Taming Transformers for High-Resolution Image Synthesis

Ronaldo Lopez
FC-UNI
Lima, Peru
ronaldo.lopez.c@uni.pe

Cristhian Sánchez Sauñe
FC-UNI
Lima, Peru
csanchez@uni.pe

Davis Alderete
FC-UNI
Lima, Peru
davis.alderete.v@uni.pe

Abstract

Diseñado para aprender interacciones de largo alcance en datos secuenciales, los Transformers continúan mostrando resultados de vanguardia en una amplia variedad de tareas. A diferencia de las CNN, no contienen sesgos inductivos que prioricen las interacciones locales. Esto los hace expresivos, pero también computacionalmente inviables para secuencias largas, como imágenes de alta resolución. Demostramos cómo la combinación de la efectividad del sesgo inductivo de las CNN con la expresividad de los Transformers les permite modelar y, por lo tanto, sintetizar imágenes de alta resolución. En este trabajo se muestra cómo (i) usar CNN para aprender un vocabulario rico en contexto de los componentes de la imagen y, a su vez, (ii) utilizar Transformers para modelar de manera eficiente su composición dentro de imágenes de alta resolución. El enfoque propuesto se aplica fácilmente a las tareas de síntesis condicional, donde tanto la información no espacial, como las clases de objetos, como la información espacial, como las segmentaciones, pueden controlar la imagen generada. En particular, se presentan los primeros resultados sobre la síntesis guiada semánticamente de imágenes megapíxeles con Transformers y se obtiene el estado del arte entre los modelos autorregresivos en ImageNet condicional de la clase. El código y los modelos previamente entrenados se pueden encontrar en github.com/HiroForYou/Image-Synthesis-with-Transformer

1 Introducción

Los Transformers van en aumento: ahora son la arquitectura estándar de facto para las tareas de lenguaje [1, 2, 3, 4] y se adaptan cada vez más en otras áreas como el audio [5] y la visión [6, 7]. En contraste con la arquitectura de visión predominante, las redes neuronales Convolucionales (CNN), la arquitectura del Transformer no contiene un previo inductivo incorporado en la localidad de interacciones y, por lo tanto, es libre de aprender relaciones complejas entre sus entradas. Sin embargo, esta generalidad también implica que tiene que aprender todas las relaciones, mientras que las CNN se han diseñado para explotar el conocimiento previo sobre las fuertes correlaciones locales dentro de las imágenes. Por lo tanto, la mayor expresividad de los Transformers viene con costos computacionales que aumentan cuadráticamente, porque se tienen en cuenta todas las interacciones por pares. Los requisitos de energía y tiempo resultantes de los modelos de Transformers de última generación plantean problemas fundamentales para escalarlos a imágenes de alta resolución con millones de píxeles.

Las observaciones de que los Transformers tienden a aprender estructuras Convolucionales [7] plantean, por tanto, la pregunta: ¿Tenemos que volver a aprender todo lo que sabemos sobre la estructura local y la regularidad de las imágenes desde cero cada vez que entrenamos un modelo de visión, o podemos codificar de manera eficiente el sesgo inductivo de la imagen mientras se conserva la flexibilidad de los Transformers?

Se presume que la estructura de la imagen de bajo nivel está bien descrita por una conectividad local, es decir, una arquitectura Convolucional, mientras que esta suposición estructural deja de ser efectiva en niveles semánticos superiores. Además, las CNN no solo exhiben un fuerte sesgo de localidad, sino también un sesgo hacia la invariancia espacial mediante el uso de pesos compartidos en todas las posiciones. Esto los hace ineficaces si se requiere una comprensión más holística de la entrada. La idea clave planteada para obtener un modelo eficaz y expresivo es que, en conjunto, las arquitecturas Convolucional y del Transformer pueden modelar la naturaleza compositiva del mundo visual [8]: se utilizó un enfoque Convolucional para aprender de manera eficiente un libro de códigos de partes visuales ricas en contexto y, posteriormente, aprenderá un modelo de sus composiciones globales. Las interacciones de largo alcance dentro de estas composiciones requieren una arquitectura Transformer expresiva para modelar distribuciones sobre sus partes visuales constituyentes. Además, se utilizó un enfoque adversario para garantizar que el diccionario de partes locales capture la estructura local perceptualmente importante para aliviar la necesidad de modelar estadísticas de bajo nivel con la arquitectura del Transformer. Permitir que los Transformers se concentren en modelar relaciones de largo alcance les permite generar imágenes de alta resolución como en la Fig. 1, una hazaña que anteriormente estaba fuera de su alcance.

La formulación planteada da control sobre las imágenes generadas mediante el condicionamiento de la información sobre las clases de objetos deseados o los diseños espaciales. Finalmente, los experimentos demuestran que este enfoque conserva las ventajas de los Transformers al superar los enfoques de vanguardia basados en libros de códigos anteriores basados en arquitecturas Convolucionales.

Figure 1: El enfoque propuesto permite a los Transformers sintetizar imágenes de alta resolución como esta, que contiene 1280x460 píxeles.

2 Trabajos Relacionados

2.1 La familia de Transformers

La característica definitoria de la arquitectura del Transformer es que modela las interacciones entre sus entradas únicamente a través de la atención que les permite manejar fielmente las interacciones entre las entradas independientemente de su posición relativa entre sí. Aplicado originalmente a las tareas de lenguaje, las entradas al Transformer se daban mediante tokens, pero se pueden utilizar otras señales, como las obtenidas a partir de audio o imágenes. Cada capa del Transformer consta de un mecanismo de atención, que permite la interacción entre las entradas en diferentes posiciones, seguido de una red completamente conectada en función de la posición, que se aplica a todas las posiciones de forma independiente.

Dado que el mecanismo de atención se basa en el cálculo de los conductos internos entre todos los pares de elementos de la secuencia, su complejidad computacional aumenta cuadráticamente con la longitud de la secuencia. Si bien la capacidad de considerar las interacciones entre todos los elementos es la razón por la que los Transformers aprenden de manera eficiente las interacciones

de largo alcance, también es la razón por la que los Transformers se vuelven inviables rápidamente, especialmente en imágenes, donde la longitud de la secuencia en sí escala cuadráticamente con la resolución.

Para abordar esta limitación, expresamos la atención propia como un producto punto lineal de los mapas de características del núcleo y hacemos uso de la propiedad de asociatividad de los productos matriciales para reducir la complejidad de $O(N^2)$ a $O(N)$, donde N es la longitud de la secuencia. Mostramos que esta formulación permite una implementación iterativa que acelera dramáticamente los Transformers autorregresivos y revela su relación con las redes neuronales recurrentes. Los Transformers lineales propuestos en este trabajo logran un rendimiento similar al de los Transformers originales y son hasta 4000 veces más rápidos en la predicción autorregresiva de secuencias muy largas.

2.2 Enfoque Convolucional

La estructura bidimensional de las imágenes sugiere que las interacciones locales son particularmente importantes. Las CNN explotan esta estructura al restringir las interacciones entre las variables de entrada a una vecindad local definida por el tamaño del kernel del kernel convolucional. Por lo tanto, la aplicación de un kernel da como resultado costos que escalan linealmente con la longitud total de la secuencia (el número de píxeles en el caso de las imágenes) y cuadráticamente con el tamaño del kernel, que, en las arquitecturas CNN modernas, a menudo se fija en una pequeña constante como 3×3. Este sesgo inductivo hacia interacciones locales conduce a cálculos eficientes, pero la amplia gama de capas especializadas que se introducen en las CNN para manejar diferentes tareas de síntesis sugiere que este sesgo es a menudo demasiado restrictivo. Se han utilizado arquitecturas convolucionales para el modelado autorregresivo de imágenes pero, para imágenes de baja resolución, trabajos anteriores demostraron que los Transformers superan consistentemente a sus homólogos convolucionales. El enfoque propuesto nos permite modelar de manera eficiente imágenes de alta resolución con Transformers al tiempo que conserva sus ventajas sobre los enfoques convolucionales de última generación.

Las redes encoder-decoder que utilizan la arquitectura de red neuronal convolucional (CNN) se han utilizado ampliamente en la literatura de aprendizaje profundo gracias a su excelente rendimiento para varios problemas inversos. Sin embargo, todavía es difícil obtener una visión geométrica coherente de por qué una arquitectura de este tipo ofrece el rendimiento deseado. Inspirándonos en la comprensión teórica reciente sobre la generalización, la expresividad y el panorama de optimización de las redes neuronales, así como en la teoría de los marcos convolucionales profundos, aquí proporcionamos un marco teórico unificado que conduce a una mejor comprensión de la geometría de las CNN de encoder-decoder. El marco unificado que se propone muestra que la arquitectura CNN del encoder-decoder está estrechamente relacionada con la representación de cuadros no lineales utilizando cuadros de convolución combinatoria, cuya expresividad aumenta exponencialmente con la profundidad. También demostramos la importancia de la conexión omitida en términos de expresividad y panorama de optimización.

2.3 Enfoque de Dos-Etapas

Entre los enfoques más cercanos al propuesto se encuentran los enfoques de dos etapas que, en primer lugar, aprenden una codificación de datos y, después, las salas aprenden, en una segunda etapa, un modelo probabilístico de esta codificación. [9] demostró evidencia tanto teórica como empírica sobre las ventajas de aprender primero una representación de datos con un Autoencoder Variacional (VAE) [10, 11] , y luego aprender nuevamente su distribución con un VAE. [12, 13] demuestran ganancias similares cuando se usa un flujo de normalización incondicional para la segunda etapa, y [14, 15] cuando se usa un flujo de normalización condicional. Para mejorar la eficiencia del entrenamiento de las Redes Generativas Adversarias (GAN), [16] aprende un GAN [17] en representaciones de un codificador automático y [18] en coeficientes de ondas de baja resolución que luego se decodifican en imágenes con un generador aprendido. [19] presenta el Autoencoder Variacional Cuantizado Vectorial (VQVAE), un enfoque para aprender representaciones discretas de imágenes, y modela su distribución de forma autorregresiva con una arquitectura convolucional. [20] amplía este enfoque para utilizar una jerarquía de representaciones aprendidas. Sin embargo, estos métodos todavía se basan en la estimación de densidad convolucional, lo que dificulta la captura de interacciones de largo alcance en imágenes de alta resolución. [6] modela imágenes de forma autorregresiva

con Transformers con el fin de evaluar la idoneidad del preentrenamiento generativo para aprender representaciones de imágenes para tareas posteriores. Dado que las resoluciones de entrada de 32 × 32 píxeles siguen siendo bastante costosas desde el punto de vista computacional [6], se utiliza un VQVAE para codificar imágenes hasta una resolución de 192 × 192. En un esfuerzo por mantener la representación discreta aprendida tan espacialmente invariante como sea posible con respecto a los píxeles, se emplea un VQVAE poco profundo con un pequeño campo receptivo. Por el contrario, demostramos que una primera etapa poderosa, que capture la mayor cantidad de contexto posible en la representación aprendida, es fundamental para permitir una síntesis de imágenes de alta resolución eficiente con Transformers.

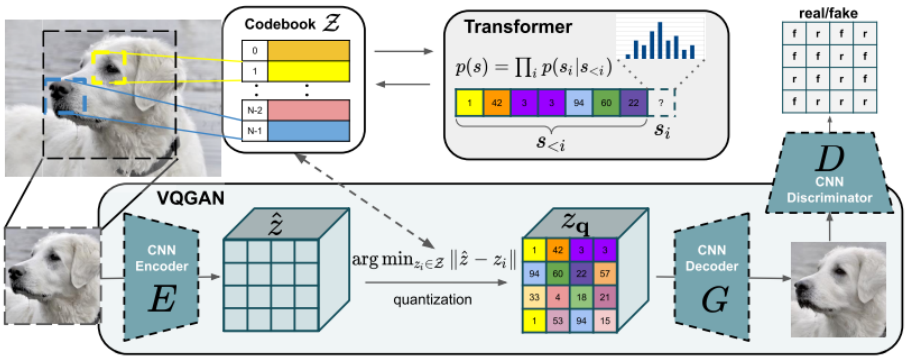

Figure 2: El enfoque propuesto utiliza un VQGAN convolucional para aprender un libro de códigos de partes visuales ricas en contexto, cuya composición se modela posteriormente con una arquitectura de Transformer autorregresivo. Un libro de códigos discreto proporciona la interfaz entre estas arquitecturas y un discriminador basado en parches permite una fuerte compresión al tiempo que conserva una alta calidad de percepción. Este método presenta la eficiencia de los enfoques convolucionales para la síntesis de imágenes de alta resolución basada en Transformers.

3 Enfoque

El objetivo principal es aprovechar las prometedoras capacidades de aprendizaje de los modelos de Transformers [1] e introducirlos en la síntesis de imágenes de alta resolución hasta el rango de megapíxeles. El trabajo anterior [21, 1] que aplicó Transformers a la generación de imágenes demostró resultados prometedores para imágenes de hasta un tamaño de 64 × 64 píxeles pero, debido al aumento cuadrático del costo en la longitud de la secuencia, no se puede simplemente escalar a resoluciones más altas. La síntesis de imágenes de alta resolución requiere un modelo que comprenda la composición global de las imágenes, lo que le permite generar patrones localmente realistas y globalmente consistentes. Por lo tanto, en lugar de representar una imagen con píxeles, la representamos como una composición de componentes de imagen perceptivamente ricos de un libro de códigos. Aprendiendo un código efectivo, como se describe en la Sec. 3.1, podemos reducir significativamente la longitud de descripción de las composiciones, lo que nos permite modelar de manera eficiente sus interrelaciones globales dentro de imágenes con una arquitectura de Transformer como se describe en la Sec. 3.2. Este enfoque, resumido en la Figura 2, es capaz de generar imágenes de alta resolución realistas y consistentes tanto en un entorno incondicional como condicional.

3.1 Aprendiendo un libro de códigos eficaz de componentes de imagen para usar en Transformers

Para utilizar la arquitectura de Transformer altamente expresiva para la síntesis de imágenes, necesitamos expresar los componentes de una imagen en forma de secuencia. En lugar de basarse en píxeles individuales, la complejidad requiere un enfoque que utilice un libro de códigos discreto de representaciones aprendidas, de modo que cualquier imagen $x \in \mathbb{R}^{H \times W \times 3}$ se puede representar mediante una colección espacial de entradas del libro de códigos $z_q \in \mathbb{R}^{h \times w \times n_z}$, donde n_z es la dimensionalidad de los códigos. Una representación equivalente es una secuencia de índices hw que especifican las entradas respectivas en el libro de códigos aprendido. Para aprender eficazmente un

libro de códigos espacial tan discreto, proponemos incorporar directamente los sesgos inductivos de las CNN e incorporar ideas del aprendizaje de la representación neural discreta [19]. Primero, aprendemos un modelo convolucional que consta de un codificador E y un descodificador G, de modo que, tomados en conjunto, aprenden a representar imágenes con códigos de un libro de códigos discreto aprendido $Z = \{z_k\}_{k=1}^{K} \subset \mathbb{R}^{n_z}$ (consulte la Figura.2 para obtener una descripción general). Más precisamente, aproximamos una imagen dada x por $\hat{x} = G(z_q)$. Nosotros obtenemos z_q usando la codificación $\hat{z} = E(x) \in \mathbb{R}^{hxwxn_z}$ y una cuantificación posterior por elementos $q(.)$ de cada código espacial $\hat{z_{ij}} \in \mathbb{R}^{n_z}$ en su entrada de libro de códigos más cercana z_k :

$$z_q = q(\hat{z}) := (argmin \|\hat{\mathbf{z_{ij}}} - \mathbf{z_k}\|) \in R^{h \times w \times n_z} \tag{1}$$

La reconstrucción $\hat{x} \approx x$ esta dado por:

$$\hat{x} = G(z_q) = G(q(E(x))). \tag{2}$$

Backpropagation a través de la operación de cuantificación no diferenciable en la Ec. (2) se logra mediante un estimador de gradiente directo, que simplemente copia los gradientes del decodificador al codificador [22], de modo que el modelo y el libro de códigos se pueden entrenar de un extremo a otro mediante la función de pérdida:

$$\pounds_{VQ}(E, G, Z) = \|x - \hat{x}\|^2 + \|sg[E(x)] - z_q\|_2^2 + \|sg[z_q] - E(x)\|_2^2. \tag{3}$$

Aquí $\zeta_{rec} = \|x - \hat{x}\|^2$ es una perdida de reconstrucción, $sg[.]$ denota la operación de parada-gradiente, y $\|sg[z_q] - E(x)\|_2^2$ es la llamada "pérdida de compromiso" [19].

3.1.1 Aprendiendo un libro de códigos rico en percepción

El uso de Transformers para representar imágenes como una distribución sobre los componentes de la imagen latente requiere que superemos los límites de la compresión y aprendamos un rico libro de códigos. Para ello, proponemos VQGAN, una variante del VQVAE original, y utilizamos un discriminador y una pérdida de percepción [23, 24, 25, 26, 27] para mantener una buena calidad de percepción a una mayor tasa de compresión. Tenga en cuenta que esto contrasta con trabajos anteriores que aplicaron modelos autorregresivos basados en píxeles [28, 20] y basados en Transformers [6] sobre un modelo de cuantificación superficial. Más específicamente, reemplazamos la pérdida L2 utilizada en [19] por una pérdida perceptiva e introducimos un procedimiento de entrenamiento adversario con un discriminador D [29] basado en parche que tiene como objetivo diferenciar entre imágenes reales y reconstruidas:

$$\pounds_{GAN}(\{E, G, Z\}, D) = [\log D(x) + log(1 - D(\hat{x}))] \tag{4}$$

El objetivo completo para encontrar el modelo de compresión óptimo $Q_* = E_*, G_*, Z_*$ luego dice:

$$Q^* = \arg_{E,G,Z} \, min \, max_D \mathbb{E}_{x \sim p(x)} [\pounds_{VQ}\{E, G, Z\} + \lambda \pounds_{GAN}(\{E, G, Z\}, D)]. \tag{5}$$

donde calculamos el peso adaptativo λ de acuerdo con

$$\lambda = \frac{\bigtriangledown_{G_L}[\pounds_{rec}]}{\bigtriangledown_{G_L}[\pounds_{GAN}] + \delta} \tag{6}$$

donde $\pounds_{rec}$ es la pérdida de reconstrucción perceptual [30], $\bigtriangledown_{G_L}$ denota el gradiente con respecto a su entrada. la última capa L del decodificador, y $\delta = 10^6$ se utiliza para la estabilidad numérica. Para agregar contexto de todas partes, aplicamos una sola capa de atención en la resolución más baja. Este procedimiento de entrenamiento reduce significativamente la longitud de la secuencia al desenrollar el código latente y, por lo tanto, permite la aplicación de potentes modelos de Transformers.

3.2 Aprendiendo la composición de imágenes con Transformers

3.2.1 Transformers Latentes

Con E y G disponibles, ahora podemos representar imágenes en términos de índices de libro de códigos de sus codificaciones. Más precisamente, la codificación cuantificada de una imagen x viene

dada por $z_q = q(E(x)) \in \mathbb{R}^{h \times w \times n_z}$ y es equivalente a una secuencia $s \in \{0, ..., |z| - 1\}^{h \times w}$ de índices del libro de códigos, que se obtiene reemplazando cada código por su índice en el libro de códigos Z:

$$s_{ij} = k \; tal \; que \; (z_q)_{ij} = z_k \tag{7}$$

Al mapear los índices de una secuencia s de nuevo a sus correspondientes entradas del libro de códigos, $z_q = (z_{q_{ij}})$ se recupera y decodifica fácilmente en una imagen $\hat{x} = G(z_q)$.

Por lo tanto, después de elegir algún orden de los índices en s, la generación de imágenes se puede formular como predicción autoregresiva del índice siguiente: Dados los índices $s_{<i}$, el Transformer aprende a predecir la distribución de los índices siguientes posibles, es decir, $p(s_i|s_{<i})$ para calcular la probabilidad de la representación completa como $p(s) = \prod_i p(s_i|s_{<i})$. Esto nos permite maximizar directamente la probabilidad logarítmica de las representaciones de datos:

$$\pounds_{Transformer} = E_{x \sim p(x)} \sim [-\log p(s)] \tag{8}$$

3.2.2 Síntesis Condicionada

En muchas tareas de síntesis de imágenes, un usuario exige control sobre el proceso de generación proporcionando información adicional a partir de la cual se sintetizará un ejemplo. Esta información, que llamaremos c, podría ser una sola etiqueta que describa la clase de imagen general o incluso otra imagen en sí. La tarea es entonces aprender la probabilidad de la secuencia dada esta información c:

$$p(s|c) = \prod_i p(s_i|s_{<i}, c) \tag{9}$$

Si la información de condicionamiento c tiene extensión espacial, primero aprendemos otro VQGAN para obtener nuevamente una representación basada en índices $r \in \{0, ..., |Z_c| - 1\}^{h_c \times w_c}$con el libro de códigos recién obtenido Z_c Debido a la estructura autorregresiva del Transformer , entonces podemos simplemente anteponer r a s y restringir el cálculo de la probabilidad logarítmica negativa a las entradas $p(s_i|s_{<i}, r)$. Esta estrategia de "solo decodificador" también se ha utilizado con éxito para tareas de resumen de texto [31].

3.2.3 Generando Imágenes en Alta Resolución

El mecanismo de atención del Transformer pone límites a la secuencia de longitud $h \cdot w$ de sus

Figure 3: Ventana de atención deslizante.

entradas s. Si bien podemos adaptar el número de bloques de submuestreo m del presente VQGAN para reducir imágenes de tamaño $H \times W \; para \; h = (H/2^m) \times w = W/2^m$, observamos una degradación de la calidad de reconstrucción más allá de un valor crítico. valor de m, que depende del conjunto de datos considerado. Para generar imágenes en el régimen de megapíxeles, tenemos que trabajar en forma de parches y recortar imágenes para restringir la longitud de sa un tamaño máximo posible durante el entrenamiento. Para muestrear imágenes, usamos el Transformer en forma de ventana deslizante como se ilustra en la Figura 3. El VQGAN asegura que el contexto disponible aún sea suficiente para modelar fielmente las imágenes, siempre que las estadísticas del conjunto de datos sean aproximadamente invariantes espacialmente o información de acondicionamiento espacial está disponible. En la práctica, este no es un requisito restrictivo, porque cuando se viola, es decir, síntesis de imagen incondicional en datos alineados, podemos simplemente condicionar en coordenadas de imagen, similar a [32].

References

[1] Ashish Vaswani et al. “Attention Is All You Need”. In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[2] Alec Radford et al. “Improving language understanding by generative pre-training”. In: (2018).

[3] Alec Radford and Jeffrey Wu. “Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019”. In: *Language models are unsupervised multitask learners. OpenAI Blog* 1.8 (2019), p. 9.

[4] Tom B Brown et al. “Language models are few-shot learners”. In: *arXiv preprint arXiv:2005.14165* (2020).

[5] Rewon Child et al. “Generating long sequences with sparse transformers”. In: *arXiv preprint arXiv:1904.10509* (2019).

[6] Mark Chen et al. “Generative pretraining from pixels”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1691–1703.

[7] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[8] Bjorn Ommer and Joachim M Buhmann. “Learning the compositional nature of visual objects”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2007, pp. 1–8.

[9] Bin Dai and David P. Wipf. “Diagnosing and Enhancing VAE Models”. In: *CoRR* abs/1903.05789 (2019). arXiv: 1903.05789. URL: http://arxiv.org/abs/1903.05789.

[10] Diederik P Kingma and Max Welling. “Stochastic gradient VB and the variational auto-encoder”. In: *Second International Conference on Learning Representations, ICLR*. Vol. 19. 2014, p. 121.

[11] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. “Stochastic backpropagation and approximate inference in deep generative models”. In: *International conference on machine learning*. PMLR. 2014, pp. 1278–1286.

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. “A disentangling invertible interpretation network for explaining latent representations”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9223–9232.

[13] Zhisheng Xiao et al. “Generative latent flow: A framework for non-adversarial image generation”. In: *arXiv preprint arXiv:1905.10485* (2019).

[14] Robin Rombach, Patrick Esser, and Björn Ommer. “Making sense of cnns: Interpreting deep representations and their invariances with inns”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer. 2020, pp. 647–664.

[15] Robin Rombach, Patrick Esser, and Björn Ommer. “Network-to-network translation with conditional invertible neural networks”. In: *arXiv preprint arXiv:2005.13580* (2020).

[16] Jinlin Liu, Yuan Yao, and Jianqiang Ren. “An acceleration framework for high resolution image synthesis”. In: *arXiv preprint arXiv:1909.03611* (2019).

[17] Ian Goodfellow et al. “Generative adversarial networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[18] Seungwook Han et al. “not-so-BigGAN: Generating High-Fidelity Images on Small Compute with Wavelet-based Super-Resolution”. In: *arXiv preprint arXiv:2009.04433* (2020).

[19] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. “Neural discrete representation learning”. In: *arXiv preprint arXiv:1711.00937* (2017).

[20] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. “Generating diverse high-fidelity images with vq-vae-2”. In: *Advances in neural information processing systems*. 2019, pp. 14866–14876.

[21] Niki Parmar et al. “Image Transformer”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 4055–4064. URL: https://proceedings.mlr.press/v80/parmar18a.html.

[22] Y Bengio. “N. L, , eonard, and A. Courville. Estimating or, propagating gradients through stochastic neurons for conditional, computation”. In: *arXiv preprint arXiv:1308.3432* 2.3 (2013).

[23] Anders Boesen Lindbo Larsen et al. “Autoencoding beyond pixels using a learned similarity metric”. In: *arXiv e-prints* (2015), arXiv–1512.

[24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. “Perceptual losses for real-time style transfer and super-resolution”. In: *European conference on computer vision*. Springer. 2016, pp. 694–711.

[25] Alex Lamb, Vincent Dumoulin, and Aaron Courville. *Discriminative Regularization for Generative Models*. 2016. arXiv: 1602.03220 [stat.ML].

[26] Alexey Dosovitskiy and Thomas Brox. “Generating images with perceptual similarity metrics based on deep networks”. In: *Advances in neural information processing systems* 29 (2016), pp. 658–666.

[27] Fabian Mentzer et al. “High-fidelity generative image compression”. In: *arXiv preprint arXiv:2006.09965* (2020).

[28] Aaron van den Oord et al. “Conditional image generation with pixelcnn decoders”. In: *arXiv preprint arXiv:1606.05328* (2016).

[29] Phillip Isola et al. “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[30] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[31] Peter J Liu et al. “Generating wikipedia by summarizing long sequences”. In: *arXiv preprint arXiv:1801.10198* (2018).

[32] Chieh Hubert Lin et al. “Coco-gan: Generation by parts via conditional coordinating”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4512–4521.