elektron ámbar

Original Article

Revista elektron, Vol. 7, No. 2, pp. 61-70 (2023)
https://doi.org/10.37537/rev.elektron.7.2.185.2023

Estudio de redes generativas de confrontación para generación de datos sintéticos y su aplicación a tomografía optoacústica

Study of Generative Adversarial Networks for Generating Synthetic Data and its Application on Optoacoustic Tomography

Alejandro D. Scopa Lopina*[1], Martín G. González *†, Matías Vera*†

*Facultad de Ingeniería, Universidad de Buenos Aires
Paseo Colon 850, C1063ACV, Buenos Aires, Argentina
†Consejo Nacional de Investigaciones Científicas y Técnicas, (CONICET)
Godoy Cruz 2290, C1425FQB, Buenos Aires, Argentina
[1]ascopa@fi.uba.ar

Recibido: 31/10/23; Aceptado: 06/12/23

***Resumen*— En este trabajo se propone el uso de una red generativa de confrontación (GAN) para efectuar un aumento de datos con el objetivo de mejorar la reconstrucción de imágenes en sistemas para tomografía optoacústica (TOA). Se utilizó el modelo denominado *FastGAN* que es una red compacta, capaz de generar imágenes de alta resolución a partir de un conjunto de datos reducidos. La calidad de los datos generados se evaluó a través de dos métodos. Por un lado, se usó la distancia de inicio de Fréchet (FID), observándose una tendencia decreciente a largo de todo el entrenamiento de la GAN. En el segundo método se entrenó una red neuronal U-Net diseñada para un sistema de TOA con y sin datos aumentados. En este caso, el modelo entrenado con los datos extras aportados por la GAN logró una mejora apreciable en las figuras de mérito asociadas a la reconstrucción.**

Palabras clave: Tomografía optoacústica; Aprendizaje profundo; Redes generativas de confrontación; Datos sintéticos.

***Abstract*— This work proposes the use of a Generative Adversarial Network (GAN) to perform data augmentation with the goal of improving image reconstruction in Optoacustic Tomography (OAT) applications. We employ the *FastGAN* model, a compact net capable of generating high resolution images from small datasets. The quality of the generated data was assessed by two methods. First, the Fréchet distance (FID) was measured, observing a decreasing trend throughout the entire GAN training. Then, a U-Net neural network designed for a OAT system with and without augmented data was trained. In this case, the model trained with the extra data generated by the GAN achieved an appreciable improvement in the figures of merit associated with the reconstruction.**

Keywords: Optoacoustic Tomography; Deep Learning; Generative Adversarial Networks; Synthetic Data.

I. Introducción

La tomografía optoacústica (TOA) es un método de obtención de imágenes médicas mediante el uso del efecto optoacústico (OA). Un pulso de luz que incide en el tejido biológico blando se esparcirá por el mismo y una parte será absorbida por moléculas presentes en la muestra biológica, conocidas como cromóforos. La energía del cromóforo excitado se convierte luego en calor, que en el marco de un proceso isocórico, termina generando un aumento de presión. Esto se detecta a través de distintos arreglos de sensores de ultrasonido, generando sinogramas. Estos son una representación gráfica de las señales acústicas en función del tiempo medido por cada detector. Finalmente, a través de un proceso de reconstrucción, es posible recuperar los datos de interés.

El proceso de reconstrucción en sistemas para TOA conlleva dos problemas de inversión: el acústico y el óptico. En el primero se desea obtener la presión acústica inicial, mientras que en el segundo se intenta recuperar el coeficiente de absorción óptico. El problema de inversión acústica se puede resolver en forma cerrada en condiciones ideales. Sin embargo, en la mayoría de los casos esto no es posible, dado las heterogeneidades en la velocidad del sonido o las limitaciones de ancho de banda en las mediciones, por ejemplo. Cuando además consideramos la inversión óptica, la tarea de reconstrucción se vuelve compleja. Existen soluciones basadas en modelos iterativos, donde se busca incorporar algún tipo de conocimiento previo en estos modelos para minimizar la complejidad. De todas maneras, estas soluciones terminan siendo lentas y computacionalmente intensivas [1]–[3].

Con al advenimiento de nuevas ideas en el campo del aprendizaje estadístico, como ser las técnicas de aprendizaje profundo o *deep learning* (DL) [4], se ha generado un cúmulo importante de métodos diversos y su aplicación a nuevos y viejos problemas. El problema de procesamiento de imágenes ha sido paradigmático en el sentido de que fue

Creative Commons License - Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)

uno de los primeros campos en donde DL ha mostrado su enorme potencialidad, generando desempeños nunca antes vistos en diversos problemas como ser clasificación, filtrado (*denoising*), segmentación, etc. En el ámbito de TOA, el estado del arte se ha destacado por el empleo de arquitecturas de aprendizaje profundo asociado a la familia de redes convolucionales [5]. Una arquitectura ampliamente reconocida en este contexto es la U-Net [6], la cual se ha convertido en la elección preferida para la reconstrucción de imágenes TOA debido a su capacidad para capturar características de alta resolución y su habilidad para tratar con problemas de imágenes médicas, como la escasez de datos y el ruido [7]. Su estructura combina una ruta de contracción y una ruta de expansión, la cual permite obtener resultados precisos y detallados. Además, se han realizado diversos avances en la mejora de la U-Net mediante la adaptación de la arquitectura para abordar desafíos específicos de la TOA. Estos avances continúan impulsando el estado del arte en la reconstrucción de imágenes de TOA, abriendo nuevas oportunidades para la mejora de diagnósticos y tratamientos médicos.

Actualmente, uno de los principales problemas es la carencia de suficiente cantidad de datos para entrenar las redes mencionadas anteriormente. Esta escasez resulta un inconveniente particular de TOA, donde hoy en día no se cuenta con un estándar certificado de imágenes médicas en gran volumen, como si sucede por ejemplo para resonancias magnéticas (MRI) o tomografías computadas (CT). Dado que las técnicas de DL suelen desempeñarse mejor o directamente requieren de un gran volumen de datos para su entrenamiento, nos encontramos frente a una problemática de interés común para muchos investigadores del campo. En este sentido, en este trabajo se propone el estudio de redes generativas de confrontación (GAN, por sus siglas en inglés) para crear muestras sintéticas (aumentación de datos) para obtener pares de entradas y salidas que sirvan para entrenar satisfactoriamente a las redes neuronales. De esta manera, se puede lograr un mejor aprovechamiento de los escasos y costosos datos experimentales para el refinamiento final de sus parámetros.

II. Marco teórico TOA

La TOA es un método que proporciona mapas de absorción óptica de alta resolución mediante la detección de ondas de ultrasonido resultantes de la expansión térmica producida por la irradiación del tejido con pulsos cortos de luz. A través del fenómeno OA se genera un pulso acústico a partir de la absorción de un pulso óptico. La incidencia de un pulso de luz en un tejido biológico se dispersa por el mismo, eventualmente abandonándolo o siendo absorbido por moléculas conocidas como cromóforos, de los cuales la hemoglobina es la más importante. La energía del cromóforo excitado luego se convierte en calor. Este proceso ocurre en la escala de los nanosegundos, un tiempo mucho más corto que lo que el tejido demora en moverse, es decir, que la densidad de su masa local cambie (escala en microsegundos). De esta manera, el calentamiento es isocórico y, por lo tanto, viene acompañado de un aumento en presión. El tejido es elástico, por lo que las regiones de alta presión terminan actuando como fuentes de ondas acústicas. Las ondas acústicas son sensibles a la velocidad del sonido y la densidad del medio y estos parámetros suelen variar con la posición. Sin embargo, en tejidos blandos, las variaciones suelen ser pequeñas y, como rara vez se conocen de antemano, el medio suele tratarse como acústicamente homogéneo. Por la diferencia en escala temporal, el incremento de presión se puede considerar instantáneo. Esto permite modelar la generación y propagación de la onda OA como un problema con condiciones iniciales conocidas [5].

Las mediciones de ondas acústicas generadas por el efecto OA se realizan en una superficie S alrededor de una región Ω que contiene el objeto a analizar. La superficie S no es un contorno, por lo que no afecta el campo acústico. Existen varios operadores de muestreo para TOA, entre los más destacables o utilizados:

- Muestreo por puntos, donde la superficie S puede ser una figura geométrica como un plano, un cilindro o una esfera.
- Mediciones de integrales espaciales del campo acústico a lo largo de planos, líneas o patrones.
- Mediciones a través de un anillo de detectores enfocados en un plano.
- Mediciones a través de un arreglo lineal de detectores enfocados en un plano.

Las señales OA son de banda ancha por naturaleza, típicamente mayor a los de un sensor de ultrasonido, por lo que el rango de detección de frecuencias es limitado. Por otro lado, debido al tamaño finito de los detectores de ultrasonidos reales, también se filtran los números de onda espaciales. Esto sucede dado que a medida que aumenta el área, los detectores se vuelven más direccionales, es decir, su ángulo de aceptación disminuye.

A. *Problema inverso*

Como se mencionó previamente, en la TOA existen dos problemas inversos, uno correspondiente al operador directo de la parte acústica y otro al operador directo de la parte óptica. En este trabajo nos centramos en el primero, donde f será la presión acústica inicial y g el sinograma.

Para resolver los problemas directos e inversos, contar con ciertos operadores facilita el trabajo. En nuestro caso, el operador de mayor relevancia es $\mathcal{A}$, que representa un mapeo lineal entre la distribución de presión acústica inicial f y las mediciones acústicas g bajo el efecto del ruido ϵ. El operador $\mathcal{A}$ mapea del espacio de imágenes al espacio de datos medidos.

$$g = \mathcal{A} \cdot f + \epsilon \tag{1}$$

Se ha demostrado que este problema converge si los datos son suficientes. A continuación se mencionan algunas de las restricciones o problemas más comunes en el problema de inversión acústica:

- Ruido siempre presente en cualquier medición real.
- La respuesta de los detectores tiene un rango de frecuencia finito.
- Los detectores sólo rodean parte de la muestra (*limited view*).
- Submuestreo en espacio o tiempo.

Por otro lado, también existen incertezas en los operadores. Si bien las distintas ecuaciones capturan los fenómenos físicos de la TOA, las soluciones numéricas de los mismos

implican ciertas diferencias. Por ejemplo, es habitual el uso de simplificaciones para reducir el costo computacional. También existe una dependencia entre algunos parámetros reales que afectan a los operadores, pero que resulta complejo definir o controlar en las mediciones.

B. *Métodos de reconstrucción*

Para la TOA existen múltiples métodos clásicos de reconstrucción, esto es, técnicas con un enfoque no basado en DL. En este trabajo, haremos uso de la técnica de *delay-and-sum* (DAS), uno de los algoritmos de *beamforming* más utilizados en reconstrucción de imágenes OA [8]. Su simplicidad permite su utilización en aplicaciones de tiempo real, pero generalmente se encuentran ciertas limitaciones. La aparición de artefactos intensos o de grandes lóbulos laterales suele ser común en las imágenes reconstruidas. En nuestro caso esto no es un problema, ya que este método es simplemente usado pasar del dominio de datos de medición (sinograma) al dominio imagen.

El algoritmo busca reconstruir una imagen a través de presiones acústicas capturadas por distintos arreglos de sensores. Para el caso en que la región imagen se encuentre contenida en el plano xy, y se use un arreglo de N_s detectores distribuidos alrededor de la muestra, se tiene la siguiente expresión [8]:

$$S_{DAS}(x,y) = \sum_{i=1}^{N_s} S(i, t(x,y,i)) \tag{2}$$

donde $S_{DAS}(x,y)$ es la señal reconstruida en la posición (x,y) y $S(i,t)$ es la señal recibida en el sensor i en el tiempo t. La función $t(x,y,i)$ representa el retraso temporal debido a la propagación de la señal OA generada en (x,y) hasta el sensor i:

$$t(x,y,i) = \frac{d(x,y,i)}{v_s} \tag{3}$$

donde v_s es la velocidad del sonido y $d(x,y,i)$ es la distancia entre el punto medido en la región imagen y el sensor i. Dividiendo la región imagen en píxeles se puede obtener la imagen reconstruida a través de (2).

III. Redes neuronales de confrontación

A. *Teoría general*

Las redes neuronales de tipo GAN nos permiten generar o sintetizar imágenes a partir de cierta familia de datos. Están compuestas de dos redes que compiten entre sí, el generador y el discriminador. La primera se encarga de generar imágenes, que luego son usadas como entrada de la segunda red. Ésta debe detectar si provienen del generador o no, es decir, distinguir entre muestras sintéticas o reales. La evaluación del discriminador es luego utilizada para mejorar la calidad del generador, dando lugar a esta competencia entre ambas redes. Este tipo de red tiene un gran potencial y rango de aplicaciones, desde procesamiento de imágenes en la forma de clasificadores o reconstructores, así como también en su habilidad de expandir conjuntos de datos existentes. Su contraparte más común es el costo computacional que requieren y la cantidad de datos necesarios para obtener resultados aceptables. Esto muchas veces termina limitando su aplicación en problemas reales, donde la información o los recursos no abundan.

B. *FastGAN*

Como se mencionó previamente, la TOA es una técnica donde la disponibilidad de conjuntos de datos de gran tamaño es escasa. La metodología llamada *transfer-learning* [9] con modelos pre-entrenados presenta una posible solución a este problema, pero no siempre se cuenta con la garantía de poder encontrar un conjunto de datos compatible con nuestro modelo. En algunos casos, el ajuste fino de este tipo de redes puede incluso decantar en un peor rendimiento.

La red GAN presentada en este trabajo, de aquí en adelante denominada FastGAN [10], busca resolver o minimizar el problema de escasez de datos para TOA u otras disciplinas similares. La idea es presentar un proceso de generación de imágenes de alta resolución a partir de conjuntos de datos acotados, que además requiera poco poder computacional. Estas condiciones de entrenamiento hacen que el modelo sea vulnerable al sobreajuste y errores por el modo colapso [11] [12]. Para evitar estos comportamientos es necesario un generador G que pueda aprender rápidamente y un discriminador D que pueda proveer información útil continuamente. Para enfrentar estos desafíos se propone:

- Un módulo de excitación por canales con *skip-layers* (SLE), que aprovecha las activaciones en mapas de baja resolución para luego reutilizarlas en las respuestas de los canales en los mapas de alta resolución [10]. SLE permite que el flujo del gradiente a través de los pesos de cada capa del modelo sea más robusto, permitiendo un entrenamiento más rápido.
- Un discriminador D auto-supervisado que es entrenado como codificador de características con un decodificador extra. Este es forzado a aprender un mapa de características más descriptivo, cubriendo así más regiones de una imagen de entrada. De esta manera podemos brindar señales más comprehensivas a G para su entrenamiento.

El diseño de la red resulta minimalista. Para cada resolución de G se utiliza una única capa de convolución. En las altas resoluciones ($\geq 512 \times 512$) se utilizan tres canales de entrada y salida para las capas convolucionales, tanto en G como D. En la Fig. 1 podemos ver la estructura general del generador.

Para la síntesis de imágenes de alta resolución, resulta inevitable la necesidad de un generador G profundo, con muchas capas de convolución. Esto lleva a un tiempo de entrenamiento más largo que modelos más superficiales, dado la cantidad de parámetros y el efecto de flujo de gradiente débil [13]. La estructura residual *ResBlock* [14] surge como propuesta a esta problemática de entrenamiento en redes profundas. Se plantea la incorporación de capas de conexión o *skip-layers*, para mejorar el flujo de gradiente entre capas. Si bien el uso de esta estructura es abundante, conlleva un aumento en el costo computacional.

El módulo SLE reformula la incorporación de *skip-layers* de dos maneras. En primer lugar, *ResBlock* implementa estas conexiones como adiciones término a término entre las distintas funciones de activación de cada capa. Esto requiere que las dimensiones espaciales de cada función

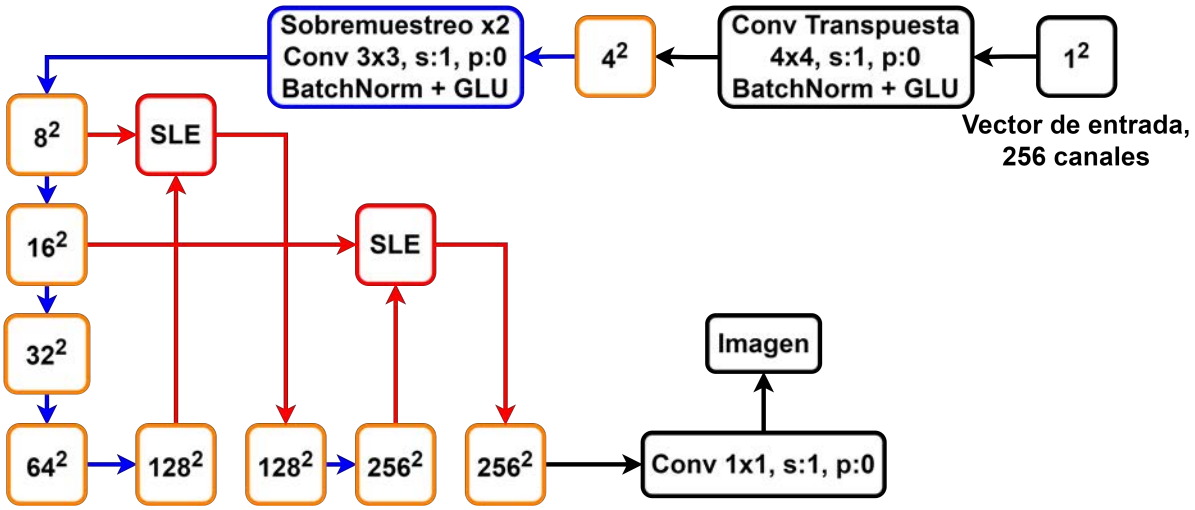


Figura 1: Estructura del generador. Los recuadros naranjas representan mapas de características, con su dimensión espacial (se omiten los canales). Los recuadros y flechas azules representan la misma estructura de sobremuestreo, los recuadros rojos representan los módulos *skip-layer excitation*.

de activación sean iguales. En cambio, con SLE se propone aplicar multiplicaciones de canal a canal entre las activaciones, eliminando así el alto costo computacional que conllevan las convoluciones (una de las activaciones tiene una dimensión espacial de 1^2). Por otro lado, en general, las *skip-layers* solo se utilizan entre capas de una misma resolución. En SLE las conexiones se realizan entre rangos mucho más amplios, por ejemplo, entre 8^2 y 128^2 o 16^2 y 256^2. Estas dos consideraciones conservan la mejora en el flujo de gradiente, minimizando el costo computacional. Formalmente, definimos al módulo SLE como:

$$y = F\left(x_{low}, \{W_i\}\right) \cdot x_{high} \tag{4}$$

donde x e y representan las entradas y salidas de los mapas de características del módulo SLE, respectivamente. La función F representa las operaciones aplicadas en x_{low} (la entrada de baja resolución) y W_i los pesos a aprender.

En la Fig. 2 podemos ver en ejemplo entre dos entradas de dimensión 8^2 y 128^2. Primero, una capa de reducción o *average-pooling* realiza un submuestreo de x_{low}, reduciendo la dimensión de salida a 4^2. Este proceso se repite a través de una capa de convolución, obteniendo una salida de 1^2. Luego pasamos por una capa *LeakyReLU* para modelar las propiedades no lineales y utilizamos otra capa de convolución para que la cantidad de canales coincida con x_{high}. Por último se aplica una función Sigmoid y su resultado se multiplica término a término a lo largo de cada canal con x_{high}. De esta manera, la dimensión de y y de x_{high} coinciden.

La estructura del discriminador D puede verse en la Fig. 3, donde la estrategia buscada es la siguiente: pensamos a la red como un codificador, que a su vez es entrenada con pequeños decodificadores. Este estilo de entrenamiento es denominado *auto-encoding* (AE) y obliga a D a extraer características de las imágenes que luego cada decodificador aprovecha para generar una buena reconstrucción. Estos decodificadores son optimizados junto con D a través de una función de pérdida simple, que solo se entrena con muestras reales:

$$\mathcal{L}_r = \mathbb{E}_{f\sim D_{encode}(x), x\sim I_{real}}[||\mathcal{D}(f,d) - \mathcal{T}(x)||] \tag{5}$$

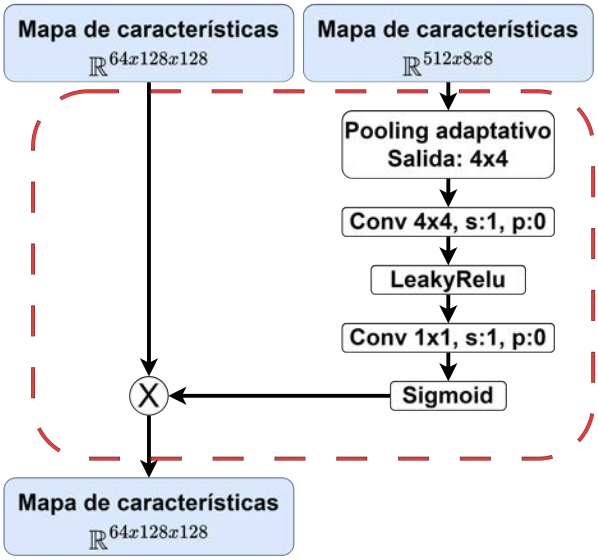


Figura 2: Estructura del módulo SLE.

La función $\mathcal{D}$ representa la transformación de los datos de entrada a través de los mapas de características intermedios del discriminador D (f) y también de los bloques de decodificación (d). Por otro lado, la función $\mathcal{T}$ representa las transformaciones aplicadas a las imágenes reales, en este caso, el submuestreo y recorte. El subíndice r hace referencia al proceso de reconstrucción llevado a cabo por los decodificadores.

En nuestro caso se emplean dos decodificadores a la salida de distintos mapas de características, de resolución 16^2 (f_1) y 8^2 (f_2). Cada decodificador está compuesto por cuatro capas de convolución, que permiten obtener una resolución final de 128^2. En cada ciclo de entrenamiento se toma un cuadrante aleatorio del mapa de características entrante a f_1. La entrada de f_2 es el último mapa de características de la red. De esta manera obtenemos I'_{part} e I' de f_1 y f_2, mientras que a través de un recorte y un submuestreo se obtienen I_{part} e I. Finalmente, D y los decodificadores son

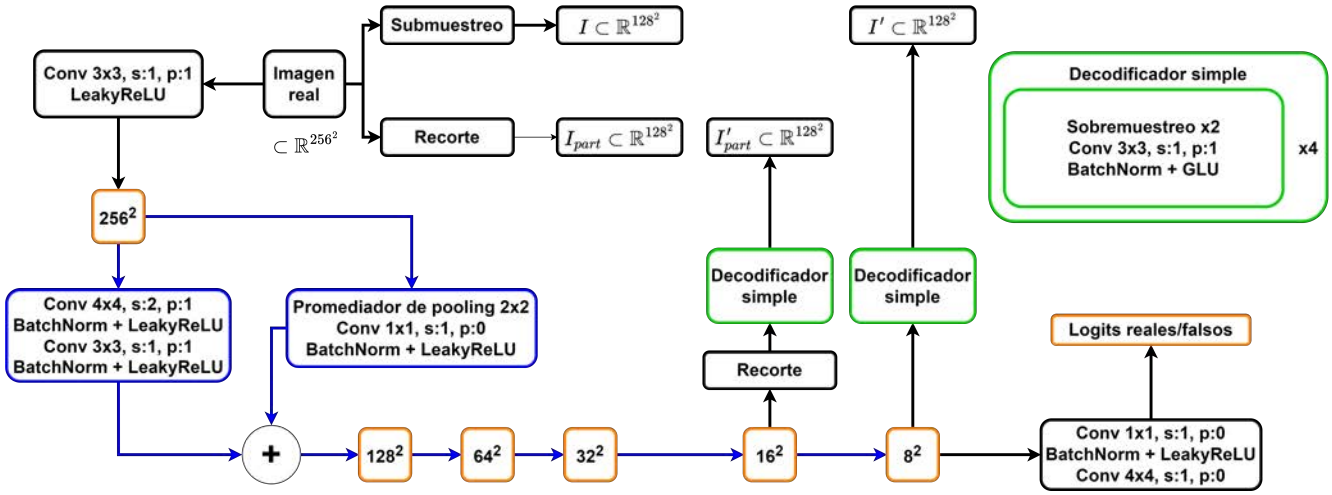


Figura 3: Estructura del discriminador. Los recuadros y flechas azules representan la misma estructura de submuestreo, los recuadros verdes el mismo decodificador.

entrenados en conjunto, buscando minimizar (5) mediante la comparación entre estos conjuntos imágenes.

Este tipo de aprendizaje se asegura de que D extraiga una representación más comprensiva de cada entrada, teniendo en cuenta la composición general a través de f_1 y detalles particulares a través de f_2. De esta manera, nuestro discriminador combina el análisis de una imagen completa por un lado y por otro el análisis de diferentes regiones, similar a la metodología utilizada en una red PatchGAN [15].

El método de AE es utilizado típicamente en aprendizaje auto-supervisado y es reconocido por mejorar la robustez de los modelos y la habilidad de generalización [16]. En el contexto de redes GAN, el hecho de contar con un discriminador D regularizado a través de estrategias de entrenamiento auto-supervisado incrementa significativamente la calidad de síntesis de G. Particularmente, AE resulta la estrategia que genera mejores resultados.

Si bien la estrategia de un entrenamiento auto-supervisado para D se lleva a cabo a través de AE, la solución propuesta es distinta a la típica combinación de GAN y esta metodología. Generalmente, G se entrena como un decodificador sobre un espacio latente de D. En este caso, el modelo propuesto es una GAN pura con un esquema de entrenamiento mucho más simple. El entrenamiento mediante AE es solo utilizado para regularizar D, donde G no está involucrado.

Como función de pérdida se utilizó una versión de la *hinge loss* adaptada para GANs para entrenar D y G de forma iterativa [17]:

$$\begin{aligned}\mathcal{L}_D = &- \mathbb{E}_{x\sim I_{real}}[min(0, -1 + D(x))] \\ &- \mathbb{E}_{\hat{x}\sim G(z)}[min(0, -1 - D(\hat{x}))] \\ &+ \mathbb{L}_r \end{aligned} \tag{6}$$

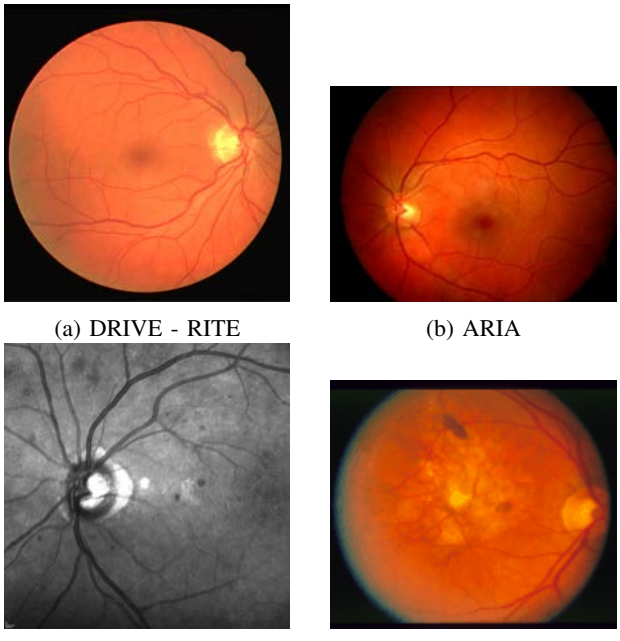
$$\mathcal{L}_G = - \mathbb{E}_{z\sim\mathcal{N}}[D(G(z))] \tag{7}$$

De acuerdo a lo mencionado en [10], la misma permite realizar el computo de pérdida de manera más rápida.

IV. Métodos

A. *Generación de datos*

La TOA se utiliza para obtener imágenes de alta resolución de tejido biológico. En muchos casos, los resultados obtenidos suelen mostrar estructuras con gran predominancia de vasos sanguíneos. Teniendo en cuenta esto y la finalidad de nuestra red, que es aumentar bases de datos para TOA, se buscaron bancos de datos con este tipo de características. Dicho esto, se recopiló información de índole médica, particularmente de vasos sanguíneos (BV, por sus siglas en inglés), compuesta por las bases de datos DRIVE [18], STARE [19], RITE [20], ARIA [21] y RAVIR [22]. En la Fig. 4 se pueden visualizar muestras para cada caso.



(a) DRIVE - RITE (b) ARIA

(c) RAVIR (d) STARE

Figura 4: Ejemplos de imágenes disponibles en las bases de datos utlizadas en este trabajo.

Las bases de datos mencionadas ya cuentan con una segmentación de las imágenes para hacer foco en los vasos

sanguíneos. Luego, se aplica una aumentación sencilla, compuesta de rotaciones verticales y horizontales. De esta manera, el conjunto resultante cuenta con 6252 imágenes disponibles para entrenar, con una resolución de 256×256 píxeles. Finalmente, se convierte la imagen a escala de grises, para asemejar a muestras de imágenes OA. La base de datos resultante, denominada de acá en más BV, se separa en conjuntos de entrenamiento e inferencia, con una relación $90\,\% - 10\,\%$ obteniendo 5626 y 626 imágenes para cada conjunto respectivamente. La Fig. 5 muestra algunos ejemplos utilizados para el entrenamiento.

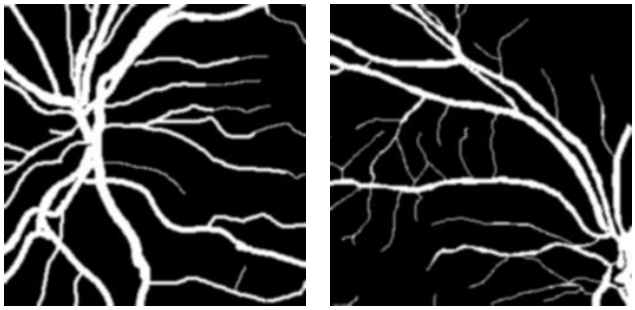


Figura 5: Ejemplo de las imágenes usadas para el entrenamiento de las redes neuronales.

B. Entrenamiento de red FastGAN

Como se mencionó previamente, los datos utilizados para el entrenamiento de nuestra red tienen una resolución de 256^2. Para ambas redes G y D se utiliza el optimizador de *Adam*, parametrizado de la siguiente manera: *lr* (tasa de aprendizaje) = 10^{-4}, $\beta_1 = 0{,}5$ y $\beta_2 = 0{,}99$. Otro atributo importante resulta el tamaño del ruido, en nuestro caso un vector aleatorio de $[0, 1]$ y dimensión 400x1. Para ambas redes se define un parámetro semilla que gobierna la relación y cantidad de filtros de cada capa convolucional. De esta manera se puede adaptar simplemente la arquitectura a necesidad. Los valores elegidos resultan $ndf = 64$ y $ngf = 64$ para D y G respectivamente. Por último, se entrena en mini-lotes de una imagen, durante $100{,}000$ iteraciones. Cada $10{,}000$ iteraciones se guardan los diccionarios que contienen los parámetros de cada red, también se calculan las métricas de rendimiento. De esta manera, solo se conservan dos modelos para ambas redes: el actual y el mejor histórico. Los valores seleccionados provienen de [10], a excepción del lr, el cuál fue ajustado empíricamente según los resultados obtenidos en distintos entrenamientos. En cada iteración se actualizan los pesos de G y D una vez. Para el caso de D, su función de pérdida conlleva un término calculado con imágenes reales y otro con imágenes sintéticas, provenientes de G. Para las reales se utiliza el proceso de AE mencionado previamente, que procesa y segmenta las mismas de distintas maneras. Estos resultados son luego utilizados para calcular los diferentes términos de la función de pérdida correspondiente a imágenes reales.

C. Estrategia de inferencia

La medición de rendimiento en generación de imágenes sintéticas con redes GAN resulta complejo [23]. En la actualidad no existe un consenso general de cuál o cuáles figuras de mérito capturan de mejor manera las fortalezas y limitaciones de distintos modelos. En muchos casos resulta común un análisis visual de las muestras generadas por estas redes, por lo menos en los entrenamientos iniciales. En nuestro caso, y de acuerdo a lo mencionado en [10], la figura elegida es la *Frétchet Inception Distance* (FID). Esta mide el realismo semántico promedio de imágenes sintéticas, realizando comparaciones contra un conjunto de datos real [24]. En primer lugar se utiliza una red *Inception* [25] pre-entrenada para extraer distintas características de las imágenes. En nuestro caso, la implementación utilizada es la que provee Pytorch [26], que utiliza *Inception V3*. Los vectores de características resultantes poseen una distribución normal multivariada. Dicho esto, se calcula la distancia de Fréchet entre ambos vectores gaussianos, de la siguiente manera:

$$d = ||\mu_r - \mu_f||^2 + T_r\left(cov_r + cov_f + 2\sqrt{cov_r \cdot cov_f}\right) \quad (8)$$

donde μ_i representa los valores medio de cada distribución, cov_i la matriz de covarianza y T_r la traza de la matriz resultante. Los subíndices r y f hacen referencia a muestras reales y ficticias. La distancia entre ambas medias es la distancia Euclídea.

Para el entrenamiento de nuestra red GAN se calculó el valor de FID cada $1{,}000$ iteraciones. Cada vez se realiza el siguiente proceso:

- Se generan 1000 imágenes con el generador G.
- Se toman $5{,}626$ imágenes del conjunto de entrenamiento.
- Se calcula la FID entre ambos conjuntos.

La certeza de FID es directamente proporcional a la cantidad de muestras utilizadas para su cálculo. Por esta razón utilizamos el conjunto de entrenamiento completo. Las $1{,}000$ imágenes generadas por G resultan un compromiso entre velocidad y calidad.

D. Reconstrucción de imágenes TOA usando DL

Si bien se utilizó FID para corroborar la calidad de las imágenes sintéticas generadas, esto no necesariamente implica una correlación real para mejoras de desempeño en aplicaciones de TOA. En este sentido, se entrenó una red neuronal con y sin los datos aumentados por nuestra GAN. Se eligió el esquema de reconstrucción descripto en [27] compuesto por un enfoque clásico y una red neuronal encargada del post-procesamiento de las imágenes OA. En nuestro caso se optó por el método de reconstrucción DAS y un modelo U-Net. El primero es el encargado de pasar del dominio de datos medidos (señales OA) al dominio imagen. Mientras que el segundo es entrenado para reducir o eliminar los artefactos u otros defectos introducidos por DAS. En la Fig. 6 se muestran los pasos seguidos para entrenar la red U-Net.

Para la simulación de obtención de sinogramas se utilizó el esquema descripto en [28] y que se muestra en la Fig. 7. Éste consiste en un sistema para TOA 2-D implementado con un sensor que rota alrededor de la región imagen, lugar donde está colocada una muestra uniformemente iluminada. Este tipo de sistemas basados en un solo detector resultan muy útiles para estudios de prueba de concepto debido

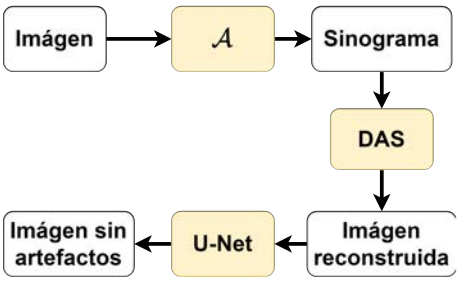


Figura 6: Esquema utilizado en este trabajo para el entrenamiento de la red neuronal U-Net.

a su simplicidad, bajo costo y efectividad [29]. En este trabajo se tomó una región imagen cuadrada con un tamaño de $12{,}8\,\mathrm{mm} \times 12{,}8\,\mathrm{mm}$ y una resolución de 128×128 píxeles. El sensor, supuesto puntual, se colocó sobre una circunferencia de $R_s = 22{,}5\,\mathrm{mm}$ de radio y las señales OA se detectaron en $N_s = 32$ ángulos. La elección de priorizar un valor pequeño de N_s y N_t se hizo en función de reducir la complejidad y el costo del sistema de detección [30]. Para la recopilación de datos, el intervalo de tiempo Δt fue de 49 ns con $N_t = 512$ muestras. La velocidad del sonido se fijó en $v_s = 1500\,\mathrm{m/s}$ y el medio se supuso homogéneo y sin absorción o dispersión del sonido. La respuesta en frecuencia del transductor se modeló utilizando un filtro pasabanda con frecuencias de corte superior e inferior de $0{,}1\,\mathrm{MHz}$ y $20\,\mathrm{MHz}$, respectivamente.

Una vez definidos los parámetros del sistema TOA, se creó la matriz del operador directo $\mathcal{A}$ siguiendo los pasos detallados en [28]. Luego, usando las imágenes del conjunto BV, se obtuvieron los sinogramas (ver Fig. 6). Por último, se agrega un ruido blanco de manera que la relación señal a ruido (SNR) resultante se encuentre en el rango entre $30\,\mathrm{dB}$ y $50\,\mathrm{dB}$. Todas las simulaciones se llevaron a cabo en Python.

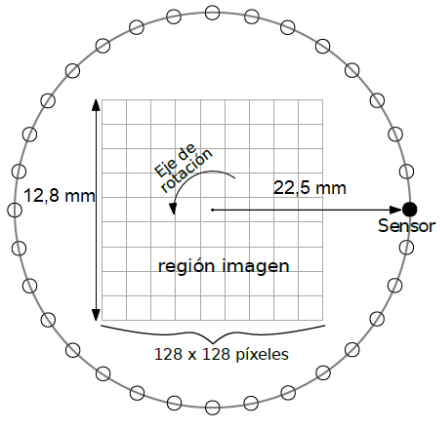


Figura 7: Esquema del sistema TOA usado en las simulaciones. [28].

Como se mencionó previamente, se utilizó una red U-Net para la etapa de post-procesamiento. Estas redes reciben su nombre por la forma de su estructura, donde poseen un camino descendiente, uno ascendiente y uno de conexión entre ambos. El primero se denomina ruta de contracción y está compuesto por distintas capas de convolución que buscan reducir la resolución de la entrada pero aumentar la cantidad de canales. De esta manera se capturan las características relevantes para cada resolución, así codificando los datos. El otro se denomina ruta de extensión y está compuesto por capas de convolución transpuesta, las necesarias para decodificar los datos hasta su resolución original. El camino que une a estos se denomina cuello de botella y es la capa que representa el mayor punto de abstracción, respecto a la entrada original. Por último, existen las *skip-connections*, conexiones entre las distintas rutas que buscan acelerar el entrenamiento y alivianar el problema del gradiente desvanecente. Existen muchos esquemas distintos de U-Net. En este trabajo fue utilizada la *Fully-Dense U-Net* (FD-UNet) [31]. Su particularidad es el uso de bloques densos convolucionales. La entrada de cada uno de estos bloques está compuesta de todas las salidas de capas anteriores concatenadas. De esta manera, cada capa aprende mapas de características adicionales basados en el "conocimiento colectivo" generado por las capas previas. Esta estrategia incrementa la capacidad de representación a través del reuso de características.

Los hiperparámetros seleccionados fueron los siguientes: $lr = 5 \cdot 10^{-4}$ y lotes de 15 muestras. La red se entrenó por 50 ciclos. El conjunto de datos utilizado se separa en entrenamiento ($64\,\%$), validación ($16\,\%$) y prueba ($20\,\%$). El segundo conjunto se utiliza para medir el desempeño de la red durante el ciclo de entrenamiento y aplicar la técnica de detención anticipada [32]. La red se entrena con la función de pérdida de error cuadrático medio (MSE). Se entrenaron dos FD-Unet idénticas con conjuntos de datos distintos: (i) usando solo las imágenes de la base de datos BV y (ii) agregando también los datos sintéticos generados por nuestra GAN. Debido a las restricciones en poder computacional y espacio, solo se generaron $5{,}626$ imágenes sintéticas. De esta manera se obtuvo un conjunto de entrenamiento con el doble de imágenes totales, al que denominamos BV'.

Finalizado los entrenamientos de ambas redes, se calculan cuatro figuras de mérito para comparar su desempeño de forma cuantitativa: la correlación de Pearson (PC), la raíz del error cuadrático medio (RMSE), la relación ruido y señal pico (PSNR) y la similitud estructural (SSIM). Las mismas son utilizadas ampliamente en el ámbito de cuantificación de imágenes y se complementan entre ellas [28] [33]. Para ello se utiliza el conjunto de prueba, que contiene aquellos datos que nunca fueron utilizados durante el entrenamiento.

V. Resultados

Como mencionamos previamente, la figura de mérito elegida para medir el desempeño de nuestra red GAN fue la distancia FID. En la Fig. 8 podemos ver los resultados obtenidos para el entrenamiento realizado.

En primer lugar podemos destacar la tendencia decreciente de los valores obtenidos. Esto indica claramente la mejora en las imágenes generadas por la red, minimizando las diferencias entre datos sintéticos y reales en cada ciclo. Los valores absolutos obtenidos no brindan una información relevante, dado que estos varían ampliamente según el dominio de los conjuntos de datos utilizados. No se encontraron trabajos o referencias donde se utilice FID como métrica de

TABLA I: Figuras de mérito para cada metodología.

	SSIM	**PC**	**RMSE**	**PSNR**
DAS	$0{,}145 \pm 0{,}035$	$0{,}478 \pm 0{,}027$	$0{,}415 \pm 0{,}033$	$7{,}670 \pm 0{,}751$
FD-UNet	$0{,}801 \pm 0{,}089$	$0{,}910 \pm 0{,}041$	$0{,}098 \pm 0{,}034$	$20{,}711 \pm 3{,}197$
FD-UNet(Aug)	$0{,}841 \pm 0{,}076$	$0{,}933 \pm 0{,}034$	$0{,}085 \pm 0{,}030$	$22{,}040 \pm 3{,}288$

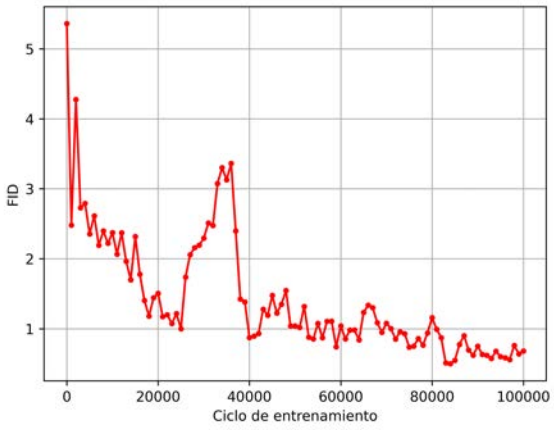


Figura 8: Valores de FID obtenidos para entrenamiento de red GAN con la base de datos BV.

síntesis para imágenes de dominio médico. Por otro lado, también podemos ver la saturación de la red llegando a los 100,000 ciclos de entrenamiento. Un mayor tiempo o cantidad de iteraciones no conllevan a mejor calidad de las imágenes generadas, demarcando así el límite empírico de esta configuración. Un barrido de los hiperparámetros de la red podría generar mejores resultados, pero esto no se llevó a cabo debido a las limitaciones de recursos computacionales. De todas maneras cabe destacar que un aumento en la cantidad de iteraciones no generó un modo colapso, demostrando la estabilidad de la red. Es posible que la red haya extraído la completitud o mayoría de la información disponible en la base de datos proporcionada, explicando así la disminución y eventual cese de mejoras en su rendimiento. Una vez entrenada la red GAN se procedió a aumentar el conjunto BV, generando BV'.

En la Fig. 9 se puede apreciar la evolución de las imágenes al pasar por las distintas etapas detalladas en la Fig. 6, para el caso donde se utilizó la base de datos BV'. Comenzamos con un dato sintético proveniente de la red GAN, con el que construimos un sinograma utilizando la matriz $\mathcal{A}$. Luego del agregado de ruido, se reconstruye la imagen usando el método DAS (pasaje del dominio de datos al dominio imagen). Cómo se observa en la imagen central de la Fig. 9, la reconstrucción obtenida posee artefactos y otros defectos mencionados previamente, que son esencialmente causados por el bajo muestreo espacial [27]. Este tipo imágenes son las entradas de la red FD-UNet. Luego de 50 iteraciones se obtiene la imagen post-procesada que presenta una notable mejora respecto a la imagen devuelta por DAS. Esto indica que la red U-Net realiza un trabajo eficiente en la eliminación de desperfectos.

De esta manera, se realizaron dos entrenamientos, uno con la base de datos BV y otro con la base de datos aumentado

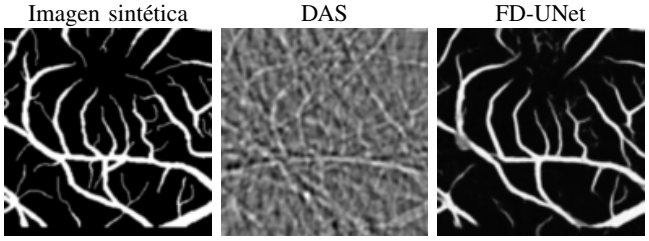


Figura 9: Imagen sintética perteneciente a la base de datos BV' (izq.), reconstrucción DAS (med.) e imagen post-procesada con FD-UNet (der.).

BV'. Finalizados ambos, se procedió a analizar la calidad del post-procesamiento de ambas redes utilizando el conjunto de datos de inferencia apartado inicialmente, compuesto por 626 imágenes no utilizadas hasta este momento. En la Fig. 10 podemos ver distintos casos de los resultados en ambas redes al utilizar el mismo.

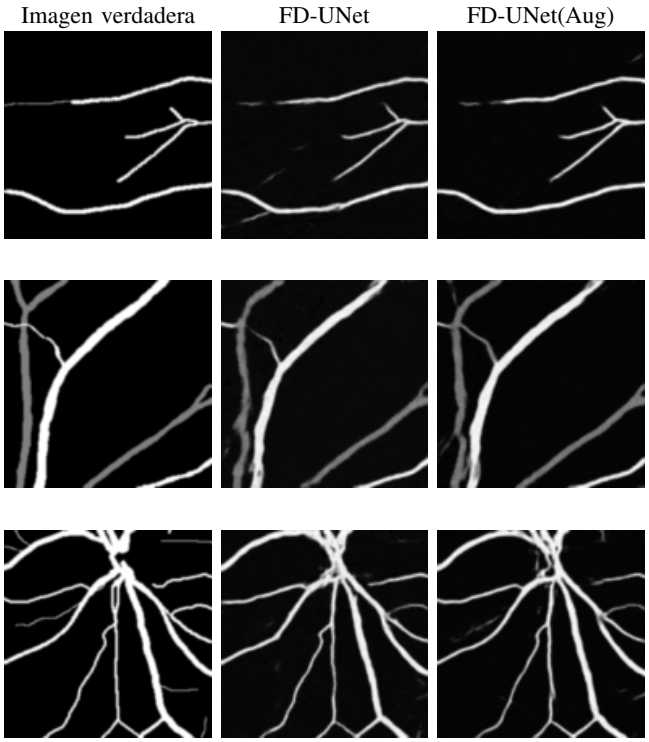


Figura 10: Imágenes verdaderas (izq.), imágenes procesadas con FD-UNet entrenada con BV (med.) e imágenes procesadas con FD-UNet entrenada con BV' (der.).

Como se puede apreciar a simple vista, los resultados obtenidos para cada iteración de la red FD-UNet resultan similares. Esto se condice con los valores obtenidos para cada figura de mérito, presentados en la Tabla I. Si bien hay una gran diferencia entre los resultados obtenidos para reconstrucción con solo DAS, las diferencias entre cada

red de post-procesamiento son pequeñas. Esto confirma dos hipótesis:

- Hay una clara mejora en el proceso de reconstrucción gracias a la etapa de post-procesamiento.
- Hay una mejora tangible en el rendimiento de la etapa de post-procesamiento debido a la aumentación de datos.

Si bien los valores obtenidos gracias al entrenamiento con BV' quedan dentro del rango de varianza de aquellos obtenidos con el entrenamiento de BV, los valores medios mejoran para todas las figuras de mérito. Es importante destacar que no se alcanzó el límite empírico de cantidad de imágenes sintéticas generadas. Se optó por generar la misma cantidad de datos que los originales debido a limitaciones de espacio de computo, pero la tendencia de las figuras de mérito fue creciente en pruebas con menos datos. De esta manera queda entonces la posibilidad en futuros trabajos de encontrar el límite de nuestra red GAN. Lo que si podemos afirmar es que las muestras generadas por nuestra red GAN tienen un impacto real en aplicaciones de TOA. Si las muestras sintéticas hubiesen sido una aumentación simple de las originales, o la red hubiera aprendido a replicar la base de datos BV, la diferencia entre entrenar a la red FD-UNet con un conjunto o el otro sería despereciable o nula. Dado que la red GAN genera nueva información a partir de un ruido gaussiano, esto permite generar datos sintéticos que ayuden a mejorar el entrenamiento de nuestra red de post-procesamiento.

VI. Conclusiones

Este trabajo demuestra la factibilidad de utilizar redes GAN para la generación de muestras sintéticas de TOA, que luego pueden ser utilizadas para aumentar conjuntos preexistentes y así mejorar el rendimiento de redes neuronales que se entrenen con los mismos. También podemos concluir que la FID es una métrica acertada para medir el comportamiento de una red GAN al momento de generar imágenes a partir de ruido, donde no se cuenta con imágenes de entrada en la red a modo de referencia para comparar.

A continuación se mencionan ciertas limitaciones que podrían ser solventadas en futuros trabajos, en búsqueda de mejores resultados. En primer lugar, el poder y espacio de computo reducido impactó en la generación de datos sintéticos. Sería deseable seguir iterando sobre la cantidad de muestras sintéticas en el conjunto BV', para encontrar el límite práctico de nuestra red GAN, aquel donde empiece a generalizar y las imágenes generadas dejen de aportar información útil en el entrenamiento de la red U-Net. Por otro lado, herramientas como barrido de parámetros para ambas redes tampoco fueron utilizadas. Por último, otro tipo de caso que podría resultar de interés es la utilización de un conjunto de datos BV' que contenga menor cantidad de datos reales que BV. Por ejemplo, se podría igualar la cantidad de imágenes en ambos conjuntos pero reducir la cantidad total de muestras reales en el conjunto aumentado. De esta manera la confianza en la calidad de las muestras generadas sería todavía mayor.

Agradecimientos

Este trabajo fue financiado por la Universidad de Buenos Aires (UBACYT 20020190100032BA), CONICET (PIP 11220200101826CO) y la Agencia I+D+i (PICT 2018-04589, PICT 2020-01336).

Referencias

[1] C. Huang, K. Wang, L. Nie, and et al., "Full-wave iterative image reconstruction in photoacoustic tomography with acoustically inhomogeneous media," *IEEE Transactions on Medical Imaging*, vol. 32, pp. 1097–1110, 2013.

[2] S. Arridge, P. Beard, M. Betcke, and et al., "Accelerated high-resolution photoacoustic tomography via compressed sensing," *Physics in medicine and biology*, vol. 61, pp. 8908–8940, 2016.

[3] Y. E. Boink, M. J. Lagerwerf, W. Steenbergen, and et al., "A framework for directional and higher-order reconstruction in photoacoustic tomography," *Physics in Medicine & Biology*, vol. 63, 2018.

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[5] A. Hauptmann and B. Cox, "Deep learning in photoacoustic tomography: Current approaches and future directions," *Journal of Biomedical Optics*, vol. 25, 09 2020.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv preprint arXiv:1505.04597*, 2015.

[7] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, "Fully Dense UNet for 2-D Sparse Photoacoustic Tomography Artifact Removal," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 568–576, 2020.

[8] X. Ma, C. Peng, J. Yuan, Q. Cheng, G. Xu, X. Wang, and P. L. Carson, "Multiple delay and sum with enveloping beamforming algorithm for photoacoustic imaging," *IEEE Trans. on Medical Imaging*, vol. 39, pp. 1812–1821, 2019.

[9] L. Torrey and J. Shavlik, "Transfer learning," *Handbook of Research on Machine Learning Applications*, 01 2009.

[10] B. Liu, Y. Zhu, K. Song, and A. Elgammal, "Towards faster and stabilized GAN training for high-fidelity few-shot image synthesis," *arXiv preprint arXiv:2101.04775*, 2021.

[11] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," *stat*, vol. 1050, 01 2017.

[12] D. Zhang and A. Khoreva, "PA-GAN: Improving gan training by progressive augmentation," *arXiv preprint arXiv:1901.10422*, 01 2019.

[13] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," 07 2017, pp. 5967–5976.

[16] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," *Advances in Neural Information Processing Systems*, pp. 15 663––15 674, 2019.

[17] J. Lim and J. C. Ye, "Geometric GAN," *arXiv preprint arXiv:1705.02894*, 05 2017.

[18] "DRIVE: Digital retinal images for vessel extraction," 2020. [Online]. Available: https://drive.grand-challenge.org/

[19] "STARE: Structured analysis of the retina," 2000. [Online]. Available: https://cecas.clemson.edu/~ahoover/stare/

[20] "RITE: Retinal images vessel tree extraction," 2013. [Online]. Available: https://medicine.uiowa.edu/eye/rite-dataset

[21] "ARIA: Automated retinal image analysis," 2006. [Online]. Available: http://www.damianjjfarnell.com/

[22] A. Hatamizadeh, H. Hosseini, N. Patel, J. Choi, C. Pole, C. Hoeferlin, S. Schwartz, and D. Terzopoulos, "RAVIR: A dataset and methodology for the semantic segmentation and quantitative analysis of retinal arteries and veins in infrared reflectance imaging," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[23] A. Borji, "Pros and cons of gan evaluation measures," *Computer Vision and Image Understanding*, vol. 1793, pp. 41–65, 2019.

[24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, pp. 6626–6637, 2017.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[27] M. G. Gonzalez, M. Vera, and L. R. Vega, "Combining band-frequency separation and deep neural networks for optoacoustic imaging," *Optics and Lasers in Engineering*, vol. 163, p. 107471, 2023.

[28] L. Hirsch, M. G. Gonzalez, and L. R. Vega, "A comparative study of time domain compressed sensing techniques for optoacoustic imaging," *IEEE Latin America Transactions*, vol. 20, pp. 1018–1024, 2022.

[29] C. Tian, M. Pei, K. Shen, S. Liu, Z. Hu, and T. Feng, "Impact of system factors on the performance of photoacoustic tomography scanners," *Phys. Rev. Applied*, vol. 13, p. 014001, 2020.

[30] M. Haltmeier, M. Sandbichler, T. Berer, J. Bauer-Marschallinger, P. Burgholzer, and L. Nguyen, "A sparsification and reconstruction strategy for compressed sensing photoacoustic tomography," *Acoust. Soc. Am.*, vol. 143, no. 6, p. 3838–3848, 2018.

[31] S. Guan, A. Khan, S. Sikdar, and P. Chitnis, "Fully dense unet for 2D sparse photoacoustic tomography artifact removal," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 568–576, 2020.

[32] W. Xing-xing and L. Jin-guo, "A new early stopping algorithm for improving neural network generalization," in *2009 Second International Conference on Intelligent Computation Technology and Automation*, vol. 1, 2009, pp. 15–18.

[33] N. Awasthi, G. Jain, S. K. Kalva, M. Pramanik, and P. Yalavarthy, "Deep neural network-based sinogram super-resolution and bandwidth enhancement for limited-data photoacoustic tomography," *IEEE Transactions on Ultrasonics Ferroelectrics and Frequency Control*, vol. PP, 02 2020.