Cross-Modal Self-Supervised Adversarial Learning for MA-XRF Super-Resolution

Herman Verinaz-Jadan[1*], Su Yan[2*], Catherine Higgitt[3], and Pier Luigi Dragotti[4]
[1]Faculty of Electrical and Computer Engineering (FIEC), Escuela Superior Politecnica del Litoral (ESPOL), Ecuador
Email: hverinaz@espol.edu.ec
[2]Department of Bioengineering, Imperial College London, UK
Email: s.yan18@imperial.ac.uk
[3]Scientific Department, The National Gallery, London, UK
Email: catherine.higgitt@nationalgallery.org.uk
[4]Department of Electrical and Electronic Engineering, Imperial College London, UK
Email: p.dragotti@imperial.ac.uk

***Abstract*—High-quality element distribution maps enable precise analysis of Old Master paintings. These maps are typically produced by Macro X-ray Fluorescence (MA-XRF) scanning, a non-invasive technique for elemental imaging of flat surfaces. However, MA-XRF faces a trade-off between resolution and acquisition time, making high-resolution (HR) scans impractical for large artworks. Super-resolution MA-XRF mitigates this by enhancing scan quality while reducing acquisition time. This paper introduces a deep learning framework for MA-XRF super-resolution that removes the need for paired HR MA-XRF training data by leveraging RGB images to model cross-modal dependencies. Our approach is specifically tailored for MA-XRF, an important feature as RGB and MA-XRF data lack a common spectral domain. We introduce self-supervised adversarial training, where the discriminator learns from patches across modalities, guiding the generator toward realistic MA-XRF reconstructions. Additionally, our method enforces physical consistency via network design and enhances training through pseudo-real data augmentation. Experiments on Old Master paintings show our method outperforms state-of-the-art MA-XRF super-resolution techniques, demonstrating the need for tailored solutions as existing approaches from other domains do not generalize effectively to this task.**

***Index Terms*—X-ray fluorescence, MA-XRF super-resolution, adversarial, self-supervised learning, deep unfolding**

I. Introduction

Macro X-ray fluorescence (MA-XRF) is widely used to analyze the material composition and artistic techniques of Old Master paintings, offering detailed maps of elemental distributions across paint layers. MA-XRF works by measuring secondary X-ray photons emitted from chemical elements when excited by a primary X-ray beam [1]–[3]. While highly effective, MA-XRF imaging is constrained by a trade-off between resolution and acquisition time: high-resolution (HR) scans require prolonged exposure, making them impractical for large paintings. MA-XRF super-resolution (SR) techniques address this limitation by reconstructing HR element maps from low-resolution (LR) acquisitions, reducing scan times while preserving fine details [4], [5]. Early work on MA-XRF SR relied on model-based techniques, such as dictionary learning, to integrate HR RGB information with LR MA-XRF images [4]. Similar strategies have been explored in hyperspectral and multispectral imaging, where sparse representation, matrix factorization, and tensor-based methods have been widely used [6]–[8]. More recently, coupled dictionary learning was proposed in [5] to explicitly distinguish shared and unique information across modalities, preventing artifacts.

Deep learning has significantly advanced SR in spectral imaging domains such as Multispectral Imaging (MSI) [9], [10], Hyperspectral imaging (HSI) [11]–[13], and RGB guided depth map SR (GDSR) [14], [15]. However, applying these methods to MA-XRF is challenging due to two key factors: (i) spectral mismatch—unlike HSI and MSI, RGB and MA-XRF data do not share a common spectral domain, resulting in the lack of a linear mapping; (ii) severe data scarcity—HR MA-XRF datasets are rare due to cultural heritage constraints, limiting training. To address these challenges, we propose a tailored deep learning framework for MA-XRF SR that effectively integrates RGB guidance. The key contributions of this work include:

- **Self-supervised multi-modal learning:** Leverages structural information from a single HR RGB image.
- **Pseudo-real data augmentation:** Fuses HR RGB and LR MA-XRF images to mitigate training data limitations.
- **Misclassification-Focused Adversarial Loss:** Targets misclassified patches to enhance training efficiency.
- **Model-inspired architecture:** Incorporates domain knowledge into the network design.

II. Problem Formulation

This problem requires reconstructing a HR MA-XRF image (an element distribution map) by leveraging information from both MA-XRF and RGB modalities. Specifically, the reconstruction process involves synthesizing an HR MA-XRF image, $\tilde{\mathbf{Y}} \in \mathbb{R}^{B\times N_{\text{h}}}$, from a given LR MA-XRF image, $\tilde{\mathbf{Y}}_{\downarrow} \in \mathbb{R}^{B\times N_{\text{l}}}$, and an HR RGB image, $\tilde{\mathbf{Z}} \in \mathbb{R}^{b\times N_{\text{h}}}$. Here, B represents the number of spectral channels in the MA-XRF data, and N_{h} and N_{l} denote the total pixels in the HR and LR images, respectively.

*Herman Verinaz-Jadan and Su Yan contributed equally to this work.
This work is in part supported by EPSRC grant EP/R032785/1.

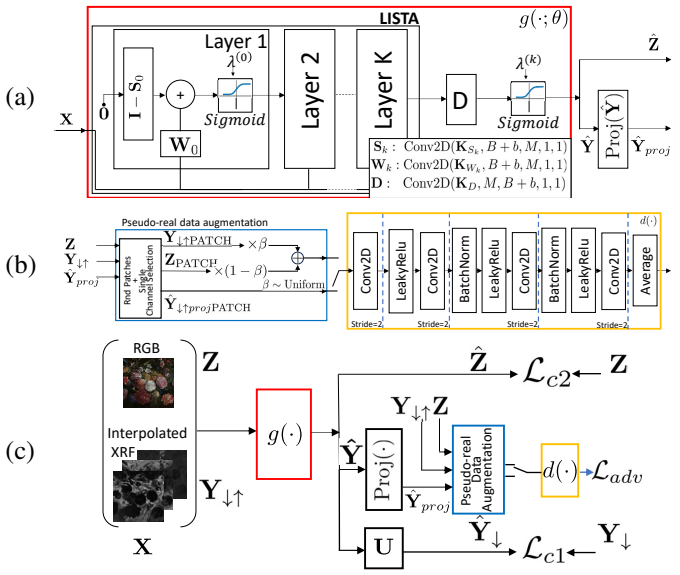


Fig. 1: Overview of the method. Part (a) shows the reconstruction network, while part (b) presents the discriminator and pseudo-real data augmentation. The adversarial framework is shown in part (c).

To facilitate reconstruction, a dictionary-based representation is commonly used to capture both the shared and unique features of MA-XRF and RGB images [5], [16]–[18]. Following the multimodal SR framework in [5], the HR MA-XRF and RGB images are represented by a set of dictionaries:

$$\begin{cases} \tilde{\mathbf{Y}} = \tilde{\mathbf{Y}}_{\mathrm{c}} + \tilde{\mathbf{Y}}_{\mathrm{u}} = \tilde{\mathbf{D}}_{\mathrm{c}}^{\mathrm{xrf}} \tilde{\mathbf{A}}_{\mathrm{c}} + \tilde{\mathbf{D}}_{\mathrm{u}}^{\mathrm{xrf}} \tilde{\mathbf{A}}_{\mathrm{u}}^{\mathrm{xrf}}, \\ \tilde{\mathbf{Z}} = \tilde{\mathbf{Z}}_{\mathrm{c}} + \tilde{\mathbf{Z}}_{\mathrm{u}} = \tilde{\mathbf{D}}_{\mathrm{c}}^{\mathrm{rgb}} \tilde{\mathbf{A}}_{\mathrm{c}} + \tilde{\mathbf{D}}_{\mathrm{u}}^{\mathrm{rgb}} \tilde{\mathbf{A}}_{\mathrm{u}}^{\mathrm{rgb}}, \end{cases} \tag{1}$$

where $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{Z}}$ denote the HR MA-XRF and RGB images, respectively. Here, the subscript c indicates the common (shared) components, while u denotes the unique components. The dictionaries $\tilde{\mathbf{D}}_{\mathrm{c}}^{\mathrm{xrf}} \in \mathbb{R}^{B \times M_c}$ and $\tilde{\mathbf{D}}_{\mathrm{u}}^{\mathrm{xrf}} \in \mathbb{R}^{B \times M_{xu}}$ (for MA-XRF) and $\tilde{\mathbf{D}}_{\mathrm{c}}^{\mathrm{rgb}} \in \mathbb{R}^{b \times M_c}$ and $\tilde{\mathbf{D}}_{\mathrm{u}}^{\mathrm{rgb}} \in \mathbb{R}^{b \times M_{ru}}$ (for RGB) contain M_c, M_{xu}, and M_{ru} dictionary components, respectively. The representation matrices are given by $\tilde{\mathbf{A}}_{\mathrm{c}} \in \mathbb{R}^{M_c \times N_h}$, $\tilde{\mathbf{A}}_{\mathrm{u}}^{\mathrm{xrf}} \in \mathbb{R}^{M_{xu} \times N_h}$, and $\tilde{\mathbf{A}}_{\mathrm{u}}^{\mathrm{rgb}} \in \mathbb{R}^{M_{ru} \times N_h}$, with $\tilde{\mathbf{A}}_{\mathrm{c}}$ shared across both modalities.

Furthermore, it is assumed that the LR MA-XRF image, $\tilde{\mathbf{Y}}_{\downarrow}$, is acquired from the HR MA-XRF image, $\tilde{\mathbf{Y}}$, through downsampling:

$$\tilde{\mathbf{Y}}_{\downarrow} = \tilde{\mathbf{Y}}\mathbf{U}. \tag{2}$$

The goal is to reconstruct $\tilde{\mathbf{Y}}$ given Equations (1) and (2), along with the available LR MA-XRF image, $\tilde{\mathbf{Y}}_{\downarrow}$, and HR RGB image, $\tilde{\mathbf{Z}}$. Note that we can also consolidate Equations (1) into a compact form:

$$\tilde{\mathbf{X}} = \tilde{\mathbf{D}}\tilde{\mathbf{A}}, \tag{3}$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix} \tilde{\mathbf{Y}} \\ \tilde{\mathbf{Z}} \end{bmatrix}, \ \tilde{\mathbf{D}} = \begin{bmatrix} \tilde{\mathbf{D}}_{\mathrm{c}}^{\mathrm{xrf}} & \tilde{\mathbf{D}}_{\mathrm{u}}^{\mathrm{xrf}} & \mathbf{0} \\ \tilde{\mathbf{D}}_{\mathrm{c}}^{\mathrm{rgb}} & \mathbf{0} & \tilde{\mathbf{D}}_{\mathrm{u}}^{\mathrm{rgb}} \end{bmatrix}, \ \tilde{\mathbf{A}} = \begin{bmatrix} \tilde{\mathbf{A}}_{\mathrm{c}} \\ \tilde{\mathbf{A}}_{\mathrm{u}}^{\mathrm{xrf}} \\ \tilde{\mathbf{A}}_{\mathrm{u}}^{\mathrm{rgb}} \end{bmatrix}. \tag{4}$$

III. Proposed Method

We propose a model-inspired deep learning framework in which a deep neural network (DNN) learns a representation $\mathbf{A}$ for the MA-XRF and RGB data in a data-driven manner. Traditional sparse representation methods aim to solve:

$$\arg\min_{\mathbf{A}} \|\mathbf{D}\mathbf{A} - \mathbf{X}\|_2^2 + \lambda \|\mathbf{A}\|_1, \quad \text{s.t.} \quad \mathbf{A} \geq 0. \tag{5}$$

where the inequality is applied element-wise, and the norms are entry-wise. Note that although matrices here are denoted without tildes to distinguish the model from the data-driven approach, they share the same dimensions as those in the problem formulation. A common approach to solving such problems is the Iterative Shrinkage-Thresholding Algorithm (ISTA) [22]. Moreover, the Learned ISTA (LISTA) framework [23] unfolds ISTA iterations into a deep network, where each layer learns to approximate an ISTA update step:

$$\mathbf{A}^{(k+1)} = \mathrm{ReLU}\left(\mathbf{A}^{(k)} - \mathbf{W}\mathbf{A}^{(k)} + \mathbf{S}\mathbf{X} - \lambda\right), \tag{6}$$

where $\mathbf{W}$, $\mathbf{S}$, and λ are optimized during training. The matrices $\mathbf{W}$ and $\mathbf{S}$ are not part of Equation (5) but are introduced as learnable parameters to mimic the structure of ISTA. However, we do not aim to directly solve the sparse representation problem in Equation (5). Instead, we are motivated by the LISTA framework and introduce modifications to better align with MA-XRF reconstruction. To meet the sparsity, non-negativity, and boundedness constraints of $\mathbf{A}$ [4], we propose a Sigmoid activation with a bias term, as follows:

$$\mathbf{A}^{(k+1)} = \mathrm{Sigmoid}\left(\mathbf{A}^{(k)} - \mathbf{W}^{(k)}\mathbf{A}^{(k)} + \mathbf{S}^{(k)}\mathbf{X} - \lambda^{(k)}\right), \tag{7}$$

where $\mathbf{W}^{(k)}$, $\mathbf{S}^{(k)}$, and $\lambda^{(k)}$ are learnable parameters for each unfolded iteration k. In our model, $\lambda^{(k)}$ is not a scalar but a vector, where each component $\lambda_i^{(k)}$ corresponds to a specific row of $\mathbf{A}^{(k)}$. Thus, each $\lambda_i^{(k)}$ is applied element-wise to each row i, before the Sigmoid function. Furthermore, note that $\mathbf{Y}$, being the desired final output, is not accessible. Hence, we set $\mathbf{X} = [\mathbf{Y}_{\downarrow\uparrow}, \mathbf{Z}]^T$ as the input of the network, where $\mathbf{Y}_{\downarrow\uparrow}$ is the bilinear upscaled version of $\mathbf{Y}_{\downarrow}$.

Finally, after K unfolded iterations of the network, the MA-XRF image is obtained via a final synthesis layer that leverages Equation (3) as follows:

$$\hat{\mathbf{X}} = \mathrm{Sigmoid}(\mathbf{D}\mathbf{A}^{(K)} - \lambda^{(K)}), \tag{8}$$

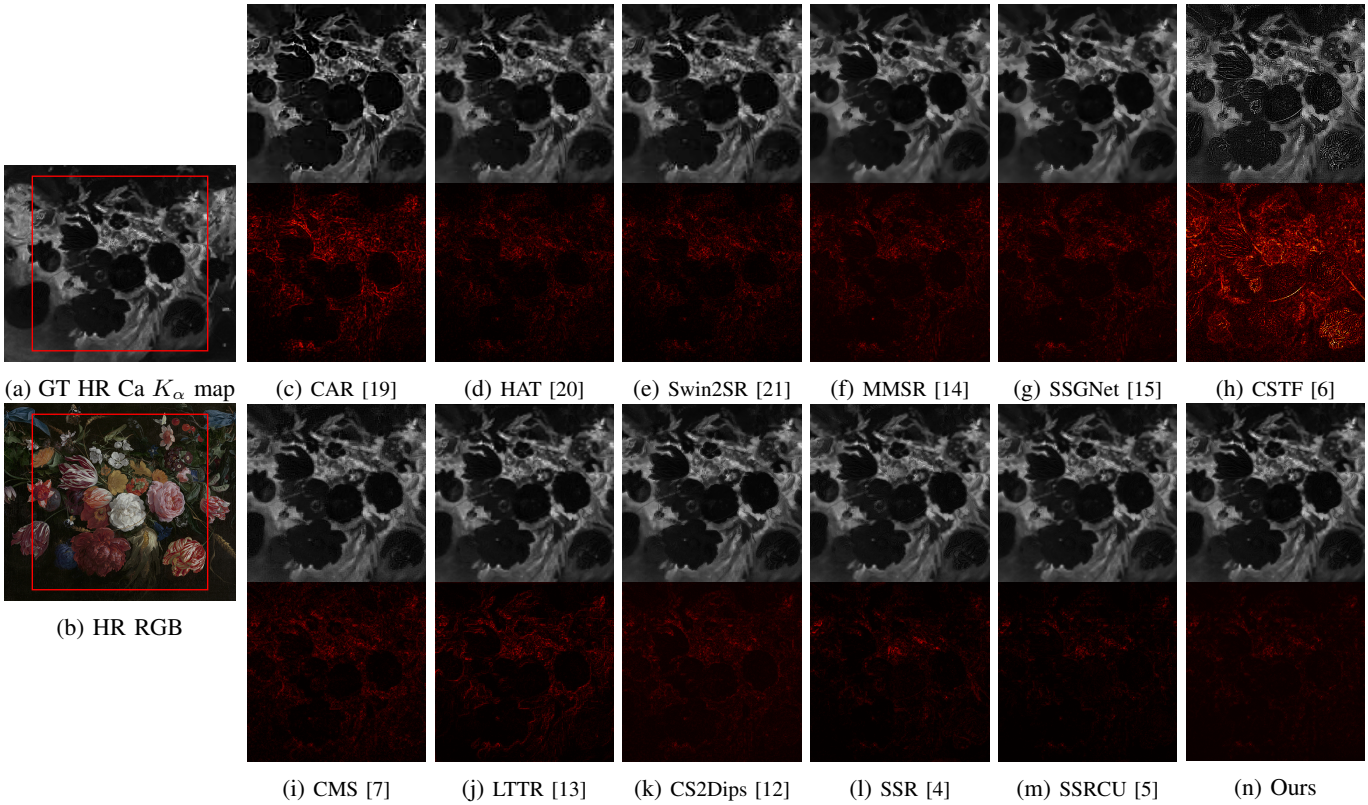
where $\hat{\mathbf{X}}$ is the concatenated reconstruction. To ensure a bounded non-negative output, we include a last non-linearity with a Sigmoid layer and a bias term $\lambda^{(K)}$. To extract the reconstructed HR MA-XRF image $\hat{\mathbf{Y}}$ and HR RGB image $\hat{\mathbf{Z}}$ from $\hat{\mathbf{X}}$, specific channels are selected via slicing:

$$\hat{\mathbf{Y}} = \hat{\mathbf{X}}_{[0:B]}, \quad \hat{\mathbf{Z}} = \hat{\mathbf{X}}_{[B:B+b]}. \tag{9}$$

Here, $\hat{\mathbf{X}}_{[0:B]}$ selects the first B channels, forming HR MA-XRF, while $\hat{\mathbf{X}}_{[B:B+b]}$ extracts the next b channels, corresponding to HR RGB. The end-to-end network $g(\cdot;\theta)$, where θ represents the learnable parameters, is shown in Figure 1 (a). Each matrix multiplication in Equations (7) and (8) is implemented with 1×1 convolutional layers.

TABLE I: Comparative results for 4× upscaling on Old Master paintings. Best results are in bold, second-best are underlined.

Dataset	Metric	Methods for SR problems											
		SISR			GDSR		HSI SR				MA-XRF SR		
		CAR [19]	HAT [20]	Swin2SR [21]	MMSR [14]	SSGNet [15]	CSTF [6]	CMS [7]	LTTR [13]	CS2DIPs [12]	SSR [4]	SSRCU [5]	Ours
Flowers and Insects	RMSE	0.0380	0.0275	0.0275	0.0231	0.0242	0.1336	0.0412	0.0582	0.0187	0.0232	0.0187	**0.0136**
	PSNR	28.42	31.22	31.22	32.72	32.34	17.48	27.70	24.71	34.56	32.69	34.55	**37.31**
The Virgin of the Rocks	RMSE	0.0281	0.0226	0.0229	0.0240	0.0247	0.0771	0.0397	0.0657	0.0211	0.0223	0.0182	**0.0161**
	PSNR	31.01	32.92	32.79	32.41	32.14	22.26	28.03	23.65	33.52	33.03	34.80	**35.84**
Doña Isabel de Porcel	RMSE	0.0388	0.0296	0.0297	0.0252	0.0264	0.0777	0.0373	0.0513	0.0247	0.0264	0.0252	**0.0207**
	PSNR	28.22	30.57	30.54	31.97	31.55	22.19	28.56	25.80	32.16	31.55	31.98	**33.66**



(a) GT HR Ca K_α map (c) CAR [19] (d) HAT [20] (e) Swin2SR [21] (f) MMSR [14] (g) SSGNet [15] (h) CSTF [6]

(b) HR RGB

(i) CMS [7] (j) LTTR [13] (k) CS2Dips [12] (l) SSR [4] (m) SSRCU [5] (n) Ours

Fig. 2: Visual comparison of HR MA-XRF data (Ca K_α element distribution maps), reconstructed using various SR methods for a 4× upscaling. Parts (a) and (b) show the ground truth HR Ca K_α map and the HR RGB image of *Flowers and Insects* ©KMSKA, respectively. Parts (c) to (n) illustrate the reconstructed maps and their corresponding error maps, each compared to the ground truth.

A. Training Strategy

Our method is fully self-supervised and does not require a separate training dataset. Instead, it learns directly from input data, consisting of an HR RGB image, LR MA-XRF image, and the known downsampling matrix. To use this information effectively, we introduce a step that projects the network output, $\hat{\mathbf{Y}}$, onto the convex set of solutions for $\mathbf{Y}$ such that $\mathbf{Y}_\downarrow = \mathbf{YU}$. Specifically, we use the following equation [24]:

$$\text{Proj}(\hat{\mathbf{Y}}) = \hat{\mathbf{Y}} - (\hat{\mathbf{Y}}\mathbf{U} - \mathbf{Y}_\downarrow)\mathbf{U}^T. \tag{10}$$

This approach ensures that the final reconstruction always matches the downsampled version of the given LR MA-XRF image $\mathbf{Y}_\downarrow$. As shown in Figure 1 (a), the output of our reconstruction method is $\text{Proj}(\hat{\mathbf{Y}})$. Furthermore, to guide the training, we employ a loss function that integrates both fidelity to the observed data and regularization, as follows:

$$\mathcal{L}_{c1}(\mathbf{Y}_\downarrow, \hat{\mathbf{Y}}_\downarrow) + \alpha_1 \mathcal{L}_{c2}(\mathbf{Z}, \hat{\mathbf{Z}}) + \alpha_2 \mathcal{L}_{adv}(\text{Proj}(g(\mathbf{X}))), \tag{11}$$

where $\mathbf{Y}_\downarrow$ is the given LR image, $\hat{\mathbf{Y}}_\downarrow$ is the downsampled network reconstruction (without the projection step), $\mathbf{Z}$ is the HR RGB image, and $\hat{\mathbf{Z}}$ is the reconstructed HR RGB image. Here, $\mathbf{X}$ represents the concatenated input, as in Equation (7). The weight for each loss component is controlled by the scalars α_1 and α_2. We adopt Mean Squared Error (MSE) for both $\mathcal{L}_{c1}(\cdot)$ and $\mathcal{L}_{c2}(\cdot)$. The adversarial loss $\mathcal{L}_{adv}(\cdot)$ encourages realistic HR MA-XRF images and is computed via the discriminator $d(\cdot)$ in Figure 1 (b), which processes single-channel MA-XRF patches.

1) Misclassification-Focused Adversarial Training: We extend least squares generative adversarial networks (LSGANs) loss [25] by restricting the training objective to only patches misclassified by the discriminator, focusing updates on the most challenging examples. A discriminator is typically trained to output 1 for real images and -1 for fakes; however, outputs beyond these thresholds are acceptable. For instance, a real sample with an output above 1 or a fake sample below -1

is not problematic. Thus, updates in our approach are driven exclusively by misclassified samples.

Formally, let $d_\phi(y)$ denote the discriminator output for a single-channel MA-XRF patch y (with parameters ϕ), and let τ be a threshold. A patch y is misclassified if

$$(\tau > 0 \text{ and } d_\phi(y) < \tau) \quad \text{or} \quad (\tau < 0 \text{ and } d_\phi(y) > \tau).$$

We set $\tau \in \{+1, -1\}$. Let $\mathbb{P}_r^\tau$ be the distribution of real misclassified patches under threshold τ, and $\mathbb{P}_\theta^\tau$ the corresponding distribution for generated patches. The generator loss is then:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{y\sim\mathbb{P}_\theta^1}\Big[\big(d_\phi(y)-1\big)^2\Big], \tag{12}$$

where correctly classified generated patches ($d_\phi(y) \geq 1$) are discarded. The discriminator is trained analogously:

$$\mathcal{L}_D = \mathbb{E}_{y\sim\mathbb{P}_\theta^{-1}}\Big[\big(d_\phi(y)+1\big)^2\Big] + \mathbb{E}_{y\sim\mathbb{P}_r^1}\Big[\big(d_\phi(y)-1\big)^2\Big], \tag{13}$$

where $\mathbb{P}_\theta^{-1}$ and $\mathbb{P}_r^1$ are the sets of generated and real misclassified patches, respectively. In practice, this is implemented by sampling a batch and discarding correctly classified patches before computing the batch mean.

2) Pseudo-real data augmentation: We exploit spatial similarities between MA-XRF and RGB images to avoid the need for real patches in training. Specifically, we form pseudo-real patches by computing a weighted average between a randomly selected channel of the interpolated MA-XRF image and the most correlated channel of the RGB image:

$$\beta\mathbf{Y}_{\downarrow\uparrow\text{patch}} + (1-\beta)\,\mathbf{Z}_{\text{patch}}, \tag{14}$$

where $\mathbf{Y}_{\downarrow\uparrow\text{patch}}$ is a single-channel patch from the interpolated MA-XRF data and $\mathbf{Z}_{\text{patch}}$ is a patch from the RGB channel that shows the highest correlation with that MA-XRF channel. See Figure 1 (b) and (c).

IV. Experiments and Results

A. Implementation details

To evaluate the effectiveness of our method, we conducted tests on datasets derived from three renowned oil paintings. These include Jan Davidsz. de Heem's *Flowers and Insects* (oil on canvas, Royal Museum of Fine Arts Antwerp, inv. no. 54) [26]; Francisco de Goya's *Doña Isabel de Porcel* (before 1805, oil on canvas, The National Gallery, London, NG1473) [27]; and Leonardo da Vinci's *The Virgin of the Rocks* (circa 1491/2-1508, oil on poplar, thinned and cradled, The National Gallery, London, NG1093) [1]. Our approach was benchmarked against several methods including SSR [4] and SSRCU [5] specifically designed for MA-XRF SR; CSTF [6], CMS [7], LTTR [13], CS2DIPs [12], targeting HSI SR; MMSR [14] and SSGNet [15] for RGB guided depth map SR; and CAR [19], HAT [20], Swin2SR [21] for single image SR (SISR). We used deconvoluted MA-XRF element maps [28] as HR ground truth, with LR versions produced by $4\times$ downsampling.

Next, we detail the configurations used in our experiments. The MA-XRF image has $B = 21$, $B = 7$, and $B = 9$ spectral channels for *Flowers and Insects*, *Doña Isabel de Porcel*, and *The Virgin of the Rocks*, respectively. All RGB images are $512 \times 512 \times 3$ ($b = 3$). The number of channels M in each layer of our DNN is set to 64, and the number of layers K is set to 5. See Figure 1 (a). The initial training phase excludes the adversarial loss component specified in Equation (11). We initialize network weights using Xavier uniform initialization, and then pretrain for 1×10^5 epochs with the Adam optimizer, setting $\alpha_1 = 0.003$ during this phase.

In the adversarial training stage, we maintain the weight parameter α_1 and introduce α_2, set at 0.5×10^{-6}. As shown in Table II, α_1 has minimal impact, while α_2 was empirically set to ensure stable training when coupling with $\mathcal{L}_{\text{adv}}$. The adversarial training uses a patch size and batch size of 32, with learning rates of 3×10^{-4} for the generator $g(\cdot)$ and 3×10^{-6} for the discriminator $d(\cdot)$, spanning 5×10^5 epochs. The scalar β in Equation (14) is randomly adjusted from 0 to 0.9 in each iteration, aiding data augmentation. We also implement random flipping of patches to further diversify the adversarial training. During this phase, the network architecture and data are as described in the previous section. Finally, the discriminator architecture follows the design shown in Figure 1 (b).

TABLE II: Ablation: Loss Terms

$\mathcal{L}_{c1}$	$\mathcal{L}_{c2}$	$\mathcal{L}_{adv}$	PSNR (dB) ↑	RMSE ↓
✓	✓	✓	**37.31**	**0.0136**
✓		✓	**37.31**	**0.0136**
✓			36.97	0.0142
✓	✓		36.97	0.0142
	✓	✓	31.85	0.0256

TABLE III: Ablation: Network Components

Configuration	PSNR (dB) ↑	RMSE ↓
Full Model	**37.31**	**0.0136**
Standard LSGAN Loss	36.98	0.0142
No Sigmoid/Bias in $\mathbf{D}$ (Linear Synthesis Layer)	36.75	0.0145
No Projection Module	36.34	0.0152
Use Leaky ReLU (All Sigmoids Removed)	35.97	0.0159

B. Results

The proposed approach shows superior performance across all evaluated datasets, notably outperforming state-of-the-art MA-XRF-specific techniques such as SSRCU [5]. Among the SISR methods, HAT [20] stands out; however, these techniques still fall short of addressing the specific challenges of MA-XRF SR. GDSR methods, MMSR [14] and SSGNet [15], further improve results but do not achieve the best performance. In the HSI SR group, CS2DIPs [12] secures second-best performance on several datasets, though overall SSRCU remains the second-best since it is designed for MA-XRF SR. Notably, our approach consistently achieves the best results in terms of both RMSE and PSNR, highlighting the importance of employing methods specifically tailored to the unique spectral ranges and physical properties of MA-XRF imaging. Further insights into these performance improvements are provided in Fig. 2, which displays the calcium (Ca K_α) element distribution maps alongside their corresponding error maps. Our method recovers finer details and sharper edges, resulting in higher-quality MA-XRF reconstructions.

C. Ablation Study

In this section, we perform ablation studies on the *Flowers and Insects* dataset. Table II evaluates different combinations of the three loss terms, $\mathcal{L}_{c1}, \mathcal{L}_{c2}, \mathcal{L}_{adv}$, while keeping the scalar weights α_1 and α_2 the same as in Section IV-A. We observe that combining $\mathcal{L}_{c1}$ and $\mathcal{L}_{adv}$ achieves the highest PSNR (37.31 dB) and the lowest RMSE (0.0136), while including $\mathcal{L}_{c2}$ does not alter these metrics for this experiment. This suggests that $\mathcal{L}_{c2}$ could be omitted in scenarios where the learned representation is not required to reconstruct both RGB and MA-XRF images. Nevertheless, removing either $\mathcal{L}_{c1}$ or $\mathcal{L}_{adv}$ degrades performance, showing that these two terms are key to achieving the best reconstruction results.

Table III examines the impact of removing or modifying various components in our approach. Each modification is tested in isolation. Specifically, we tested: (1) replacing our focused adversarial loss with a standard LSGAN objective, (2) removing the sigmoid and bias after $\mathbf{D}$ (making the synthesis layer linear), (3) eliminating the projection module, and (4) substituting all sigmoid activations with leaky ReLU. In each case, we observed a drop in performance (lower PSNR and higher RMSE). The complete configuration delivers the best results, underscoring the importance of each component.

V. Conclusion

This paper introduces the first deep learning approach tailored to MA-XRF SR, specifically designed for Old Master paintings. Our approach requires only a single HR RGB image and LR MA-XRF data for training, assuming a known downsampling model, removing dependency on extensive datasets or pre-trained architectures. Both qualitative assessments of image quality and quantitative evaluations using PSNR and RMSE confirm that our method surpasses state-of-the-art techniques. Its efficiency, outlined by minimal data requirements, enables broader application in Old Master painting analysis, including Macro X-ray Powder Diffraction and Macro Fourier Transform Infrared Scanning in reflection mode.

VI. Acknowledgment

The authors thank the Royal Museum of Fine Arts Antwerp (KMSKA) and Prof. Koen Janssens and Nouchka De Keyser (University of Antwerp) for the *Flowers and Insects* data.

References

[1] M. Spring, et al., "Leonardo's Virgin of the Rocks in the National Gallery, London; New Discoveries from Macro X-ray Fluorescence Scanning and Reflectance Imaging Spectroscopy," *National Gallery Technical Bulletin*, vol. 41, pp. 68–117, 2021.

[2] S. Yan, et al., "Revealing Hidden Drawings in Leonardo's 'the Virgin of the Rocks' from Macro X-Ray Fluorescence Scanning Data through Element Line Localisation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*. IEEE, 2020, pp. 1444–1448.

[3] S. Yan, et al., "When de Prony Met Leonardo: An Automatic Algorithm for Chemical Element Extraction From Macro X-Ray Fluorescence Data," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 908–924, 2021.

[4] Q. Dai, et al., "Spatial-spectral representation for x-ray fluorescence image super-resolution," *IEEE Transactions on Computational Imaging*, vol. 3, no. 3, pp. 432–444, 2017.

[5] S. Yan, et al., "Super-resolution for macro x-ray fluorescence data collected from old master paintings," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[6] S. Li, et al., "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4118–4130, 2018.

[7] L. Zhang, et al., "Exploiting clustering manifold structure for hyperspectral imagery super-resolution," *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 5969–5982, 2018.

[8] H.-F. Yan, et al., "Hyperspectral and multispectral image fusion: When model-driven meet data-driven strategies," *Information Fusion*, vol. 116, pp. 102803, 2025.

[9] C. Wang, et al., "Mswagan: multi-spectral remote sensing image super resolution based on multi-scale window attention transformer," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[10] F. Zhao, et al., "High-frequency feature transfer for multispectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[11] Q. Zhu, et al., "Spectral correlation-based fusion network for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[12] Y. Fang, et al., "Cs2dips: Unsupervised hsi super-resolution using coupled spatial and spectral dips," *IEEE Transactions on Image Processing*, vol. 33, pp. 3090–3101, 2024.

[13] R. Dian, et al., "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2672–2683, 2019.

[14] X. Dong, et al., "Learning mutual modulation for self-supervised cross-modal super-resolution," in *European Conference on Computer Vision*. Springer, 2022, pp. 1–18.

[15] J. Shin, et al., "Task-specific scene structure representations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 2272–2281.

[16] F. Gao, et al., "Multi-modal convolutional dictionary learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 1325–1339, 2022.

[17] J. Bioucas-Dias, et al., "Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based Approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[18] M. Elad and M. Aharon, "Image Denoising Via Sparse and Redundant Representations Over Learned Dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[19] W. Sun and Z. Chen, "Learned image downscaling for upscaling using content adaptive resampler," *IEEE Transactions on Image Processing*, vol. 29, pp. 4027–4040, 2020.

[20] X. Chen, et al., "Activating more pixels in image super-resolution transformer," *arXiv preprint arXiv:2205.04437*, 2022.

[21] M. V. Conde, et al., "Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration," *arXiv preprint arXiv:2209.11345*, 2022.

[22] I. Daubechies, et al., "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 57, no. 11, pp. 1413–1457, 2004.

[23] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th international conference on international conference on machine learning*, Madison, WI, USA, 2010, ICML'10, pp. 399–406, Omnipress.

[24] C. D. Meyer, *Matrix analysis and applied linear algebra*, SIAM, 2023.

[25] X. Mao, et al., "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[26] N. De Keyser, et al., "Jan Davidsz. de Heem (1606–1684): a technical examination of fruit and flower still lifes combining MA-XRF scanning, cross-section analysis and technical historical sources," *Heritage Science*, vol. 5, no. 1, pp. 1–13, 2017.

[27] M. Spring, et al., "Goya's Portraits in the National Gallery: their Technique, Materials and Development," *National Gallery Technical Bulletin*, vol. 37, pp. 78–104, 2016.

[28] S. Yan, et al., "A fast automatic method for deconvoluting macro x-ray fluorescence data collected from easel paintings," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 649–664, 2023.