a generative network, the *generator*, learned in parallel. Essentially, the generator synthesizes images based on a latent noise vector input, ultimately learning the true distribution of the dataset. The goal is to generate images that are indistinguishable, but not identical, to the images constituting the true dataset, i.e. unique images sampled from the distribution of the true dataset. The discriminator governs the learning process by learning to classify whether its input images, which are either generated by the generator or drawn from the true dataset, are synthesized or sampled from the true data distribution. This is a simple binary classification problem. Thus, the objectives of the generator and the discriminator are adversarial, and the two networks are trained in parallel, continuously competing with each other and improving based on gradient feedback from the other network. Convergence is reached when the generator is able to produce image samples that the discriminator is unable to distinguish from samples drawn from the true dataset. Many successful applications followed, including generation of sharp and realistic pseudo-celebrity faces (Karras et al., 2018) and other classes of synthetic photographs (Brock et al., 2019).

In 2017, Isola et al. presented a *conditional* extension of the generative adversarial network, cGAN, in which a conditional generative model of the true data is learned. Here, the output images synthesized by the generator are conditioned on an input image. The network, known as *pix2pix*, has proven to be highly generalizable, showing promising results in a wide range of image translation tasks. The area is accelerating quickly, with interesting early developments including unpaired translations between image domains using CycleGans (Zhu et al., 2017), temporally coherent video-to-video translation (Wang et al., 2018) as well as generating photo-realistic images conditioned on text descriptions (Zhang et al., 2017).

In this project, the feasibility of using this approach to generate gripability heatmaps given rendered RGB images of piles of logs in a simulated environment is investigated. The following subsections provide an overview of the cGAN architecture, and covers previous applications to similar image translation tasks.

2.2 Conditional GAN

The conditional GAN is an extended version of the original GAN (Goodfellow et al., 2014), where the generator learns a conditional distribution of the data. The following description follows the notation of Isola et al. (2017). The generator of the pix2pix network learns a mapping from the noise vector z and the input image x on which the output is conditioned, to the output image y. In the case of general conditional GANs, x is not necessarily an image, but can be constituted by any auxiliary information such as text sequences or class labels. Here, we focus on image-conditioned GANs used for paired image-to-image translation.

Equation (1) shows the objective function (the adversarial loss) of the conditional GAN using this notation. The goal of the generator is to minimize the log of the inverse probability predicted by the discriminator for synthesized images, and the goal of the discriminator is to maximize the average of the log probability of real images and the log of the inverse probability for images synthesized by the generator. The two networks therefore compete in a two-player *minmax*-game according to Equation (2). As the training progresses, the generator learns to generate images with low probability of being classified as synthesized by the discriminator, and the discriminator improves its ability to correctly filter out synthetic images from images generated from the true target space. The learning process generally does not converge, and finding the right balance between the training process of the discriminator and the generator can be challenging.

$$\mathcal{L}_{cGAN}(G,D) = \mathop{\mathbb{E}}_{x,y}[logD(x,y)] + \mathop{\mathbb{E}}_{x,y}[log(1-D(x,G(x,z)))] \tag{1}$$

$$G^{*} = \underset{G}{\operatorname{argmin}}\,\underset{D}{\operatorname{argmax}}\,\mathcal{L}_{cGAN}(G,D) + \lambda\mathcal{L}_{L1}(G) \tag{2}$$

In the conditional case, the discriminator takes both the image x and the output image y (the true output or the generated output) as input, and predicts the probability that y is a true translation of x. As seen in (2), the objective function is combined with a *L1*-loss term, such that the generator is encouraged to also minimize the L1-distance (mean absolute error) between the generated images and the ground-truth output images. λ is a hyperparameter weighting the pixel reconstruction loss against the adversarial loss.

In the original pix2pix network, the generator has a *'U-net'* architecture; an encoder-decoder architecture with so called skip-connections, allowing low-level information to travel between the input and the output images. This framework involves downsampling the input image to a bottleneck representation, followed by upsampling the image to the target output size.

The discriminator architecture is called *PatchGAN*, and relies on classification of $N \times N$ $= 70 \times 70$ patches across its input images. The final classification is an average of the local classifications. This approach assumes that the image can be modelled as a *Markov random field*. More details can be found in Isola et al. (2017).

2.3 Previous Work

By using a spatio-temporal adversarial objective, the vid2vid-network proposed by Wang et al. (2018) learns a mapping between for example sequences of semantic segmentation maps and photo-realistic videos, displaying impressive temporally coherency between frames. The pix2pix-network (Isola et al., 2017) produces impressive results on similar, but static, image translation tasks, including translation from maps to aerial photos, predicting photo-realistic street-views from static semantic segmentation maps, and generating RGB-photos from 2D sketches of 3D objects. In all of these applications, the network is required to learn a mapping that encodes information not present in the input, as opposed to the opposite translation tasks in which information is strictly removed. This property is ultimately required for heatmap prediction and other tasks related to high-level scene understanding and machine intuition.

Generative adversarial networks have been used in several heatmap generation contexts, displaying this ability. If, for example, it is possible for generative models to produce accurate monocular depth map predictions from RGB images, this implies the incorporation of some level of spatial intuition in the final model, i.e. the model learns to conceive the physical concept of distance from flat images. Using the pix2pix conditional generative adversarial architecture, this has been demonstrated by for example Lore et al. (2018). Here, the depth map prediction problem is framed as an image translation problem where single RGB frames are mapped to grayscale pixel intensity representations of the corresponding LiDAR depth maps. The model generalizes well to unseen data and performance is comparable to current state-of-the-art methods for monocular depth prediction that are not GAN-based. This is encouraging for our purposes.

Other examples of successful pix2pix applications in similar contexts include human pose heatmap prediction (Matsuzaki et al., 2017), and a similar architecture is used by Chou et al. (2018) to learn a mapping from RGB images to keypoint heatmaps for human pose estimation. The pix2pix architecture has also been successfully applied to uncertainty map estimation in deep learning-based optical flow determination methods (Lee et al., 2020), producing uncertainty maps similar to those obtained using conventional methods while significantly reducing the processing time.

Recently, cGANs have also been used for medical image-to-image translation tasks, for example to predict diffusion weighted MR scans from multi-modal CT perfusion maps (Rubin and Abulnaga, 2019) to assist in the identification of infarcted brain tissue in stroke patients. A more general framework, *MedGAN* (Armanious et al., 2020), has also been proposed for medical image translation tasks, successfully demonstrating how advanced derivations of cGANs can be used in for example PET-CT scan translation contexts. Again, this requires the network to produce more information than what is provided by the input images, since CT-scans are more detailed than PET-scans.

In (Ma et al., 2020) a conditional GAN is used to predict probability distributions of possible efficient paths between two locations on a map. The predicted map showing the distribution of feasible paths between the two locations is combined with a modification of the RRT path planning algorithm for optimal path planning. This is another example where generative adversarial networks have successfully incorporated high-level scene understanding and been able to extract additional information based solely on exposure to training data consisting of flat images.

Uricár et al. (2019) gives an overview of how generative adversarial networks have and can be used in the context of autonomous driving. For example, the approach has been used for segmentation and generation of the occluded parts of objects as well as for domain adaptation for simulation-to-reality transfer. Bousmalis et al. (2018) uses this method in the context of simulation-to-reality transfer for robotic grasping. Pedersen et al. (2020) combines this approach with a deep reinforcement learning-controlled agent trained to perform robotic grasping tasks. Using a CycleGAN (Zhu et al., 2017), designed for unpaired image translation between domains, the agent's observations in the real world can be translated back into the simulation domain. Other GAN-inspired applications in the field of robotic grasping include using a conditional WGAN to generate multi-fingered robotic grasp candidates from depth information (Patzelt et al., 2019).

3 Problem Statement

Accurate depth map predictions using conditional GANs (Lore et al., 2018) force models to learn spatial relationships between objects in a scene based purely on the input image. The power of conditional GANs in similar image-to-heatmap translation tasks has been demonstrated in many contexts, as discussed in the previous section. This project extends this approach to gripability map prediction based on the idea that inferring gripability, which also depends on the spatial relationships between objects given a specific grapple and grasping strategy, should be possible given RGB- or RGBD data. To produce accurate gripability maps, the model is required to understand these relationships, as well as how they relate to gripability depending on the chosen definition of the latter.

For our purposes, a *gripability map* refers to a heatmap visualization of the gripability variation across an image. The gripability has to be defined according to a set of predefined criteria, such as a well-defined grapple and a gripability score depending on the definition of a successful grasp. Given raw RGB- or RGBD-data capturing a forest scene containing a pile of logs, we investigate the possibility of predicting the corresponding RGB gripability maps using supervised machine learning techniques. The problem is therefore framed as an image-to-image translation problem, where the input and output images consists of three or four channels, respectively. This project constitutes an initial evaluation of the feasibility of using *Conditional Generative Adversarial Networks* for automatic gripability map generation, with data generated entirely through physics-based simulation. Future research directions are investigated, and recommendations are provided based on available scanned data from Komatsu Forest.

4 Gripability Maps

In the initial approach, binary ground-truth gripability maps are generated from a simple physics-based simulation environment modelled using the AGX Dynamics simulation software. The scene consists of a static base plane ($10 \times 10 \times 0.2$ m collide-box geometry) and a pile of 7 logs modelled as rigid bodies in the shape of $3 \times 0.1 \times 0.1$ sticks (cuboids) or cylinders of length 3 m and diameter 0.1 m. The log density is 800 $\frac{kg}{m^3}$ (roughly corresponding to wood).

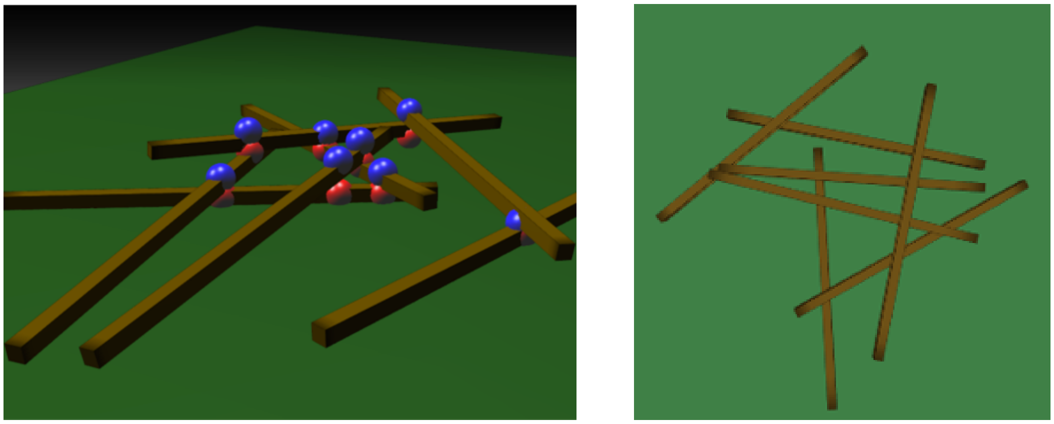
The centre of mass of the first log is located 1 mm above the ground plane, and each of the remaining logs are spawned at a height of 1 mm above the preceding log, to ensure collision free simulation initiation. The Cartesian orientation of each log is drawn from a uniform random distribution.

A RenderToImage-camera is created to render depth and color buffers, which are used to generate RGB- and depth images of the scene as viewed from above. When the simulation is initiated, the logs fall into a pile and once the simulation has reached its final state (i.e. when the average speed of all logs reaches below some limit, RGBD-data is collected and saved to file.

Finally, the gripability maps are generated by analysing the gripability using a straightforward approach. Here, a log is classified as graspable if at no point on its geometry it is occluded by another log. We use a simple algorithm to determine the gripability based on a step-by-step approach to determine which logs are *not* graspable. Each log is approximated by three straight lines in the horizontal plane. The lines approximating each log are then analyzed to the lines approximating every other logs in a pair-wise manner. When an intersection is detected, the height of each of the corresponding objects at the intersection point is approximated. The log corresponding to the lowest height at the intersection point is then determined to be not graspable.

Looping through all logs pair-wise, only graspable logs remain and a simple binary color heatmap constituting the resulting gripability map is generated. 2 shows the preprocessed input data (bottom left), where the preprocessed depth image (top right) has replaced the third color channel (see Chapter 5). The processed depth image is normalized to cover the entire pixel range, compared to the raw depth map (top left). The resulting gripability map is also depicted (bottom right).

For simplicity, and due to time-constraints, some approximations have been made in the gripability determination using this algorithm. For example, the approach relies on determining the intersection points in a 2D view of the scene. Each log is assumed to be of 3 m in length in this view, and as a result false intersection points may be detected. By inspection, the rate of erroneous gripability maps is estimated to be low. Most of these are borderline cases that are often hard to detect even by manual inspection. The impact on model performance should be analysed in detail.



(a) (b)

Figure 1: a) Side-view of the simulated scene marking the intersection points used by our algorithm to determine which logs are graspable. Red and blue spheres mark the detected intersection points, and blue spheres correspond to logs detected as graspable. b) Top-view of the simulated scene corresponding to the input RGB images used in the network.

5 Implementation

We follow the Pix2Pix architecture proposed by Isola et al. (2017), taking 8-bit 3-channel *256* × *256* pixel frames (RGB) as input and generating 8-bit images of the same shape. These are normalized to floating point values between 0 and 1. This architecture can be modified to perform image-to-image translation between images of different dimensions, but for simplicity we process our data to fit the original dimensions. This means that, in the case where greyscale depth images are added as an additional channel to the input images, these replace one of the color channels. This is not expected to affect the heatmap prediction quality due to the very high dimensionality of the original 8-bit RGB frames. Thus, two color channels (and additional depth information) are assumed to contain more than enough information to perform the desired gripability heatmap predictions. Summarizing, the input images consist of two color channels and a greyscale depth channel replacing the third color channel. Analyses using pure RGB inputs are also conducted.

The discriminator and the generator are optimized in parallel, alternating the gradient descent step between the two networks. The discriminator is optimized using binary cross entropy. A weight is incorporated to enforce a lower learning rate in the discriminator compared to the generator, but finding the right balance between the learning rates of the two networks can be a delicate task. The original paper recommends using a weight of $\lambda = 100$ between the adversarial loss, $\mathcal{L}_{cGAN}$, and the $L1$-loss, $\mathcal{L}_{L1}$, of the combined objective function in Equation (2), emphasizing correct heatmap predictions relative to the input image over realistic heatmap predictions. This encourages the generator to synthesize

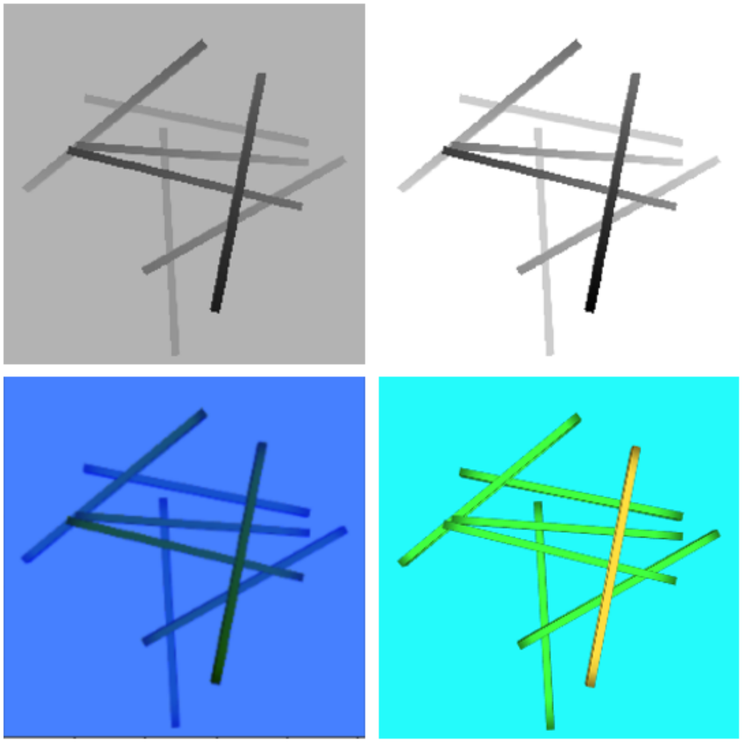


Figure 2: *Top left:* Raw depth image from a simulated scene showing a pile of sticks. *Top right: Processed depth map. Bottom left: Input image. Bottom right: Ground-truth gripability heatmap.*

images sharing the global structures of the input images.

As in the original paper, we use the Adam-optimizer with a learning rate of $\alpha = 0.0002$ and identical momentum parameters. Initial analyses indicate that the training stability and performance is not very sensitive to these parameters.

The generator and the discriminator networks consists of blocks of convolutional layers, batch normalization layers, and leakyReLU-activation layers. The tanh-activation function is used in the output layer.

6 Evaluation Metrics

In many contexts where generative adversarial networks are applied, assessing the performance of the generator is difficult. In image-to-image translation tasks, the problem boils down to finding a proper metric to evaluate the similarity or dissimilarity between images, comparing the output of the generator to the ground-truth. Here, we use two common evaluation metrics: the *mean-squared error* (MSE), which compares the pixel-wise intensity difference between the two images, and the *structural similarity index* (SSIM) (Wang et al., 2004), which measures structural similarity based on groups of pixels.

The MSE quality measure is calculated as the average of the squared pixel-wise intensity differences between the ground-truth image $\mathcal{I}_{gt}$ and the generated image $\mathcal{I}_{gen}$, see Equation (3) where N is the number of pixels in the images. The MSE metric efficiently captures absolute differences between images, but often fails to measure perceived differences since it does not take structural information into account. The SSIM metric, on the other hand, does not assume independence between neighboring pixel intensities, and allows for structural quality assessment by measuring the degradation of structural information in the image. It also includes luminance and contrast masking terms, and is often a better measure of perceived differences and more intuitive to human perception.

The structural similarity index SSIM $\in [-1, 1]$ is calculated across an image based on local pixel windows. To obtain a global quality measure, the mean SSIM is calculated according to Equation (4), where M denotes the total number of windows used to obtain local SSIM evaluations. μ and σ denotes the average pixel intensities and variance (or covariance if twice-indexed) in the local windows of each image. c_1 and c_2 are stabilizing constants based on the dynamic range of the pixel intensities.

$e_{SSIM} = 1$ for identical images and $e_{SSIM} = -1$ for complete dissimilarity. In our case, the SSIM is always positive, and we use a modified (inverted) version where a perfectly reconstructed image generates a modified mean SSIM of 0. Thus, we want to minimize both e_{MSE} and e_{SSIM} between the generated images and the corresponding ground-truth.

$$e_{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathcal{I}_{gt,i} - \mathcal{I}_{gen,i}\right)^2 \tag{3}$$

$$e_{SSIM} = \frac{1}{M}\sum_{j=1}^{M}\frac{(2\mu_{\mathcal{I}_{gen,j}}\mu_{\mathcal{I}_{gt,j}} + c_1)(2\sigma_{\mathcal{I}_{gen,j}\mathcal{I}_{gt,j}} + c_2)}{(\mu^2_{\mathcal{I}_{gen,j}} + \mu^2_{\mathcal{I}_{gt,j}} + c_1)(\sigma^2_{\mathcal{I}_{gen,j}} + \sigma^2_{\mathcal{I}_{gt,j}} + c_2)} \tag{4}$$

Figure 3 exemplifies the visual deviation between the predicted gripability map and the ground truth, corresponding to four different values of the modified SSIM. A modified SSIM of ~ 0.03 corresponds to approximately one misclassified log using the binary gripability map in our analyses.

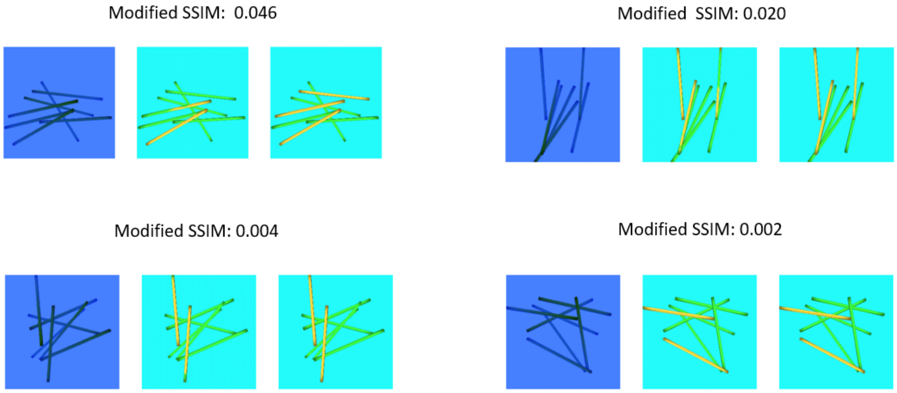


Figure 3: *Top left:* Predicted gripability has classified between one and two logs that are graspable according to the ground truth as ungraspable. This yields a modified SSIM of ~ 0.05. *Top right:* Half a log is classified as ungraspable, while the entire log is graspable according to the ground truth. This yields a modified SSIM of 0.02. *Bottom left:* A small fraction of a log is classified as ungraspable, while the entire log is graspable according to the ground truth. This yields a modified SSIM of 0.004. *Bottom right:* The predicted gripability map is almost identical to the ground truth. This yields a modified SSIM of 0.002.

7 Preliminary Results

7.1 Experiments

Several experiments are conducted using two different kinds of log piles: the *jack-straw pile*, where seven sticks are placed in a pile that is not well-aligned, and *well-aligned piles* where the sticks are placed in more well-aligned piles that are more similar to the piles of timber prepared by the harvester in real forwarding scenarios. We also investigate the impact on object segmentation in the preprocessing chain. Using real forest data, object-specific properties such as the texture, color, lighting and shades varies between each log, so this does not necessarily make the simulation setup more ideal. Most analyses are performed combining the RGB input data with a depth map. A performance increase following this approach is recorded. In more advanced settings with varying weather- and lighting conditions, the depth map can be more stable than RGB images of a scene, as the depth map is not affected by these external factors.

For each experiment, we show the raw RGB data from simulation (no depth), together with a three-image sequence showing the preprocessed network input (left), the predicted gripability map (middle) and the corresponding ground truth (right). The samples shown are chosen randomly from the dataset. We also show a kernel density estimate (KDE) of the probability density function of the modified SSIM as evaluated on the training dataset and the validation dataset, respectively. The total dataset consists of 6400 samples, of which 10% constitutes the validation set. Thus, each experiment is carried out using 5760 training samples and 640 validation samples. For training data performance evaluation, the modified SSIM is calculated on 640 random training samples.

Figure 4 shows the training- and validation performance of models trained using jack-straw piles generated from simulation. The corresponding training- and validation performance of models trained on segmented jack-straw piles is presented in Figure 5. Figure 6 shows the model performance for a dataset based on rod-shaped logs, as opposed to the cuboid sticks used in the previous analysis.

The initial results indicate a promising potential for gripability map prediction using conditional GANs. As we can see, the model performance on training data is very high for models trained on datasets consisting of non-segmented jack-straw piles. Essentially, every prediction yields a modified SSIM of less than 0.025 compared to the ground truth. This suggests an error in prediction of less than one graspable object.

Figures 7 and 8 shows nine examples from the evaluation samples from the training and validation dataset, respectively. The model used for evaluation is trained on non-segmented jack-straw piles. Each sample consists of a sequence of three images: the preprocessed input data (left), the predicted gripability map (middle) and the corresponding ground truth (right). The results illustrate that the model is able to make very accurate gripability predictions on training data, which is consistent with the modified SSIM distribution shown in Figure 4). Model inference on the validation dataset also yields promising results, but the model performance is low compared to the performance on training data. As can be seen in the modified SSIM distribution over validation data samples, most predictions still correspond to a modified SSIM of less than 0.03, but the variance is larger and we need a modified SSIM of up to 0.1 to capture the bulk of the distribution. This suggests some overfitting to the training data. The results show that the training performance improves as the number of iterations increases, while this is not true for the validation performance, which starts to decrease after $108k$ iterations.

The top row of Figure **??**b) shows successful gripability map predictions on the validation data. The middle row shows partly successful predictions, which illustrate the reason for the generally larger modified SSIM compared to the corresponding evaluation on training data. As we can see, the graspable objects are generally correctly classified, but the model is not able to correctly color the entire object according to the defined gripability heatmap. The bottom row shows situations where the algorithm has produced inaccurate ground truth gripability maps, and the model does a better job at predicting the true gripability. This is not captured by the current evaluation metrics. It is possible that this constitutes part of the the reason for the lower performance on the evaluation data, if the model is able to adjust to the specifics of the algorithm when fitting to the training data. This follows from the fact that the fraction of erroneous gripability maps should be identical in both datasets. Further analyses should be performed to conclude whether or not the model performs better on data in which the ground truth is not accurate when evaluated on the training data as opposed to the validation data.

The model performance on rod-shaped logs presented in Figure 6 shows that this significantly reduces the prediction accuracy on the training data according to the current evaluation metric. The main difference between this model and the non-segmented jack-straw model is that there are no thin black lines separating the objects in the RGB image. The performance on validation data is more similar to that of the original jack-straw model. Based on the modified SSIM distribution, we can also conclude that the model trained on rod-shaped logs appear to be less prone to overfitting, and the model performance generalizes well to the validation data. We also note that the model performance on both the training and validation data decreases as the training continues after $36k$ iterations.

The performance of the model trained on rod-shaped log compared to sticks suggest that the model performance can possibly increase with increased object segmentation in the preprocessing chain. Figure 5 shows the model performance of a model trained on segmented jack-straw piles. These results do not show higher performance compared to the non-segmented logs on the training and validation data, but the model performance is comparable to that of the validation performance using the non-segmented jack-straw piles. Again, we observe improved generalization to validation data, as the performance on validation data unseen during training is consistent with the performance on training data.

Figure 9 and 10 shows the corresponding training- and validation performance for models trained using non-segmented and segmented well-aligned piles. As we can see, the performance of the model trained on non-segmented piles is similar to the validation performance of non-segmented jack-straw piles, and the model generalizes well to unseen data. It is worth noting that with the current algorithm, well-aligned piles generally result in a larger number of graspable logs than the more spread out jack-straw piles. This exposes a weakness of our evaluation metric; for example, a model that classifies all logs as non-graspable will appear to be better when evaluated on jack-straw piles than when evaluated on well-aligned piles with more graspable objects to detect. The exact implications on the performance comparisons need to be investigated, but we note that a similar distribution of the modified SSIM over the evaluation data possibly suggests a better model performance for models trained on well-aligned piles.

The perhaps most interesting results are those presented in Figure 10, which shows

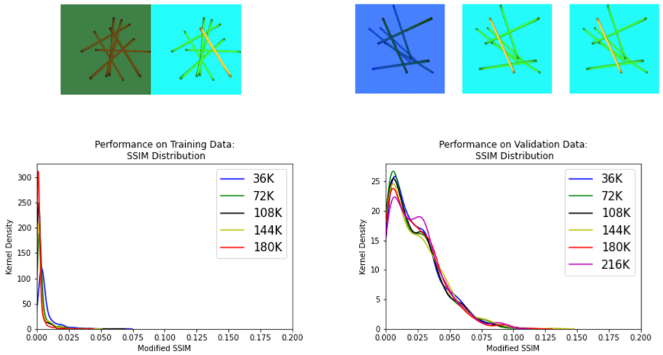


Figure 4: Jack-straw piles.

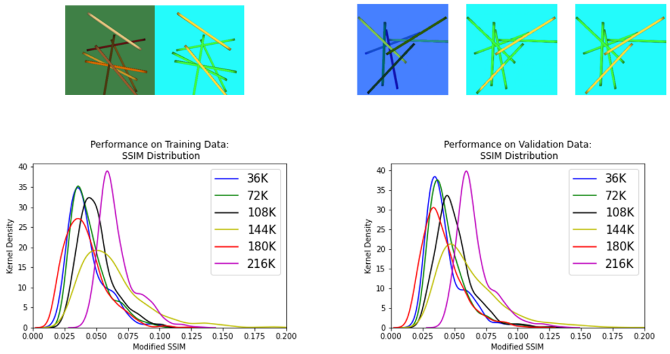


Figure 5: Segmented jack-straw piles.

excellent performance on the segmented well-aligned piles. The performance on training data is essentially perfect after 108k training steps. The performance on validation data is slightly decreased compared to the performance on training data, but the model still shows superior validation performance compared to the previously analysed models. After 108k training steps, almost every prediction on the validation dataset generates a modified SSIM of less than0.025. These results are very promising and demonstrates the potential in using conditional GANs to predict gripability.

Initial results show that using only RGB representations of the scene as input to the network decreases the model performance compared to including depth information in a network of identical size. For comparison, the model performance is evaluated using only RGB images as input for two datasets previously analyzed: the non-segmented jack-straw piles and the segmented well-aligned piles. The modified SSIM distribution obtained from model inference on training- and validation data is shown in Figure 11 and 12. In both cases, the performance improves when depth information is included, but the effect is smaller

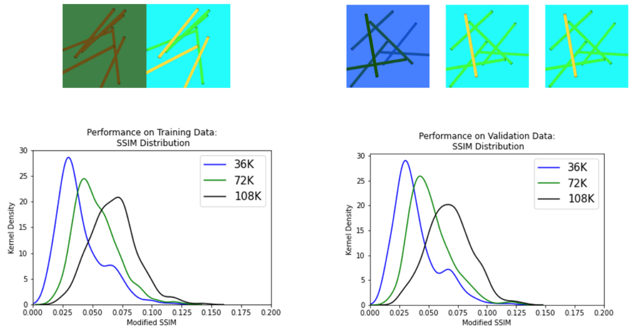


Figure 6: Jack-straw piles with rod-shaped logs.

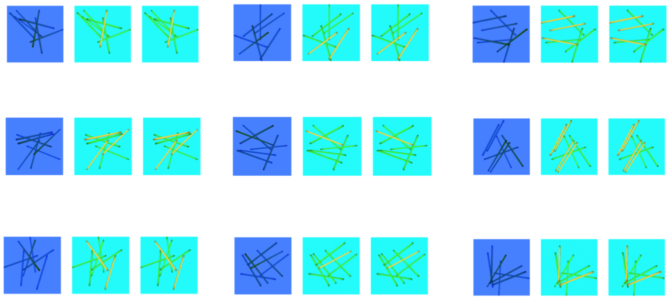


Figure 7: Training evaluation samples

for the model trained to predict gripability from segmented well-aligned piles. This is reasonable, since the depth map also induces a form of pixel-wise object segmentation that can be captured by the network. These results confirm the assumption that including depth maps will improve gripability map predictions using this approach. It is also likely that the simulation-to-reality gap is reduced when depth information is fed to the network, since the depth map does not depend on for example lighting and weather conditions; effects that are non-negligible in pure RGB frames.

To investigate how sensitive this approach is to increasing heatmap complexity, two additional analyses are conducted. Figure 13 shows the model performance using a dataset in which only the central parts of graspable logs are highlighted as graspable in the gripability maps. In Figure 14, the model performance using a slightly more complex heatmap is presented. Here, a third gripability level is added, with the central region of graspable logs marked as highly graspable, and the remaining parts of the graspable logs highlighted as a third, medium-gripability level.

Drawing conclusions from the modified SSIM distribution in Figure 13 is difficult, since misclassification of graspable objects results in such a small difference between the predicted

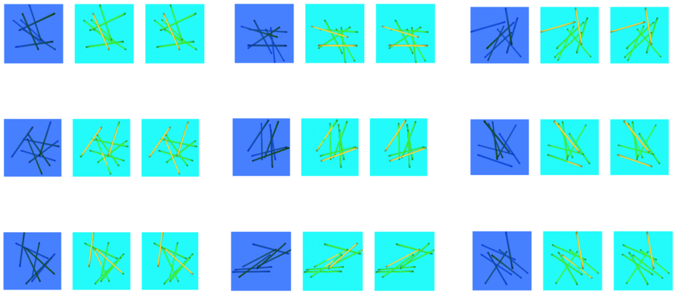


Figure 8: Validation evaluation samples.

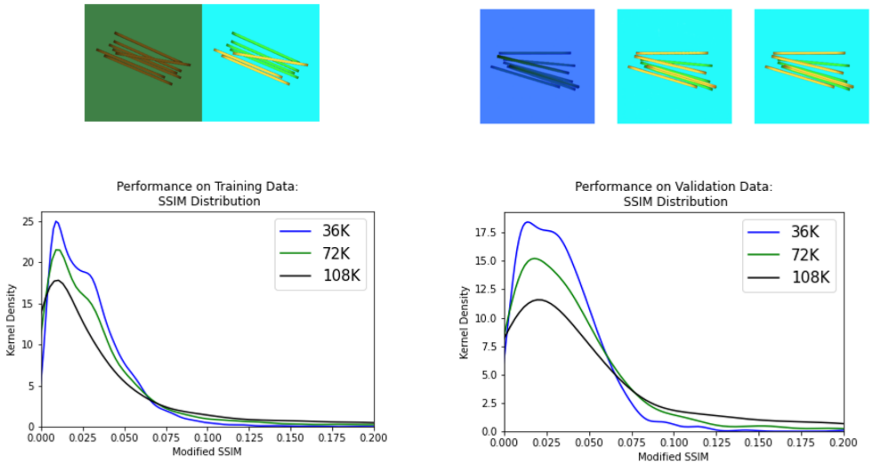


Figure 9: Well-aligned piles.

gripability map and the ground truth, which results in a small modified SSIM even when all graspable objects are misclassified. Based on manual inspection, however, the model performance appears to be very similar to that of the more complex heatmap model. Both models generalize well to the validation data, and the performance does not decrease compared to models trained using the less complex heatmaps of previous analyses. An interesting note is that the models in these cases appear to classify only one graspable object as graspable to a significantly larger extent than the previous models, omitting the other graspable objects in the heatmap. This likely explains the spikes in the modified SSIM distribution seen in Figure 14. Overall, however, the model performance is promising.

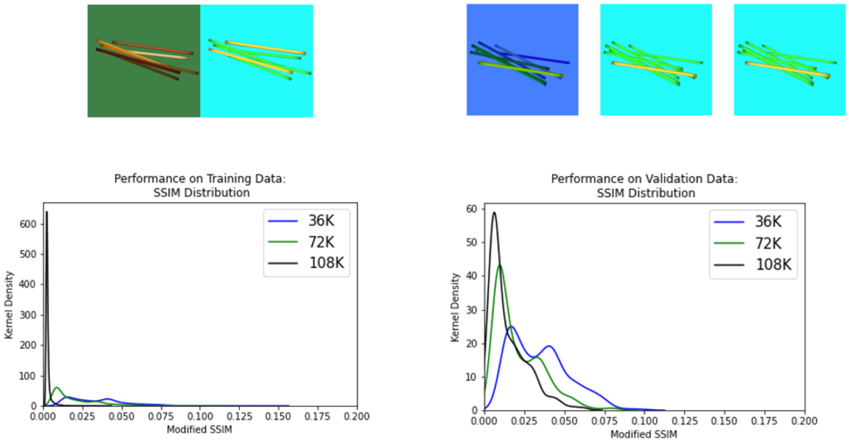


Figure 10: Segmented well-aligned piles.

7.1.1 Forest Terrains

It is currently possible to obtain terrain data with a resolution comparable to that required by an operating forestry machine. At the time of this research, a high-density laser scan and corresponding LiDAR-measurements of a 341 × 272 m forest site is available through the courtesy of Komatsu Forest. These provide top-view RGB images of log piles prepared by the harvester, as well as large patches of the surrounding area. A .tif-file with RGB data and a .las file containing height field information (and redundant RGB data and additional information) is provided. Figure 15 a) shows a top-view RGB representation of the available data.

To extend the analysis to less ideal environments, the terrain is divided into 90 × 90 images, each with a dimensionality of 3.8 × 4.2m. These are combined in neighboring quadruples resulting in 7.6 × 8.4m terrain slices. Terrain slices containing white pixels, logs, dense forest areas or other terrains unlikely to contain piles of logs are removed, and a final dataset containing 850 terrain slices is selected. This is done in order to test the model response to background noise and unstructured plane textures. When the training and validation datasets are generated, the base plane is covered by a terrain slice chosen randomly during each simulation. Since the top surface of the base plane is 10 × 10m, the terrain texture is slightly distorted. Following the current terrain slice generation procedure, terrain slices of precisely 10 × 10m are obtained by initially dividing the terrain into 68 × 54 images, but for the current analysis our approximation is sufficient. Figure 15 b) shows examples of the terrain slices (top) together with the input RGB image generated from simulation and the corresponding gripability map (bottom).

Natural future directions include representing the LiDAR point cloud as a height field

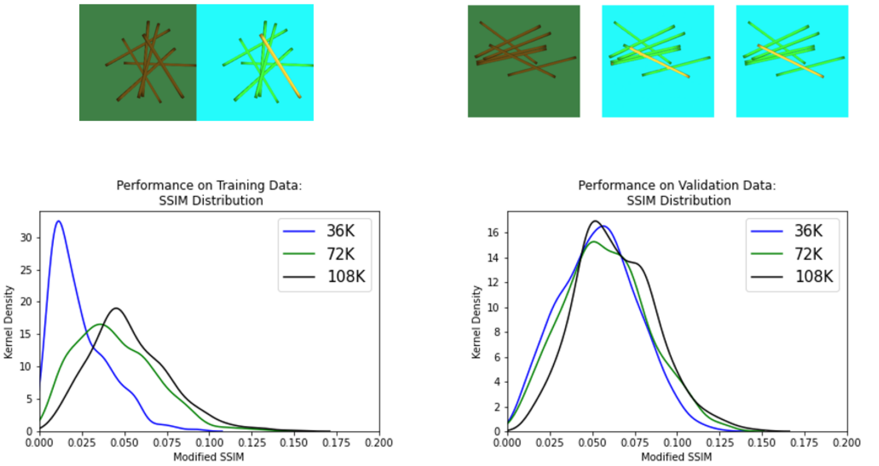


Figure 11: Jack-straw piles with only RGB input.

in order to obtain a more accurate depth map representation of the unstructured forest environment. This can be done through terrain reconstruction in AGX Dynamics, by generating an agxCollide.Heightfield directly from the LiDAR point clouds.

Figure 16 and 17 shows the model performance using this terrain texture together with segmented well-aligned piles, which generated the best performance using the ideal simulation environment. In the dataset used to train models whose performance is shown in Figure 16, the logs are segmented using RGB colors from logs present in the available data from Komatsu Forest. This gives a small segmentation effect similar to that of real piles of logs. Figure 17 shows the corresponding model performance for a model trained using logs segmented according to previous segmentation analyses.

The results show very high model performance on training data for both models, but contrary to the previous analysis using well-aligned piles, the best training and validation results are obtained for the model trained on non-segmented logs. The top performance of the two models are essentially comparable, however, and the impact of segmentation is not as large as detected in previous analyses.

These models do not generalize as well as other models to the validation data, but the performance on training data is very high and the performance on validation data is comparable to the validation performances of previous models. This could partly account to the fact that the non-segmented logs in this case are slightly segmented according to the color differences between logs in the real forest data. This could suggest that even conservative segmentation in the the preprocessing chain could have a positive effect on the model performance.

In the case of segmented logs on forest background, the model performance actually

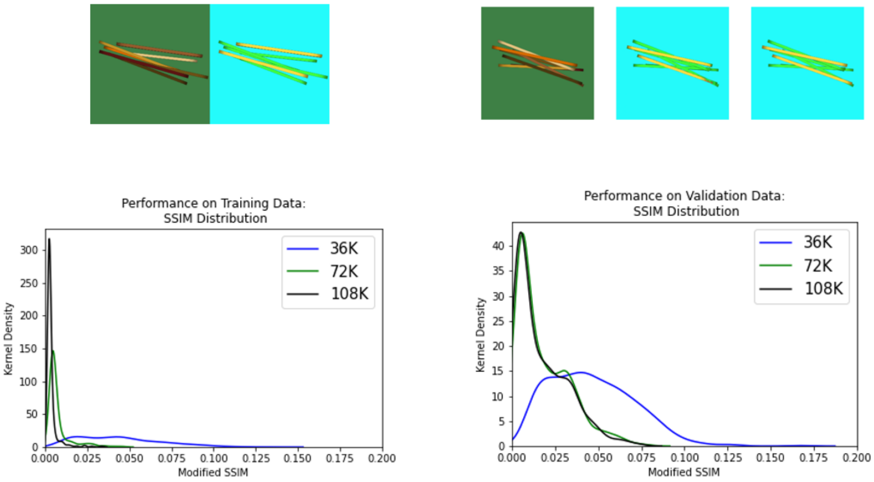


Figure 12: Segmented well-aligned piles with only RGB input.

increases compared to the corresponding model trained with a plain background, and the conclusion is that the added background noise does not decrease the potential for gripability map prediction using these kinds of networks. This is encouraging for the potential to apply this method to even more realistic forest environments.

8 Discussion & Future Work

This study has investigated the potential application of cGANs for gripability heatmap prediction in the context of forestry log grasping. The network, based on the *Pix2Pix* network developed by (Isola et al., 2017), is observed to be capable of producing generally accurate gripability heatmaps, and we conclude that the initial analysis show potential for using a similar approach to generate more advanced gripability heatmaps from less ideal sensor data.

Analyses have been carried out for very simple gripability map predictions using RGB input as well as three-channel images where one color channel is replaced by a depth map captured by a built-in depth camera. We observe that feeding depth information to the input increases the model performance, but the extent varies depending on pile configuration and the level of object segmentation in the preprocessing chain. The exact impact of including depth information should be subject to further investigation.

Models are evaluated using a modified structural similarity index (modified SSIM). The best results are achieved for well-aligned piles with segmented logs in our entirely simulated setting. The corresponding results using forest background from scanned forest data suggests promising generalization of the method to less ideal settings. The forest

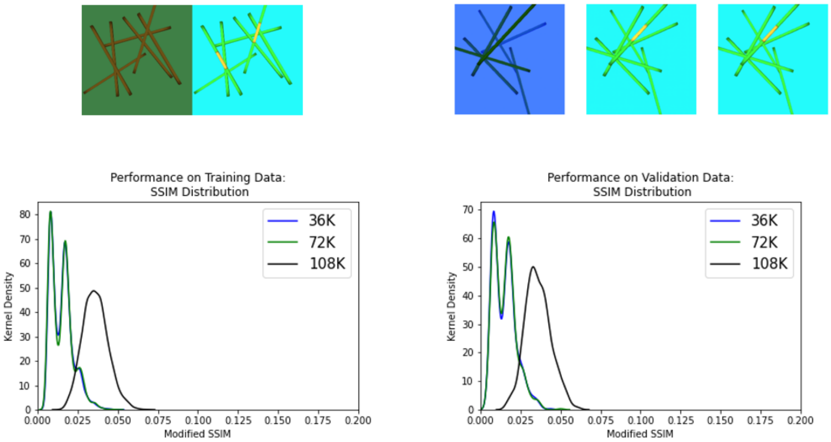


Figure 13: Jack-straw piles with the central region highlighted as graspable in the gripability maps.

background texture induces noise in the RGB representation of the image but does not include height variations in the terrain. Important future steps include investigating the potential for gripability map prediction with cGAN derivatives using log piles on simulated terrains reconstructed from high-density laser scans, including local height maps. An ongoing discussion regards the point-of-view of the collected sensor data. A suggestion is that the gripability map aids an intelligent agent in determining what logs to grasp. In the current analysis, a drone is required to collect aerial data, but it is likely possible for sensors to equip the driver's cabin of forwarders to obtain camera observations following the motion of the vehicle more closely. Ideally, it is possible to perform real-time inference on the incoming sensor data. In this case, a high-level intelligent agent views the scene through a gripability map representation-lens as the forwarder navigates the environment.

Our approach shows very good predictability if only one graspable object is required. This is encouraging, as an intelligent agent is only required to pick one grasping pose at each time. Moreover, the false positive rate (FPR) is generally low according to inspection (i.e. non-graspable objects are seldom predicted as graspable in the predicted gripability heatmaps). Gripability heatmaps with high modified SSIM compared to the corresponding ground truth is usually due to the model neglecting or half-detecting some graspable objects, i.e. the false negative rate (FNR) is sometimes high.

Despite our gripability maps being very simple, the network is not able to produce highly correct gripability maps with an accuracy that generalizes very well to validation data. The bulk of the modified SSIM distribution is generally at scales of < 0.1 for the training and validation data, and this should be reduced to < 0.03 to ensure that the deviation between the predicted gripability maps and the corresponding ground truth amounts to less than

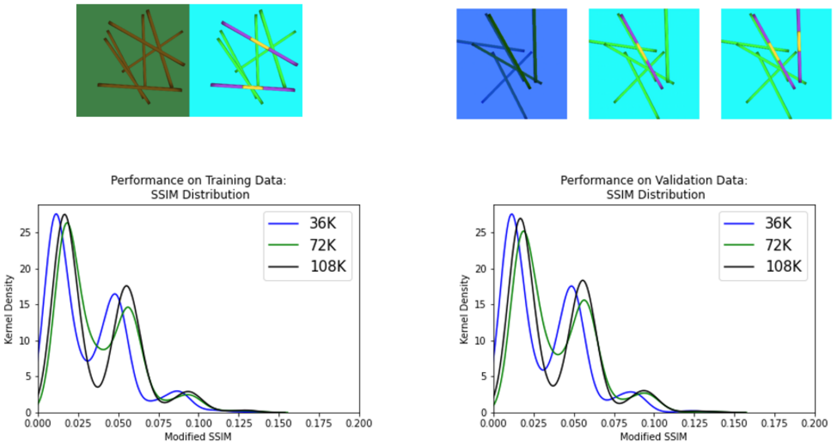


Figure 14: Jack-straw piles with added gripability heatmap complexity.

one graspable object. This is indeed the case for several models, and is particularly true for the model trained on data consisting of segmented well-aligned piles based on inference on both the training and validation data, respectively.

Most models show promising results and many models show good generalization to validation data. While the best results are obtained for segmented logs or sticks in well-aligned piles, the model performance on validation data is often similar between the different experiments and it is difficult to conclude where the generalization and performance bottlenecks lie. Some results appear contradictory; for example, for well-aligned piles, segmentation increases accuracy, but not for jack-straw piles. This could be reasonable, though, as it is easier to separate objects in jackstraw piles even without segmentation, so the segmentation effect is expected to be limited compared to well-aligned piles. However, the training is generally stable and the method shows potential for gripability heatmap prediction on different kinds of datasets. The initial study also indicates a possibility of maintaining good predictions when more complex heatmaps are necessary, but tailored loss functions or more advanced network architectures may be required to increase performance further.

Observed prediction problems may be caused by under-training or overfitting, and may also be related to the suboptimal quality of the ground-truth gripability maps used during training. In several cases, the method is able to produce very accurate gripability maps on training data. These models do not generalize as well to the validation data, and efforts to understand the reason for this should be made. Over-fitting is not generally an issue for cGANs, but applications where the produced images have actual physical meaning are rare. Known issues include modal collapse, where the generator fools the discriminator by outputting a limited number of images regardless of its input, and issues in stabilizing

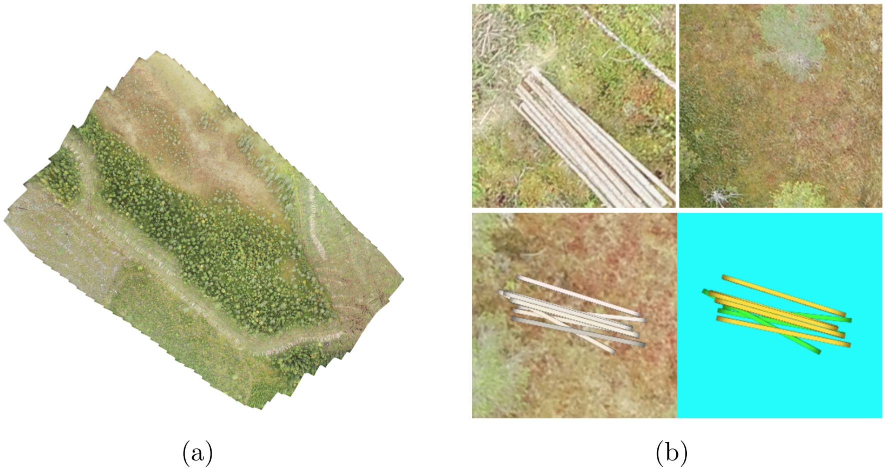


Figure 15: a) Scan of the forest harvesting site provided by Komatsu Forest. b) *Top left:* 3.8 × 4.2 m terrain slice including a pile of logs. *Top right:* 7.6 × 8.4 m terrain slice. *Bottom left:* Input RGB data where a pile of cuboid logs or sticks are imaged on a base plane wrapped in a terrain slice texture. *Bottom right:* Corresponding gripability heat map.

the joint training of the discriminator and the generator. These are not observed in our experiments. Modal collapse also refer to the generator generating outputs independent on the latent code, which is observed in the original *Pix2Pix*-network as well, but in our case a deterministic output is desired.

The impact of using a larger dataset should be investigated, as this may reduce potential overfitting and increase generalization to validation data. Similar effects can be obtained using training data augmentation, subjecting the complete training dataset to affine transformations such as flipping and rotating images.

More sophisticated loss functions should be considered. The *Pix2Pix* network is very general and focuses on images looking *realistic*, as opposed to accurately making predictions about the physical world. Pixel-reconstruction losses often leads to blurry results, and may not capture detailed differences of the images. For example, the MedGAN network uses additional losses such as *perceptual loss* and *style transfer losses.* They show successful results in image-to-image translation tasks such as PET-to-CT scan translations and MR motion artefact correction, where the target image contains detailed soft-tissue and bone structure information that is not present in the input image. Using the current architecture and loss functions, performance sensitivity to varying the weight parameters for the adversarial loss and the $\mathcal{L}_1$-loss should also be investigated.

The possibility of increasing accuracy in the gripability map prediction using more advanced network architectures should also be investigated. In fact, in many other applications the *Pix2Pix* network is used as a baseline for more complex and application specific

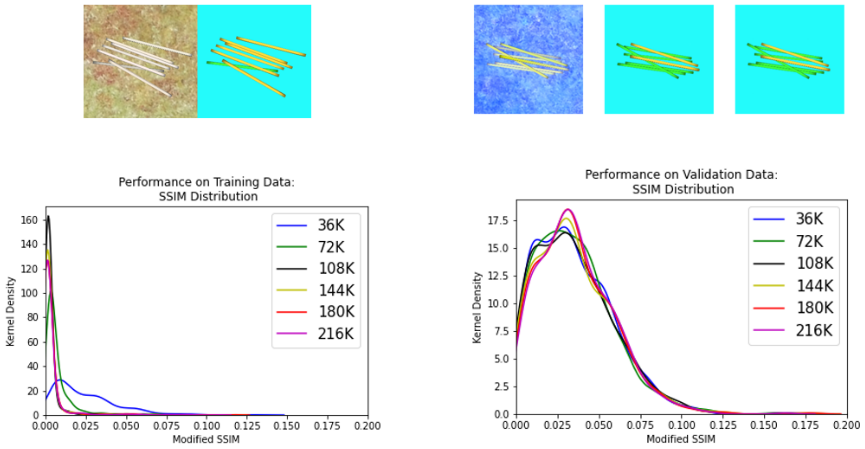


Figure 16: Well-aligned piles on a forest background.

architectures. For example, (Lore et al., 2018) uses two conditional GANs to improve depth map estimations from RGB image frames. This is done by concatenating the original RGB input and the generated depth map prediction, and training a second GAN to map the concatenation to a more refined depth map. In our case, more advanced derivatives of the cGAN family may be able to generate more complex predictions, which is necessary since our data is very ideal compared to real world sensor data.

Another important follow-up to this study is to improve the evaluation metric. Here, the MSE did not provide useful information adding to the modified SSIM, but a combination of evaluation metrics is likely necessary to obtain more accurate evaluations, especially if more complex heatmaps are evaluated. For example, MSE might better capture differences in scale of such heatmaps, while the SSIM captures structures are correctly. In previous sections, some drawbacks of using only the modified SSIM metric has been discussed, such as the SSIM immediately giving a stronger response for a bad network if more logs are graspable in the input data. One solution is to use a specific number of graspable logs for validation, but this is specific to the current heatmaps in which there is no gripability gradient over individual objects. It would also be beneficial to separate false positives (FP) from false negatives (FN), since grasping objects that are not graspable is of more concern than grasping a graspable object that is not the currently most optimal.

A crucial next step is also to address what kind of gripability maps are necessary in the real-world application, and feasible for ground-truth generation from available data. Should they be more or less complex, i.e. should they consist of a continuous heatmap over the entire view or highlight the most graspable object immediately? What metric should be used to define gripability? This is not trivial, as gripability depends on many external criteria, including how such a metric conveys the capacity of the grapple and what defines a

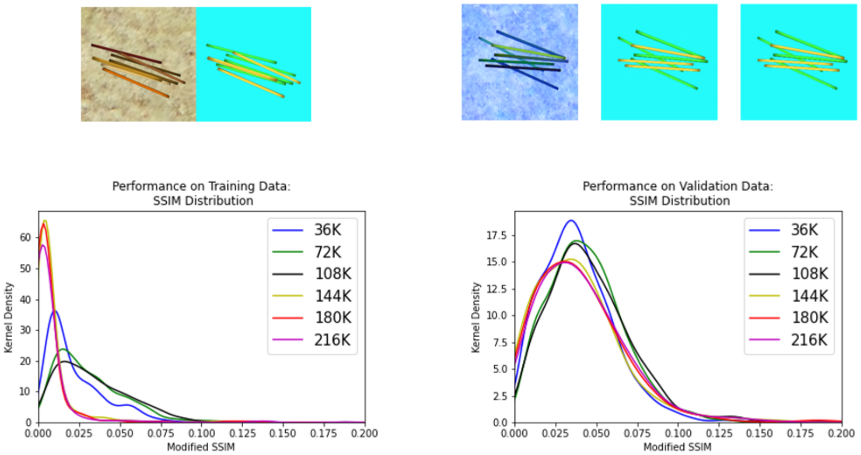


Figure 17: Segmented, well-aligned piles on a forest background.

successful grasp, which will depend on details that may not be visible from pure RGB data - such as the density or mass variations between individual logs, or other physical properties. This is not taken into account by our approach, but using realistic data and a sophisticated gripability metric it should be possible for an intelligent system to build intuition for the specific task. Other questions concern how the solution to this subtask fit into the overall grasping pipeline. For example, it is important that the gripability map is not dependent on the particular sensor view, since what is graspable cannot change arbitrarily from the point of view of the agent if the sensors move along with it during operation.

Many questions are left to be answered, but the current investigation has shown that there is potential in accurately predicting gripability in a scene solely based on RGB and depth data, successfully posing the problem of gripability map prediction as an image-to-image translation problem.

9 Bibliography

Armanious, K., Yang, C., Fischer, M., Küstner, T., Nikolaou, K., Gatidis, S., and Yang, B. (2020). Medgan: Medical image translation using gans. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 79:101684.

Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., Downs, L., Ibarz, J., Pastor, P., Konolige, K., Levine, S., and Vanhoucke, V. (2018). Using simulation and domain adaptation to improve efficiency of deep robotic grasping. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4243–4250.

Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale gan training for high fidelity natural image synthesis. *ArXiv*, abs/1809.11096.

Chou, C.-J., Chien, J.-T., and Chen, H.-T. (2018). Self adversarial training for human pose estimation. *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 17–30.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2:2672—-2680.

Isola, P., Zhu, J., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. *ArXiv*, abs/1710.10196.

Lee, S., Capuano, V., Harvard, A., and Chung, S.-J. (2020). Fast uncertainty estimation for deep learning based optical flow. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10138–10144.

Lore, K. G., Reddy, K., Giering, M., and Bernal, E. (2018). Generative adversarial networks for depth map estimation from rgb video. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1258–1288.

Ma, N., Wang, J., and Meng, M. (2020). Conditional generative adversarial networks for optimal path planning. *ArXiv*, abs/2012.03166.

Matsuzaki, Y., Okayasu, K., Nakamura, A., and Kataoka, H. (2017). Generated motion maps.

Patzelt, F., Haschke, R., and Ritter, H. (2019). Conditional wgan for grasp generation. In *ESANN*.

Pedersen, O.-M., Misimi, E., and Chaumette, F. (2020). Grasping unknown objects by coupling deep reinforcement learning, generative adversarial networks, and visual servoing. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5655–5662.

Rubin, J. and Abulnaga, S. M. (2019). Ct-to-mr conditional generative adversarial networks for ischemic stroke lesion segmentation. *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–7.

Uricár, M., Krizek, P., Hurych, D., Sobh, I., Yogamani, S., and Denny, P. (2019). Yes, we gan: Applying adversarial techniques for autonomous driving. *ArXiv*, abs/1902.03442.

Wang, T., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. (2018). Video-to-video synthesis. In *NeurIPS*.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916.

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