Tạo hình ảnh từ văn bản

Phạm Trung Sơn
Trường đại học Công nghiệp thành phố Hồ Chí Minh
MSSV: 19518291

Nguyễn Tuấn Sinh
Trường đại học Công nghiệp thành phố Hồ Chí Minh
MSSV: 19477821

Trần Hồ Phi Long
Trường đại học Công nghiệp thành phố Hồ Chí Minh
MSSV: 19500551

Nguyễn Tiến Sỹ
Trường đại học Công nghiệp thành phố Hồ Chí Minh
MSSV: 19440181

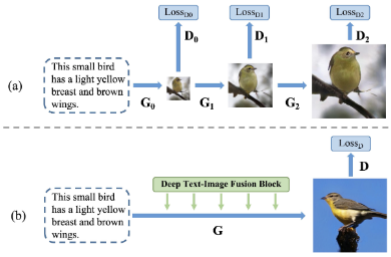
Tóm tắt nội dung*—Tạo hình ảnh thật có chất lượng cao từ vản bản mô tả là một nhiệm vụ thách thức. Các mạng Generative Adversarial Networks generally hiện có cho bài toán text-to-image sử dụng kiến trúc stack (xếp chồng) nhưng vẫn còn ba lỗi. Đầu tiên kiến trúc stack giới thiệu sự vướng víu giữa các Generator các tỷ lệ hình ảnh khác nhau. Thứ hai các nghiên cứu hiện có thích áp dụng và khắc phục thêm mạng để mô hình có tính nhất quán giữa ngữ nghĩa và hình ảnh, điều này làm hạn chế khả năng quan sát của mô hình. Thứ ba, sự kết hợp văn bản và hình ảnh dựa trên cross-modal attetion-base (sự chú ý đa phương thức) đã áp dụng rộng rãi giữa các công trình trước đó bị giới hạn ở một số tỉ lệ hình ảnh đặc biệt do chi phí tính toán. Để giải quyết vấn đề này, chúng tôi đề xuất một cách đơn giản hơn nhưng hiệu quả hơn là Deep Fusion Generative Adversarial Networks (DF-GAN). Cụ thể, (i) kiến trúc backbone one-stage text-to-image để tạo hình ảnh có độ phân giải cao trực tiếp không có sự vướng víu các mạng Generator khác nhau. (ii) novel target-Aware Discriminator composed of Matching- Aware Gradient Penalty and One-Way Output giúp tăng cường tính nhất quán ngữ nghĩa của hình ảnh, văn bản mà không cần thêm các mạng bổ sung. (iii) a novel deep text-image fusion block, làm sâu sắc thêm quá trình hợp nhất để tạo sự hợp nhất đầy đủ giữa văn bản và các đặc trưng ảnh.So với các phương pháp tiên tiến hiện nay, DF-GAN được đề xuất của chúng tôi đơn giản hơn nhưng hiệu quả hơn để tạo các hình ảnh thực tế và khớp với văn bản và đạt được hiệu suất tốt hơn trên các bộ dữ liệu được sử dụng rộng rãi

1 GIỚI THIỆU

Vài năm qua đã chứng kiến sự thành công rực rỡ của Generative Adversarial Networks (GANs)[8] cho nhiều ứng dụng. Trong số đó, bài toán text-to-image là một trong những ứng dụng quan trọng. Mục đích là tạo ra các hình ảnh thực phù hợp với văn bản từ các mô tả tự nhiên nhất định.Do giá trị thực tế mà text-to-image đã trở thành một lĩnh vực nghiên cứu tích cực gần đây.[3, 9, 13, 19–21, 32, 33, 35, 51, 53, 60].

Hai thách thức lớn đối với bài toán text-to-image synthethis là tính xác thực của hình ảnh được tạo và sự nhất quán về ngữ nghĩa giữa văn bản đã cho và hình ảnh được tạo ra. Do tính không ổn định của mô hình Gan hầu hết những mô hình gần đây đều không áp dụng kiến trúc stack (xếp chồng) [56,57] như backbone để tạo ra hình ảnh có độ phân giải cao. Họ sử dụng cross-modal attention để kết hợp các feature(đặc trưng) của hình ảnh và văn bản [37, 50, 56, 57, 60].

Mặc dù có những kết quả ấn tượng đã được trình bày trong các công trình trước đó nhưng vẫn có ba vấn đề. Đầu tiên là kiến trúc Stack thì có sự vướng víu giữa các Generator Networks và điều này làm cho hình ảnh sinh ra cuối cùng trông giống như sự kết hợp đơn giản giữa hình mờ và một số chi tiết. Như hình 1 đã cho thấy hình ảnh mờ được tạo ra từ G_0 các thuộc tính thô như (mắt, mỏ) được tạo ra từ G_1 và các chi tiết tinh vi được thêm bởi G_2. Hình ảnh cuối cùng trông giống như sự kết hợp đơn giản của các đặc điểm từ các tỉ lệ hình ảnh khác nhau.



Hình 1. (a) Mô hình stack Gan. (b) Mô hình DF-Gan

Thứ hai các nghiên cứu hiện tại thường khắc phục bằng cách bổ sung thêm các mạng. Trong suốt quá trình adversarial training làm cho các mạng trở nên kém đi. Thứ ba cross-modal attention không tận dụng hết thông tin của văn bản. Họ chỉ có thể áp dụng hai lần trên image 64x64 và 128x128 do chi phí tính toán của nó. Nó làm hạn chế hiệu xuất của model và làm cho mô hình khó mở rộng để áp dụng cho hình ảnh có độ phân giải cao hơn.

Đẻ giải quyết những vấn đề trên. chúng tôi đề xuất một phương pháp mới có tên là Deep Fusion Genera-tive Adversarial Network (DF-GAN). Đối với vấn đề đầu tiên, chúng tôi thay thế kiến trúc stack backbone vơi one-stage backbone. Nó bao gồm hinge loss [54] và residual network [11] giúp ổn định quá trình đào tạo GAN để tổng hợp hình ảnh có độ phân giải cao một cách trực tiếp. Vì chỉ có một Generator network trong one-stage nên tránh được sự vướng víu giữa các generator khác nhau.

Để giải quyết vấn đề 2 chúng tôi thiết kế a Target-Aware Discriminator composed of Matching-Aware Gradient Penalty (MA-GP) và One-Way Output để tăng cường tính nhất quán giữa văn bản và hình ảnh. MA-GP là chiến lượng chính quy

hóa trên mạng Discriminator. Nó theo Gradient của Discriminator trên dữ liêu target (hình ảnh thực và text-matching image) là không. Do đó MA-GP làm cho hàm loss (mất mát) trở nên mịn hơn tại các điểm dữ liệu thực và phù hợp, điều này làm thúc đẩy cho Generator sinh ra ảnh khớp với văn bản. Ngoài ra Two-Way output làm chậm quá trình hội tụ của Generator nên chúng tôi sử dụng One Way output.

Còn vấn đề 3 chúng tôi đề xuất Deep text-image Fusion Block (DFBlock) để kết hợp thông tin của văn bản vào các đặc trưng của ảnh hiệu quả hơn. DFBlock bao gồm các Affine Transformations[31]. Affine Transformations là mô hình vận dụng các đặc trưng thông qua chanel-wise scaling và shifting.

Tóm lại : chúng tôi đề xuất one-stage text-to-image để có thể sinh ra ảnh có độ phân giải cao một cách trực tiếp mà không qua các mạng Generator khác nhau.

Chúng tôi đề xuất a novel Target-Aware Discriminator composed of Matching-Aware Gradient Penalty (MA-GP) và One-Way Output nó tăng cường tính nhất quán giữa ngữ nghĩa và hình ảnh mà không cần thêm các mạng bổ sung.

Chúng tôi đề xuất một Deep text-image Fusion Block (DFBlock), để kết hợp đầy đủ giữa văn bản và các đặc trưng ảnh một cách hiệu quả và sâu hơn

2 Nghiên cứu liên quan

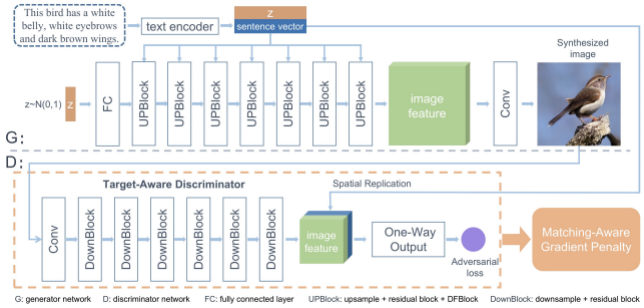
Mô hình CGan là mô hình đầu tiên được áp dụng để giải quyết bài toán text-to-image [37, 38]. StackGan [56,57] sinh ra các ảnh có độ phân giải cao bằng nhiều mạng Generator và Discriminator và cung cấp thông tin cho mạng Generator bằng cách nối các vector text và input noise vector. Tiếp theo, attnGan [50] sử dụng cơ chế cross-modal attention(chú ý đa phương thức) giúp mô hình sinh ra ảnh với nhiều chi tiết hơn. Mirror Gan [33] khôi phục lại văn bản từ hình ảnh được tạo cho tính nhất quán giữa ngữ nghĩa và hình ảnh.SD-Gam [51] sử dụng cấu trúc Siamese [45,46] để chắt lọc những điểm chung về ngũ nghĩa từ các văn bản để tạo ra hình ảnh nhất quán . DM Gan[60] sử dụng memory network [10, 49] để tinh chỉnh hình ảnh mờ khi hình ảnh ban đầu không được tạo tốt trong kiến trúc stack.

Mô hình Gan của chúng tôi khác nhiều so với các kiến trúc trước đây. Đầu tiên nó sinh ra hình ảnh có độ phân giải cao trực tiếp thông qua One-stage backbone. Thứ hai, nó thông qua một Target-Aware Discriminator nâng cao tính nhất quán giữa ngữ nghĩa và hình ảnh mà không cần thêm các mạng. Thứ ba nó hợp nhất các đặc trưng của văn bản và hình ảnh một cách sâu sắc hơn và hiệu quả hơn thông qua một chuỗi DFBlocks. So sánh với mô hình trước thì mô hình DF-Gan đơn giản nhưng hiệu quả hơn.

3 Method

3.1 *Kiến trúc DF-GAN*

Trong bài báo này, nhóm đã đề xuất một mô hình đơn giản để tổng hợp văn bản thành hình ảnh có tên là Deep Fusion GAN(DF-GAN). Để tổng hợp các hình ảnh phù hợp với văn bản và thực tế hơn, chúng tôi đề xuất: (i) một đường trục chuyển văn bản thành hình ảnh một giai đoạn có thể tổng hợp trực tiếp các hình ảnh có độ phân giải cao mà không vướng víu về tính năng thực tế. (ii) một Discriminator nhận biết mục tiêu mới bao gồm Matching-Aware Gradient Penalty (MA-GP), giúp tăng cường tính nhất quán ngữ nghĩa hình ảnh văn bản mà không cần giới thiệu thêm mạng. (iii) Một khối kết hợp hình ảnh văn bản sâu mới(DFBlock), kết hợp đầy đủ hơn các tính năng hình ảnh và văn bản.



Hình 2. Kiến trúc của DF-GAN được đề xuất để tổng hợp văn bản thành hình ảnh. DF-GAN tạo hình ảnh có độ phân giải cao trực tiếp bằng cách một cặp bộ tạo và bộ phân biệt và kết hợp thông tin văn bản và bản đồ đặc điểm trực quan thông qua nhiều Kết hợp hình ảnh văn bản sâu Khối (DFBlock) trong UPBlocks. Được trang bị Hình phạt chuyển màu nhận biết phù hợp (MA-GP) và Đầu ra một chiều, mô hình của nhóm có thể tổng hợp hình ảnh thực tế hơn và phù hợp với văn bản.

3.2 *Tổng quan mô hình*

DF-GAN được đề xuất bao gồm một bộ generator, một bộ discriminator và bộ mã hóa văn bản được pretrain như trong hình trên. Trình tạo có hai đầu vào, một vectơ câu được mã hóa bằng bộ mã hóa văn bản và một vectơ nhiễu được lấy mẫu từ phân phối Gaussian để đảm bảo tính đa dạng của hình ảnh được tạo. Đầu tiên, vectơ nhiễu được đưa vào một lớp được kết nối đầy đủ và được định hình lại. Sau đó, chúng tôi áp dụng một loạt UPBlocks để lấy mẫu lại các đặc điểm của hình ảnh. UPBlock là bao gồm một lớp lấy mẫu, một khối dư và DFBlocks để hợp nhất các tính năng văn bản và hình ảnh. Cuối cùng, một lớp tích chập chuyển đổi đặc điểm hình ảnh thành hình ảnh.

Bộ discriminator chuyển đổi hình ảnh thành các tính năng hình ảnh thông qua một loạt DownBlocks. Khi đó vectơ câu sẽ được sao chép và kết nối với các tính năng hình ảnh. Một mất đối thủ(adversarial loss) sẽ được sử dụng để đánh giá tính hiện thực trực quan và tính nhất quán về ngữ nghĩa của đầu vào. Bằng cách phân biệt hình ảnh được tạo từ các mẫu thực, bộ discriminator thúc đẩy trình tạo tổng hợp hình ảnh với chất lượng cao hơn và nhất quán ngữ nghĩa hình ảnh văn bản.

Bộ mã hóa văn bản là Long Short-Term Memory (LSTM) [41] trích xuất các vectơ ngữ nghĩa từ mô tả văn bản. Nhóm trực tiếp sử dụng mô hình được đào tạo trước được cung cấp bởi AttnGAN [50].

3.3 *Trục chuyển văn bản thành hình ảnh một giai đoạn(One-Stage Text-to-Image Backbone)*

Do mô hình GAN không ổn định, các GAN chuyển văn bản thành hình ảnh trước đây thường sử dụng kiến trúc xếp chồng [56,57] để tạo hình ảnh có độ phân giải cao từ những

hình ảnh có độ phân giải thấp. Tuy nhiên, kiến trúc xếp chồng lên nhau gây vướng víu giữa các máy phát điện khác nhau và nó tạo ra sản phẩm tinh chế cuối cùng hình ảnh trông giống như một sự kết hợp đơn giản của hình dạng mờ và một số chi tiết(Hình 1.a).

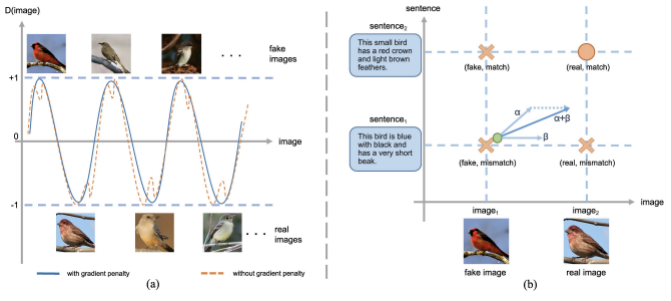
Lấy cảm hứng từ các nghiên cứu gần đây về tạo hình ảnh vô điều kiện [23, 54], chúng tôi đề xuất một đường trục chuyển văn bản thành hình ảnh một giai đoạn có thể tổng hợp hình ảnh có độ phân giải cao trực tiếp bằng cách một cặp generator và discriminator. Chúng tôi sử dụng mất bản lề [23] để ổn định quá trình đào tạo đối thủ. Vì chỉ có một trình generator trong đường trục một tầng, nó tránh được sự vướng víu giữa các generator khác nhau. Là trình generator duy nhất trong khuôn khổ một giai đoạn của chúng tôi cần để tổng hợp trực tiếp các hình ảnh có độ phân giải cao từ các vectơ nhiễu, nó phải chứa nhiều lớp hơn các trình generator trước đó trong kiến trúc xếp chồng lên nhau. Để đào tạo các lớp này một cách hiệu quả, chúng tôi giới thiệu các mạng còn lại [11] để ổn định việc đào tạo mạng sâu hơn. Việc xây dựng phương pháp một giai đoạn của chúng tôi được tính toán loss như sau:

$$
\begin{aligned}
L_D = & -\mathbb{E}_{x\sim\mathbb{P}_r}[\min(0,-1+D(x,e))] \\
& -(1/2)\mathbb{E}_{G(z)\sim\mathbb{P}_g}[\min(0,-1-D(G(z),e))] \\
& -(1/2)\mathbb{E}_{x\sim\mathbb{P}_{\text{mis}}}[\min(0,-1-D(x,e))] \\
L_G = & -\mathbb{E}_{G(z)\sim\mathbb{P}_g}[D(G(z),e)]
\end{aligned}
$$

trong đó z là vectơ nhiễu được lấy mẫu từ phân phối Gaussian; e là vectơ câu; Pg, Pr, Pmis lần lượt biểu thị phân phối dữ liệu tổng hợp, phân phối dữ liệu thực và phân phối dữ liệu không khớp.

3.4 *Công cụ Discriminator nhận biết mục tiêu*

Trong phần này, chúng tôi trình bày chi tiết về Target-Aware được đề xuất. Discriminator, Matching-Aware Gradient Penalty (MA-GP) và Đầu ra một chiều. Target-Aware Discriminator thúc đẩy trình tạo tổng hợp hình ảnh thực tế hơn và văn bản phù hợp với ngữ nghĩa hình ảnh.



Hình 3. (a) So sánh cảnh quan mất mát trước và sau khi áp dụng hình phạt độ dốc. Hình phạt gradient làm mịn bộ discriminator bề mặt mất mát hữu ích cho sự hội tụ của máy phát. (b) Một sơ đồ của MA-GP. Điểm dữ liệu (thực, khớp) nên được áp dụng MA-GP.

3.4.1 ***Hình phạt chuyển màu nhận biết phù hợp:*** Hình phạt Gradient không tập trung vào Matching-Aware (MA-GP) là chiến lược mới được thiết kế của nhóm để nâng cao hình ảnh văn bản nhất quán ngữ nghĩa. Trong tiểu mục này, trước tiên nhóm chỉ ra hình phạt gradient vô điều kiện [28] từ một cuốn tiểu thuyết và rõ ràng phối cảnh, sau đó mở rộng nó sang MA-GP của nhóm cho tác vụ tạo văn bản thành hình ảnh.

Như được hiển thị trong Hình 3(a), trong quá trình tạo ảnh vô điều kiện, dữ liệu đích (ảnh thực) tương ứng với mức thấp discriminator loss. Tương ứng, các hình ảnh tổng hợp tương ứng với tổn thất phân biệt đối xử cao. Discriminator giới hạn phạm vi tổn thất phân biệt giữa -1 và 1. hình phạt gradient trên dữ liệu thực sẽ làm giảm độ dốc của điểm dữ liệu thực và vùng lân cận của nó. Bề mặt của hàm mất mát xung quanh điểm dữ liệu thực sau đó được làm nhẵn, đó là hữu ích cho điểm dữ liệu tổng hợp để hội tụ với thực điểm dữ liệu

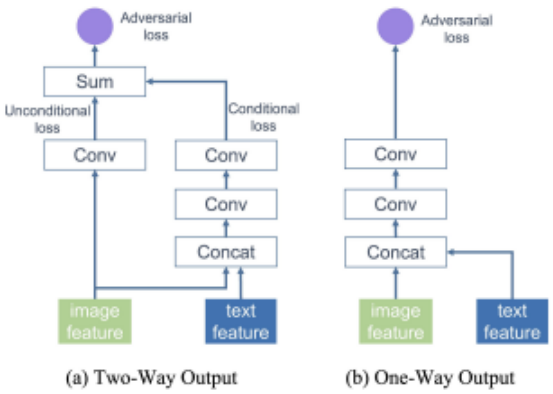
Dựa trên phân tích ở trên, chúng tôi thấy rằng hình phạt độ dốc trên dữ liệu đích sẽ xây dựng cảnh quan tổn thất tốt hơn để giúp trình tạo hội tụ. Bằng cách tận dụng xem vào quá trình tạo văn bản thành hình ảnh. Như được hiển thị trong Hình 3(b), trong quá trình tạo văn bản thành hình ảnh, phân biệt quan sát bốn loại đầu vào: hình ảnh tổng hợp có văn bản khớp (giả, khớp), hình ảnh tổng hợp không khớp văn bản (giả, không khớp), hình ảnh thực với văn bản phù hợp (thật, khớp), hình ảnh thực với văn bản không khớp (thực, không khớp). Đối với tính nhất quán ngữ nghĩa văn bản-hình ảnh, nhóm có xu hướng áp dụng hình phạt gradient trên dữ liệu thực phù hợp với văn bản, mục tiêu của tổng hợp văn bản thành hình ảnh. Do đó, trong MA-GP, hình phạt gradient nên được áp dụng trên các hình ảnh thực có khớp chữ. Toàn bộ công thức của mô hình của chúng tôi với MA-GP là như sau:

$$
\begin{aligned}
L_D = & -\mathbb{E}_{x\sim\mathbb{P}_r}[\min(0,-1+D(x,e))] \\
& -(1/2)\mathbb{E}_{G(z)\sim\mathbb{P}_g}[\min(0,-1-D(G(z),e))] \\
& -(1/2)\mathbb{E}_{x\sim\mathbb{P}_{mis}}[\min(0,-1-D(x,e))] \\
& +k\mathbb{E}_{x\sim\mathbb{P}_r}[(\|\nabla_x D(x,e)\|+\|\nabla_e D(x,e)\|)^p] \\
L_G = & -\mathbb{E}_{G(z)\sim\mathbb{P}_g}[D(G(z),e)]
\end{aligned}
$$

trong đó k và p là hai siêu tham số để cân bằng hiệu quả của hình phạt độ dốc.

Bằng cách sử dụng tổn thất MA-GP như một quy tắc hóa trên discriminator, mô hình của chúng tôi có thể hội tụ tốt hơn với dữ liệu thực khớp văn bản, do đó tổng hợp được nhiều hình ảnh khớp văn bản hơn. Bên cạnh đó, kể từ khi discriminator là cùng được đào tạo trong mạng của chúng tôi, nó ngăn trình tạo tổng hợp các tính năng bất lợi của mạng bổ sung cố định. Ngoài ra, do MA-GP không kết hợp bất kỳ mạng bổ sung nào để đảm bảo tính nhất quán của hình ảnh văn bản và độ dốc đã được tính toán bằng quy trình lan truyền ngược, tính toán duy nhất được giới thiệu bởi MA-GP được đề xuất của chúng tôi là độ dốc tổng kết, thân thiện với tính toán hơn các mạng bổ sung.

3.4.2 ***Đầu ra một chiều(One-Way Output):*** Trong các GAN chuyển văn bản thành hình ảnh trước đó [50,56,57], các đặc điểm hình ảnh được trích xuất bởi bộ phân biệt thường được sử dụng trong hai cách (Hình 4(a)): người ta xác định xem hình ảnh là thật hay giả, cái kia nối đặc điểm hình ảnh và vectơ câu để đánh giá tính nhất quán ngữ nghĩa của văn



Hình 4. So sánh giữa Đầu ra hai chiều và Đầu ra một chiều của chúng tôi. (a) Đầu ra hai chiều dự đoán tổn thất có điều kiện và mất mát vô điều kiện và tổng hợp chúng như là đối thủ cuối cùng sự mất mát. (b) Đầu ra một chiều của chúng tôi dự đoán toàn bộ tổn thất đối thủ trực tiếp.

bản-hình ảnh. Tương ứng, tổn thất vô điều kiện và tổn thất có điều kiện được tính toán trong các mô hình này.

Tuy nhiên, nó cho thấy rằng Đầu ra hai chiều yếu đi hiệu quả của MA-GP và làm chậm quá trình hội tụ của máy phát. Cụ thể, như được mô tả trong Hình 3(b), tổn thất có điều kiện tạo ra một gradient và chỉ vào đầu vào thực và phù hợp sau khi lan truyền ngược, trong khi mất mát vô điều kiện mang lại một gradient chỉ trỏ đến hình ảnh thực tế. Tuy nhiên, hướng của gradient cuối cùng mà chỉ đơn giản là tổng hợp và không trỏ đến thực và khớp các điểm dữ liệu như chúng tôi mong đợi. Kể từ khi mục tiêu của trình tạo là tổng hợp các hình ảnh thực và khớp với văn bản, độ dốc cuối cùng với độ lệch không thể đạt được nhất quán ngữ nghĩa văn bản-hình ảnh và làm chậm quá trình hội tụ của trình generator.

Do đó, chúng tôi đề xuất Đầu ra một chiều để tổng hợp văn bản thành hình ảnh. Như được hiển thị trong Hình 4(b), bộ phân biệt đối xử của chúng tôi nối đặc điểm hình ảnh và vectơ câu, sau đó chỉ xuất ra một tổn thất đối thủ thông qua hai lớp tích chập. Thông qua Đầu ra một chiều, chúng tôi có thể tạo độ dốc đơn trỏ đến các điểm dữ liệu đích (thực và khớp) trực tiếp, giúp tối ưu hóa và tăng tốc hội tụ của generator.

Bằng cách kết hợp MA-GP và Đầu ra một chiều, chúng tôi Target-Aware Discriminator có thể hướng dẫn trình tạo tổng hợp các hình ảnh thực và phù hợp với văn bản hơn.

*3.5 **Kết hợp văn bản - hình ảnh hiệu quả(Efficient Text-Image Fusion)***

Để hợp nhất các tính năng văn bản và hình ảnh một cách hiệu quả, chúng tôi đề xuất một Khối kết hợp hình ảnh văn bản sâu (DFBlock) mới lạ. So với các mô-đun hợp nhất hình ảnh văn bản trước đây, DFBlock của chúng tôi tăng cường quá trình hợp nhất hình ảnh văn bản để tạo ra một kết hợp văn bản-hình ảnh

Như được hiển thị trong Hình 2, trình tạo DF-GAN của chúng tôi bao gồm 7 UPBlocks. Một UPBlock chứa hai Văn bản-Hình ảnh khối hợp nhất. Để sử dụng đầy đủ thông tin văn bản trong quá trình hợp nhất, chúng tôi đề xuất Khối hợp nhất hình ảnh văn bản sâu (DFBlock) ngăn xếp nhiều Biến đổi Affine và Các lớp ReLU trong Fusion Block. Đối với phép biến đổi Affine, như được hiển thị trong Hình 5 (c), chúng tôi áp dụng hai MLP (Multilayer Perceptron) để dự đoán kênh thông minh có điều kiện ngôn ngữ tham số chia tỷ lệ và thay đổi tham số từ vectơ câu e, tương ứng:

$$\gamma = MLP_1(\boldsymbol{e}), \quad \boldsymbol{\theta} = MLP_2(\boldsymbol{e}).$$

Đối với một bản đồ tính năng đầu vào nhất định XR B×C×H×W , chúng tôi đầu tiên tiến hành thao tác mở rộng kênh thông minh trên X với tham số tỷ lệ , sau đó áp dụng dịch chuyển kênh khôn ngoan hoạt động với tham số dịch chuyển . Một quá trình như vậy có thể được thể hiện như sau:

$$AFF\,(\boldsymbol{x_i} \mid \boldsymbol{e}) = \gamma_i \cdot \boldsymbol{x_i} + \theta_i,$$

trong đó AFF biểu thị Biến đổi Affine; xi là tôi kênh bản đồ đặc trưng trực quan; e là vectơ câu; i và i là tham số tỷ lệ và tham số dịch chuyển cho cái tôi kênh bản đồ đặc trưng trực quan

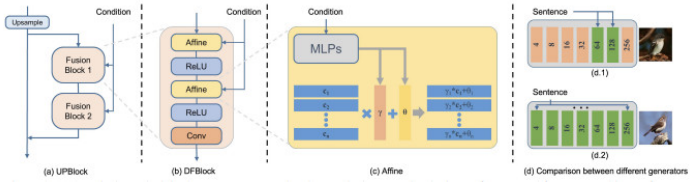
Lớp Affine mở rộng biểu diễn có điều kiện không gian của máy phát điện. Tuy nhiên, phép biến đổi Affine là một biến đổi tuyến tính cho mỗi kênh. Nó hạn chế hiệu quả của quá trình hợp nhất văn bản-hình ảnh. Qua đó, chúng tôi bổ sung một lớp ReLU giữa hai lớp Affine mang lại phi tuyến tính vào quá trình nhiệt hạch. Nó mở rộng không gian biểu diễn có điều kiện so với chỉ một Affine lớp. Không gian biểu diễn lớn hơn sẽ hữu ích cho trình tạo ánh xạ các hình ảnh khác nhau thành các biểu diễn khác nhau theo văn bản miêu tả

DFBlock của chúng tôi một phần được lấy cảm hứng từ Lô có điều kiện Chuẩn hóa (CBN) [5] và Chuẩn hóa phiên bản thích ứng (AdaIN) [14, 16] có chứa phép biến đổi Affine. Tuy nhiên, cả CBN và AdaIN đều sử dụng các lớp chuẩn hóa [15,44] để biến các bản đồ đối tượng thành phân phối bình thường. Nó tạo ra một hiệu ứng ngược lại với Biến đổi Affine dự kiến sẽ tăng khoảng cách giữa các mẫu khác nhau. Sau đó, nó là vô ích cho quá trình tạo có điều kiện. Để kết thúc này, chúng tôi loại bỏ các quá trình bình thường hóa. Hơn nữa, DFBlock của chúng tôi tăng cường quá trình hợp nhất văn bản-hình ảnh. Chúng tôi xếp chồng nhiều lớp Affine và thêm một lớp ReLU ở giữa. Nó thúc đẩy sự đa dạng của các tính năng trực quan và mở rộng không gian biểu diễn để thể hiện các tính năng trực quan khác nhau theo văn bản khác nhau mô tả.

Với việc đào sâu quá trình hợp nhất, DFBlock mang lại hai lợi ích chính cho việc tạo văn bản thành hình ảnh: Thứ nhất, nó làm cho trình generator khai thác đầy đủ hơn thông tin văn bản khi kết hợp các tính năng văn bản và hình ảnh. Thứ hai, đào sâu quá trình hợp nhất sẽ mở rộng không gian biểu diễn của mô-đun hợp nhất, có lợi để tạo ngữ nghĩa hình ảnh nhất quán từ các mô tả văn bản khác nhau.

Hơn nữa, so với chuyển văn bản thành hình ảnh trước đây GAN [50, 56, 57, 60], DFBlock được đề xuất làm cho chúng tôi mô hình không còn xem xét giới hạn từ tỷ lệ hình ảnh khi hợp nhất các tính năng văn bản và hình ảnh. Điều này là do các GAN chuyển văn bản thành hình ảnh hiện tại thường sử

dụng cơ chế chú ý đa phương thức, cơ chế này chịu sự tăng trưởng nhanh chóng của chi phí tính toán cùng với sự gia tăng kích thước hình ảnh.



Hình 5. (a) Một UPBlock điển hình trong mạng generator. UPBlock lấy mẫu các tính năng hình ảnh và hợp nhất các tính năng văn bản và hình ảnh bởi hai Khối Fusion. (b) DFBlock bao gồm hai lớp Affine, hai lớp kích hoạt ReLU và một lớp Convolution. (c) Các minh họa của Biến đổi Affine. (d) So sánh giữa (d.1) trình generator với sự chú ý đa phương thức [50, 60] và (d.2) của generator với DFBlock.

4 THỰC NGHIỆM

Trong phần này chúng tôi giới thiệu bộ dữ liệu, huấn luyện chi tiết và chỉ số đánh giá được sử dụng trong các thử nghiệm của chúng tôi, sau đó đánh giá DF-GAN và các biến thể của nó một cách định lượng và định tính. Bộ dữ liệu chúng tôi theo dõi đó là chim CUB [47]. Các Bộ dữ liệu CUB chứa 11.788 hình ảnh thuộc về 200 loài chim. Mỗi hình ảnh con chim có mười ngôn ngữ mô tả. Chi Tiết Đào Tạo. Chúng tôi tối ưu hóa mạng của mình bằng Adam [18] với 1=0,0 và 2=0,9. Tỷ lệ học tập được đặt thành 0,0001 cho bộ generator và 0,0004 cho bộ discriminator theo Quy tắc cập nhật hai thang thời gian (TTUR) [12].

Chung tôi lựa chọn đánh giá bằng Inception Score (IS) [40] và Fr´echet Inception Distance (FID)[12].

Cụ thể, IS tính Kullback-Leibler (KL) sự khác biệt giữa phân phối có điều kiện và cận biên phân bổ. IS cao hơn có nghĩa là chất lượng hình ảnh được tạo ra cao hơn và mỗi hình ảnh rõ ràng thuộc về một hình ảnh cụ thể.

lớp. FID [12] tính toán khoảng cách Fr´echet giữa phân phối hình ảnh tổng hợp và hình ảnh trong thế giới thực trong không gian tính năng của mạng Inception v3 được đào tạo trước.

Trái ngược với IS, hình ảnh chân thực hơn có FID thấp hơn. Đến tính toán cả IS và FID, mỗi mô hình tạo ra 30.000 hình ảnh (độ phân giải 256×256) từ các mô tả văn bản một cách ngẫu nhiên được chọn từ tập dữ liệu thử nghiệm.

Hơn nữa, chúng tôi đánh giá số lượng tham số (NoP) để so sánh kích thước mô hình với các phương pháp hiện tại.

4.1 Đánh giá định lượng

Chúng tôi so sánh phương pháp được đề xuất với một số phương pháp tiên tiến nhất, bao gồm StackGAN [56], StackGAN++ [57], AttnGAN [50], MirrorGAN [33], SD-GAN [51] và DM-GAN [60], đã đạt được thành công đáng kể trong việc tổng hợp văn bản thành hình ảnh bằng cách sử dụng các cấu trúc xếp chồng lên nhau. Chúng tôi cũng so sánh với các mô hình gần đây hơn [26,39]. Cần phải chỉ ra rằng các mô hình gần đây luôn sử dụng thêm kiến thức hoặc giám sát. Ví dụ: The results of IS, FID and NoP compared with the state of-the-art methods on the test set of CUBDAE GAN [39] sử dụng thêm tính năng gắn thẻ NLTK POS và quy tắc ký tên theo cách thủ công cho các bộ dữ liệu khác nhau và TIME [26] sử dụng thêm Mã hóa vị trí 2-D.

Hình 6. Hình 6. Ví dụ về hình ảnh được tổng hợp bởi DF-GAN được đề xuất của chúng tôi dựa trên mô tả văn bản từ bộ dữ liệu thử nghiệm CUB.

Bảng I
KẾT QUẢ IS, FID SO VỚI CÁC PHƯƠNG PHÁP HIỆN ĐẠI TRÊN BỘ THỬ NGHIỆM CỦA CUB

Model	IS ↑	FID ↓
StackGAN [56]	3.70	-
StackGAN++ [57]	3.84	-
AttnGAN [50]	4.36	23.98
MirrorGAN [33]	4.56	18.34
SD-GAN [51]	4.67	-
DM-GAN [60]	4.75	16.09
DAE-GAN [39]	4.42	15.19
TIME [26]	4.91	14.30
DF-GAN (author)	5.10	14.81
DF-GAN (Ours)	4.31	18.27

Như thể hiện trong Bảng 1, so với các mô hình hàng đầu khác, DF-GAN của chúng tôi có số lượng nhỏ hơn đáng kể về hiệu suất cạnh tranh. So với StackGAN [56] và StackGAN++ [57] thì DF-GAN của chúng tôi cải thiện chỉ số IS từ 3.70 lên 4.31 trên bộ dữ liệu CUB. So với AttnGAN [50] và MirrorGAN [33] thì DF-GAN của chúng tôi cải thiện FID từ 23.98 xuống 18.27 tương tự trên bộ dữ liệu CUB. So sánh

với các mô hình thì DF-GAN của chúng tôi vẫn chưa hoàn toàn tốt và vẫn chưa đạt được kết quả như kết quả của bài báo

4.2 Đánh giá định tính

Chúng tôi cũng so sánh các kết quả trực quan được tổng hợp bởi AttnGAN [50], DM-GAN [60] và DF-GAN được đề xuất. Có thể thấy hình ảnh do AttnGAN tổng hợp [50] và DM-GAN [60] trong Hình 6 trông giống như một sự kết hợp đơn giản giữa hình dạng mờ và một số chi tiết trực quan (số 1, 2, 5 và 8). Như được hiển thị số 1 và số 2, những con chim được tổng hợp bởi AttnGAN [50] và DM-GAN [60] chứa hình dạng sai. Hơn nữa, các hình ảnh được tổng hợp bởi DF-GAN của chúng tôi có hình dạng đối tượng tốt hơn và các chi tiết chi tiết thực tế (ví dụ: số 10 và số 12). Bên cạnh đó, tư thế của con chim trong DF-GAN của chúng tôi kết quả cũng tự nhiên hơn (ví dụ: số 11 và số 12). do đó DF-GAN được đề xuất có thể tổng hợp chúng một cách chính xác hơn.

5 KẾT LUẬN

Trong bài báo này, chúng tôi đề xuất mô hình DF-Gan cho bài toán text-to-image chúng tôi thực hiện One stage để sinh ra hình ảnh có độ phân giải cao trực tiếp mà không có sự vướng víu giữa các mạng Generator khác nhau. Chúng tôi thực hiện Target-Aware Discrimina- tor composed of Matching-Aware Gradient Penalty (MA- GP) và One-Way Output. Nó có thể tăng cường hơn tính nhất quán ngữ nghĩa của hình ảnh và văn bản mà không cần thêm các mạng. Bên cạnh đó chúng tôi thực hiện Deep text-image Fusion Block (DFBlock) để hợp nhất các đặc trưng vằn bản vầ hình ảnh một cách hiệu quả và sâu sắc hơn. Các kết quả thử nghiệm mở rộng chứng minh rằng DF-GAN được đề xuất của chúng tôi vượt trội đáng kể so với các mô hình hiện đại nhất trên bộ dữ liệu CUB và bộ dữ liệu COCO đầy thách thức hơn

6 REFERENCES

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthes int arXiv:2005.14165, 2020. 2

[3] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. Rifegan: Rich feature generation for text-toimage synthesis from prior knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10911-10920, 2020. 1

[4] Wen-Huang Cheng, Sijie Song, Chieh-Yun Chen, Shintami Chusnul Hidayati, and Jiaying Liu. Fashion meets computer vision: A survey. ACM Computing Surveys (CSUR), 54(4):1-41, 2021. 1

[5] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. In Advances in Neural Information Processing Systems, pages 6594-6604, 2017. 5,8

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. 6, 8

[7] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. arXiv preprint arXiv:2105.13290, 2021. 2, 6

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672-2680, 2014. 1, 2

[9] Yuchuan Gou, Qiancheng Wu, Minghao Li, Bo Gong, and Mei Han. Segattngan: Text to image generation with segmentation attention. arXiv preprint arXiv:2005.12444, 2020. 1

[10] Caglar Gulcehre, Sarath Chandar, Kyunghyun Cho, and Yoshua Bengio. Dynamic neural turing machine with continuous and discrete addressing schemes. Neural computation, 30(4):857-884, 2018. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770-778, 2016. 2, 3

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in neural information processing systems, pages 6626-6637, 2017. 6

[13] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text- to-image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 79867994, 2018. 1

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, pages 1501-1510, 2017. 5

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning, 2015. 5

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4401-4410, 2019. 2, 5, 8

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8110-8119, 2020. 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015. 6

[19] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In Advances

in Neural Information Processing Systems, pages 2065-2075, 2019. 1

[20] Ruifan Li, Ning Wang, Fangxiang Feng, Guangwei Zhang, and Xiaojie Wang. Exploring global and local linguistic representation for text-to-image synthesis. IEEE Transactions on Multimedia, 2020. 1

[21] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12174-12182, 2019. 1, 6

[22] Jiadong Liang, Wenjie Pei, and Feng Lu. Cpgan: Contentparsing generative adversarial networks for text-to-image synthesis. In European Conference on Computer Vision, pages 491-508. Springer, 2020. 6, 7

[23] Jae Hyun Lim and Jong Chul Ye. Geometric gan. arXiv preprint arXiv: 1705.02894, 2017. 3

[24] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. arXiv preprint arXiv:2103.00823, 2021. 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740-755. Springer, 2014. 6

[26] Bingchen Liu, Kunpeng Song, Yizhe Zhu, Gerard de Melo, and Ahmed Elgammal. Time: text and image mutual-translation adversarial networks. arXiv preprint arXiv:2005.13192, 2020. 6, 7

[27] Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya. Generative adversarial networks for image and video synthesis: Algorithms and applications. Proceedings of the IEEE, 109(5):839-862, 2021. 1

[28] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In International Conference on Machine Learning, pages 3481-3490, 2018. 4

[29] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. arXiv preprint arXiv:1802.05637, 2018. 8

[30] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-toimage synthesis. 2021. 2

[31] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018. 2

[32] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Learn, imagine and create: Text-to-image generation from prior knowledge. In Advances in Neural Information Processing Systems, pages 887-897, 2019. 1

[33] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1505-1514, 2019. 1, 2, 6, 7

[34] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019. 2, 8

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092, 2021. 1, 2, 6

[36] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018. 6

[37] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Proceedings of the International Conference on Machine Learning, pages 1060 1069, 2016. 1, 2

[38] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In Advances in neural information processing systems, pages 217-225, 2016. 2

[39] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. Dae-gan: Dynamic aspectaware gan for text-to-image synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13960 – 13969, 2021.6, 7

[40] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Advances in neural information processing systems, pages 2234-2242, 2016. 6

[41] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 45(11):2673-2681, 1997. 3

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 6 [43] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *ECCV*, 2020. 2

[44] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022, 2016. 5

[45] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In European conference on computer vision, pages 791-808. Springer, 2016. 2

[46] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In European conference on computer vision, pages 135-153. Springer, 2016. 2

[47] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.6

[48] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. IEEE

transactions on pattern analysis and machine intelligence, 43(10):33653387, 2020. 1

[49] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In International Conference on Learning Representations, 2015. 2

[50] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Finegrained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1316 1324, 2018. 1, 2, 3, 5, 6, 7, 8

[51] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-toimage generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2327–2336, 2019. 1, 2, 6, 7, 8

[52] Fangchao Yu, Li Wang, Xianjin Fang, and Youwen Zhang. The defense of adversarial example with conditional generative adversarial networks. Security and Communication Networks, 2020, 2020. 2

[53] Mingkuan Yuan and Yuxin Peng. Ckd: Cross-task knowledge distillation for text-to-image synthesis. IEEE Transactions on Multimedia, 2019. 1

[54] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In International conference on machine learning, pages 73547363. PMLR, 2019. 2, 3

[55] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-toimage generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 833-842, 2021. 6, 7

[56] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 59075915, 2017. 1, 2, 3, 5, 6, 7, 8 [57] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE TPAMI, 41(8):1947–1962, 2018. 1, 2, 3, 5, 6, 7

[58] Zhenxing Zhang and Lambert Schomaker. Dtgan: Dual attention generative adversarial networks for text-to-image generation. In 2021 International Joint Conference on Neural Networks (IJCNN), pages 1-8. IEEE, 2021. 6

[59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 22232232, 2017. 2

[60] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for textto-image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 58025810, 2019. 1, 2, 6, 7,8