

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN TIN-HỌC



KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC

MULTI-SCALE LEARNING A GAN
FROM A SINGLE REPRESENTATIVE

Thực hiện: SV Lý Phi Long

Hướng dẫn: TS. Huỳnh Thế Đăng

Tp. Hồ Chí Minh, 2020

Lời cảm ơn

Để hoàn thành tốt đề tài tốt nghiệp này, ngoài sự nỗ lực của bản thân, em còn nhận được sự quan tâm giúp đỡ của nhiều tập thể và cá nhân.

Đặc biệt, em xin bày tỏ lòng biết ơn sâu sắc đến Thầy hướng dẫn của em, TS. Huỳnh Thế Đăng. Thầy không những hỗ trợ em về mặt chuyên môn cũng như ý tưởng mà còn hướng dẫn em về tư duy nghiên cứu khoa học. Em rất tự hào khi trở thành sinh viên của thầy.

Bên cạnh đó, em xin gửi lời cảm ơn chân thành đến quý Thầy Cô trong Bộ môn Ứng dụng Tin học đã tạo môi trường học tập thuận lợi để em có thể hoàn thành khóa luận này. Đồng thời, em xin cảm ơn đến quý Thầy Cô trong khoa Toán-Tin học, trường ĐH Khoa học Tự nhiên, ĐHQG-HCM đã dạy bảo em tận tình trong suốt quá trình em học tập tại khoa.

Do thời gian thực hiện khóa luận không nhiều, kiến thức còn hạn chế nên khi làm khóa luận em không tránh khỏi những sai sót. Em mong nhận được sự góp ý từ quý Thầy, Cô và bạn đọc. Xin chân thành cảm ơn!

Cuối cùng, em xin kính chúc quý Thầy, Cô dồi dào sức khỏe và thành công trong sự nghiệp cao quý của mình.

Tp.HCM, ngày 25 tháng 08 năm 2020

Sinh viên

Lý Phi Long

Tóm tắt

Bài toán *từ hình ảnh sang hình ảnh (image-to-image)* (I2I) bằng phương pháp học *không giám sát (unsupervise learning)* từ lâu đã trở thành một ứng dụng không thể thiếu của các phương pháp học sâu. Tuy nhiên, để thực hiện được bài toán trên mô hình cần một lượng dữ liệu nhất định để huấn luyện. Nghiên cứu này được xây dựng dựa trên mô hình ConSinGAN [14] và nghiên cứu của Benaim cùng cộng sự [4] nhằm thực hiện chuyển đổi một video sang một video khác theo dạng của một hình ảnh cho trước, đưa về bài toán chuyển đổi từ hình ảnh sang hình ảnh và chỉ dùng hai hình ảnh A và B cho quá trình huấn luyện trong thời gian ngắn nhất. Mô hình sẽ học được ánh xạ để chuyển đổi từ tầng thô sơ nhất đến tầng chi tiết nhất ứng với từng độ phân giải khác nhau của hình ảnh. Nhờ đó, hình ảnh sẽ được chuyển đổi từ cấu trúc tổng quát nhất đến những chi tiết nhỏ nhất trong hình. Tuy nhiên, việc lựa chọn kết hợp hai mô hình [4] và [14] đã tạo ra những kết quả không tốt và điều đó trở thành tiền đề nên tránh cho các nghiên cứu sau này.

Mục lục

Lời cảm ơn	1
Tóm tắt	2
1 Mở đầu	9
1.1 Lý do chọn đề tài	9
1.2 Mục tiêu nghiên cứu	10
2 Phương pháp nghiên cứu gần đây	11
2.1 Mạng đối sinh	11
2.2 Mô hình sinh được huấn luyện trên một ảnh	13
2.3 Bài toán chuyển đổi hình ảnh sang hình ảnh	14
3 Phương pháp nghiên cứu đề xuất	15
3.1 Kiến trúc mô hình	16
3.1.1 Kiến trúc đa tầng	16
3.1.2 Cấu trúc bộ sinh	19
3.1.3 Cấu trúc bộ phân biệt	20
3.1.4 Kiến trúc đa tầng cho hình ảnh	20
3.2 Hàm măt mát	22
3.2.1 Măt mát đối kháng	22
3.2.2 Măt mát tái tạo	23
3.2.3 Măt mát chu kỳ nhất quán	23
4 Kết quả và thảo luận	25
4.1 Kết quả	25
4.2 Thảo luận	28

5	Kết luận	29
6	Đề nghị nghiên cứu thêm	30

Danh sách hình vẽ

2.1	Cấu trúc tổng quát của mạng đối sinh	12
3.1	So sánh kiến trúc hai mô hình	17
3.2	So sánh kiến trúc bộ sinh của hai mô hình	19
3.3	Mát mát chu kỳ nhất quán	24
4.1	Video chuyển đổi	26

Danh sách bảng

4.1 So sánh kiến trúc giữa nghiên cứu này và Benaim cùng cộng sự . . . 25

Thuật ngữ

từ hình ảnh sang hình ảnh	image-to-image	. . .	2
không giám sát	unsupervise learning	. . .	2
mạng đối sinh	generative adversarial network	. . .	9
đầu vào	input	. . .	9
bộ sinh	generator	. . .	11
bộ phân biệt	discriminator	. . .	11
mẫu	sample	. . .	11
hàm mất mát	loss function	. . .	11
cân bằng nash	nash equilibrium	. . .	12
vùng nhỏ	patch	. . .	13
nội suy ảnh	inpainting	. . .	13
khử nhiễu	denoise	. . .	13
siêu phân giải	super-resolution	. . .	13
phân đoạn ảnh	image segmentation	. . .	13
trọng số	weight	. . .	13
đầu ra	output	. . .	13
tăng mẫu	upsamling	. . .	13
học giám sát	supervise learning	. . .	14
mất mát chu kỳ nhất quán	cycle-consistent loss	. . .	14
mô hình tiền huấn luyện	pre-train model	. . .	14
hình ảnh mục tiêu	target image	. . .	15
giảm mẫu	downsample	. . .	15
quá khớp	overfiting	. . .	16
nâng độ phân giải	upscale	. . .	16
đặc trưng	feature	. . .	18
nhân	kernel	. . .	19

chuẩn hóa theo cụm dữ liệu	batch norm 19
hàm kích hoạt	activation function 19
liên kết dư	residual connection 19
mất mát đối kháng	adversarial loss 22
lượng phạt đạo hàm	gradient penalty 22
mất mát tái tạo	reconstruction loss 23
dòng quang	optical flow 30

Chương 1

Mở đầu

1.1 Lý do chọn đề tài

Ngày nay, các *mạng đối sinh* (*Generative Adversarial Network*) (GAN) đã đạt được nhiều kết quả nổi bật trong nghiên cứu và cả trong ứng dụng, có thể kể đến như tạo ảnh [20], xóa những vật thể không mong muốn trong ảnh [17], tự động tạo ra một đoạn văn bản [29], xử lý ảnh y khoa, phân đoạn ảnh, tự động tô màu ảnh, chuyển đổi ảnh [2] và tạo ảnh nghệ thuật [8]. Không những thế, GAN còn được ứng dụng rộng rãi trong việc tổng hợp và chỉnh sửa ảnh mặt người như đoán số tuổi [36, 40], chuyển đổi giới tính [35].

Tuy nhiên, để có thể huấn luyện được GAN cần phải có một số lượng dữ liệu khổng lồ và điều này đã gây bất lợi cho các mô hình học sâu nói chung và GAN nói riêng, nên việc giảm số lượng dữ liệu *đầu vào* (*input*) cho GAN dường như là một trong những hướng nghiên cứu trong tương lai. Vào năm 2019, Shaham cùng nhóm cộng sự [30] đã giới thiệu SinGAN, đây là mô hình GAN có rất nhiều ứng dụng như tạo ảnh vô điều kiện, hòa trộn ảnh, chuyển đổi hình vẽ sang hình thật, tăng độ phân giải, làm chuyển động cho ảnh. Mô hình SinGAN đã đạt được nhiều ứng dụng như vậy nhưng quá trình huấn luyện mô hình chỉ diễn ra trên một tấm ảnh duy nhất, điều này đã tạo nên một bước đệm trong nghiên cứu huấn luyện GAN trên một hình ảnh.

1.2 Mục tiêu nghiên cứu

Vào năm 2020, Hinz cùng cộng sự [14] đã đề xuất mô hình ConSinGAN nhằm cải tiến thời gian huấn luyện SinGAN và Benaim cùng cộng sự [4] đã đưa ra những ứng dụng mới cho SinGAN trong đó có ứng dụng chuyển đổi từ video này sang video khác mang kiểu dáng của hình ảnh cho trước. Điều này đã thôi thúc sự ra đời của nghiên cứu này nhằm kết hợp cả hai mô hình trên lại với mong muốn giữ được các đặc tính tốt nhất của các nghiên cứu [4, 14]. Cụ thể, các nội dung như sau:

- Mô hình sẽ tận dụng sức mạnh của kiến trúc đa tầng nhằm giải quyết được bài chuyển đổi từ hình ảnh sang hình ảnh.
- Bằng cách kết hợp mô hình ConSinGAN [14] và mô hình của Benaim và cộng sự [4], ta có thể mong đợi việc xây dựng mô hình nhằm giải quyết bài toán chuyển đổi từ hình ảnh sang hình ảnh với mô hình ít tham số nhất.

Chương 2

Phương pháp nghiên cứu gần đây

2.1 Mạng đối sinh

Mạng đối sinh lần đầu tiên được giới thiệu bởi Goodfellow [11] và được cấu tạo bởi hai phần: *bộ sinh (generator)* (ký hiệu G) và *bộ phân biệt (discriminator)* (ký hiệu D). Cấu trúc tổng quát được miêu tả như hình 2.1.

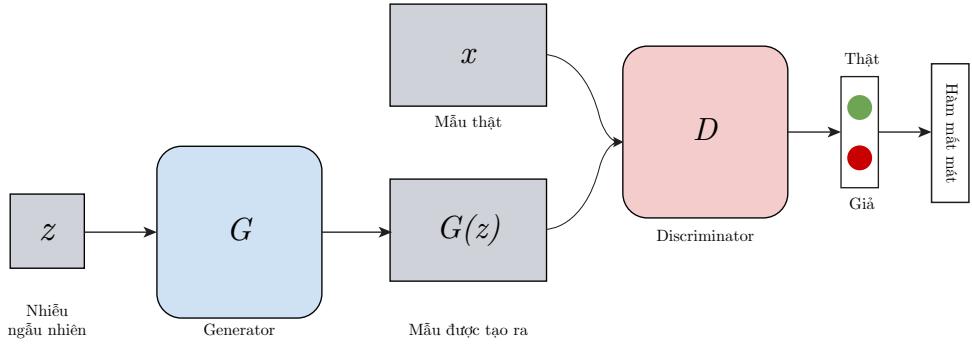
Trong đó, G sẽ tạo ra những *mẫu (sample)* tương tự với dữ liệu thật, trong khi D sẽ cố gắng phân biệt đâu là mẫu từ phân phối dữ liệu thật, đâu là mẫu do G tạo ra. Có thể tưởng tượng, G là một nhóm người làm hàng giả, còn D là cảnh sát cố gắng phát hiện ra được hàng giả. Quá trình huấn luyện diễn ra như một cuộc cạnh tranh nhằm bắt buộc cả G và D cải thiện dần dần phương pháp học tốt hơn.

D sẽ cho ra giá trị dự đoán $D(x)$ cho biết x có phải là ảnh thật hay không. Mục tiêu của D là tối đa hóa khả năng phân biệt được hình ảnh nào là hình ảnh thật và hình ảnh nào là hình ảnh do G tạo ra. D sẽ được huấn luyện để giá trị $D(x) \rightarrow 1$ còn $D(G(x)) \rightarrow 0$. Nghĩa là, *hàm mất mát (loss function)* muốn tối đa hóa giá trị $D(x)$ và tối thiểu hóa giá trị $D(G(x))$. Quá trình tối thiểu hóa giá trị $D(G(x))$ tương đương với việc tối đa hóa $(1 - D(G(x)))$. Vì vậy, hàm mất mát của D là:

$$\max_D V(D) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (2.1)$$

Trong đó, $p_{data}(x)$ là ký hiệu cho phân phối dữ liệu thật, $p_z(z)$ là ký hiệu cho phân phối nhiễu.

G sẽ học cách đánh lừa D bằng cách sinh ra ảnh gần giống thật, hay giá trị $D(G(x)) \rightarrow 1$. Tức là, hàm mất mát mong muốn tối đa hóa giá trị $D(G(x))$, tương



Hình 2.1: Cấu trúc tổng quát của mạng đối sinh. Đầu vào của G là z , với z là nhiễu được khởi tạo ngẫu nhiên. D sẽ lấy mẫu được sinh ra từ G là $G(z)$ và mẫu thật x từ tập huấn luyện. D sẽ giải quyết bài toán phân loại nhị phân, trả về kết quả từ 0 đến 1, với xác suất đầu ra càng cao thì khả năng mẫu đó là thật càng lớn và ngược lại.

đương với việc tối thiểu hóa giá trị $(1 - D(G(x)))$. Vì vậy, hàm mất mát của G là:

$$\min_G V(G) = \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (2.2)$$

Do đó, hàm mất mát tổng quát của GAN được định nghĩa:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))] \quad (2.3)$$

Từ hàm mất mát của GAN, ta có thể thấy là quá trình huấn luyện G và D là quá trình đối nghịch nhau, trong khi D cố gắng tối đa hóa hàm mất mát thì G cố gắng tối thiểu hóa hàm mất mát. Quá trình huấn luyện GAN kết thúc khi GAN đạt đến trạng thái cân bằng của cả hai mô hình G và D , được gọi là *cân bằng Nash (Nash Equilibrium)*.

2.2 Mô hình sinh được huấn luyện trên một ảnh

Quá trình huấn luyện để học được phân phối của các *vùng nhỏ* (*patch*) trên một tấm ảnh từ lâu đã được chứng minh rất tốt. Bằng hướng tiếp cận này, mà nhiều bài toán như *nội suy ảnh* (*inpainting*) [33, 39], *khử nhiễu* (*denoise*) [43], *siêu phân giải* (*super-resolution*) [28], *phân đoạn ảnh* (*image segmentation*) [9] đã được giải quyết mà chỉ cần huấn luyện một hình ảnh duy nhất. Cụ thể, bài toán siêu phân giải hình ảnh [37, 15, 10, 3] và chỉnh sửa hình ảnh [6, 27, 32] đã được tập trung vào nghiên cứu theo hướng huấn luyện bằng một ảnh duy nhất và đã đạt được kết quả tốt. Hơn nữa, vài nghiên cứu gần đây cũng chỉ ra việc sử dụng một tấm ảnh cũng đủ thông tin để giúp huấn luyện mô hình [1].

Tuy nhiên, các nghiên cứu về huấn luyện GAN trên tấm ảnh duy nhất vẫn còn chưa nhiều. Hầu hết, các mô hình này đều không sử dụng hình ảnh "tự nhiên", mà thay vào đó họ chỉ huấn luyện mô hình học các thông tin họa tiết trên hình [19, 41, 5, 24]. Trong thời gian gần đây, hai mô hình GAN đã được đề xuất mà quá huấn luyện chỉ diễn ra trên một hình ảnh "tự nhiên", cụ thể là InGAN [31] và SinGAN [30]. Bên cạnh đó, còn có mô hình ConSinGAN [14] nhằm cải tiến mô hình SinGAN để giúp việc huấn luyện trở nên dễ dàng và nhanh hơn.

SinGAN xây dựng kiến trúc mô hình như một kim tự tháp nhằm huấn luyện cả bộ sinh và bộ phân biệt qua nhiều tầng, mỗi tầng ứng với một độ phân giải khác nhau của hình ảnh. Tại mỗi tầng, mô hình được huấn luyện riêng biệt và *trọng số* (*weight*) của mô hình ở các tầng phía trước sẽ được giữ cố định. *Dầu ra* (*output*) của mỗi tầng là một hình ảnh và hình ảnh đó sẽ được *tăng mẫu* (*upsampling*) theo hệ số r^s ($r, s \in \mathbb{R}$) và trở thành đầu vào cho tầng tiếp theo. Mô hình ConSinGAN cũng sử dụng cấu trúc tương tự như SinGAN, tuy nhiên thay vì huấn luyện mỗi tầng là độc lập, thì ConSinGAN sẽ huấn luyện 3 tầng với nhau để đạt được kết quả tốt hơn. Trong nghiên cứu này, mô hình sẽ được xây dựng dựa trên ConSinGAN [14] và nghiên cứu của Benaim cùng cộng sự [4].

2.3 Bài toán chuyển đổi hình ảnh sang hình ảnh

Bài toán chuyển đổi hình ảnh sang hình ảnh là bài toán rất phổ biến trong GAN với mục tiêu giúp mô hình học được ánh xạ giữa ảnh đầu vào và ảnh đầu ra [13]. Nhờ sự phát triển mạnh mẽ của mạng học sâu mà các mạng đối sinh đã đạt nhiều kết quả khá tốt trong bài toán I2I. Isola cùng với các cộng sự [18] đã đề xuất mô hình "pix2pix" - mạng GAN có điều kiện - nhằm thực hiện một số ứng dụng trong bài toán I2I theo phương pháp *học giám sát (supervise learning)* , tức là dữ liệu sẽ được đi theo từng đôi. Tuy nhiên, việc tìm dữ liệu theo từng đôi là rất khó, thậm chí là bất khả thi trong một số trường hợp. Ví dụ, việc chuyển đổi qua lại giữa hình ảnh đối thường và hình ảnh do các nghệ sĩ vẽ. Các mô hình DiscoGAN [22], CycleGAN [42] và DualGAN [38] đã được đề xuất nhằm giải quyết bài toán I2I với phương pháp học không giám sát bằng cách sử dụng *mất mát chu kỳ nhất quán (cycle-consistent loss)* . Liu cùng cộng sự [26] đề xuất mô hình FUNIT có thể chuyển đổi được nhiều miền ảnh hơn. Tuy nhiên, FUNIT cần rất nhiều dữ liệu để huấn luyện, đồng thời dữ liệu giữa hai miền cũng phải tương tự nhau. Trong nghiên cứu này không cần bất kỳ một mô hình tiền huấn luyện (*pre-train model*) nào cũng như không cần bất kỳ một yêu cầu đặc biệt nào về dữ liệu, trong đó bao gồm việc sử dụng hình ảnh "tự nhiên", từ một video và một hình ảnh làm gốc, sẽ chuyển đổi thành một video khác mang kiểu dáng giống như hình ảnh đó.

Chương 3

Phương pháp nghiên cứu đề xuất

Dầu vào của bài toán là một video và hình ảnh mục tiêu. Mô hình sẽ chuyển đổi video đầu vào thành video đầu ra có kiểu dáng giống *hình ảnh mục tiêu (target image)*. Video đầu vào gồm chuỗi các khung hình f_0, f_1, \dots, f_k với $k \in N$. Do đó, tại mỗi vòng lặp, ta sẽ chọn ngẫu nhiên một khung hình f_i bất kỳ làm hình ảnh gốc I_A và hình ảnh mục tiêu đặt là I_B . Điều này đã đưa bài toán đang làm thành bài toán chuyển đổi hai hình ảnh $I_A \in A$ và $I_B \in B$, trong đó A và B là hai miền ảnh tương ứng, bằng cách chuyển đổi I_A sang $I_{AB} \in B$ và I_B sang $I_{BA} \in A$ mà không cần bất kỳ dữ liệu nào khác. Như được đề cập ở trên, dữ liệu đầu vào rất hạn chế, nên mô hình sẽ huấn luyện G_A và G_B qua nhiều tầng, từ tầng thô sơ nhất đến chi tiết nhất. Ban đầu, mô hình tiến hành *giảm mẫu (downsample)* I_A và I_B thành hai kim tự tháp gồm N tầng ($N \in \mathbb{N}$): $\mathcal{I}_A = \{I_A^n | n = 0, 1, \dots, N - 1\}$ và $\mathcal{I}_B = \{I_B^n | n = 0, 1, \dots, N - 1\}$.

Các nghiên cứu trước đây cho thấy được sức mạnh của kiến trúc đa tầng trong việc tạo ảnh vô điều kiện [21, 16, 7], tạo ảnh điều kiện [34] hay dùng để huấn luyện mô hình bằng một ảnh [30, 14]. Trong nghiên cứu này sẽ tận dụng kiến trúc đa tầng vào ứng dụng chuyển đổi hình ảnh không giám sát như trong nghiên cứu của Benaim cùng cộng sự [4] và đồng thời sử dụng các cải tiến của mô hình ConSinGAN [14] cho SinGAN [30] với hy vọng đạt được mong muốn có thể giữ được các đặc điểm tốt của mô hình mà vẫn có thể giải được bài toán.

3.1 Kiến trúc mô hình

Kiến trúc của mô hình được minh họa trong hình 3.1 (a). Trong đó, G_A và G_B là chuỗi các bộ sinh, $\{G_A^n\}_{n=0}^N$ và $\{G_B^n\}_{n=0}^N$ được dùng để chuyển đổi hình ảnh ứng với tầng n . Tại mỗi tầng, ta cần có hai bộ phân biệt D_A^n và D_B^n ($n \in \{0, 1, \dots, N\}$) được dùng để phân biệt hình ảnh đầu vào có phải là hình ảnh thật trong miền A và B hay không. Kiến trúc của mô hình được chia làm hai phần chính: Trình tạo ảnh vô điều kiện và trình tạo ảnh điều kiện.

3.1.1 Kiến trúc đa tầng

Trình tạo ảnh vô điều kiện Tại bất kỳ tầng n , để tìm ảnh xạ giữa hai miền A và B , đầu tiên, mô hình sẽ giảm mẫu hình ảnh I_A và I_B xuống thành I_A^n và I_B^n theo một hệ số r^s ($r, s \in \mathbb{R}$) (xem phần 3.1.4).

Tương tự mô hình của Benaim cùng cộng sự, mô hình cũng sẽ được huấn luyện từ tầng thô sơ nhất đến tầng tốt nhất. Sau đó, được huấn luyện xuyên suốt qua các tầng từ đầu đến cuối thay vì chỉ huấn luyện độc lập các tầng với nhau. Khác với, nghiên cứu của Benaim cùng cộng sự là mô hình chỉ được huấn luyện tại tầng hiện tại (cao nhất) và các tham số của tất cả các tầng phía trước sẽ được đóng băng. Tuy nhiên, mô hình trong nghiên cứu này sẽ không huấn luyện xuyên suốt qua các tầng vì sẽ rất dễ dẫn đến hiện tượng *quá khớp* (*overfitting*) nên mô hình sẽ được huấn luyện từng tầng một và các tham số của tầng trước đó sẽ không bị đóng băng.

Tại tầng thô sơ nhất $n = 0$, cả mô hình của Benaim cùng với cộng sự và mô hình trong nghiên cứu này đều sẽ học được ảnh xạ từ nhiều z đến hình ảnh có độ phân giải thấp nhất:

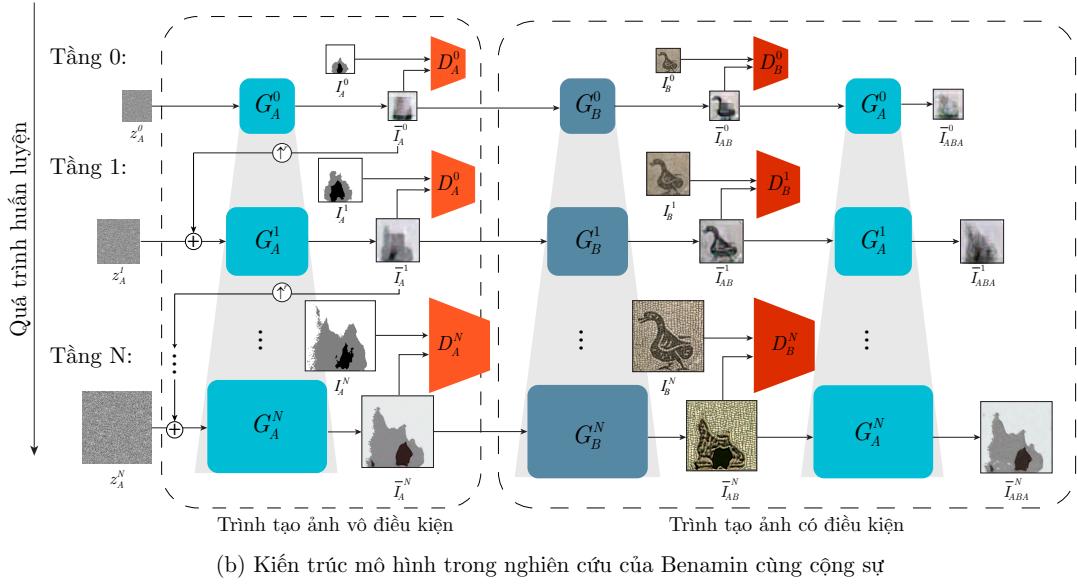
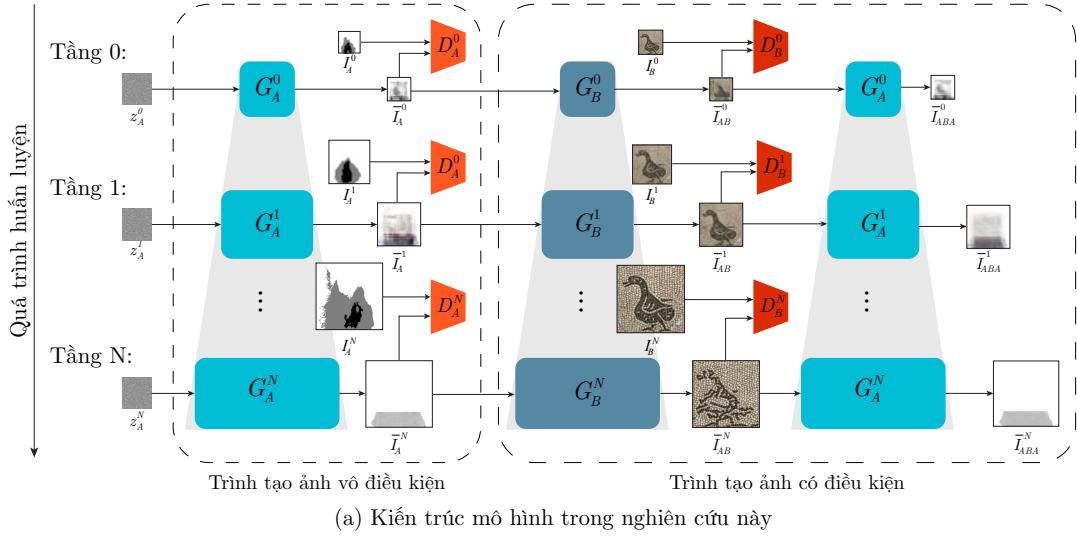
$$\bar{I}_A^0 = G_A^0(z_A^0) \quad (3.1)$$

$$\bar{I}_B^0 = G_B^0(z_B^0) \quad (3.2)$$

Trong đó, z_A^n là nhiều được khởi tạo theo chuẩn Gaussian.

Khi lên tầng cao hơn $n > 0$, mô hình của Benaim và cộng sự sẽ đưa hình ảnh được tạo ra ở tầng trước được *nâng độ phân giải* (*upscale*) và cộng thêm vào nhiều z_n để đưa đến tầng tiếp theo:

$$\bar{I}_A^n = G_A^n(z_A^n + \uparrow \bar{I}_A^{n-1}) \quad (3.3)$$



Hình 3.1: So sánh kiến trúc hai mô hình. Hình vẽ miêu tả quá trình chuyển đổi từ ngọn lửa sang vịt đá trong mô hình nghiên cứu này và mô hình của Benaim cùng cộng sự. Trong đó, I_A^n là hình ảnh ngọn lửa, I_B^n là hình ảnh vịt đá tại tầng n với $0 < n < N$. Đầu tiên đối với cả hai mô hình, bộ sinh G_A^n tạo ra hình ảnh ngọn lửa giả tại tầng n , bằng cách đưa vào nhiều ngẫu nhiên z_n được khởi tạo theo phân phối Gaussian. Sau đó, D_A^n được dùng để phân biệt \bar{I}_A^n với hình ảnh gốc I_A^n . Bộ sinh G_B^n sẽ ánh xạ \bar{I}_A^n thành \bar{I}_{AB}^n . Khi đó, D_B^n sẽ phân biệt \bar{I}_{AB}^n với hình ảnh gốc I_B^n . Bộ sinh G_A^n tiếp tục ánh xạ \bar{I}_{AB}^n sang \bar{I}_{ABA}^n để chắc chắn khoảng cách giữa \bar{I}_{ABA}^n và \bar{I}_A^n gần nhau. Hình (a) là kiến trúc của mô hình được xây dựng trong nghiên cứu này: Đầu vào ở các tầng là đặc trưng được trích xuất từ tầng trước. Hình (b) là kiến trúc được đề xuất bởi Benaim [4]: Đầu vào ở các tầng tiếp theo sẽ là hình ảnh được nâng độ phân giải ở tầng trước được thêm nhiễu.

$$\bar{I}_B^n = G_B^n \left(z_B^0 + \uparrow \bar{I}_B^{n-1} \right) \quad (3.4)$$

Khác với mô hình Benaim cùng cộng sự, khi lên tầng cao hơn $n > 0$, thay vì mô hình nhận đầu vào là hình ảnh được nâng độ phân giải từ tầng trước thì trong nghiên cứu này mô hình nhận đầu vào là các *đặc trưng* (*feature*) được trích xuất ra từ tầng trước và truyền vào tầng tiếp theo. Đồng thời, nhiều cũng sẽ được thêm ở tất cả các tầng của mô hình:

$$\bar{I}_A^n = G_A^n \left(z_A^n + G_A^{n-1} \left(z_A^{n-1} \right) \right) \quad (3.5)$$

$$\bar{I}_B^n = G_B^n \left(z_B^n + G_B^{n-1} \left(z_B^{n-1} \right) \right) \quad (3.6)$$

Việc truyền trực tiếp đặc trưng được trích xuất ra từ tầng trước giúp cho mô hình tiết kiệm được thời gian huấn luyện, đồng thời sẽ không tốn nhiều chi phí tính toán để huấn luyện lại các tầng trước giống như mô hình của Benaim cùng cộng sự.

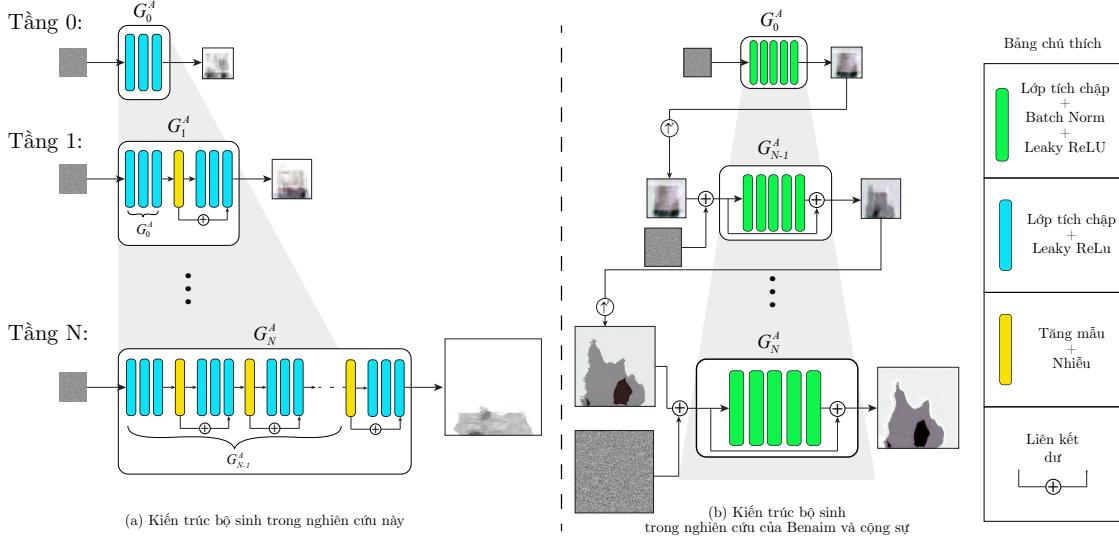
Trình tạo ảnh điều kiện Tại bất kỳ tầng n , kiến trúc đa tầng cả hai mô hình như nhau, đầu vào là một mẫu $\bar{I}_A^n \in A$ và $\bar{I}_B^n \in B$ trong trình tạo ảnh vô điều kiện. Quá trình chuyển đổi hình ảnh sẽ được đi từ tầng thô sơ nhất đến tầng chi tiết nhất. Hai bộ sinh G_B^n và G_A^n sẽ ánh xạ \bar{I}_A^n và \bar{I}_B^n sang miền ảnh mục tiêu tương ứng. Có thể hiểu một cách toán học như sau:

$$\bar{I}_{AB}^n = G_B^n \left(\bar{I}_A^n \right) \quad (3.7)$$

$$\bar{I}_{BA}^n = G_A^n \left(\bar{I}_B^n \right) \quad (3.8)$$

Bộ sinh được xem là chuyển đổi tốt khi và chỉ khi đổi với mỗi tầng, từng vùng trong ảnh I_A được chuyển đổi tương ứng sang từng vùng trong ảnh I_B và ngược lại. Do đó, tại tầng thô sơ nhất $n = 0$, ánh xạ giữa hai hình ảnh I_A^0 và I_B^0 nên ánh xạ được cấu trúc tổng quát nhất từ I_A sang I_B . Ví dụ, mặt đất nằm ở phía đáy của hình I_A nên được ánh xạ sang mặt cỏ ở dưới đáy của hình I_B , ánh sáng ban ngày trong I_A nên được ánh xạ sang ánh sáng ban đêm bên I_B . Đối với những tầng cao hơn $n > 0$, bộ sinh G_B^n và G_A^n sẽ học được những chi tiết nhỏ hơn trong tấm hình I_B^n và I_A^n . Tuy nhiên, ánh xạ mà bộ sinh G_B^n và G_A^n học được vẫn phải giữ được cấu trúc tổng thể được học từ các tầng trước.

3.1.2 Cấu trúc bộ sinh



Hình 3.2: So sánh kiến trúc bộ sinh của hai mô hình. Hình (a) là kiến trúc của mô hình trong nghiên cứu này. Tại mỗi tầng sẽ được thêm mới vào 1 lớp tăng mẫu và 3 khối tích chập. Đồng thời, đầu vào của mỗi tầng là đặc trưng được trích xuất từ tầng trước được cộng thêm nhiễu. Hình (b) là kiến trúc bộ sinh của mô hình Benaim và cộng sự [4] gồm 5 khối tích chập. Đồng thời, đầu vào của mỗi tầng là hình ảnh từ tầng trước được nâng độ phân giải và cộng thêm nhiễu.

Tương tự như SinGAN, cấu trúc bộ sinh của mô hình Benaim và cộng sự sẽ là một mạng tích chập gồm 5 khối tích chập. Ban đầu mỗi khối tích chập gồm 32 *nhân* (*kernel*) có kích thước (3×3) nối đến một lớp *chuẩn hóa theo cụm dữ liệu* (*batch norm*) và cuối cùng cho đi qua một *hàm kích hoạt* (*activation function*) LeakyReLU như hình 3.2 (b). Đầu tiên tại tầng thô sơ nhất $n = 0$, nhóm tác giả bắt đầu với 32 nhân tại mỗi khối và sẽ được tăng số lượng nhân lên gấp đôi sau 4 tầng và số lượng lớp sẽ được giữ nguyên xuyên suốt các tầng.

Trong nghiên cứu này, cấu trúc của bộ sinh là một mạng nơ-ron tích chập được miêu tả như hình 3.2 (a). Khác biệt với mô hình của Benaim cùng cộng sự, mô hình bộ sinh trong nghiên cứu này được lấy ý tưởng từ mô hình bộ sinh trong PGGAN [21]. Cụ thể, tại tầng thô sơ nhất $n = 0$, bộ sinh gồm 3 khối tích chập, mỗi khối tích chập sẽ 64 nhân kích thước (3×3) lớp tích chập sau đó sẽ được đi qua hàm kích hoạt Leaky ReLU. Sau khi huấn luyện xong tầng đầu tiên, bộ sinh sẽ được thêm vào 1 lớp tăng mẫu và 3 khối tích chập. Đặc biệt, đối với các tầng $n > 0$, bộ sinh sẽ được thêm vào *liên kết dư* (*residual connection*) từ đặc trưng gốc được tăng mẫu

đến đầu ra của lớp tích chập mới được thêm vào (xem "bộ sinh: Tầng 1" trong hình 3.2). Cũng giống như mô hình của Benaim cùng cộng sự, tại mỗi tầng $n > 0$, trước khi thêm vào 3 khối tích chập, đặc trưng gốc từng tầng trước sẽ được thêm vào nhiều nhằm tạo được độ đa dạng của hình ảnh. Quá trình này sẽ lặp đi lặp lại đến khi nào bộ sinh tạo ra được hình ảnh có độ phân giải như ban đầu. Vì với những tầng có độ phân giải càng nhỏ thì mô hình không cần quá nhiều lớp tích chập để có thể học được các đặc trưng của tấm hình. Đồng thời, điều này giúp giảm được số lượng trọng số của mô hình giúp cho mô hình được huấn luyện nhanh hơn.

3.1.3 Cấu trúc bộ phân biệt

Tương tự như mô hình của Benaim cùng với cộng sự, bộ phân biệt được sử dụng trong nghiên cứu này là một bộ phân biệt đã được đề xuất trong PatchGAN [18] gồm 5 khối tích chập, mỗi khối tích chập sẽ có dạng 1 lớp tích chập (3×3) và hàm kích hoạt Leaky-ReLU. Bộ phân biệt truyền thống sẽ nhận đầu vào là một hình ảnh và đầu ra là một giá trị xác suất để xem ảnh đầu vào là thật hay giả. Đối với PatchGAN, bộ phân biệt cũng sẽ nhận hình ảnh làm đầu vào nhưng thay vì đầu ra là một giá trị duy nhất thì đầu ra sẽ là một ma trận X với từng phần tử X_{ij} là kết quả xác suất trên từng vùng ảnh nhỏ của hình ảnh lớn được đưa qua bộ phân biệt.

Bộ phân biệt PatchGAN sẽ phân loại từng vùng nhỏ trên tấm ảnh thay vì trên toàn bộ ảnh nên sẽ có kết quả tốt hơn. Tất nhiên khi huấn luyện bộ phân biệt của PatchGAN với ảnh thật mô hình mong muốn đầu ra của tất cả các vùng ảnh là 1, còn giá trị đầu ra trên tất cả các vùng ảnh giả sẽ là 0. Ngược lại, khi huấn luyện bộ sinh, giá trị mà bộ phân biệt nên trả ra trên tất cả các vùng ảnh là 1.

3.1.4 Kiến trúc đa tầng cho hình ảnh

Trong xây dựng kiến trúc đa tầng, có một thông số rất quan trọng được dùng để tạo ra kim tự tháp, đó chính là hệ số giảm mẫu hình ảnh. Mô hình của Benaim cùng cộng sự cũng sử dụng phương pháp giảm mẫu hình ảnh x tương tự như SinGAN bằng phép nhân với hệ số r^{N-n} với N là tổng số lượng tầng của mô hình, n là tầng hiện tại, r là hệ số thay đổi với giá trị r mặc định là 0.75. Với cách trên, mô hình phải mất từ 8 đến 10 tầng mới có thể đạt được ảnh có chiều dài hoặc chiều rộng là 250px. Nếu thay đổi giá trị r (ví dụ như $r = 0.50$) để giúp việc giảm mẫu diễn ra

nhanh hơn, dùng ít tầng hơn thì dẫn đến việc hình ảnh được tạo ra sẽ mất đi cấu trúc tổng quát của chúng.

Trong trường hợp mô hình không đủ tầng để huấn luyện với hình ảnh có độ phân giải thấp thì khi lên những tầng cao hơn, mô hình sẽ không giữ được cấu trúc tổng thể của hình ảnh mà thay vào đó mô hình chỉ có thể học được các thông tin về hoa văn. Do đó, để hình ảnh đạt được bối cảnh tổng thể mô hình cần được huấn luyện đủ lâu ở các tầng mà hình ảnh có độ phân giải thấp và không cần quá nhiều tầng với hình ảnh có độ phân giải cao. Vì vậy, phương pháp thay đổi kích thước hình ảnh có thể điều chỉnh để hình ảnh không bị biến đổi hình học quá thô cứng ($x_n = x_N \times r^{N-n}$) và đồng thời giúp cho mô hình trong nghiên cứu này có thể học được đủ lâu ở các tầng hình ảnh có độ phân giải thấp nhằm giúp cho hình ảnh tạo ra vẫn giữa được cấu trúc tổng quát như ban đầu:

$$x_n = x_N \times r^{\frac{N-1}{\log(N)} \times \log(N-n)+1} \text{ với } n = 0, 1, \dots N-1 \quad (3.9)$$

Trong đó, $r = \left(\frac{25}{\min(H,W)} \right)^{\frac{1}{N-1}}$ (với H, W lần lượt là chiều cao và chiều ngang của ảnh đầu vào), N là tầng cao nhất, n là tầng hiện tại, x_N là kích thước ban đầu của ảnh.

Ví dụ, phương pháp thay đổi kích thước trong mô hình của Benaim cùng cộng sự với hệ số $r = 0.67$ lên hình ảnh có độ phân giải ban đầu là 188×250 px sẽ tạo ra một cấu trúc đa tầng gồm 6 tầng với độ phân giải tương ứng là 25×34 px, 38×50 px, 57×75 px, 84×112 px, 126×167 px, 188×250 px. Tuy nhiên, với hệ số thay đổi $r = 0.67$ khi sử dụng phương pháp đề xuất ở phương trình 3.9 thì mô hình cũng sẽ được huấn luyện với 6 tầng nhưng độ phân giải tương ứng là 25×34 px, 32×42 px, 42×56 px, 63×84 px, 126×167 px, 188×250 px. Có thể thấy được độ phân giải thấp của hình ảnh được tạo ra với phương trình 3.9 có mật độ dày hơn phương pháp thay đổi hình ảnh gốc.

3.2 Hàm măt măt

3.2.1 Măt măt đói kháng

Măt măt đói kháng (Adversarial loss) được xây dựng nhằm để bộ phân biệt cő găng phân biệt được hình ảnh thật từ hình ảnh tổng hợp được trong khi đó bộ sinh sē cő găng đánh lừa bộ phân biệt bằng cách tạo ra ảnh giống thật nhất. Tại tầng thứ n , hai bộ phân biệt sē nhận được hình ảnh làm đầu vào. Đầu ra sē là điểm đánh giá của hình ảnh đầu vào là thật hay giả so với từng miền hình ảnh tương ứng. Cụ thể tại tầng n , D_A^n sē cő găng phân biệt hình ảnh đầu vào có phải là hình ảnh thật trong miền ảnh A hay không và D_B^n sē cő găng xác định xem hình ảnh đầu vào có phải là hình ảnh thật trong miền ảnh B hay không. Trong bài này, mô hình sē sử dụng hàm măt măt của WGAN-GP [12] làm măt măt đói kháng để giúp cho quá trình huấn luyện ổn định hơn. Măt măt đói kháng cho trình tạo ảnh vô điều kiện với chiều tạo ảnh từ A sang B :

$$\mathcal{L}_{adv}^1(D_A^n, G_A^n) = D_A^n(\bar{I}_A^n) - D_A^n(I_A^n) + \lambda_{pen} \left(\left\| \nabla_{\hat{I}_A^n} D_A^n(\hat{I}_A^n) \right\|_2 - 1 \right)^2 \quad (3.10)$$

Trong đó $\hat{I}_A^n = \alpha I_A^n + (1 - \alpha) \bar{I}_A^n$ với $\alpha \sim U(0, 1)$ và λ_{pen} là hệ số của *lượng phạt đạo hàm (gradient penalty)*. Hình ảnh \bar{I}_A^n được tạo ra bằng cách đưa một nhiễu ngẫu nhiên z_A^n qua bộ sinh G_A^n theo phương trình (3.5).

Tương tự như vậy với trình tạo ảnh điều kiện chiều từ A sang B , D_B^n cũng cő găng phân biệt đầu vào \bar{I}_{AB}^n là thật hay giả trong miền ảnh B :

$$\mathcal{L}_{adv}^2(D_B^n, G_B^n) = D_B^n(\bar{I}_{AB}^n) - D_B^n(I_B^n) + \lambda_{pen} \left(\left\| \nabla_{\hat{I}_{AB}^n} D_B^n(\hat{I}_{AB}^n) \right\|_2 - 1 \right)^2 \quad (3.11)$$

Trong đó $\hat{I}_{AB}^n = \alpha I_B^n + (1 - \alpha) \bar{I}_{AB}^n$ với $\alpha \sim U(0, 1)$. Hình ảnh \bar{I}_{AB}^n được tạo ra theo phương trình (3.7).

Tương tự như vậy với chiều chuyển đổi hình ảnh ngược lại từ miền B sang miền A là:

$$\mathcal{L}_{adv}^1(D_B^n, G_B^n) = D_B^n(\bar{I}_B^n) - D_B^n(I_B^n) + \lambda_{pen} \left(\left\| \nabla_{\hat{I}_B^n} D_B^n(\hat{I}_B^n) \right\|_2 - 1 \right)^2 \quad (3.12)$$

$$\mathcal{L}_{adv}^2(D_A^n, G_A^n) = D_A^n(\bar{I}_{BA}^n) - D_A^n(I_A^n) + \lambda_{pen} \left(\left\| \nabla_{\hat{I}_{BA}^n} D_A^n(\hat{I}_{BA}^n) \right\|_2 - 1 \right)^2 \quad (3.13)$$

trong đó $\hat{I}_B^n = \alpha I_B^n + (1 - \alpha) \bar{I}_B^n$, $\hat{I}_{BA}^n = \alpha I_A^n + (1 - \alpha) \bar{I}_{BA}^n$ với $\alpha \sim U(0, 1)$. Hình ảnh \bar{I}_B^n và \bar{I}_{BA}^n được tạo theo phương trình (3.6), (3.8).

Tổng tất cả các mất mát đối kháng của mô hình:

$$\mathcal{L}_{adv_n} = \mathcal{L}_{adv_n}^1(D_n^A, G_n^A) + \mathcal{L}_{adv_n}^1(D_n^B, G_n^B) + \mathcal{L}_{adv_n}^2(D_n^A, G_n^A) + \mathcal{L}_{adv_n}^2(D_n^B, G_n^B). \quad (3.14)$$

3.2.2 Mất mát tái tạo

Tương tự như mô hình Begamin và cộng sự, mô hình trong nghiên cứu này cũng sử dụng *mất mát tái tạo (reconstruction loss)*, tuy nhiên mô hình Begamin và cộng sự sử dụng hàm mất mát tái tạo trên hình ảnh được nâng độ phân giải từ tầng trước:

$$\mathcal{L}_{rec_n}(G_A^n, G_B^n) = \|G_A^n(\uparrow \bar{I}_A^{n-1}) - I_A^n\|_2^2 + \|G_B^n(\uparrow \bar{I}_B^{n-1}) - I_B^n\|_2^2 \quad (3.15)$$

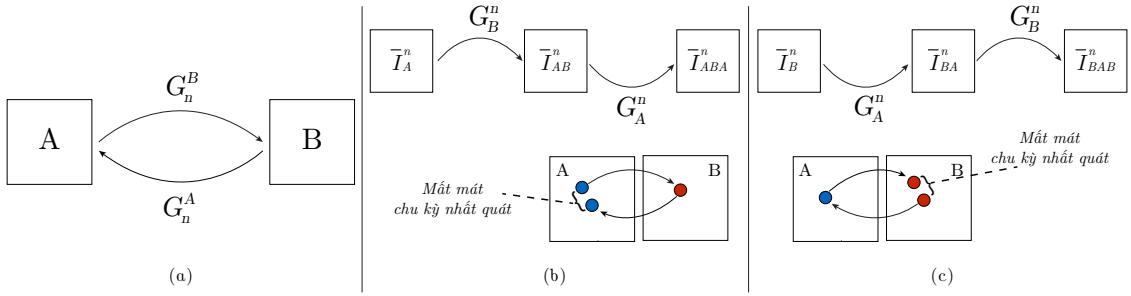
Tuy nhiên, với hàm mất mát tái tạo trong nghiên cứu này có một chút khác biệt nhằm để cải thiện quá trình huấn luyện được ổn định hơn. Tại tầng n , bộ sinh G_A^n sẽ nhận đầu vào là tấm hình được giảm mầu xuống tầng thô sơ nhất (I_A^0) của tấm hình gốc (I_A). Bộ sinh G_A^n sẽ được huấn luyện để có thể tạo lại tấm hình I_A^n có độ phân giải ứng với tầng n hiện tại. Tương tự với bộ sinh G_B^n , định nghĩa như sau:

$$\mathcal{L}_{rec_n}(G_A^n, G_B^n) = \|G_A^n(I_A^0) - I_A^n\|_2^2 + \|G_B^n(I_B^0) - I_B^n\|_2^2 \quad (3.16)$$

Mục tiêu của mất mát tái tạo này, giúp cho mạng quyết định xem tại mỗi tầng nên thêm những chi tiết mới nào so với tầng trước vào tấm hình. Đồng thời, còn giúp cho hình ảnh được tạo ra không bị thay đổi quá nhiều so với tầng trước và có thể giữ được cấu trúc tổng thể của tấm hình.

3.2.3 Mất mát chu kỳ nhất quán

Vào năm 2017, Zhu cùng cộng sự [42] đã đề xuất mất mát chu kỳ nhất quán trong CycleGAN để giải quyết vấn đề dữ liệu không cùng một cặp. Về mặt lý thuyết, quá trình học dựa trên mất mát đối kháng có thể giúp mô hình tìm ra được hai ánh xạ G_A và G_B tạo được đầu ra tương ứng với miền ảnh mục tiêu A và B . Tuy nhiên, để có thể đạt được điều này mô hình cần phải có rất nhiều dữ liệu. Do đó, nếu chỉ dựa vào mỗi hàm học đối kháng thì sẽ khó có thể ánh xạ được I_A sang I_B và chiều ngược lại.



Hình 3.3: Mất mát chu kỳ nhất quán. Mô hình bao gồm hai ánh xạ $G_n^A : A \rightarrow B$ và $G_n^B : B \rightarrow A$, Để đảm bảo ánh xạ chính xác hơn mô hình sử dụng mất mát chu kỳ nhất quán với mục tiêu chuyển đổi hình ảnh từ vùng này sang vùng khác thì kết quả chuyển đổi phải có khả năng chuyển ngược lại về hình ảnh ban đầu, nên mất mát chu kỳ nhất quán gồm hai phần: (b) mất mát chu kỳ nhất quán truyền xuôi: $\bar{I}_A \rightarrow G_B(\bar{I}_A) \rightarrow G_A(G_B(\bar{I}_A)) \approx \bar{I}_A$, (c) mất mát chu kỳ nhất quán truyền ngược: $\bar{I}_B \rightarrow G_A(\bar{I}_B) \rightarrow G_B(G_A(\bar{I}_B)) \approx \bar{I}_B$

Để có thể giải quyết bài toán ít dữ liệu, Zhu cùng cộng sự đã chỉ ra rằng ánh xạ mà mô hình học được nên là một chu trình nhất quán (như trong hình 3.3(b)). Tức là, mỗi tám hình $\bar{I}_A \in A$ sẽ được chuyển đổi sang $\bar{I}_{AB} \in B$ và sau đó sẽ được chuyển đổi ngược lại về $\bar{I}_{ABA} \in A$ và khoảng cách \bar{I}_A và \bar{I}_{ABA} nên được ngắn nhất, $\bar{I}_A \rightarrow G_B(\bar{I}_A) \rightarrow G_A(G_B(\bar{I}_A)) \approx \bar{I}_A$. Tương tự như hình 3.3(c), với mỗi hình ảnh $I_B \in B$ sẽ được chuyển đổi sang $\bar{I}_{BA} \in A$, và sau đó sẽ được chuyển đổi ngược lại về $\bar{I}_{BAB} \in B$ và khoảng cách \bar{I}_B và \bar{I}_{BAB} nên được ngắn nhất, $\bar{I}_B \rightarrow G_A(\bar{I}_B) \rightarrow G_B(G_A(\bar{I}_B)) \approx \bar{I}_B$. Mô hình sẽ được triển khai mất mát chu kỳ nhất quán cho tất cả các tầng $n = 0, 1, \dots, N$ nhằm làm cho cấu trúc của hai hình ảnh có thể được cân chỉnh phù hợp:

$$\mathcal{L}_{cyc_n}(G_A^n, G_B^n) = \|\bar{I}_A^n - \bar{I}_{ABA}^n\|_2^2 + \|\bar{I}_B^n - \bar{I}_{BAB}^n\|_2^2 \quad (3.17)$$

trong đó $\bar{I}_{ABA}^n = G_A^n(\bar{I}_{AB}^n)$, $\bar{I}_{BAB}^n = G_B^n(\bar{I}_{BA}^n)$.

Từ phương trình (3.14), (3.16) và (3.17) hàm mất mát cho mô hình tại tầng n được định nghĩa:

$$\mathcal{L}_n = \min_{G_A^n, G_B^n, D_A^n, D_B^n} \max \mathcal{L}_{adv_n} + \lambda_{rec} \mathcal{L}_{rec_n} + \lambda_{cyc} \mathcal{L}_{cyc_n} \quad (3.18)$$

Trong đó, λ_{rec} và λ_{cyc} là hai siêu tham số.

Chương 4

Kết quả và thảo luận

4.1 Kết quả

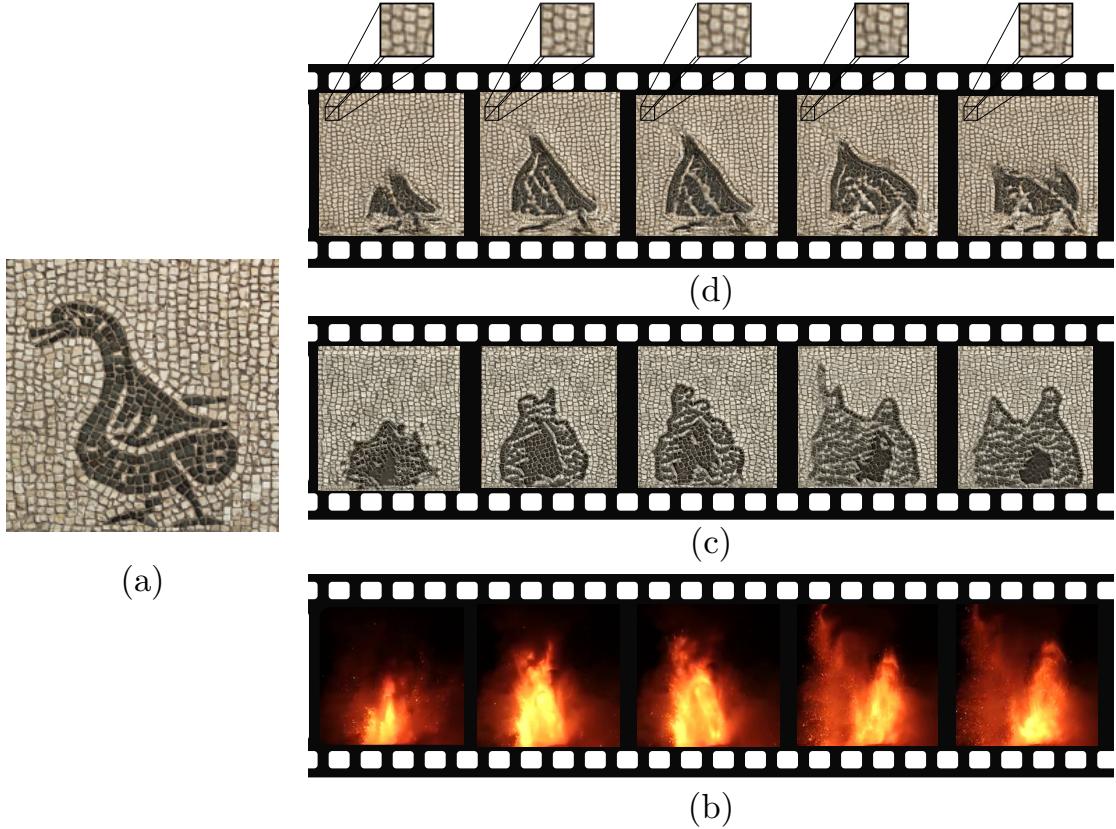
Cho trước một video và hình ảnh mục tiêu, mô hình có gắng chuyển đổi thành một video có chuyển động như video ban đầu nhưng có kiểu dáng, cấu trúc như hình ảnh mục tiêu. Video gốc bao gồm một chuỗi các khung hình f_1, \dots, f_k . Mô hình sẽ chuyển đổi từng khung hình f_i với $i = \overline{1, k}$ của video gốc sang từng khung hình v_i với $i = \overline{1, k}$ của video đầu ra. Khung hình v_i được định nghĩa như sau:

$$v_i = G_B^N(f_i + z) \quad (4.1)$$

Trong đó z là nhiễu Gauss được khởi tạo ngẫu nhiên và giữ cố định xuyên suốt cho tất cả các khung hình f_i . Nhieu z giúp cho các khung hình tạo ra liên tục về mặt thời gian, kết quả được trình bày trong hình 4.1. Nhờ vào cấu trúc đa tầng và các phương pháp thay đổi trên (được đề cập ở chương 3) đã giúp cho mô hình giảm được số lượng tham số khoảng 50.77% (xem bảng 4.1). Mô hình được huấn luyện với kỹ thuật tối ưu Adam [23], tốc độ học ban đầu là 0.0005, cứ sau mỗi tầng mô

Bảng 4.1: So sánh kiến trúc giữa nghiên cứu này và Benaim cùng cộng sự

Mô hình	Số lượng bộ sinh	Số lượng bộ phân biệt	Số tầng	Số lượng tham số
Nghiên cứu này	2	2	6	$\sim 1,320,000$
Benaim và cộng sự [4]	2	2	10	$\sim 2,600,000$



Hình 4.1: Video chuyển đổi. Mô hình chuyển đổi giữa video gốc là video núi lửa và hình ảnh mục tiêu là hình ảnh con vịt đá. (a) Hình ảnh mục tiêu. (b) Video gốc với các khung hình đầu vào f_i . (c) Kết quả các khung hình được chuyển đổi bằng mô hình của Benaim cùng nhóm cộng sự [4]. (d) Kết quả của các khung hình được chuyển đổi v_i bằng mô hình trong nghiên cứu này. Hình ảnh phóng to của (d) là một phần trong các khung hình v_i để thấy được các khung hình chuyển đổi xuyên suốt đều không có chuyển động.

hình sẽ giảm tốc độ học xuống 10 lần. Tại bất kỳ một tầng n , thay vì sử dụng cùng một tốc độ học cho tất cả các tầng $(n, n - 1, n - 2, \dots)$ thì dùng tốc độ học nhỏ hơn cho những tầng thấp hơn $(n - 1, n - 2, \dots)$ sẽ giúp giảm được hiện tượng quá khớp. Do đó, tốc độ học η với hệ số δ sẽ thay đổi ngày càng chậm. Cụ thể, tại mỗi tầng n , mô hình sẽ huấn luyện với tốc độ học là $\delta\eta$. Ví dụ, tầng thứ $n - 1$ sẽ được huấn luyện với tốc độ học $\delta^1\eta$ tầng thứ $n - 2$ sẽ huấn luyện với tốc độ học $\delta^2\eta$. Trong nghiên cứu này, cho thấy được là $\delta = 0.1$ là điểm cân bằng giữa chất lượng hình ảnh và độ đa dạng của hình ảnh tạo ra.

Mô hình được huấn luyện 3000 vòng lặp tại mỗi tầng, có tất cả 6 tầng. Các tham số trọng số của mô hình được cài đặt như mô hình của Benaim và cộng sự [4]

$\lambda_{pen} = 0.1$, $\lambda_{rec} = 1$ và $\lambda_{cyc} = 10$. Kết quả được trình bày trong hình 4.1. Khi kết quả không tốt, mô hình được thay đổi bằng cách tăng $\lambda_{cyc} = 20$ vì để giúp cho việc ánh xạ qua lại giữa hai miền ảnh tốt hơn. Tuy nhiên, điều này dẫn đến quá trình huấn luyện mô hình không thể hội tụ hoặc hội tụ với kết quả cũng không tốt. Nghiên cứu này đã thử trên rất bộ siêu tham số khác nhau nhưng kết quả đều không tốt.

Các khung hình được tạo ra vẫn giữ được tính liên tục về mặt thời gian, tức là những vùng cần chuyển động thì mới chuyển động còn những vùng nên cố định thì mô hình vẫn giữ được (xem phần phóng to của hình 4.1 (d)). Các thông tin về hoa văn và họa tiết của hình ảnh 4.1 (a) được mô hình học rất tốt. Tuy nhiên, khi chuyển đổi sang các khung hình thì kết quả mô hình tạo ra hoàn toàn không tốt. Mô hình không thể chuyển đổi được cấu trúc ngọn lửa theo được kiểu dáng của hình 4.1 (a). Có thể thấy các viền của ngọn lửa không được mượt (xem hình 4.1 (d)), các khung hình được tạo ra vẫn còn sót lại cấu trúc tổng quát của hình mục tiêu 4.1 (a) mà không thể chuyển đổi hoàn toàn. Trong khi kết quả của mô hình trong nghiên cứu Benaim và cộng sự (xem hình 4.1 (c)) thật sự rất tốt, mô hình của nhóm tác giả đã học được cả cấu trúc của ngọn lửa dưới kiểu dáng của tấm hình mục tiêu.

4.2 Thảo luận

Mô hình ConSinGAN giảm thời gian huấn luyện SinGAN xuống còn 20–30 phút và nghiên cứu của Benaim và cộng sự ứng dụng được SinGAN trong bài toán chuyển đổi hình ảnh sang hình ảnh. Do đó, khi kết hợp 2 mô hình lại với nhau, với mong muốn mô hình mới sẽ có thể tạo ra được kết quả tốt và vẫn có thể giữ lại những điểm mạnh về thời gian huấn luyện cũng như khả năng ứng dụng của mô hình. Mô hình trong nghiên cứu này đã giảm được thời gian huấn luyện đáng kể. Tuy nhiên, kết quả tạo ra lại đi ngược lại với những điều mong đợi từ ban đầu. Dưới đây là một số nguyên nhân có thể dẫn đến việc tạo kết quả trong mô hình không tốt:

- Cấu trúc mô hình đã được rút gọn dẫn đến việc học các thông tin trở nên vừa đủ làm cho mô hình không thể có nhiều thông tin để chuyển đổi.
- Trong quá trình thay đổi kích thước hình ảnh, độ phân giải chủ yếu là độ phân giải thấp, dẫn đến mô hình không đủ tầng để có thể học được các đặc trưng khác của hình ảnh.
- Bằng cách chỉ truyền vào mô hình các đặc trưng được học ở tầng trước đã làm cho mô hình học rất tốt cấu trúc tổng quát của tấm hình gốc nên dẫn đến việc mô hình chỉ có thể tạo ra được các kết quả giống với tấm hình gốc ban đầu (có thể nhìn nhận đây là hiện tượng sụp đổ mô hình).
- Tốc độ huấn luyện bộ sinh và bộ phân biệt trên hai miền ảnh là khác nhau, nghĩa là, quá trình huấn luyện G_A, D_A trên miền ảnh A đã hội tụ nhưng quá trình huấn luyện G_B, D_B trên miền ảnh B vẫn chưa hội tụ mà lại vội đưa kết quả qua tầng kế tiếp. Điều này, có thể hiểu như lỗi chồng lỗi khiến cho mô hình bị thiên về một miền ảnh, còn miền ảnh kia thì không học thêm được thông tin gì.

Chương 5

Kết luận

Tính đến hiện tại, số lượng các mô hình có thể giải quyết được bài toán chuyển đổi hình ảnh mà huấn luyện trên một cặp dữ liệu là rất ít [4, 25]. Trong khuôn khổ của khóa luận tốt nghiệp, nghiên cứu này cố gắng đưa ra hướng giải quyết bài toán chuyển đổi hình ảnh sang hình ảnh với thời gian huấn luyện ngắn nhất, nhưng kết quả mang lại không tốt. Vì để huấn luyện được mô hình GAN thông thường cần rất nhiều dữ liệu, chưa kể đến việc các mô hình chuyển đổi hình ảnh thường sẽ huấn luyện trên cả hai bộ sinh và hai bộ phân biệt nên sẽ rất dễ dẫn đến các hiện tượng như sụp đổ mô hình, dưới khớp, quá khớp,... Đồng thời, các mô hình GAN rất dễ nhạy cảm với các siêu tham số. Tóm lại, cách kết hợp trong mô hình của nghiên cứu này đưa ra là không phù hợp. Tuy nhiên, điều này đã làm tiền đề giúp cho các nghiên cứu sau có thể tránh đi theo vết xe đổ, nhờ đó sẽ không gây lãng phí về thời gian cũng như công sức vào hướng đi này.

Chương 6

Đề nghị nghiên cứu thêm

Với những kết quả như được trình bày trong chương 4 đã tạo ra động lực để có thể tiếp tục nghiên cứu phát triển mô hình này. Có thể nói, trong tương lai mô hình đã tránh được một trong các hướng đi thất bại. Trong thời gian sắp tới, nghiên cứu vẫn tiếp tục giải quyết bài toán chuyển đổi hình ảnh trên một cặp hình ảnh duy nhất với thời gian huấn luyện ngắn bằng cách đưa vào một video chứ không đơn thuần là một tấm ảnh như nghiên cứu này. Việc đưa vào một video có thể giúp cho mô hình học được nhiều thông tin hơn như *dòng quang* (*optical flow*) của một video. Đồng thời, có thể thử sử dụng phương pháp thay đổi kích thước hình ảnh như mô hình [4] để giúp cho mô hình có thể học được nhiều tầng hơn. Ngoài ra, còn có thể sử dụng thêm các hàm măt măt khác để làm cho mô hình phức tạp hơn nhằm học được các cấu trúc tổng quát của video . Hy vọng rằng, mô hình sẽ không chỉ dừng lại giải quyết bài chuyển đổi hình ảnh sang hình ảnh mà còn có thể mở rộng thêm các ứng dụng khác cho mô hình như siêu phân giải hình ảnh, hòa trộn ảnh, chỉnh sửa hình ảnh,...

Tài liệu tham khảo

- [1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. “A critical analysis of self-supervision, or what we can learn from a single image”. In: *8th International Conference on Learning Representations*. 2020.
- [2] Samaneh Azadi et al. “Multi-Content GAN for Few-Shot Font Style Transfer”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7564–7573.
- [3] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. “Blind Super-Resolution Kernel Estimation using an Internal-GAN”. In: *Advances in Neural Information Processing Systems 32*. 2019, pp. 284–293.
- [4] Sagie Benaim et al. “Structural-analogy from a Single Image Pair”. In: *CoRR* abs/2004.02222 (2020).
- [5] Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf. “Learning Texture Manifolds with the Periodic Spatial GAN”. In: *Proceedings of the 34th International Conference on Machine Learning*. 2017, pp. 469–477.
- [6] Taeg Sang Cho et al. “The patch transform and its applications to image editing”. In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2008.
- [7] Emily L Denton, Soumith Chintala, Rob Fergus, et al. “Deep generative image models using a Laplacian pyramid of adversarial networks”. In: *Advances in neural information processing systems*. 2015, pp. 1486–1494.
- [8] Ahmed M. Elgammal et al. “CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms”. In: *Proceedings of the Eighth International Conference on Computational Creativity*. 2017, pp. 96–103.

- [9] Yossi Gandelsman, Assaf Shocher, and Michal Irani. ““Double-DIP”: Unsupervised Image Decomposition via Coupled Deep-Image-Priors”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 11018–11027.
- [10] Daniel Glasner, Shai Bagon, and Michal Irani. “Super-resolution from a single image”. In: *IEEE 12th International Conference on Computer Vision*. 2009, pp. 349–356.
- [11] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in Neural Information Processing Systems*. 2014, pp. 2672–2680.
- [12] Ishaan Gulrajani et al. “Improved training of wasserstein GANs”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 5767–5777.
- [13] Aaron Hertzmann et al. “Image analogies”. In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM. 2001, pp. 327–340.
- [14] Tobias Hinz et al. “Improved Techniques for Training Single-Image GANs”. In: *CoRR* abs/2003.11512 (2020).
- [15] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. “Single Image Super-Resolution From Transformed Self-Exemplars”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.
- [16] Xun Huang et al. “Stacked generative adversarial networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5077–5086.
- [17] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. “Globally and Locally Consistent Image Completion”. In: *ACM Trans. Graph.* (2017), 107:1–107:14.
- [18] Phillip Isola et al. “Image-to-Image Translation with Conditional Adversarial Networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5967–5976.
- [19] Nikolay Jetchev, Urs Bergmann, and Roland Vollgraf. “Texture synthesis with spatial generative adversarial networks”. In: *Workshop on Adversarial Training, NIPS* (2016).

- [20] Donggyu Joo, Doyeon Kim, and Junmo Kim. “Generating a Fusion Image: One’s Identity and Another’s Shape”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1635–1643.
- [21] Tero Karras et al. “Progressive Growing of GANs for Improved Quality, Stability, and Variation”. In: *6th International Conference on Learning Representations*. 2018.
- [22] Taeksoo Kim et al. “Learning to Discover Cross-Domain Relations with Generative Adversarial Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. 2017, pp. 1857–1865.
- [23] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations*. 2015.
- [24] Chuan Li and Michael Wand. “Precomputed real-time texture synthesis with markovian generative adversarial networks”. In: *European Conference on Computer Vision*. 2016, pp. 702–716.
- [25] Jianxin Lin et al. “TuiGAN: Learning Versatile Image-to-Image Translation with Two Unpaired Images”. In: *CoRR* abs/2004.04634 (2020).
- [26] Ming-Yu Liu et al. “Few-Shot Unsupervised Image-to-Image Translation”. In: *2019 IEEE/CVF International Conference on Computer Vision*. 2019, pp. 10550–10559.
- [27] Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. “Saliency Driven Image Manipulation”. In: *2018 IEEE Winter Conference on Applications of Computer Vision*. 2018, pp. 1368–1376.
- [28] Tomer Michaeli and Michal Irani. “Blind deblurring using internal patch recurrence”. In: *European Conference on Computer Vision*. Springer. 2014, pp. 783–798.
- [29] Weili Nie, Nina Narodytska, and Ankit Patel. “RelGAN: Relational Generative Adversarial Networks for Text Generation”. In: *International Conference on Learning Representations*. 2019.
- [30] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. “SinGAN: Learning a Generative Model From a Single Natural Image”. In: *2019 IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4569–4579.

- [31] Assaf Shocher, Nadav Cohen, and Michal Irani. ““Zero-Shot” Super-Resolution using Deep Internal Learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3118–3126.
- [32] Tal Tlusty et al. “Modifying Non-local Variations Across Multiple Views”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 6276–6285.
- [33] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. “Deep Image Prior”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 9446–9454.
- [34] Ting-Chun Wang et al. “High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.
- [35] Yaxing Wang et al. “SDIT: Scalable and Diverse Cross-domain Image Translation”. In: *Proceedings of the 27th ACM International Conference on Multimedia*. 2019, pp. 1267–1276.
- [36] Hongyu Yang et al. “Learning Face Age Progression: A Pyramid Architecture of GANs”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 31–39.
- [37] Wenming Yang et al. “Deep Learning for Single Image Super-Resolution: A Brief Review”. In: *IEEE Transactions on Multimedia* 21 (2019), pp. 3106–3121.
- [38] Zili Yi et al. “DualGAN: Unsupervised Dual Learning for Image-to-Image Translation”. In: *IEEE International Conference on Computer Vision*. 2017, pp. 2868–2876.
- [39] Haotian Zhang et al. “An Internal Learning Approach to Video Inpainting”. In: *2019 IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2720–2729.
- [40] Zhifei Zhang, Yang Song, and Hairong Qi. “Age Progression/Regression by Conditional Adversarial Autoencoder”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4352–4360.

- [41] Yang Zhou et al. “Non-stationary texture synthesis by adversarial expansion”. In: *ACM Trans. Graph.* (2018), 49:1–49:13.
- [42] Jun-Yan Zhu et al. “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”. In: *IEEE International Conference on Computer Vision*. 2017.
- [43] Maria Zontak, Inbar Mosseri, and Michal Irani. “Separating Signal from Noise Using Patch Recurrence across Scales”. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 1195–1202.