

Thermal Single Image Super-Resolution

Andrea J. Parra A. Guillermo Pinto Nicolás A. Ramírez C.
Universidad Industrial de Santander
Bucaramanga, Colombia

Abstract

La super-resolución de imágenes térmicas (TISR) es crucial para mejorar imágenes de sensores infrarrojos de bajo costo, pero su baja resolución, bajo contraste y ruido dificultan los métodos tradicionales. En el presente trabajo proponemos adaptar el Progressive Focused Transformer (PFT-light), originalmente entrenado para imágenes visibles, mediante ajuste fino con pares de imágenes térmico-RGB. Introducimos una pérdida cruzada de espectro que transfiere texturas de la modalidad RGB y una regularización de variación total (TV) que conserva la suavidad típica de las imágenes térmicas. En los resultados con el dataset CIDIS (1000 pares), nuestro método alcanza PSNR=32,0dB y SSIM=0,904, superando la interpolación cúbica y el modelo SwinFuSR. Los estudios de ablación confirman que ambas pérdidas son complementarias, y se observa una capacidad emergente de colorear imágenes térmicas cuando se usa solo la pérdida cruzada. Así, el ajuste fino de transformers diseñados para el espacio visible brinda una solución eficaz y computacionalmente ligera para la super-resolución de imágenes térmicas.

1. Introducción

En el procesamiento de imágenes, la tarea de superresolución de una sola imagen (SISR, por sus siglas en inglés *Single Image Super-Resolution*) tiene como objetivo reconstruir detalles y texturas de alta resolución a partir de imágenes de baja resolución (LR, por sus siglas en inglés *Low Resolution*) [15]. Tradicionalmente, este problema se ha abordado utilizando métodos basados en modelos matemáticos, tales como la interpolación dirigida por bordes, la codificación dispersa y el *gradient profile prior* [10]. Sin embargo, estos métodos presentan tres inconvenientes: En primer lugar, requieren el diseño manual de la función de mapeo del espacio LR al HR (por sus siglas en inglés, *High Resolution*); en segundo lugar, estos métodos implican altos costos computacionales, lo que los hace poco prácticos en aplicaciones de tiempo real y en tercer lugar, la calidad de las imágenes reconstruidas no es suficiente para las

aplicaciones requeridas; los valores PSNR (del inglés, *Peak Signal-to-Noise Ratio*) y LPIPS (del inglés, *Learned Perceptual Image Patch Similarity* [23]) no se ajustan a los valores esperados [22].

Con la llegada del *Deep Learning*, los enfoques basados en el aprendizaje profundo se han convertido en la forma común de abordar la mayoría de los problemas de la visión por computadora, entre ellas, la superresolución. Modelos como SRCNN [21], aplican redes neuronales convolucionales (CNN, por sus siglas en inglés *Convolutional Neural Networks*) para aprender el mapeo entre una imagen LR y su versión HR. SwinIR [11], es un modelo basado en *Transformers* diseñado para resolver tareas de restauración de imágenes (incluyendo SISR) que ha conseguido buenos resultados. PFT (del inglés, *Progressive Focused Transformer*), se trata de un modelo *Transformer* con cambios en la modalidad de atención que le permiten tener un menor costo computacional y mayor desempeño superando al estado del arte [15]. Estos enfoques representan algunos de los avances más destacados en la evolución de los métodos de superresolución basados en redes neuronales profundas (DNN, por sus siglas en inglés *Deep Neural Networks*).

Las imágenes infrarrojas (IR), que capturan la radiación térmica emitida por los objetos, permiten la detección de características invisibles al ojo humano, como la monitorización sin contacto de signos vitales [4] y aplicaciones de visión en ausencia de luz (visible). Aunque existen sensores infrarrojos de alta definición con resoluciones espaciales de hasta 1024×768 píxeles, su elevado costo –que puede alcanzar decenas de miles de dólares– limita su uso, siendo más comunes los sensores IR de baja resolución [2].

La superresolución de imágenes térmicas (TISR, por sus siglas en inglés *Thermal Image Super-Resolution*) aborda este problema. Sin embargo, un desafío clave reside en las diferencias inherentes entre las degradaciones que sufren las imágenes IR frente a las RGB (*Red, Green and Blue*, por sus siglas en inglés); a diferencia de las cámaras visibles, los sensores térmicos no capturan luz reflejada, sino la radiación térmica (calor) emitida por los objetos, la cual depende de la emisividad y temperatura de las superficies. Este proceso se describe por la ley de Planck, que relaciona

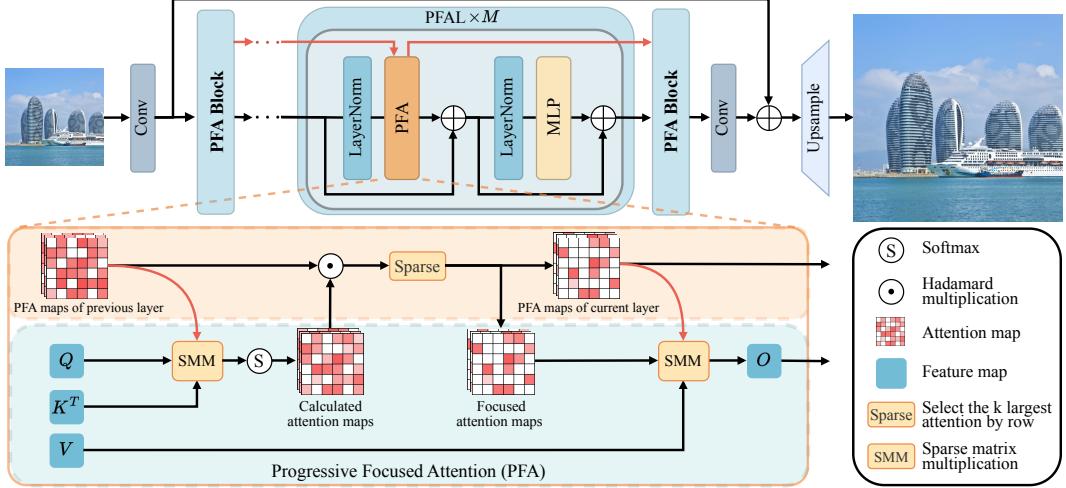


Figure 1. **Arquitectura del PFT.** El bloque PFA consta de M Progressive Focused Attention Layers (PFAL). Cada PFA toma como entrada tanto las características de la imagen como los mapas PFA de capas anteriores. *Sparse Matrix Multiplication* (SMM) se asegura que cada fila de Q interactúe únicamente con las columnas necesarias de K^T , generando los mapas de atención calculados. Los mapas PFA de la capa actual se obtienen luego de aplicar el producto Hadamard y *Sparse Focusing*, y se usan con la matriz V en una operación SMM para generar las características de atención agregadas. Tomado de [15].

la radiancia espectral con la temperatura del cuerpo [8].

En las cámaras en el espectro visible, la textura y el contraste provienen de la luz reflejada bajo una fuente de iluminación externa, lo que permite distinguir detalles geométricos gracias a las variaciones en los ángulos de las superficies. En cambio, en las cámaras térmicas, la emisión directa domina la señal, ocultando las variaciones locales que aportan textura (fenómeno conocido como *ghosting effect* o pérdida de textura geométrica). Además, los sensores infrarrojos operan a longitudes de onda mayores y con detectores de menor densidad, lo que genera menor resolución espacial, bajo contraste radiométrico y mayor ruido térmico [3].

Los métodos basados en DNNs para TISR han explorado varias arquitecturas: inicialmente CNNs para combinar sus capacidades de reconocimiento de patrones con métodos de reconstrucción tradicionales [24]. Posteriormente, modelos *end-to-end*, que reconstruyen directamente las imágenes térmicas, incorporando módulos especializados y, más recientemente, *Transformers* para capturar dependencias a largo alcance y mejorar la consistencia estructural mediante atención global [8, 26]. Dado que las imágenes visibles presentan mayor riqueza de textura y contraste, también han surgido enfoques multimodales que utilizan imágenes RGB como guía en la reconstrucción infrarroja [2, 8]. Algunos de estos modelos que han conseguido resultados importantes en el área del TISR, como lo es *SwinFuSR* [2], también relevan otros cambios importantes en su arquitectura, como lo es la incorporación de *Swin Transformers* [14], que son *Transformers* específicamente diseñados para tareas de visión por computador, este tipo de modelos utiliza

tamaños de ventana que cambian a través de las capas para conseguir conexiones entre ventanas sin solaparlas entre sí [2]. Un trabajo inspirado en SwinFuSR añadió mejoras a las propuestas ya mencionadas para hacer al modelo más robusto mediante un mejor *data mixing* y con una nueva función de pérdida que permite medir la calidad de la superresolución en varias escalas (x2, x4 y x8) [25].

En el presente proyecto abordamos el problema de TISR utilizando un enfoque novedoso que ha demostrado grandes resultados en el campo del SISR para imágenes en el espectro visible. Se trata de utilizar un modelo basado en PFT [15] para realizar superresolución con imágenes IR. En concreto, se planea ajustar finamente un modelo pre-entrenado en el dominio RGB para adaptarlo a un conjunto de imágenes IR usando una función de pérdida multimodal para aprovechar la información RGB. Este enfoque propone responder la pregunta de investigación: ¿Puede un modelo pre-entrenado con imágenes en el dominio visible ajustado finamente en el dominio térmico guiado por una función de pérdida de *espectro cruzado* llevar a resultados competentes con respecto a las técnicas actuales en el campo de TISR?

2. Método

Para responder a esta pregunta, basamos nuestro método en el ajuste fino del modelo PFT [15], supervisado mediante las funciones de pérdida propuestas en [12], adaptadas a la tarea de superresolución de imágenes térmicas, apoyándose de las características de las imágenes a color. Primero introduciremos la arquitectura empleada, posteriormente la función de pérdida de espectro cruzado, donde utilizando el

gradiente espacial para cada uno de los canales en la imagen a color, podemos extraer sus texturas y patrones para imponerlos en la imagen térmica superresuelta. Sin embargo, como esto puede introducir artefactos, por ejemplo, alucinando bordes no presentes en la imagen térmica, nos apoyamos de la función de pérdida de variación total (TV, del inglés, *Total Variation*), la cual impone suavidad, condición característica de las imágenes térmicas.

2.1. Arquitectura

La arquitectura del modelo, véase la Figura 1, sigue la estructura de otras soluciones del estado del arte [5, 11], incorporando el enfoque *shifted window* del *Swin Transformer* [13]: se divide la imagen en ventanas de 32x32 píxeles y en cada capa estas ventanas se desplazan ligeramente respecto a la capa anterior. Esto permite que la atención no quede enfocada sólo dentro de una ventana fija, sino que pueda captar relaciones entre regiones cercanas. Se incorpora además como codificación posicional LePE (*Locally-enhanced Positional Encoding*) para el cálculo de atención, pues maneja la información posicional local mejor que los esquemas de codificación existentes [6].

La novedad del PFT es su módulo de atención PFA (del inglés, *Progressive Focused Attention*), que se resume en las siguientes ecuaciones:

$$\mathbf{A}_{sc}^l = \text{Softmax}(\Psi(\mathbf{Q}^l, (\mathbf{K}^l)^T, \mathbf{I}^{l-1})), \quad (1)$$

$$\mathbf{A}^l = S_{K^l}(\text{Norm}(\mathbf{A}_{sc}^l \odot \mathbf{A}^{l-1})), \quad (2)$$

$$\mathbf{I}^l = \text{Sign}(\mathbf{A}^l), \quad (3)$$

$$\mathbf{O}^l = \Psi(\mathbf{A}^l, \mathbf{V}^l, \mathbf{I}^l). \quad (4)$$

Primero, en la Ecuación 1 se calculan los mapas de atención en la capa actual l , Ψ denotando la operación SMM (*Sparse Matrix Multiplication*); \mathbf{I}^{l-1} es una matriz de índices dispersa que determina si calculamos $\mathbf{A}_{sc}^l(i, j)$ con $\mathbf{Q}^l(i, :)$ y $\mathbf{K}^l(j, :)^T$ sólo si $\mathbf{I}^{l-1}(i, j) = 1$, de lo contrario $\mathbf{A}_{sc}^l(i, j) = 0$. Esto permite que la atención se enfoque únicamente en los elementos relevantes, evitando cálculos innecesarios.

Luego, en la Ecuación 2, se realiza la normalización del mapa de atención por filas combinado con la capa anterior, $\mathbf{A}_{sc}^l \odot \mathbf{A}^{l-1}$; así, los *tokens* menos relevantes se eliminarán durante el proceso de multiplicación, mientras que a los *tokens* importantes con altos valores de similitud entre capas se les asignarán pesos mayores después de la normalización. Tras la normalización, se aplica la operación S_K que mantiene únicamente los top K^l valores por fila, filtrando los *tokens* menos importantes. Este mecanismo progresivo garantiza que cada capa se concentre gradualmente en los *tokens* más relevantes, reduciendo la carga computacional. La Ecuación 3 genera la matriz binaria de índices \mathbf{I}^l , que indica las posiciones activas que se considerarán en la

siguiente capa. Finalmente, la Ecuación 4 calcula la salida de la capa, \mathbf{O}^l , utilizando únicamente los valores de atención activos. En la primera capa de atención, como no existe un mapa de atención previo, se inicializan \mathbf{A}^0 e \mathbf{I}^0 como matrices de unos, y \mathbf{A}^1 se calcula como una atención estándar. En las capas posteriores, PFA utiliza los mapas de atención anteriores para filtrar progresivamente *tokens* irrelevantes, ajustando el número de valores retenidos con $K^l = \alpha K^{l-1}$ y $0 < \alpha < 1$. Para evitar filtrar *tokens* importantes en etapas tempranas, se establece $K^1 = N$, siendo N el número de *tokens* inicial.

2.2. Pérdida de espectro cruzado

Para optimizar el modelo, se empleo la pérdida de espectro cruzado, la cual se define como el valor absoluto de la diferencia de los gradientes espaciales entre cada uno de los canales de la imagen a color y los gradientes de la imagen térmica predicha por el modelo:

$$\mathcal{L}_{cc} = \frac{1}{HW} \sum_{i=1}^{HW} \left| \frac{1}{3} (\nabla r_i + \nabla g_i + \nabla b_i) - \nabla th_i \right|. \quad (5)$$

En la función de pérdida \mathcal{L}_{cc} , H y W representan el alto y ancho de la imagen, respectivamente. Mientras que ∇r , ∇g y ∇b , iniciales de *red*, *green* y *blue*, son los gradientes espaciales por canal para el pixel i dado. Por su parte, ∇th representa los mismos valores para la imagen térmica. La intuición detrás de esta función de pérdida viene de [12], donde se propone utilizar la riqueza en los detalles de las imágenes a color para recuperar de mejor manera la forma y texturas de los objetos que se encuentran en la escena. Esto es particularmente especial en el caso de la superresolución, ya que, de por sí, las imágenes térmicas tienen pocas texturas, y adicionalmente, si la imagen está en muy baja resolución, como es el caso común de las imágenes capturadas por las cámaras térmicas, en ellas no se podrán distinguir siluetas, bordes, o detalles finos que si pueden estar presentes en una imagen a color capturada en la misma escena.

De esta manera integramos la información complementaria de la imagen RGB sin incorporar directamente dicha modalidad en el flujo de la red. Sin embargo, el exceso de transferencia de texturas de una imagen a color a la imagen térmica puede ser perjudicial porque podría inducir artefactos no deseados, como bordes no presentes en una imagen térmica natural. Para mitigar este efecto, promover la suavidad intrínseca de las imágenes térmicas es importante, por lo tanto, introducimos la siguiente función de pérdida.

2.3. Pérdida de variación total

La pérdida de variación total (TV, del inglés, *Total Variation*) es una función de pérdida que extrae los gradientes espaciales de una imagen, en la práctica, mediante las diferencias en el eje horizontal y vertical. Posteriormente, se

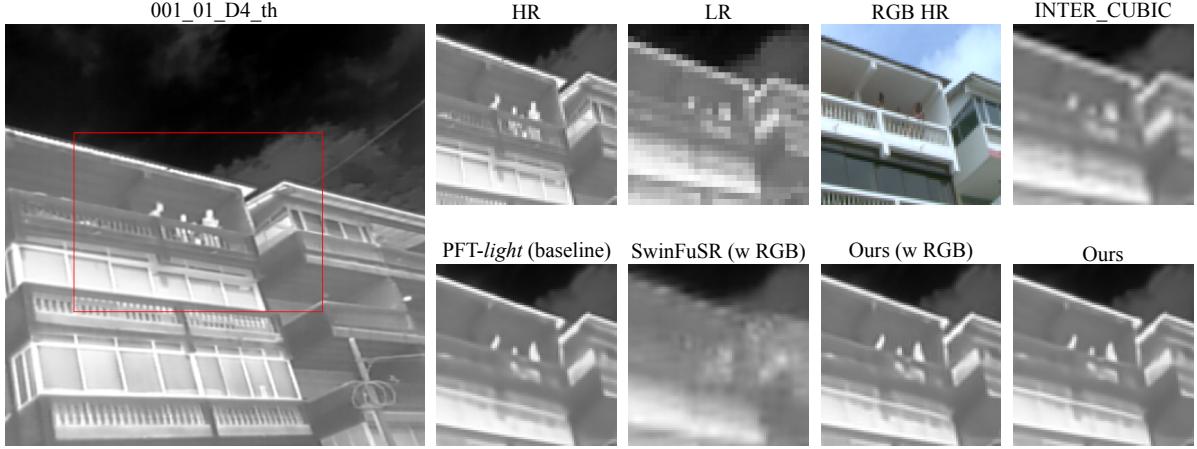


Figure 2. **Comparación cualitativa.** En la primera columna la imagen original. Luego, en la primera fila, las imágenes en alta resolución, en baja resolución, a color en alta resolución y el resultado del método INTER_CUBIC [9]. En la segunda fila, los resultados de los modelos PFT-light [15], SwinFuSR [2], el nuestro con RGB y el nuestro solamente en ajuste fino.

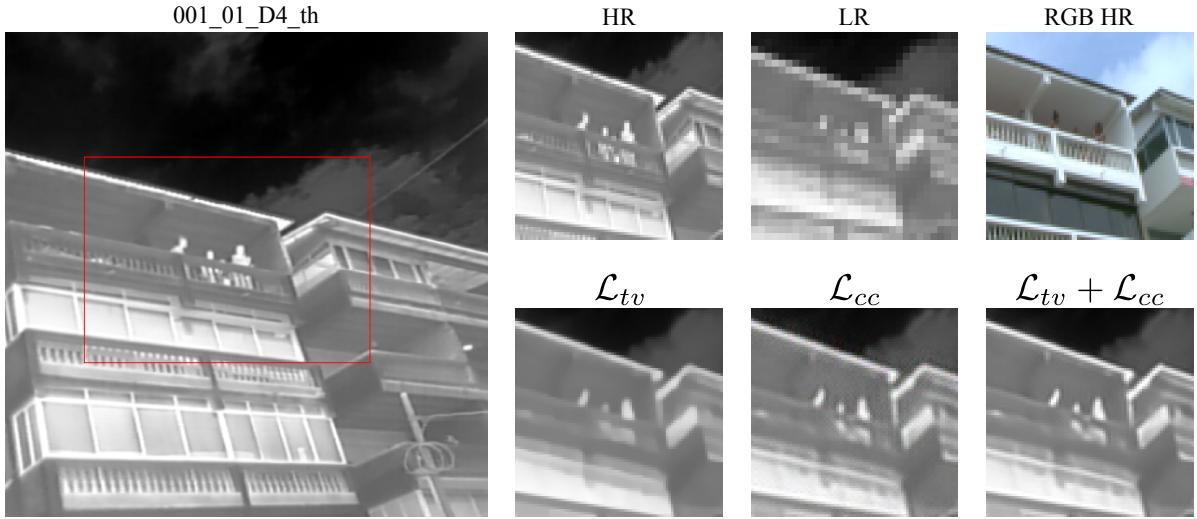


Figure 3. **Comparación cualitativa de los estudios de ablación.** En la primera columna la imagen original. Luego, en la primera fila la imagen en alta resolución, en baja resolución y la imagen RGB en alta resolución. En la segunda, el resultado con la \mathcal{L}_{tv} , \mathcal{L}_{cc} y la combinación $\mathcal{L}_{tv} + \mathcal{L}_{cc}$, respectivamente.

toma el promedio del valor absoluto de estas diferencias, descrita matemáticamente de la siguiente forma:

$$\mathcal{L}_{tv} = \frac{1}{HW} \sum_{i=1}^{HW} |\nabla th_i|. \quad (6)$$

Con esto, al querer minimizar las diferencias espaciales en nuestra imagen, estamos imponiendo suavidad, lo que es particularmente especial en las imágenes térmicas, ya que por naturaleza tienden a carecer de texturas debido a que lo que se ve en estas imágenes es la emisión térmica de los objetos en lugar de la luz solar reflejada.

3. Experimentos

En esta sección primero se describe la configuración experimental utilizada para obtener los modelos desarrollados, luego, se exponen los estudios de ablación donde se analiza la contribución de cada componente de la propuesta, posteriormente, se compara nuestro método frente a enfoques de referencia y por último, se presenta una propiedad emergente encontrada.

3.1. Configuración experimental

Conjunto de datos. Para comparar el rendimiento de la propuesta empleamos el conjunto de datos recientemente usado para evaluar el desempeño de diferentes métodos en

la tarea de TISR [18], CIDIS [19]. Este conjunto de datos está compuesto de 1000 imágenes térmicas (700, 200 y 100 para entrenamiento, validación y prueba, respectivamente) en resolución 640×448 (ancho y alto) cada una registrada con su par a color, RGB. Primero redujimos la resolución de las imágenes originales a 64×64 , en adelante notada como LR. Este proceso se llevó a cabo usando una implementación en Python de la función *imresize* del entorno de computación numérica MATLAB, método ampliamente usado en los retos de SISR [1, 7].

Arquitectura. Para SISR clásica, la red PFT está formada por 6 bloques, con un número de capas de atención por bloque de [4, 4, 4, 6, 6, 6], con 6 cabezas y un total de 240 canales. El tamaño de ventana es de 32×32 , y la cantidad de valores de atención retenidos en cada bloque es [1024, 256, 128, 64, 32, 16], respectivamente. Para SISR ligera (versión más liviana del modelo original), la red PFT-light tiene cinco bloques, con un número de capas de atención por bloque de [2, 4, 6, 6, 6], 4 cabezas de atención y un total de 52 canales. La cantidad de valores de atención retenidos en cada bloque sigue el patrón [1024, 256, 128, 64, 32]. En esta propuesta se utilizó como modelo base la versión PFT-light.

Entrenamiento. Cada modelo se entrenó durante 50000 iteraciones utilizando el conjunto de entrenamiento del conjunto de datos CIDIS [19]. Se empleó el optimizador AdamW [16] con una tasa de aprendizaje inicial de 2×10^{-4} , un *warmup* lineal durante 10000 iteraciones, el *scheduler* MultistepLR, el cual reduce la tasa de aprendizaje a la mitad en iteraciones especificadas [25000, 40000] y un *batch size* igual a 4. Por defecto, se supervisa usando la norma L1 entre la imagen HR, que en este caso es el resultado de re-escalar la imagen original al tamaño 196×196 , y la predicción. Por limitaciones de cómputo, solamente se experimentó en la escala $\times 3$.

3.2. Estudios de ablación

Resultados cuantitativos. Para analizar la contribución de cada función de pérdida propuesta realizamos unos estudios de ablación. Los modelos se evaluaron ejecutando inferencia en el conjunto de validación del conjunto de datos CIDIS [19], y las predicciones se compararon frente a la imagen HR. Los resultados cuantitativos, enseñados en la Tabla 1, demuestran que el mejor rendimiento se obtiene al ajustar finamente solamente con la norma L1. Usar solamente la función de pérdida \mathcal{L}_{tv} reduce el desempeño probablemente debido a que se pierden en mayor medida las texturas, en contraste, al solo utilizar la función de pérdida \mathcal{L}_{cc} se obtienen mejores resultados. Por último, al combinar ambas funciones de pérdida se consiguen resultados competitivos.

Resultados cualitativos. Para complementar los resultados cuantitativos, en la Figura 3 se enseña el desempeño cuali-

Table 1. Comparación cuantitativa de los estudios de ablación (PSNR/SSIM), en la tarea de superresolución *light* en el conjunto de datos CIDIS [19]. El mejor se encuentra en **negrita**, el segundo mejor sobrayado.

L1	\mathcal{L}_{tv}	\mathcal{L}_{cc}	PSNR ↑	SSIM ↑
✓	✗	✗	32.0038	0.9044
✓	✓	✗	30.7954	0.8696
✓	✗	✓	31.0192	0.8865
✓	✓	✓	<u>31.6725</u>	<u>0.8998</u>

tativo de cada estudio de ablación, donde podemos observar el efecto de cada función de pérdida. Cuando se añade solamente la función de pérdida \mathcal{L}_{tv} , nos encontramos con un resultado más plano, que carece de texturas, por ejemplo, los marcos de las ventanas no se alcanzan a percibir. Por otro lado, cuando solo tenemos la función de pérdida \mathcal{L}_{cc} , podemos imponer mejores texturas desde la imagen a color, en este caso, es posible notar los detalles de las ventanas.

Sin embargo, encontramos que también se induce un mosaico de bayer, el cual creemos se debe a dos razones: (i) Para poder ajustar finamente el modelo PFT-light tuvimos que triplicar el canal de la imagen térmica monocromática al numero de canales de una imagen RGB, luego, esto provoca que la \mathcal{L}_{cc} se compute sobre tres canales, en lugar de solo uno, lo que propicia el mosaico bayer, característico de las imágenes a color. (ii) El modelo PFT por defecto realiza el *upsampling* de los mapas de características mediante el modulo *PixelShuffle* [20], el cual reorganiza los mapas de características $(*, C \times r^2, H, W)$ (es decir, la imagen en baja resolución, pero con gran numero de canales), agrupándolos de manera espacial para obtener una imagen más grande y con menos canales $(*, C, H \times r, W \times r)$, modulado por el factor de escala r ; seguido de unas capas de convolución para definir los detalles de la predicción, sin embargo, la versión *light* del modelo PFT no utiliza dichas capas de convolución, con lo cual, hipotetizamos que ese reagrupamiento, sin su debido refinamiento, se convierte en los artefactos observados en la imagen superresuelta.

Finalmente, al combinar ambas funciones de pérdida $\mathcal{L}_{tv} + \mathcal{L}_{cc}$, se puede observar un balance entre la suavidad característica de la imagen térmica mientras se inducen los detalles en las texturas mejor definidos en la imagen a color, a su vez mitigando el artefacto del mosaico bayer. Debido a esto, en adelante definimos al modelo ajustado finamente con la función de pérdida L1 como nuestra propuesta base (ours) y a la versión que combina las tres funciones de pérdida, como nuestra propuesta con RGB (ours (w RGB)).

3.3. Comparación con los métodos de referencia

Resultados cuantitativos. Equiparamos nuestra propuesta frente a varios métodos. Primero con la función de OpenCV

para realizar el re-escalado de imágenes, *resize*, con el método de interpolación INTER_CUBIC [9]. Luego, con el método del estado del arte SwinFuSR [2], haciendo inferencia con los pesos pre-entrenados disponibles. Elegimos este modelo debido a que también utiliza la imagen a color en el entrenamiento, sin embargo, ellos la incorporan en el flujo de la red para extraer y aprovechar sus características, a diferencia de nosotros, que solo la empleamos en el cálculo de la función de pérdida \mathcal{L}_{cc} . Por último, con la inferencia por el modelo pre-entrenado PFT-light, es decir, nuestra referencia base (baseline).

Los resultados, presentados en la Tabla 2, demuestran que el modelo propuesto es superior a los método de referencia. Especialmente, supera con creces a SwinFuSR, uno de los métodos del estado del arte en los *benchmarks* de TISR [18]. No obstante, aunque la propuesta con RGB obtiene un desempeño competitivo no logra superar el baseline, con lo que aun hay oportunidades de mejora para trabajo futuro.

Table 2. Comparación cuantitativa frente a los métodos de referencia (PSNR/SSIM), en la tarea de superresolución *light* en el conjunto de datos CIDIS [19]. El mejor se encuentra en **negrita**, el segundo mejor subrayado.

Method	Scale	PSNR \uparrow	SSIM \uparrow
INTER_CUBIC		27.7669	0.8332
SwinFuSR (w RGB)		24.9115	0.7144
PFT-light (baseline)	$\times 3$	<u>31.8215</u>	<u>0.9013</u>
Ours (w RGB)		31.6725	0.8998
Ours		32.0038	0.9044

Resultados cualitativos. Para comprobar la calidad del modelo propuesto, mostramos una comparación visual con varios métodos de referencia en la Figura 2. Esta comparación expone el potencial de nuestra propuesta para recuperar bordes finos a partir de la imagen LR. Por ejemplo, nuestra propuesta con RGB es capaz de definir con un poco más de detalle los marcos de las ventanas, que el baseline. Por último, en comparación con el modelo del estado del arte SwinFuSR, las diferencias son bastante notorias, pues este último ofrece pobres capacidades de generalización.

3.4. Propiedad emergente

Durante la experimentación con diferentes funciones de pérdida, planteamos una en la que la norma L1 se calculaba directamente comparando la imagen térmica con la imagen a color, es decir, asumiendo que el *ground truth* en este caso era la información contenida en la vista RGB. A esta función de pérdida la denominamos \mathcal{L}_{rgb} . En la Figura 4 se observan los resultados obtenidos mediante esta configuración. Sorprendentemente, cuando usamos solamente dicha función de pérdida, encontramos que el modelo intenta recrear una imagen a color a partir de la imagen LR

térmica, por ejemplo, podemos notar como intenta “colorear”, las ventanas de color negro, las barandillas de color blanco y el cielo de color azul. Por otro lado, cuando la combinamos con la función de pérdida \mathcal{L}_{tv} notamos que, aunque preserva los colores, las texturas se pierden, confirmado el efecto esperado debido a que esta última impone suavidad. Esta propiedad emergente tiene el potencial de desarrollarse como un método más robusto para la tarea de traducción de imágenes, un enfoque ampliamente usado para mejorar las capacidades de los sistemas de visión [17].



Figure 4. **Propiedad emergente** del modelo para la traducción de imágenes. En la primera columna, la imagen a color original. Luego, en la primera fila la imagen a color en alta resolución y la imagen térmica en baja resolución. En la segunda fila, las predicciones con la pérdida \mathcal{L}_{rgb} y su combinación con la \mathcal{L}_{tv} .

4. Conclusiones

Se observan dos efectos visuales relevantes: (i) la aparición de un patrón Bayer al trabajar con imágenes RGB en las *losses*, lo cual se debe a triplicar los canales de la imagen térmica y la forma en que PixelShuffle reorganiza y compacta los canales al incrementar la resolución espacial; y (ii) una reconstrucción de imágenes RGB de alta resolución a partir de imágenes térmicas de baja resolución. Finalmente, el modelo muestra un rendimiento competitivo frente al estado del arte; sin embargo, se requiere más experimentación para superar el baseline con técnicas RGB como las funciones de pérdida propuestas.

References

- [1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 126–135, 2017. 5
- [2] Cyprien Arnold, Philippe Jouvet, and Lama Seoud. Swinfusr: An image fusion-inspired model for rgb-guided thermal image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3027–3036, 2024. 1, 2, 4, 6
- [3] Fanglin Bao, Shubhankar Jape, Andrew Schramka, Junjie

- Wang, Tim E. McGraw, and Zubin Jacob. Why thermal images are blurry. *Opt. Express*, 32(3):3852–3865, 2024. 2
- [4] A. Bridier, M. Shcherbakova, A. Kawaguchi, N. Poirier, C. Said, R. Noumeir, and P. Jouvet. Hemodynamic assessment in children after cardiac surgery: A pilot study on the value of infrared thermography. *Frontiers in Pediatrics*, 11: 1083962, 2023. 1
- [5] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22367–22377, 2023. 3
- [6] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12114–12124, 2022. 3
- [7] Fatheral. Python implementation of matlab imresize function. https://github.com/fatheral/matlab_imresize, 2020. 5
- [8] Yongsong Huang, Tomo Miyazaki, Xiaofeng Liu, and Shinichiro Omachi. Infrared image super-resolution: A systematic review and future trends. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 1–26, 2025. 2
- [9] Itseez. Open source computer vision library. <https://github.com/itseez/opencv>, 2015. 4, 6
- [10] Xin Li, Weisheng Dong, Jinjian Wu, Leida Li, and Guangming Shi. Superresolution image reconstruction: Selective milestones and open problems. *IEEE Signal Processing Magazine*, 40(5):54–66, 2023. 1
- [11] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021. 1, 3
- [12] Yvette Y Lin, Xin-Yi Pan, Sara Fridovich-Keil, and Gordon Wetzstein. Thermalnerf: Thermal radiance fields. In *2024 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2024. 2, 3
- [13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 3
- [14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. 2
- [15] Wei Long, Xingyu Zhou, Leheng Zhang, and Shuhang Gu. Progressive focused transformer for single image super-resolution. *arXiv preprint arXiv:2503.20337*, 2025. 1, 2, 4
- [16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [17] D Manjunath, Aniruddh Sikdar, Prajwal Gurunath, Sumanth Udupa, and Suresh Sundaram. Saga: Semantic-aware gray color augmentation for visible-to-thermal domain adaptation across multi-view drone and ground-based vision systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4578–4588. IEEE, 2025. 6
- [18] Rafael E Rivadeneira, Angel D Sappa, Chenyang Wang, Junjun Jiang, Zhiwei Zhong, Peilin Chen, and Shiqi Wang. Thermal image super-resolution challenge results-pvbs 2024. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3113–3122, 2024. 5, 6
- [19] Rafael E Rivadeneira, Henry O Velesaca, and Angel Sappa. Cross-spectral image registration: a comparative study and a new benchmark dataset. In *International Conference on Innovations in Computational Intelligence and Computer Vision*, pages 1–12. Springer, 2024. 5, 6
- [20] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition (CVPR)*, pages 1874–1883, 2016. 5
- [21] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 57–65, 2015. 1
- [22] Le Zhang, Ao Li, Qibin Hou, Ce Zhu, and Yonina C. Eldar. Deep-learning-empowered super resolution: A comprehensive survey and future prospects. *Proceedings of the IEEE*, pages 1–41, 2025.
- [23] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 1
- [24] Xudong Zhang, Chunlai Li, Qingpeng Meng, Shijie Liu, Yue Zhang, and Jianyu Wang. Infrared image super resolution by combining compressive sensing and deep learning. *Sensors*, 18:2587, 2018. 2
- [25] Hang Zhong, Yu Wang, and Shengjie Zhao. Swinpaste: A swin transformer-based framework for rgb-guided thermal image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4628–4633, 2025. 2
- [26] Yan Zou, Linfei Zhang, Chengqian Liu, Bowen Wang, Yan Hu, and Qian Chen. Super-resolution reconstruction of infrared images based on a convolutional neural network with skip connections. *Optics and Lasers in Engineering*, 146: 106717, 2021. 2