Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра інформатики факультету інформатики

ВИДАЛЕННЯ ТІНЕЙ ІЗ ЗОБРАЖЕННЯ ЗА ДОПОМОГОЮ ГЕНЕРАТИВНИХ ЗМАГАЛЬНИХ МЕРЕЖ ТА НАВЧАННЯ БЕЗ УЧИТЕЛЯ

Текстова частина до курсової роботи
за спеціальністю „Комп'ютерні науки" 122

Керівник курсової роботи
с.в. _Бучко О.А.__
(прізвище та ініціали)
_____
(підпис)
"____" _____ 2020 р.

Виконав студент _____
___Андронік В.П._____
(прізвище та ініціали)
"____" _____ 2020 р.

Київ 2020

Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра інформатики факультету інформатики

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ
на курсову роботу

студенту Андроніку В.П. факультету інформатики третього курсу
ТЕМА ВИДАЛЕННЯ ТІНЕЙ ІЗ ЗОБРАЖЕННЯ ЗА ДОПОМОГОЮ ГЕНЕРАТИВНИХ ЗМАГАЛЬНИХ МЕРЕЖ ТА НАВЧАННЯ БЕЗ УЧИТЕЛЯ

Вихідні дані:
-
-
Зміст ТЧ до курсової роботи:

Індивідуальне завдання
Вступ
1. Аналіз задачі та огляд існуючих робіт
2. Аналіз компонентів та схожі роботи
3. Розбір та опис розглянутого алгоритму. Практичне застосування.
4. Експерименти та аналіз результатів

Висновки
Список літератури
Додатки (за необхідністю)

Дата видачі „___” _____ 2020 р. Керівник Бучко О.А.
(підпис)

Завдання отримав _____
(підпис)

Тема: Видалення тіней із зображення за допомогою генеративних змагальних мереж та навчання без учителя

Календарний план виконання роботи:

№ п/п	Назва етапу дипломного проекту (роботи)	Термін виконання етапу	Примітка
1.	Отримання завдання на курсову роботу.	01.11.2019	
2.	Аналіз технічних матеріалів за темою.	01.01.2020	
3.	Розробка та програмування алгоритму.	15.02.2020	
4.	Виконання порівнялього аналізу представлених гіпотез та існуючих методів з видалення тіней за допомогою генеративних нейронних мереж.	01.04.2020	
5.	Написання пояснювальної роботи.	04.04.2020	
6.	Коригування виконаної роботи.	20.04.2020	
7.	Написання курсової роботи та відповідної презентації для доповіді.	01.05.2020	
8.	Остаточне оформлення роботи та слайдів.	05.05.2020	
9.	Захист курсової роботи.	20.05.2020	

Студент Андронік В.П.
Керівник Бучко О.А.
“ ” _____ _____

Table of contents

Abstract

This material presents the solution for shadow removal task using generative adversarial networks. Our approach is trained in unsupervised fashion which means it does not depend on time-consuming data collection and annotation. This together with training in a single end-to-end framework significantly raises its practical relevance.

Taking the existing method for unsupervised image transfer between different domains we researched its applicability to the shadow removal problem. By exploiting attention modules and multi context feature aggregation using dilated convolutions our method gives significant results compared to existing solutions in the field.

Keywords: generative adversarial networks, unsupervised learning, shadow removal, shadow generation, attention module, dilated convolutions.

Introduction

Shadow is a common visual phenomenon when the object overlaps illumination source. Detected shadows can provide the important clues for better visual scene understanding[1,2]. However, they can degrade the performance of algorithms in several computer vision spheres as object detection[3], tracking[4] and intrinsic image decomposition[5]. Therefore, effective shadow removal could give a performance boost for these tasks.

Previous works can be divided into two groups: classical and deep learning-based. Classical solutions used user input or hand-crafted features for shadow detection[6,7] after which they tried to make the shadow region match the background. Meanwhile, deep learning-based approaches use neural networks for extracting high level features and background filling. One of the early works[8] used three different neural networks operating in different contexts for more quality features extraction. Later, Hu et al. [9] explored direction-aware spatial context for this task to compensate the lack of data. After that, Wang et al.[10] used adversarial learning by stacking two networks together where one is used for shadow detection and one for shadow removal. More recently, Ding et al.[11] constructed the framework which used attention maps and recurrent learning. Other approach[12] argued that earlier works were not directly constructed for shadow removal task and proposed the novel architecture with hierarchical features aggregation.

However, all these methods used the supervised data to train and thus demanded the tedious collection and annotation processes. For the similar reason these approaches are also constrained with the complexity of the scenes. It is also argued that such approach may lead to change in illumination between shadow and shadow-free images[13].

Recently, the unsupervised solution was presented[13] where the problem was formulated in unsupervised mapping learning between two domains - shadow and shadow-free - using CycleGAN[14] framework. Hence, this approach only demands the dataset with shadow/shadow-free pairs of images which are not aligned. The

method shows capacity reaching the competitive results with its supervised counterparts.

To the contrary, this approach is not directly designed for shadow removal purposes so it has problems with artifacts on generated images and also with complex scenes due to binary masks usage.

The aim of this work is to construct the end-to-end framework for solving the shadow removal task in unsupervised manner using generative adversarial networks. We will explore how different architectural and training improvements affect the generation results both quantitatively and qualitatively.

The works consists of four sections.

The first one is analyzing the current state of research together with a problem itself. It also provides a brief overview on how others deal with this task from the traditional and deep learning-based points.

The second section describes different components of the solution and discusses the related works. Its main aim is to introduce the reader to different concepts and uncover the reasoning of how they are connected to the solution.

Third part describes the particular method together with network architectures and specific details on training.

Last section shows the experimental results and compares the presented solution with others in the sphere. Moreover, analysis on the results is also provided.

Overall, the main tasks are:

1. Analyze the domain and existing solutions.
2. Construct an algorithm for unsupervised shadow removal with the use of generative adversarial networks.
3. Provide experimental results and comparison with other methods.

Main part

Section 1. Shadows detection and removal analysis

1.1 Problem analysis and research overview

Shadows are permanent visual phenomenon which is described with interactions among objects and materials. Shadows could both help with scene interpreting or confuse it if we ignore them. Shadows detection can help with more accurate object localization, elaborating where object touches the surface and also determine object shape. They also give important clues about illumination conditions[15] and scene geometry[16]. Meanwhile, ignoring the shadows confounds the scene interpretation due to spurious edges at the boundaries of the shadows and the confusion between shading and albedo.

Shadow removal is a very challenging task because it is not enough to detect and remove the shadow, we also need to fill the background so it looks naturally for both human and a computer system.

Traditionally, this task is solved in a pipeline with a shadow detected first and then its region filled. Shadow detection is a complex task due to sophisticated scene geometry, illumination and albedo. Thus, it is hard to say from the local perspective if the dark surface is shaded or it is an albedo effect(Figure 1.1). So researchers must compare the shadow region with the others of the same material by looking at the differences in pixels intensity, texture and color. For this task physical models of color and illumination are often used[17,18].

After the shadow is detected we need to somehow use the background to fill the shadow region. In some works, this problem was posed as matting problem[6,19,20], where they tried to define how much of light is occluded and using this information to relight the whole image.

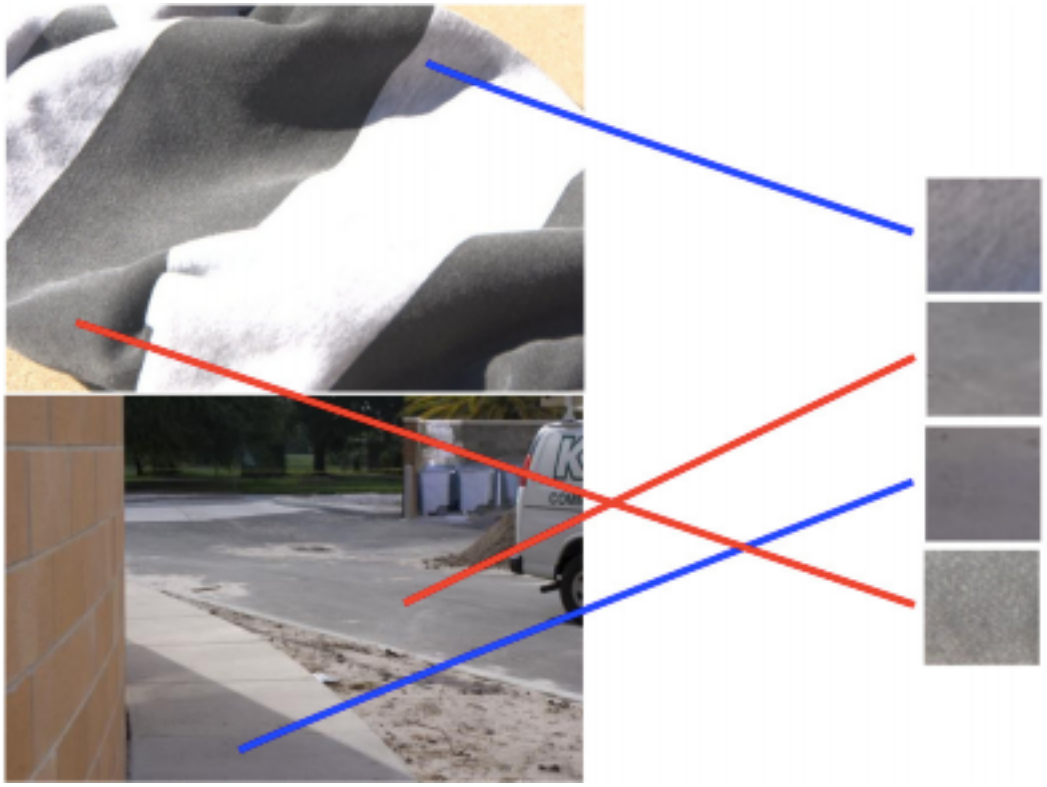

Figure 1.1: Hard to tell where the shadow is from the local region perspective. Illustration from [6].

Meanwhile, nowadays de-facto standard for those tasks are deep learning-based methods. With their ability to extract semantic features from images they outperform the traditional algorithms by a significant margin.

These algorithms solve the task in an end-to-end manner learning the mapping between shadow and shadow-free images domains. Qu et al.[8] uses three different subnetworks for features extraction from different contexts. They use one global network which is a fine-tuned pre-trained VGG16 network[21] which extracts features from the initial image. Then, there are two networks to operate in parallel given the outputs from global network shallow layers. Each of them is meant to adjust the coarse features from the global context to derive more quality result. Their results provided the insights on the applicability of neural networks to this problem.

Other work[9] that is worth mentioning used direction-aware spatial context for shadow removal and detection. Their approach is based on the idea that was outlined

above that shadow detection should be tackled by comparing the neighboring regions of the image. The results of this work shows the capacity of using recurrent modules as they give consistent results by adjusting the previous results each time. This is important especially for boundaries removal as they might not be successfully tackled at once. They also admit that supervised datasets have changes in illuminations and color between shadow and shadow-free images that is why they adjusted the loss function additionally. We will return to this issue in the next section as it justifies the practical value of unsupervised approach.

The next two approaches are particularly relevant for our own research as soon as they use adversarial learning. Wang et al.[10] were first to solve the detection/ removal tasks in a single framework. They stacked two generative adversarial networks (GAN)[22] where the first one solved the detection task while the other used the shadow matte from the first one and ground-truth image to remove the shadow. They showed how Conditional GANs (CGAN)[23] could be effectively used for shadow domain and how multi-task approach outperform the preceding ones where the problem was divided into parts. However, their method demands the dataset of triplets - shadow, shadow binary mask and shadow-free images - which significantly complicates and constrains its applicability.

Meanwhile, next work[11] combined the recurrent modules and adversarial learning for more effective training where recurrent training is used for progressive attention map updating and improving the generation result. That was a second work which successfully used the recurrent learning which may tell us about its potential. One of the major contributions of their work provides the framework of using semi-supervised learning for tackling complex scenes and unannotated images.

One more work[12] which achieved state-of-the-art argued that previous works' architectures were not correctly constructed to solve the removal task though we have already seen that much of research in the field was conducted. They proposed the novel architecture which effectively exploits hierarchically connected attention mechanisms together with multi context feature aggregation. This method also uses synthetic shadow images for complementing the existing datasets and it helps to generalize quite well. The key thing about their work is the use of multiple skip

connections which together with exponentially growing receptive field helped them to outperform previous works by a large margin.

All the works we have outlined above are solving the problem with supervised mapping learning between the domains. Meanwhile, there are few works in the field which conducted the research on the use of unsupervised learning. The only research paper[14] we have found used adversarial learning with cycle consistency loss for unsupervised mapping between domains. This approach is fully based on CycleGAN[13] with modifications undertaken to adjust it to shadow domain: first CGAN detects and removes the shadow while the other tries to generate it given the image and shadow mask as the input. Shadow mask, therefore, is received by running the Otsu's algorithm[24] on the difference between generated and input image. The final results are competitive with those outlined above and significantly higher than general image-to-image translation with CycleGAN.

However, this approach is not directly constructed to fit the shadow removal task, thus it has problems such as leaving the artifacts on the shadow boundaries and using binary masks for detected shadows. One more problem that may have consequences in a real-world application is that shadow generator network chooses the mask at random, so we could receive the image with inappropriate shadow on it.

1.2 Problems with current state of research

To conclude, the works we have outlined in the previous section are not the only in the field but seems to be essential to tell about. These methods have one thing in common - they are trained in a supervised manner and depend on tediously annotated datasets. One should fix the camera and then manually remove/add objects to get a pair of shadow and shadow free images. This approach also constrains the number of possible scenes to compose. For example, Wang et al.[10] collected the dataset with 1870 images under 135 scenarios by adding/removing objects like umbrella, tree branch etc. For this reason we cannot add the images with big objects like buildings, trees or construct the scenes with a complex geometry. This is also an issue in dynamic systems where it is hard to obtain the same scene twice as in self-driving cars computer vision systems.

One more thing to add is that such datasets could have the difference in luminosity and color between shadow and shadow free images(Figure 1.2). This effect was admitted in earlier works[9,14] where this issue was adjusted[9] or used as the reason for looking for another method[14].

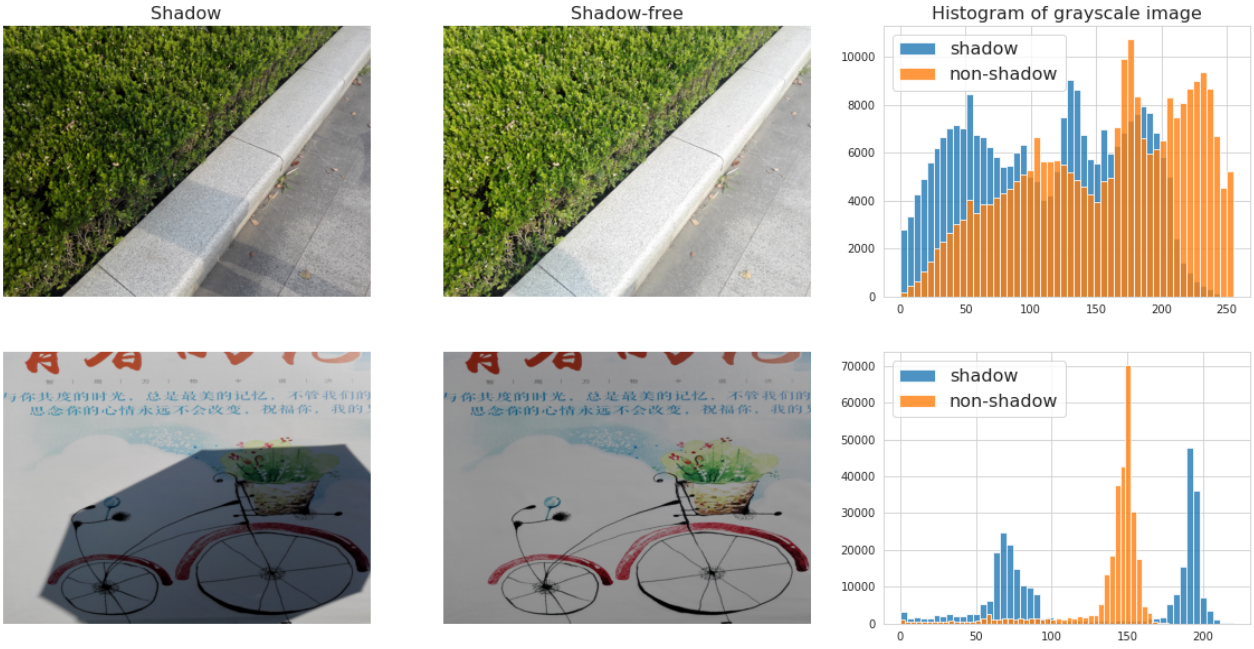

Figure 1.2. Change in the color between shadow and shadow-free images comparing non-shadow regions. ISTD[12] dataset.

1.3 Conclusion

In this section we have stated the problem of shadow removal and detection tasks, analyzed it and overviewed different computer vision methods to solve it. We have divided the previous works into traditional and deep learning-based where the latter is widely used now whereas the former conducted the relevant research into it.

Deep learning methods outperform the traditional ones by a large margin, however, at the current state of research, they lack efficient unsupervised solution. For that reason, most of community datasets are collected in a supervised manner by fixing the camera and adding/removing the objects to obtain a pair of shadow, shadow free images. We pointed out how this approach complicates the preprocessing phase and constrains the possible number of scenes. Figure 1.2. also illustrates the

issues in color change between the shadow and shadow free images that might degrade the performance of corresponding algorithms.

With that in mind, we showed the need for further research in unsupervised solution of this problem which will be covered in more details further.

Section 2. Related work

This section will briefly cover the main components of the solution for better understanding of subsequent work and for acquiring the reasoning of the research.

2.1 Generative adversarial networks (GAN)

GAN is a framework for estimating the generative models that was firstly proposed by Ian J. Goodfellow et al.[22]. It consists of two networks where generator network G captures the data distribution and discriminator network D estimates the probability of the sample came from data distribution. In an initial variant G took the noise vector(usually, from uniform distribution) as the input trying to generate a sample that will «trick» the discriminator D. This framework corresponds to a minimax game(two-player non-cooperative game) where G is maximizing the probability of D making mistake(see Figure 2.1 for illustration). To bring it formal we can consider such minimax game in which we should optimize a following loss function:

$$\min_G \max_D L(D,G) = E_{x\sim p_r(x)}[logD(x)] + E_{z\sim p_z(z)}[1 - logD(G(z))] \quad (2.1)$$

where p_r is real data distribution over sample x while p_z is a noise distribution over noise z.

This work provoked a significant amount of research due to its capacity to generate high-quality samples. It was further improved by introducing Conditional GAN[23] which uses the label information to present the multi-modal solution. Thus, researcher can tune what kind of sample the network should generate.

CGANs were successfully used in learning the mappings between the domains, for example pix2pix[25] approach can handle multiple vision tasks as day to night, summer to winter or aerial to map in a single framework by introducing the pairs of images from each domain. However, it cannot handle the domains with no one-to-one mappings as for style transfer etc.

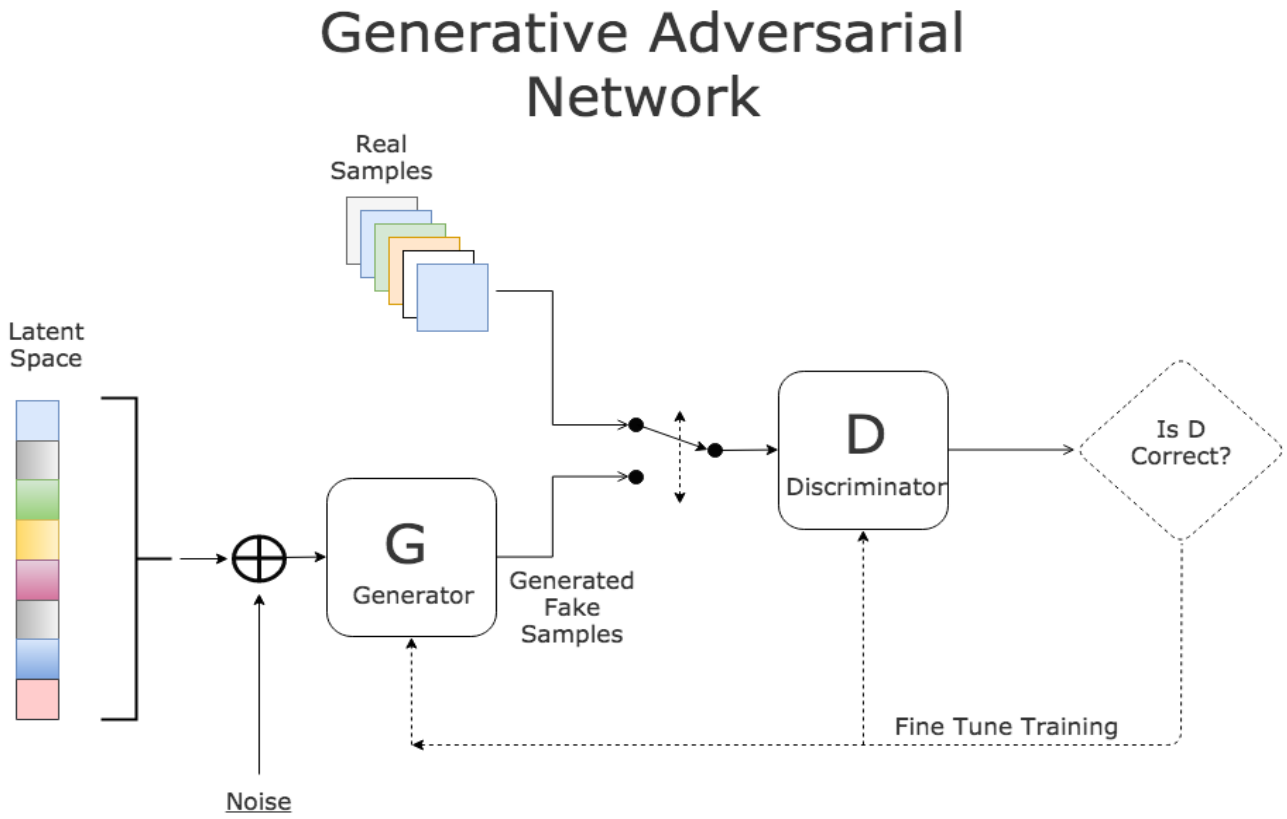

Figure 2.1: Illustration of GAN working process[39].

Hence, for unsupervised domain mapping CycleGAN[13] was firstly proposed by using the cycle consistency loss between domains. Still, it suffered from mode collapse generating one sample for different inputs. For that reason there was a research conducted[31,32] to extend the initial solution to cope with the «many-to-many» mapping with the use of latent variables.

There has been a significant amount of work done and now GANs can generate high-quality images that are hardly distinguishable from the real ones. They are particularly good at face generation[26,27], style transfer[13,28], inpainting[29,30], domain transfer/adaptation[13, 31, 32] and are also used for shadow removal/ detection[10,11].

One more thing to cover is how GANs are trained and what challenges one may encounter doing it. GANs' training requires finding a Nash equilibrium in a high-dimensional parameter space(so that *D* assigns equal probabilities to generated and real samples). GANs are typically trained with gradient descent which is not directly constructed for finding a Nash equilibrium, that is why they frequently suffer from stability issues and non-convergence[33]. For that reason researchers use dozens of

heuristics and architectural improvements to stabilize the training and solve the problems with non-convergence[34,35,36,37,40]. These works and heuristics are aggregated[38] so one could reference it when tackling a problem.

2.2 Unsupervised domain adaptation

Domain adaptation is a field of machine learning where we are learning the mapping between source and target distributions. We will consider a deep domain adaptation with the use of adversarial training because this approach is of particular interest to current work.

This method is learning from unpaired data so one can collect a reasonable amount of images for each domain and the algorithm would learn the underlying mapping between them. The idea behind that was firstly formulated in CycleGAN work[13]. Our goal is to learn the mapping G between source X and target Y domains: $G : X \rightarrow Y$ with the use of adversarial learning (2.1). However, such mapping is highly under-constrained so they proposed to learn an inverse mapping $F : Y \rightarrow X$ and to add a cycle consistency loss:

$$L_{cyc}(G, F) = E_{x \sim p_r(x)}[\|F(G(x)) - x\|_1] + E_{y \sim p_r(y)}[\|G(F(y)) - y\|_1] \quad (2.2)$$

For each image x from domain X we want to bring it back using the inverse mapping: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. That is called a forward cycle consistency, the same situation is for a target domain: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$ which is called a backward cycle consistency(see Figure 2.2).

It is argued that adversarial loss alone cannot map the individual input to a desired output. For that reason cycle consistency is intended to further reduce the possible mapping functions and is a key component for this framework.

However, such architecture is vulnerable to mode collapse when the generator cannot represent complex real-world distribution and get stuck in a low variable space. It also learns the one-to-one mapping, so the network will generate the only sample for each input image that may constrain variable mappings like style transfer.

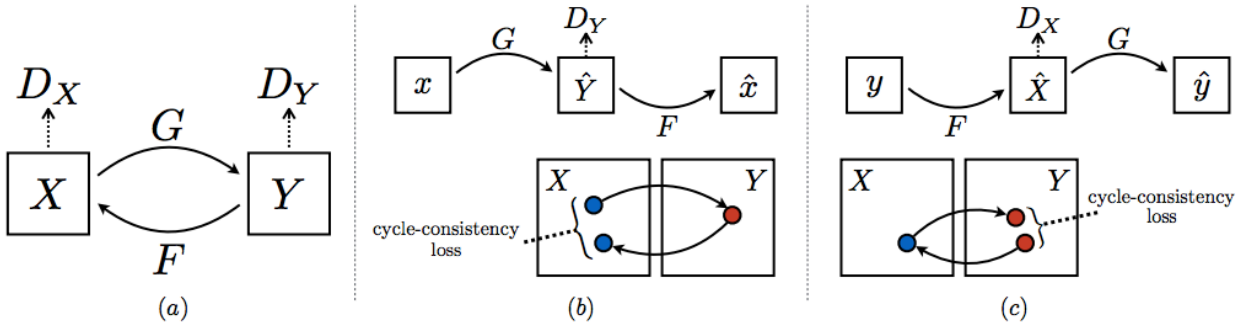

Figure 2.2: We can see the mapping between functions where Dx and Dy are corresponding discriminator networks (a). (b) and (c) are forward and backward cycle consistency correspondingly[13].

There are solutions[31,32] which extend CycleGAN to many-to-many mapping by projecting on latent variables, but for shadow removal the base approach fits just fine.

This approach proved to be a successful solution for the domains with the tractable mappings. While the existing methods for unsupervised image-to-image[41,42] are domain specific with explicit distance metrics, CycleGAN is a general-use framework. This together with existing application of MaskShadowGAN[14] are the main reasons why we decided to use this framework.

2.3 Dilated convolutions

Dilated convolution was firstly proposed for the use in wavelet decomposition algorithm[44], then it was used for multi-scale context aggregation[43] to compose a more appropriate architecture for dense prediction.

Dilated convolution operator is a more general concept than a discrete convolution. If we bring it formal, let $F : \mathbb{Z}^2 -> \mathbb{R}$ be a discrete function. $\Omega_r = [-r, r]^2 \cap \mathbb{Z}^2$ and $k : \Omega_r \to \mathbb{R}$ is a discrete filter of size $(2r + 1)^2$. Then discrete convolution operator is defined as:

$$(F * k)(p) = \sum_{s+t=p} F(s)k(t). \quad (2.3)$$

While the dilated convolution is defined as follows:

$$(F *_l k)(p) = \sum_{s+lt=p} F(s)k(t). \quad (2.4)$$

We can see that discrete convolution is simply an 1-dilated convolution. Dilated convolutions support an exponential growth of receptive field with no loss in resolution or coverage. It was firstly proposed as the substitution for pyramidal networks for dense prediction task(i.e. segmentation)[43].

Informally, dilated convolutions are inflating the kernel by inserting zero-values between its elements(Figure 2.3). Usually, l-dilated convolution means that there are $l - 1$ spaces between kernel elements. That is why 1-dilated convolution means a regular discrete convolution (2.3). Dilated convolutions are generally used for increasing the receptive field «cheaply»(number of parameters grows linearly while receptive field - exponentially, see Figure 2.4). We argue that dilated convolutions are also well suited for shadow removal task because we need to estimate the non-shadow background that is usually the major part of the image.

For that reason we divided our bottleneck layer into two parts where the first one is operating on the local level with small receptive field — presumably for shadow detection — while the other one is used for exponential receptive field growth(by using linearly growing dilation factors) to estimate the background and fill the detected shadow region in a more efficient way.

Our assumption is also reinforced by the fact that dilated convolutions are actively used for image inpainting task[30], where one needs to restore the deleted patch of the image using the remaining part. These two tasks seem to have much in common and this is a reasonable factor to consider.

Moreover, state-of-the-art work[12] in shadow removal uses the dilated convolutions inside with «extreme» receptive field growth. Together with multiple skip connections their approach reaches high qualitative level of generated images.

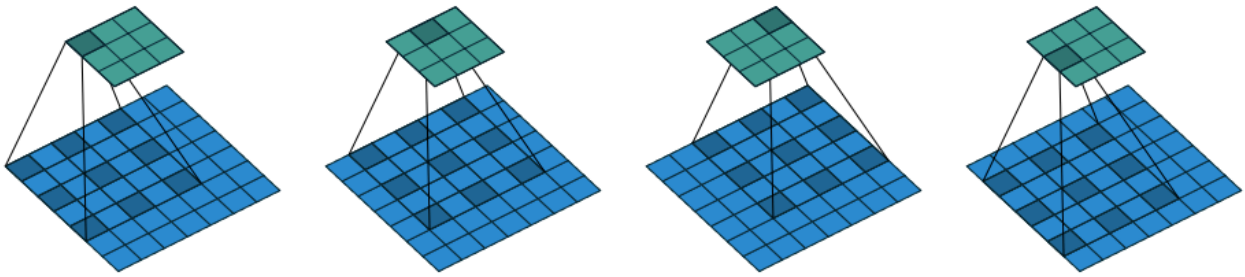

Figure 2.3: Convolving a 3x3 kernel over a 7x7 input with a dilation factor of 2[45].

2.4 Attention models

Attention mechanisms are now used in most of deep learning architectures, especially in natural language processing[46,47]. They are particularly useful for models which must capture the global context. We want to present a self-attention mechanism as the relevant example of attention models and the one which is used in GANs[37]. Self-attention helps with estimating the internal model states and in the context of GANs can help to efficiently learn the global, long-range dependencies[37]. There are also many different applications of it reaching state-of-the-art results[48,49,50].

Self-attention[37] is consisted of three layers operating in parallel and aggregating the results with the use of softmax function. This approach, involving

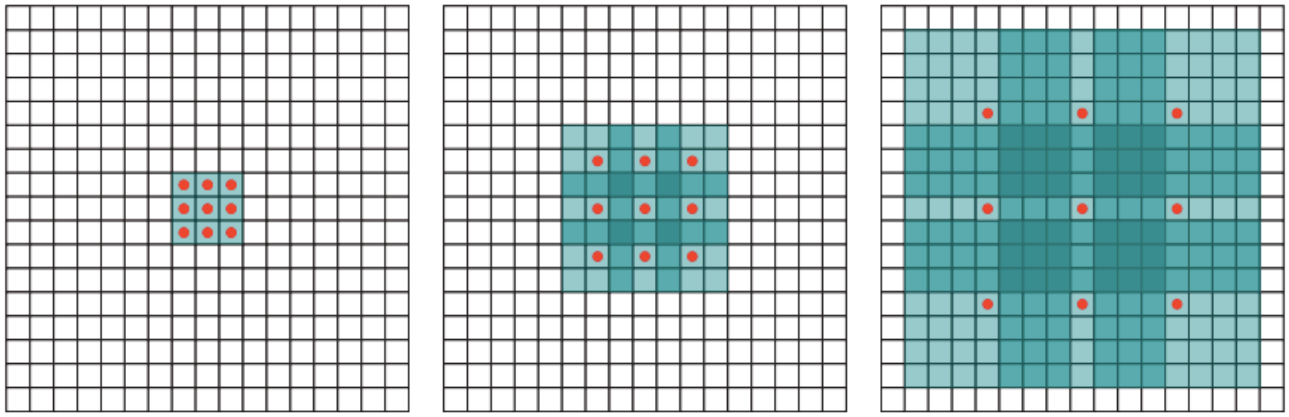

Figure 2.4: Sequential dilated convolutions and the corresponding receptive field growth given the same number of parameters.
l=1(left), l=2(middle), l=4(right)[43]

both discriminator and generator networks, helps to efficiently model the relationships between widely separated regions.

Another method[51] is directly linked to our work constructing a framework for unsupervised domain adaptation with cycle consistency and Class Activation Maps(CAM)[52]. The CAM uses the global pooling in CNN to obtain the discriminative features for different classes. In the context of domain adaptation it is used to distinguish two domains by weighting the specific channels of the image features inside the network. The CAM weights are directly optimized to distinguish images from each domain and are integrated into both generator and discriminator networks[51].

In our work we tried both self-attention[37] and CAM[52] where the latter gave more flexibility and is a more natural to use in the context of unsupervised domain adaptation. This approach is meant to complement the binary mask and to help with shadow detection for more complex scenes. The main problem with binary masks is the binary component because the shadow intensity is not uniform and tends to decline near the boundaries while increasing at the intersection of multiple shadows. We tested many hypothesis about the specific location this module should be integrated, which weight to assign and also researched how the learned attention map might be transferred between networks. The training process also should be tuned to troubleshoot the issues with stability and non-convergence problem.

Section 3. Method

This section will formally present the method for shadow removal task, network architectures and specific practical tricks for training and tackling the corresponding instability and non-convergence issues. We divide training into two parts: one to learn from shadow images(Section 3.1) and one to learn from shadow-free ones(Section 3.2).

3.1 Learning from shadow images

3.1.1 Adversarial learning

Let I_s be an image from shadow X_s domain. We use a generator network $G_{s\to f}$ to translate an image to shadow-free X_f domain and obtain $\widetilde{I}_f$. $G_{s\to f}$ network also includes an auxiliary classifier η_s, where $\eta_s(x)$ represents the probability of x taken from shadow domain[51] (see Section 3.1.2 for more details).

Then, we use the corresponding discriminator D_f to discriminate whether the data comes from X_f or $G_{s\to f}(X_s)$. This network also consists the auxiliary classifier η_{D_f} that is aimed to solve the same task as the discriminator itself. We should also notice that Least Squares GAN[36] objectives are used for more stable training, thus the adversarial loss will look like this:

$$L^s_{GAN} = E_{x\sim X_f}[(D_f(x))^2] + E_{x\sim X_s}[(1 - D_f(G_{s\to f}(x)))^2] \quad (3.1)$$

3.1.2 CAM attention

Auxiliary classifier η_s is used in $G_{s\to f}$ to distinguish between domains and is inspired by CAM[52] that is described in Section 2.4.

Let C^k_s be a k-th feature map of $G^l_{s\to f}$ output from l-th layer. Then, $C^{k_{ij}}_s$ is the value at (i, j) position and we want to learn the importance weights for each feature map by using the global pooling layers(i.e. average, max). Thus, we obtain:

$$a_s = w_s * C_s = \{w_s^k C_s^k \mid 1 \leq k \leq n\} \quad (3.2)$$

where n is a number of feature maps and w_s^k is an importance weight for the k-th feature map.

We are learning those weights from:

$$\eta_s = \sigma(\sum_k w_s^k \sum_{ij} C_s^{k_{ij}}) \quad (3.3)$$

Formula 3.3 is an example of the global average pooling, but one have to test different pooling layers for their own task.

To make η_s distinguish between domains the corresponding cross-entropy loss is optimized:

$$L_{cam}^{s \to f} = -(E_{x \sim X_s}[log(\eta_s(x))] + E_{x \sim X_f}[log(1 - \eta_s(x))]) \quad (3.4)$$

Then, a_s (3.2) is transferred as the input to the following layer of the network and the learning continues.

Attention a_s is aggregated to be transferred as attention map A_s to $G_{f \to s}$:

$$A_s = \sum_{hw} \sum_c a_s^{hwc} \quad (3.5)$$

where we sum the values over the channels c.

Auxiliary classifier η_{D_f} is also integrated in D_f to decide whether the data comes from X_f or $G_{s \to f}(X_s)$:

$$L_{cam}^{D_f} = E_{x \sim X_f}[(\eta_{D_f}(x))^2] + E_{x \sim X_s}[(1 - \eta_{D_f}(G_{s \to f}(x)))^2] \quad (3.6)$$

3.1.3 Cycle consistency and identity loss

If we only use adversarial loss (3.1) for learning then the mapping is highly under-constrained. That is why we present the inverse transformation $G_{f\to s}$ to transform the images back and encourage the contents to be the same.

As we outlined above generator network $G_{f\to s}$ additionally takes attention map A_s(3.5) and generated mask M_s[14] as the input(concatenating to the image as additional channels). To preserve the consistency between the generated shadow image and the original one we take the same attention map and shadow mask extracted from shadow removal generator $G_{s\to f}$. This allows to produce multiple shadow images from one shadow-free raising the generalization capacity. Shadow mask is a binary map where -1 indicates non-shadow region while 1 — the shadow region. Attention map is also normalized to $[-1,1]$ and is used for complementing the shadow mask.

Then, we formulate following cycle-consistency loss:

$$L_{cycle}^{s\to f} = E_{x\sim X_s}[\,||\,G_{f\to s}(G_{s\to f}(x), A_s, M_s) - x\,||_1\,] \quad (3.7)$$

We are using an L1 norm as it is stated[25] to be efficient in comparing the low frequency signals in images.

However, using only adversarial and cycle-consistency losses gives the generators freedom to change colors on images without being penalized. That is why researches in original work[13] introduced an identity loss to regularize the generators to be near an identity mapping when the inputs from target domain are provided. Furthermore, this approach allows our solution to remove/generate shadows only when the image from proper domain is given:

$$L_{idt}^{s} = E_{x\sim X_s}[\,||\,G_{f\to s}(x, A_n, M_n) - x\,||_1\,] \quad (3.8)$$

where A_n, M_n are constructed only from -1(non-shadow) which penalizes the network for generating the shadows on images where the shadow is already presented.

3.2 Learning from non-shadow images

Given the generator network $G_{f\to s}$ described above and also attention map A_s (3.5) together with shadow mask M_s we can define corresponding losses for inverse transformation. We have the same adversarial loss where generator is maximizing the probability of discriminator to make mistake:

$$L_{GAN}^{f} = E_{x\sim X_s}[(D_s(x))^2] + E_{x\sim X_f}[(1 - D_s(G_{f\to s}(x)))^2] \quad (3.9)$$

However, we do not integrate the CAM module into the inverse transformation networks due to stability issues and because this approach gives better results in experiments.

The cycle-consistency constraint also stays the same: we generate shadow image from shadow-free X_f and then using $G_{s\to f}$ to restore the image back and optimize the networks:

$$L_{cycle}^{f\to s} = E_{x\sim X_f}[||G_{s\to f}(G_{f\to s}(x, A_s, M_s)) - x||_1] \quad (3.10)$$

Finally, we adopt the $G_{s\to f}$ to produce shadow free image given the real shadow free image from X_f. That means that we encourage the network to not remove anything if there is nothing to remove.

$$L_{idt}^{f} = E_{x\sim X_f}[||G_{s\to f}(x) - x||_1] \quad (3.11)$$

3.3 Maps generation

3.3.1 Shadow mask generation

Our generator $G_{f\to s}$ uses the shadow mask as the input, so we can condition network with it and generate multiple shadows from one shadow-free image. We follow the same approach as [14] and construct the threshold binarizer B between generated shadow free image $\widetilde{I}_f$ and original image I_s:

$$M_l = B(\widetilde{I}_f, I_s) \quad (3.12)$$

Thus, when we obtain a pair of images we compute the difference $\widetilde{I}_f - I_s$ and compute the threshold to assign the values greater that it as 1 and those less - with -1. The threshold is computed using Otsu's[24] algorithm which separates the shadow from non-shadow regions by maximizing the intra-class variance.

During the training process we add those masks to the Queue and save last shadow masks to input them to generator network. Hence, to get a mask we select random sample and pushing the new ones by replacing the older ones. This approach helps to reduce the model oscillation while making more diverse results.

3.3.2 Attention map generation

As was described above in Section 3.1.2, attention map A_s is received from auxiliary classifier η_s by applying the pooling operation to feature maps. In our approach we are using average(GAP) together with max pooling(GMP) layers to get the complete picture. GAP is able to find *all* discriminative regions on the image while GMP is encouraged to find only one[52]. So we decided to combine the best from two worlds by applying both GAP and GMP and concatenating the corresponding results.

So, additionally to (3.3) we have:

$$\eta_m = \sigma(\sum_k w_s^k \, max(C_s^k)) \quad (3.13)$$

After that we concatenate the outputs of η_s and η_m and feed them to $1x1$ convolutional layer with the following ReLU non-linearity to restore the input dimensions due to channel axis concatenation.

The output from these operation is followed by aggregation (3.5) to obtain an attention map A_s which we additionally scale to $[-1,1]$ range for invariance.

During the training process those maps are generated for each shadow image in the same way as binary shadow masks. They are also added to the Queue data structure and are saved there in pair with the corresponding masks.

3.4 Losses

To conclude, we present the final loss function which is a weighted sum of adversarial, CAM, cycle-consistency and identity losses outlined above in both architectures:

$$\min_{G_{s\to f},G_{f\to s},\eta_s} \max_{D_s,D_f,\eta_{D_f}} = \lambda_{adv}(L^s_{GAN} + L^f_{GAN}) + \lambda_{cam}(L^{s\to f}_{cam} + L^{D_f}_{cam}) + \lambda_{cycle}(L^{s\to f}_{cycle} + L^{f\to s}_{cycle}) + \lambda_{idt}(L^s_{idt} + L^s_{idt}) \tag{3.14}$$

where $\lambda_{adv} = 1, \lambda_{cam} = 500, \lambda_{cycle} = 10, \lambda_{idt} = 5$.

3.5 Network architecture

The following subsection will present each network in separate. For schematic illustration one can refer to Figure 3.1.

3.5.1 Shadow removal generator network

We want to first discuss how the generator network is constructed. The architecture is following a Johnson et al.[53] and reminds the encoder-decoder architecture without skip-connections. Encoder is constructed from two downsample convolutional layers, it is important that there are no pooling layers and downsampling is implemented using convolutions with stride 2.

Then, we have a bottleneck layer where most of work takes place. It includes nine residual blocks with linear dilation growth starting from the sixth layer. Dilation factors should be tuned depending on the receptive field size. In our experiments, we have made an assumption that the first part of the network would extract the shadow region operating on the local level while the second part will be responsible for filling this region, thus it will need a background information. For that reason we added receptive field growth at the end but in the way it does not exceed the input image size.

Bottleneck layer also integrates an attention CAM module that we described above. We inserted it before the receptive field growth(i.e. shadow removal process takes place) so it would help to localize the shadow in a more efficient way.

Finally, decoder is here to restore the image back to initial size by the use of transposed convolutions, it is important that network is learning to make the downsample and upsample operations itself(see Figure 3.1(a))

3.5.2 Shadow free discriminator

We will remind that discriminator D_f network is used for discriminating the real shadow free images from those generated. Architecture for it is following the idea of PatchGAN[25] where the network is not looking at the whole image but on patches(usually, 70x70) of it deciding whether the patch is real or not. We additionally complemented it with CAM attention module which is trained to solve the same task as the discriminator itself. CAM attention is operating before the final layer. Discriminators are not using dilated convolutions(see Figure 3.1(b))

3.5.3 Shadow generator network

This network is also following the Johnson et al.[53] architecture and has dilated convolutions in it which may help in shadow generation, however we have not seen any difference in experiments. We did not use an attention module here because it exposes an unstable training.

Generator uses shadow free image together with attention map and shadow mask(binary map) where three of them are concatenated by channel axis. Attention and mask are scaled to $[-1,1]$ to improve invariance.

3.5.4 Shadow discriminator

Discriminator D_s is also a PatchGAN with 70x70 patches with no attention module in it.

3.6 Training strategy

We used ReLU non-linearity and reflection padding for generator networks. Discriminators exploited LeakyReLU with 0.2 slope as activation function. Instance normalization(IN) is utilized for all networks just after the convolutional layer. The exceptions are input and output layers where we want to encourage the networks to learn the normalization by themselves. We are adding Tanh function in generator network to output the values from $[-1,1]$ range.

All networks are using spectral normalization[54] as this proved to improve GAN learning. It constrains the Lipschitz constant by restricting the spectral norm of each layer. We added spectral normalization also to the generator network because it is argued to reduce the computational cost of training[37].

As soon as spectral normalization is regularizing learning and, thus, makes it slower the TTUR[55] is utilized. We also have used different heuristics to stabilize training by smoothing the target label from 1 to 0.9[34] and dividing the discriminator loss by 2 to further address slow learning. Moreover, we are updating the discriminator with the history of previously generated samples as it is stated to reduce the model oscillations[13]. The same approach is used for shadow mask, attention pairs which are added to a separate Queue.

All the parameters are initialized using zero-centered Gaussian distribution with 0.02 standard deviation. Adam[56] is used for optimization with $\beta_1 = 0.5, \beta_2 = 0.999$. Learning rates are set as $r = 1^{-4}$ for generator and $r = 4^{-4}$ for discriminator network according to TTUR.

All the parameters are updated on 1 mini-batch and the whole training takes 200 epochs. Learning rates are linearly decaying after 100th epoch.

For data augmentation, we resize the images to $286x286$ cropping them randomly to be $256x256$ and flipping them horizontally with 0.5 probability. The method is implemented using PyTorch framework.

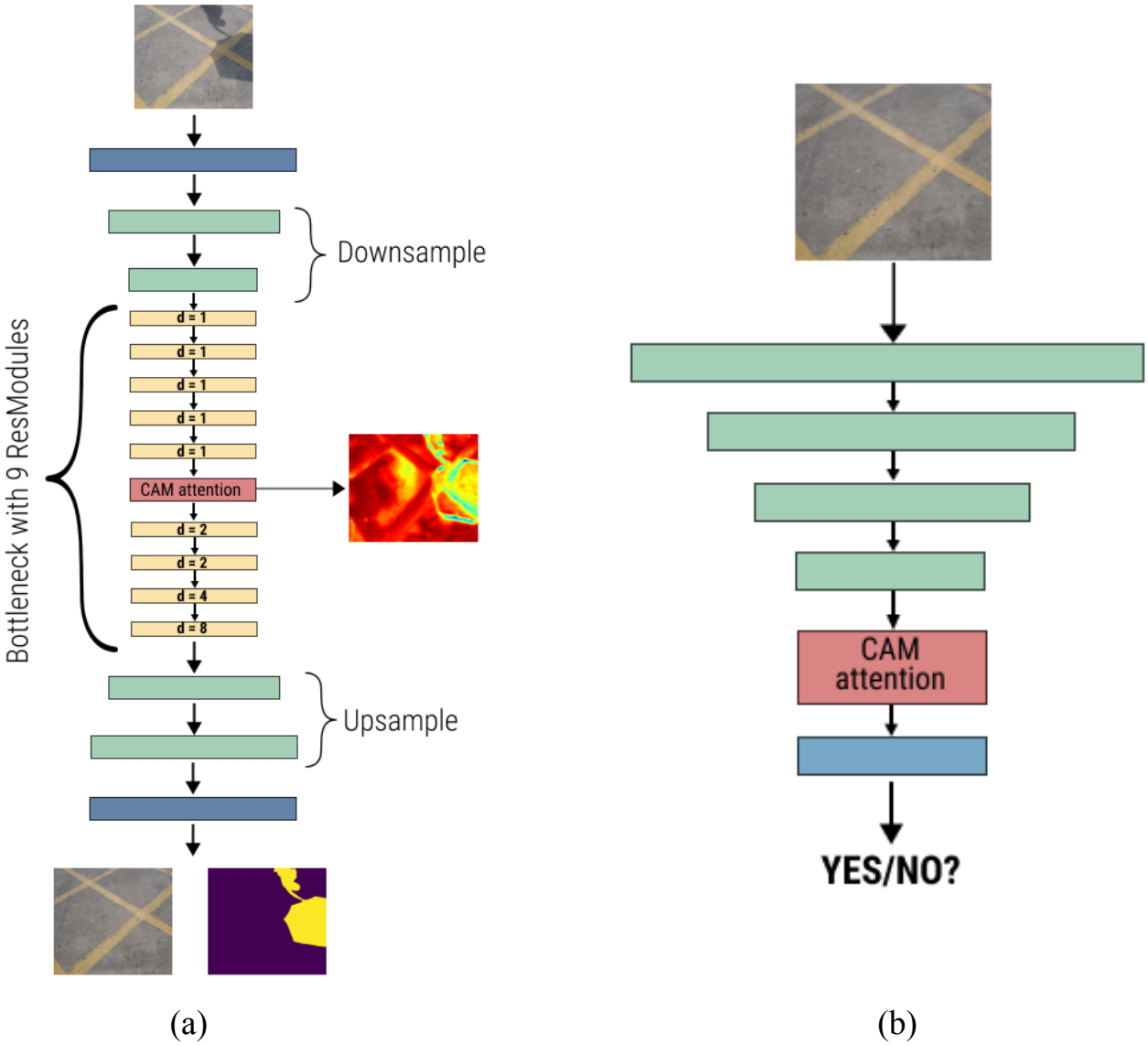

Figure 3.1. Schematic illustration of shadow removal network (a) and shadow free discriminator (b). The opposite networks are implemented in the same way except the corresponding CAM attention module.

Section 4 Experiments

4.1 Evaluation

Dataset. There are many community datasets for shadow detection and removal. Mostly, they are supervised and not large enough for deep learning solution but there are some which fits just well: ISTD[10], SRD[8]. There is also an unpaired one where more complex scenes are presented USR[14].

In our experiments we only use ISTD dataset for evaluation clarity because it has ground truth shadow free images as well as shadow masks. We does not correct our evaluation method to cover the issues with illumination and color change between shadow and shadow free images. In future, we will extend our solution to USR dataset as being more appropriate to our solution.

Metrics. In our evaluations we aim to estimate how good our network is in removing the shadows as well as detecting them. We are also concerned about the global image consistency. That is why we present three metrics.

For shadow removal we follow recent works [8,9,14] and use RMSE(Root Mean Squared Error) between generated and real shadow free images in LAB color space. Evaluation is divided into region and global where the former is applied to shadow regions while the latter to the whole images. In general, lower RMSE score tells about better results.

Shadow detection is evaluated with the use of IOU(Intersection over Union), also known as the Jaccard index, which is a widely used metric for image segmentation and object detection tasks. It is computed between the generated and ground-truth binary shadow masks by dividing the area of overlap by the union of those two. Greater IOU indicates better shadow detection.

4.2 Experimental results

Our method is trained on supervised dataset that is why the unpaired strategy is used: the first one is sampled from shadow domain while the second one is randomly chosen from the shadow free. We also selected random 100 images for validation purposes. During the experiments multiple hypothesis were tested and the most

successful of them are shown in Table 4.1. Other solutions from the field except the MaskShadowGAN[14] are not included but would be tested in the future.

As soon as we added two components to this work we wanted to test how they affect the generation quality.

We saw that method with dilated convolutions stays on roughly the same level as the one without them. However, we decided to take the one with dilated convolution as we expected it may improve the results after the attention would be added.

The majority of tests we conducted were about how to use an attention module and where to localize it. At first, we have added the attention to all networks following the original approach[51]. The results improved drastically given by RMSE score in both global and local contexts. However, IOU showed lower results for the reason we will go into details below. Adding attention module presented new instability issues while converging faster. We also encountered with messy outputs compared to initial methods

where the shadow removal generator not only identified the shadow but the background behind it. Thus, it could be seen that shadow detection quality decreased significantly indicated by lower IOU.

For these reasons, we removed an attention module from $G_{f\to s}$ and D_s which helped to suppress the instability but still it was a way higher than solutions without attention. The qualitative results improved a little but still there was a problem with over-detection.

Hence, the next solution added an attention map transfer from shadow removal generator to the input of the opposite one. We aimed to complement the binary mask with an attention map to make the learning more consistent. As the result, it reduced the model oscillations and raised the generated samples quality. However the shadow detection performance declined even more.

Attention map transfer showed the capacity to improve the results that is why we researched other ways to share this information with shadow generator network. For instance, shadow mask M_s was removed from the input of $G_{f\to s}$ so the network

Methods	Global RMSE	Shadow region RMSE	IOU
MaskShadowGAN[14] with our training strategy	3.0099	23.6703	**80.0894**
+dilation in $G_{s\to f}$ and $G_{f\to s}$	3.1253	26.9753	75.5382
+CAM attention in all G and D	2.3902	**15.0150**	73.8277
+CAM attention in $G_{s\to f}$ and D_f giving A_s as input to D_s (*)	2.3139	15.7047	71.1142
+A_s with no M_s to (*)	2.3261	15.5398	70.3951
+CAM weights from $G_{s\to f}$ to $G_{f\to s}$	**2.2436**	15.6087	70.7965

Table 4.1: Quantitative comparison of algorithms. Global RMSE tells the quality of shadow removal coupled with background restoration. Shadow region RMSE directly estimates the removal comparing shadow regions. IOU is aimed to evaluate the shadow detection quality comparing with ground-truth shadow mask. All scores are evaluated after the models are trained on validation set of images.

should have used the information from attention map A_s only. This resulted in again messy results with roughly the same metric values.

Rather than integrating the attention information to the input of the network, we researched the ways to insert it inside. After solid amount of experimenting we came up with transferring the CAM weights directly to rebalance the corresponding feature maps in $G_{f\to s}$. This improved the shadow removal performance, however increased the problem with over-detection.

4.3 Analysis and future work

Given the results in the Table 4.1 and the experiments provided in the previous section one could see a strong evidence that there is a trade-off between shadow removal and detection performances. Initial solutions do not use the attention map having the stable performance coupled with higher shadow detection quality, however they struggle to use the background information efficiently to fill up the

removed regions. For that reason they expose worse shadow removal quality. On the other hand, we have multiple solutions with attention showing much better shadow removal results while over-detecting the shadows. This leads to corrupting the generated samples which is particularly undesirable behavior in this task(see Figure 4.1 and 4.2 for details).

That is why in future work we are to modify the current dilated-attention architecture by complementing it with skip-connections[57], adaptive normalization[26,27,58] and style losses[28]. These all proved to be successful in the context of generative methods. At first, skip connections may help to solve the problem with messy results that might have been due to signal vanishing in the bottleneck layer. This could be a crucial point because latter layers(responsible for shadow removal) will use the information from the earlier ones(where the shadow is detected) in a more efficient way. This was proved to be the case in shadow removal too[12].

In our case the robust network architecture[53] was used as the backbone but we modified it adding the dilated convolutions and by rebalancing the inner features, this may have led to a faulty normalization that is relevant component for GANs.

Thus, by presenting the Adaptive normalization[58] we could end up with a more robust learning.

Last but not least, style losses[28] and style-generators[26,27] proved to be very successful for variety of applications. For that reason, we could try to use them for shadow removal problem, for instance, to encourage the networks to save the global context on the generated samples.

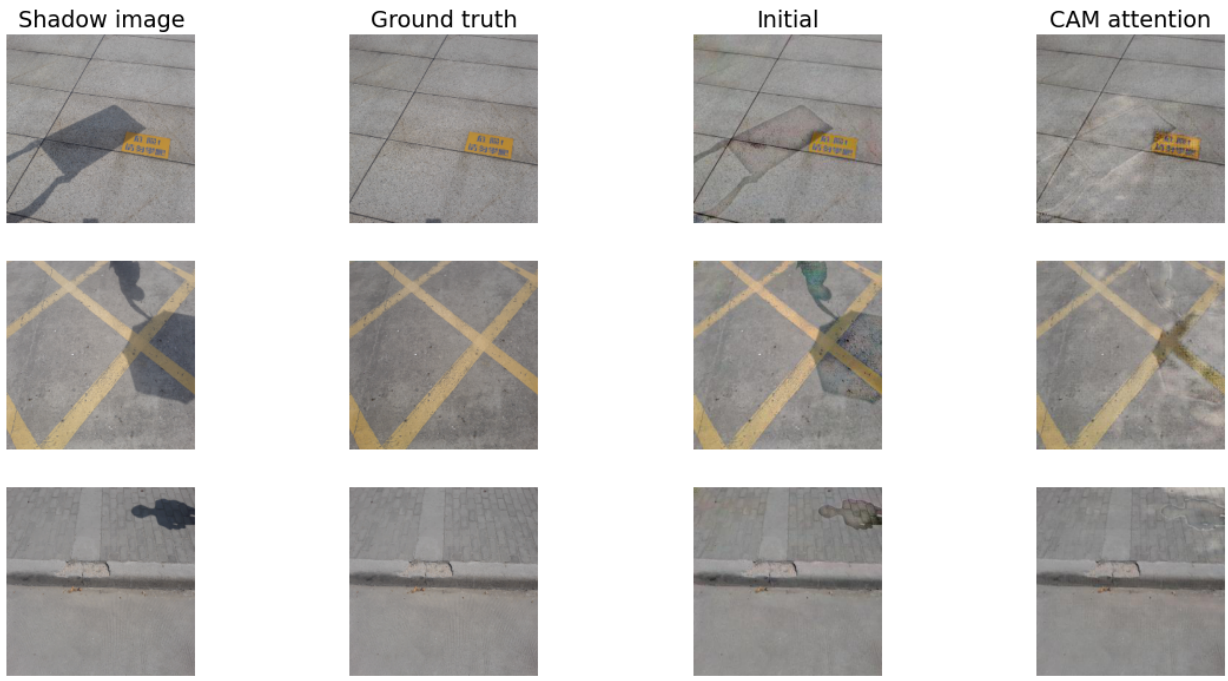

Figure 4.1. The qualitative comparison of results. **Initial** indicates solution without attention and multi context aggregation. **CAM attention** is our final solution.

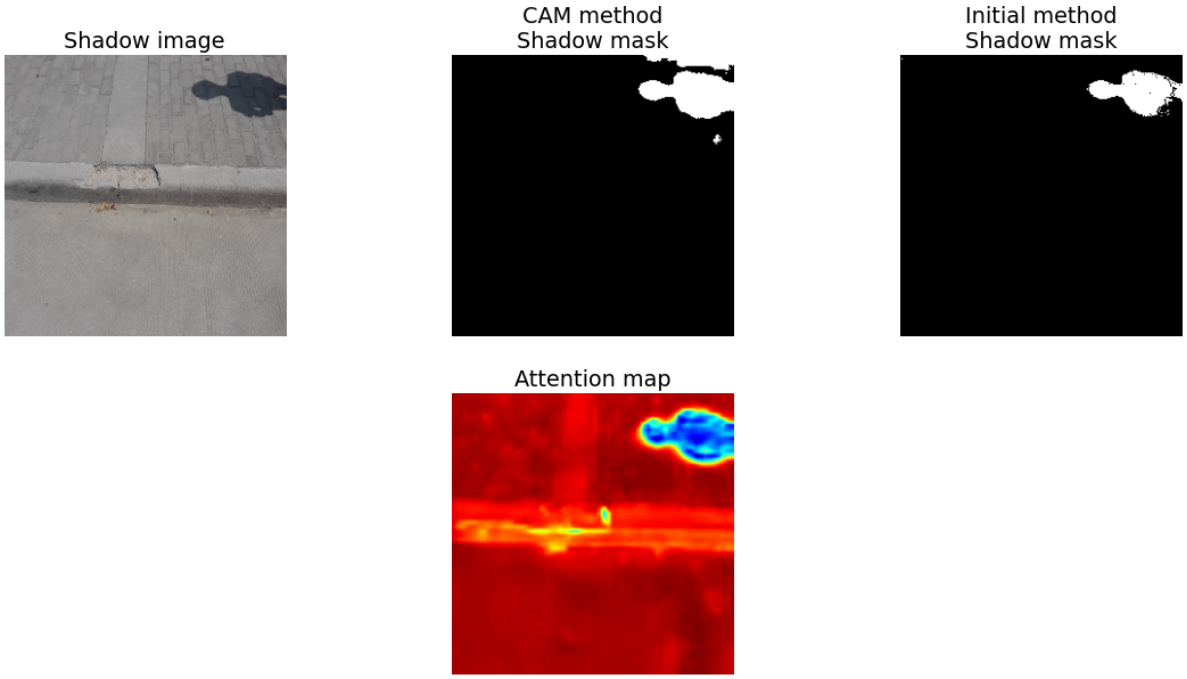

Figure 4.2. Comparison of shadow masks between our method (CAM method) and the one without attention and multi context aggregation(Initial method).

Conclusions

This work presented a solution to unsupervised shadow removal problem with the use of generative adversarial networks with attention modules and multi context aggregation. Our network produces better results compared to the existing approach in the field. Analysis showed that attention maps obtained from auxiliary classifier encourage the networks to concentrate on more distinctive regions between domains. However, GANs demand more accurate and consistent architecture to solve the problem in a more efficient way. We have also showed how attention modules can improve the quality of shadow removal while introducing the problems with the shadow over-detection.

For that reason we will research further to address the problem of more consistent architecture in the future work.

References

[1] D. L. Waltz, “Generating semantic descriptions from drawings of scenes with shadows,” Cambridge, MA, USA, Tech. Rep., 1972.
[2] H. Barrow and J. Tenenbaum, “Recovering intrinsic scene characteristics from images,” in Comp. Vision Systems, 1978.
[3] -I. Mikic, P. C. Cosman, G. T. Kogut and M. M. Trivedi, "Moving shadow and object detection in traffic scenes," *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, Barcelona, Spain, 2000, pp. 321-324 vol.1, doi: 10.1109/ICPR.2000.905341.
[4] - R. Cucchiara, C. Grana, M. Piccardi, A. Prati and S. Sirotti, "Improving shadow suppression in moving object detection with HSV color information," ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No.01TH8585), Oakland, CA, 2001, pp. 334-339, doi: 10.1109/ITSC.2001.948679.
[5] Z. Li and N. Snavely, "Learning Intrinsic Image Decomposition from Watching the World," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 9039-9048, doi: 10.1109/CVPR.2018.00942.
[6] R. Guo, Q. Dai and D. Hoiem, "Paired Regions for Shadow Detection and Removal," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2956-2967, Dec. 2013, doi: 10.1109/TPAMI.2012.214.
[7] Y.-Y. Chuang, D. B. Goldman, B. Curless, D. Salesin, and R. Szeliski, “Shadow matting and compositing,” ACM ToG, vol. 22, no. 3, pp. 494–500, 2003.
[8] - L. Qu, J. Tian, S. He, Y. Tang and R. W. H. Lau, "DeshadowNet: A Multi-context Embedding Deep Network for Shadow Removal," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 2308-2316, doi: 10.1109/CVPR.2017.248.
[9] - X. Hu, C. Fu, L. Zhu, J. Qin and P. Heng, "Direction-aware Spatial Context Features for Shadow Detection and Removal," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2019.2919616. - direction aware
[10] - J. Wang, X. Li and J. Yang, "Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 1788-1797, doi: 10.1109/CVPR.2018.00192.
[11] - B. Ding, C. Long, L. Zhang and C. Xiao, "ARGAN: Attentive Recurrent Generative Adversarial Network for Shadow Detection and Removal," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 10212-10221, doi: 10.1109/ICCV.2019.01031.
[12] - Xiaodong Cun, Chi-Man Pun, Cheng Shi. «Towards Ghost-free Shadow Removal via Dual Hierarchical Aggregation Network and Shadow Matting GAN», 2020 AAAI.

[13] - J. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2242-2251, doi: 10.1109/ICCV.2017.244.
[14] - X. Hu, Y. Jiang, C. Fu and P. Heng, "Mask-ShadowGAN: Learning to Remove Shadows From Unpaired Data," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 2472-2481, doi: 10.1109/ICCV.2019.00256.
[15] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan, “Estimating natural illumination from a single outdoor image,” in ICCV, 2009.
[16] K. Karsch, V. Hedau, D. Forsyth, and D. Hoiem, “Rendering synthetic objects into legacy photographs,” in SIGGraph Asia, 2011.
[17] E. H. Land, John, and J. Mccann, “Lightness and retinex theory,” Journal of the Optical Society of America, pp. 1–11, 1971
[18] B. A. Maxwell, R. M. Friedhoff, and C. A. Smith, “A biilluminant dichromatic reflection model for understanding images,” in CVPR, 2008.
[19] Y.-Y. Chuang, D. B. Goldman, B. Curless, D. Salesin, and R. Szeliski, “Shadow matting and compositing,” ACM ToG, vol. 22, no. 3, pp. 494–500, 2003.
[20] T.-P. Wu, C.-K. Tang, M. S. Brown, and H.-Y. Shum, “Natural shadow matting,” ACM Trans. Graph., vol. 26, no. 2, 2007.
[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
[22] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In NIPS’2014.
[23] - M. Mirza and S. Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014
[24] Nobuyuki Otsu. A threshold selection method from graylevel histograms. IEEE Transactions on Systems, Man, and Cybernetics, 9(1):62–66, 1979
[25] - P. Isola, J. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 5967-5976, doi: 10.1109/CVPR.2017.632.
[26] - T. Karras, S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 4396-4405, doi: 10.1109/CVPR.2019.00453.
[27] - Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In CVPR, 2020.
[28] - L. A. Gatys, A. S. Ecker and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2414-2423, doi: 10.1109/CVPR.2016.265.
[29] - D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[30] - Y. Shin, M. Sagong, Y. Yeo, S. Kim and S. Ko, "PEPSI++: Fast and Lightweight Network for Image Inpainting," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2020.2978501.
[31] - Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron C. Courville. Augmented CycleGAN: Learning many-to-many mappings from unpaired data. In ICML, pages 195–204, 2018.
[32] - Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In ECCV, pages 35–51, 2018.
[33] - Ian J Goodfellow. On distinguishability criteria for estimating generative models. arXiv preprint arXiv:1412.6515, 2014.
[34] - Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In Advances in neural information processing systems (pp. 2234– 2242). Schroff, F., Kal.
[35] - M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. ArXiv e-prints, Jan. 2017.
[36] - X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang and S. P. Smolley, "Least Squares Generative Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2813-2821, doi: 10.1109/ICCV. 2017.304.
[37] - Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In arXiv preprint arXiv:1805.08318, 2018.
[38] - https://github.com/soumith/ganhacks
[39] - https://www.kdnuggets.com/2017/01/generative-adversarial-networks-hot-topic-machine-learning.html
[40] - M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In ICLR, 2017.
[41] - A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In CVPR, 2017.
[42] - Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. In ICLR, 2017.
[43] - F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In ICLR, 2016.
[44] - Holschneider, M., Kronland-Martinet, R., Morlet, J., and Tchamitchian, Ph. A real-time algorithm for signal analysis with the help of the wavelet transform. In Wavelets: Time-Frequency Methods and Phase Space. Proceedings of the International Conference, 1987.
[45] - Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:1603.07285, 2016.
[46] - Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473, 2014.
[47] - Chen, X., Mishra, N., Rohaninejad, M., and Abbeel, P. Pixelsnail: An improved autoregressive generative model. In ICML, 2018.

[48] - Cheng, J., Dong, L., and Lapata, M. Long short-term memory-networks for machine reading. In EMNLP, 2016.
[49] - Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. arXiv:1706.03762, 2017.
[50] - Wang, X., Girshick, R., Gupta, A., and He, K. Non-local neural networks. In CVPR, 2018.
[51] - J. Kim, M. Kim, H. Kang, and K. Lee, “U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation,” in ICLR, 2020.
[52] - B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning Deep Features for Discriminative Localization," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2921-2929, doi: 10.1109/CVPR.2016.319.
[53] - Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In ECCV, pages 694–711, 2016.
[54] - Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In ICLR, 2018.
[55] - Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In NIPS, pp. 6629–6640, 2017.
[56] - Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
[57] - Olaf Ronneberger, Philipp Fischer, and Thomas Brox. UNet: Convolutional networks for biomedical image segmentation. In Proc. Medical Image Computing and ComputerAssisted Intervention (MICCAI), pages 234–241, 2015.
[58] - X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. CoRR, abs/1703.06868, 2:3, 2017.