Self-Attention Driven Tensor Representation for High-Order Data Recovery

Supplementary Material

1. Proof of Theorem 1

Let $\mathcal{G} = \mathcal{H}_\Phi(\mathcal{Z})$ be the core tensor parameterized by MLPs with parameters Φ, where $\mathcal{Z} \sim \mathcal{N}(0,1)$ sampled from the continuous Gaussian distribution. Assume that the weight matrices of MLPs satisfy $\|\mathbf{W}_i\|_1 \leq \gamma$ for $i = 1, 2, \cdots, l$, and the activation function $\sigma(\cdot)$ is δ-Lipschitz continuous. Since $\sigma(\cdot)$ is Lipschitz continuous, thus, for any a, b, we have

$$\|\sigma(a) - \sigma(b)\|_{\ell_1} \leq \delta \|a - b\|_1. \tag{1}$$

Then, for any $a, b \in \mathcal{Z}$, the following inequality holds:

$$\begin{aligned}
&\|\mathcal{H}_\Phi(a) - \mathcal{H}_\Phi(b)\|_1 \\
&= \|\mathbf{W}_l(\delta(\mathbf{W}_{l-1} \cdots \delta(\mathbf{W}_1 a))) - \mathbf{W}_l(\delta(\mathbf{W}_{l-1} \cdots \delta(\mathbf{W}_1 b)))\|_1 \\
&= \|\mathbf{W}_l(\delta(\mathbf{W}_{l-1} \cdots \delta(\mathbf{W}_1 a)) - \delta(\mathbf{W}_{l-1} \cdots \delta(\mathbf{W}_1 b)))\|_1 \\
&\leq \|\mathbf{W}_l\|_1 \|\delta(\mathbf{W}_{l-1} \cdots \delta(\mathbf{W}_1 a)) - \delta(\mathbf{W}_{l-1} \cdots \delta(\mathbf{W}_1 b))\|_1 \\
&\leq \gamma \|\delta(\mathbf{W}_{l-1} \cdots \delta(\mathbf{W}_1 a)) - \sigma(\mathbf{W}_{l-1} \cdots \delta(\mathbf{W}_1 b))\|_1 \\
&\leq \gamma\delta \|\mathbf{W}_{l-1} \cdots \delta(\mathbf{W}_1 a) - \mathbf{W}_{l-1} \cdots \delta(\mathbf{W}_1 b)\|_1 \\
&\quad \cdots \\
&\leq \gamma^l \delta^{l-1} \|a - b\|_1.
\end{aligned} \tag{2}$$

Therefore, we have

$$\|\mathcal{H}_\Phi(a) - \mathcal{H}_\Phi(b)\|_1 \leq \gamma^l \delta^{l-1} \|a - b\|_1. \tag{3}$$

Since $\mathcal{Z} \sim \mathcal{N}(0, I)$ follows a Gaussian distribution, $\|a - b\|_1$ is likely to be close to a constant in high-dimensional space, and thus the output difference of $\mathcal{G}$ is compressed to a finite upper bound. This means that the elements of $\mathcal{G}$ statistically tend to be close to zero, forming an implicit sparse distribution. This completes the proof.

2. Proof of Lemma 1

We first show the following lemma:

Lemma 3 ([2]). *Let $\boldsymbol{Z} = \{Z \in \mathbb{R}^D, \|Z\|_2 \leq q\}$ be the set of the input, and $\mathcal{D}_\theta$ be a deep generative model with network parameter $\theta = \{W_1, W_2, \cdots, W_L\}$, where $\prod_{i=1}^L \|W_i\|_2 \leq p$. Also let $\boldsymbol{X} = \{\mathcal{D}_\theta(Z) \mid \forall Z \in \boldsymbol{Z} \subseteq \mathbb{R}^D\}$. Then, for the set $\boldsymbol{X}$, the following upper bound holds:*

$$\mathsf{N}(\boldsymbol{X}, \varepsilon) \leq \left(\frac{3pq}{\varepsilon}\right)^D, \tag{4}$$

where $\mathsf{N}(\cdot, \cdot)$ is the covering number and $0 < \varepsilon < 1$.

Consider the $\boldsymbol{\mathcal{G}} = \{\mathcal{G} = \mathcal{H}_\Phi(\mathcal{Z}) \mid \forall\, \mathcal{Z} \in \boldsymbol{\mathcal{Z}} \subseteq \mathbb{R}^{r_1 \times r_2 \times r_3}\}$ to be a set of core tensors, where $\mathcal{H}_\Phi(\cdot)$ is an L-layers MLPs network with $\prod_{i=1}^L T_i \|\mathbf{W}_i\|_2 \leq \gamma$ and T_i is the Lipschitz constant of the i-th activation function. Since $\boldsymbol{\mathcal{Z}} = \{\mathcal{Z} \in \mathbb{R}^{r_1 \times r_2 \times r_3}, \|\mathcal{Z}\|_F \leq \alpha\}$ is a Euclidean ball that has a radius of α in the R-dimensional space. Then, according to Lemma 3, the covering number of $\boldsymbol{\mathcal{G}}$ is upper bounded by

$$\mathsf{N}(\boldsymbol{\mathcal{G}}, \varepsilon) \leq \left(\frac{3\alpha\gamma}{\varepsilon}\right)^R, \tag{5}$$

where $R = \prod_{n=1}^3 r_n$. Let $\lambda = r^{\frac{3}{2}}\beta^3\eta^3 + 3\alpha\gamma r\beta^2\eta^2$, where $r = \max\{r_n\}_{n=1}^3$. The above also implies the following upper bounds:

$$\mathsf{N}(\boldsymbol{\mathcal{G}}, \frac{\varepsilon}{\lambda}) \leq \left(\frac{3\lambda\alpha\gamma}{\varepsilon}\right)^R. \tag{6}$$

Similarly, consider the set $\boldsymbol{a} = \{a = \mathrm{FSR}_\Theta(e) \mid \forall\ e \in \boldsymbol{e} \subseteq \mathbb{R}^d\}$, where $\mathrm{FSR}_\Theta(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ is the factor self-representation with parameters $\Psi = \{\mathbf{W}^q, \mathbf{W}^k, \mathbf{W}^v, \|\mathbf{W}^q\|_2 \|\mathbf{W}^k\|_2 \|\mathbf{W}^v\|_2 \leq \eta\}$. Thus, the covering number of the set of $\boldsymbol{a}$ is upper bounded by

$$\mathsf{N}(\boldsymbol{a}, \varepsilon) \leq \left(\frac{3\beta\eta}{\varepsilon}\right)^d, \tag{7}$$

where $\boldsymbol{e} = \{e \in \mathbb{R}^d, \|e\|_2 \leq \beta\}$ is a Euclidean ball that has a radius of β in the d-dimensional space. The above also implies the following upper bounds:

$$\mathsf{N}(\boldsymbol{a}, \frac{\varepsilon}{\sqrt{r_n}}) \leq \left(\frac{3\sqrt{r_n}\beta\eta}{\varepsilon}\right)^d, \tag{8}$$

where $n = 1, 2, 3$.

Now, we consider $\boldsymbol{A}_n = \{A_n \in \mathbb{R}^{I_n \times r_n}, A_n = [a_1 \dots a_{r_n}]^\top, a_i = \mathrm{FSR}_\Theta(z_i), e_i \in \boldsymbol{e}, i \in [1, 2, \cdots, r_n]\}$ to be a set of the n-th factor matrix. Let $\bar{\boldsymbol{a}}$ be a $\frac{\varepsilon}{\sqrt{r_n}}$-net of $\boldsymbol{a}$. For any $A_n \in \boldsymbol{A}_n$, we can construct $\bar{\boldsymbol{A}}_n = \{\bar{A}_n \in \mathbb{R}^{I_n \times r_n}, \bar{A}_n = [\bar{a}_1 \dots \bar{a}_{r_n}]^\top, \bar{a}_i \in \bar{\boldsymbol{a}}, i \in [1, 2, \cdots, r_n]\}$. Then, for any $a_i \in \boldsymbol{a}$ and $\bar{a}_i \in \bar{\boldsymbol{a}}$ such that $\|a_i - \bar{a}_i\|_2 \leq \frac{\varepsilon}{\sqrt{r_n}}$ for $i = 1, 2, \cdots, r_n$. Thus, we have

$$\begin{aligned}
\|A_n - \bar{A}_n\|_2 &= \left(\sum_{i=1}^{r_n} \|a_i - \bar{a}_i\|_2^2\right)^{1/2} \\
&\leq \left(r_n \left(\frac{\varepsilon}{\sqrt{r_n}}\right)^2\right)^{1/2} \\
&= \varepsilon.
\end{aligned} \tag{9}$$

Hence, the set $\bar{\boldsymbol{A}}_n$ is an ε-net of $\boldsymbol{A}_n$. Note that the cardinality of such set is upper bounded by $(\mathsf{N}(\boldsymbol{a}, \frac{\varepsilon}{\sqrt{r_n}}))^{r_n}$.

Thus, the covering number of the set of the factor matrices $\boldsymbol{A}_n$ is upper bounded by

$$\mathsf{N}(\boldsymbol{A}_n, \varepsilon) \leq \left(\frac{3\sqrt{r_n}\beta\eta}{\varepsilon}\right)^{I_n r_n}, \tag{10}$$

where $n = 1, 2, 3$. In practical applications, $I_n \gg r_n$. Let $r = \max\{r_n\}_{n=1}^3$, the following upper bound also holds:

$$\mathsf{N}(\boldsymbol{A}_n, \varepsilon) \leq \left(\frac{3\sqrt{r}\beta\eta}{\varepsilon}\right)^{I_n r_n}. \tag{11}$$

Let $\lambda = r^{\frac{3}{2}}\beta^3\eta^3 + 3\alpha\gamma r\beta^2\eta^2$, the above also implies the following upper bounds:

$$\mathsf{N}(\boldsymbol{A}_n, \frac{\varepsilon}{\lambda}) \leq \left(\frac{3\sqrt{r}\lambda\beta\eta}{\varepsilon}\right)^{I_n r_n}. \tag{12}$$

Let $\bar{\boldsymbol{\mathcal{G}}}$ be a $\frac{\varepsilon}{\lambda}$-net of $\boldsymbol{\mathcal{G}}$ and $\bar{\boldsymbol{A}}_n$ be a $\frac{\varepsilon}{\lambda}$-net of $\boldsymbol{A}_n$ with $n = 1, 2, 3$. Now, we consider $\bar{\boldsymbol{\mathcal{X}}}_{\text{SR}} = \{\bar{\mathcal{X}} = \bar{\mathcal{G}} \times_1 \bar{A}_1 \times_2 \bar{A}_2 \times_3 \bar{A}_3 \mid \bar{\mathcal{G}} \in \bar{\boldsymbol{\mathcal{G}}}, \{\bar{A}_n \in \bar{\boldsymbol{A}}_n\}_{n=1}^3\}$ to be the set of the reconstructed tensors. Thus, the cardinality of $\bar{\boldsymbol{\mathcal{X}}}_{\text{SR}}$ is upper bounded by

$$\begin{aligned}|\bar{\boldsymbol{\mathcal{X}}}_{\text{SR}}| \leq & |\bar{\boldsymbol{\mathcal{G}}}| \cdot |\bar{\boldsymbol{A}}_1| \cdot |\bar{\boldsymbol{A}}_2| \cdot |\bar{\boldsymbol{A}}_3| \\ \leq & \left(\frac{3\lambda\alpha\gamma}{\varepsilon}\right)^R \prod_{n=1}^3 \left(\frac{3\sqrt{r}\lambda\beta\eta}{\varepsilon}\right)^{I_n r_n},\end{aligned} \tag{13}$$

where $\lambda = r^{\frac{3}{2}}\beta^3\eta^3 + 3\alpha\gamma r\beta^2\eta^2$.

Next, we will show that $\bar{\boldsymbol{\mathcal{X}}}_{\text{SR}}$ is an ε-net of $\boldsymbol{\mathcal{X}}_{\text{SR}}$, where $\boldsymbol{\mathcal{X}}_{\text{SR}} = \{\mathcal{X} = \mathcal{G} \times_1 A_1 \times_2 A_2 \times_3 A_3 \mid \mathcal{G} \in \boldsymbol{\mathcal{G}}, \{A_n \in \boldsymbol{A}_n\}_{n=1}^3\}$. For any $\mathcal{X} \in \boldsymbol{\mathcal{X}}_{\text{SR}}$ and $\bar{\mathcal{X}} \in \bar{\boldsymbol{\mathcal{X}}}_{\text{SR}}$, there exist $\mathcal{G} \in \boldsymbol{\mathcal{G}}$ and $\bar{\mathcal{G}} \in \bar{\boldsymbol{\mathcal{G}}}$, and $A_n \in \boldsymbol{A}_n$ and $\bar{A}_n \in \bar{\boldsymbol{A}}_n$, such that $\|\mathcal{G} - \bar{\mathcal{G}}\|_F \leq \frac{\varepsilon}{\lambda}$ and $\|A_n - \bar{A}_n\|_F \leq \frac{\varepsilon}{\lambda}$ for $n = 1, 2, 3$. Then, we have

$$\begin{aligned}&\|\mathcal{X} - \bar{\mathcal{X}}\|_F \\ =&\|\mathcal{G} \times_1 A_1 \times_2 A_2 \times_3 A_3 - \bar{\mathcal{G}} \times_1 \bar{A}_1 \times_2 \bar{A}_2 \times_3 \bar{A}_3\|_F \\ =&\|\mathcal{G} \times_1 A_1 \times_2 A_2 \times_3 A_3 - \mathcal{G} \times_1 \bar{A}_1 \times_2 A_2 \times_3 A_3 + \\ &\mathcal{G} \times_1 \bar{A}_1 \times_2 A_2 \times_3 A_3 - \mathcal{G} \times_1 \bar{A}_1 \times_2 \bar{A}_2 \times_3 A_3 + \\ &\mathcal{G} \times_1 \bar{A}_1 \times_2 \bar{A}_2 \times_3 A_3 - \mathcal{G} \times_1 \bar{A}_1 \times_2 \bar{A}_2 \times_3 \bar{A}_3 + \\ &\mathcal{G} \times_1 \bar{A}_1 \times_2 \bar{A}_2 \times_3 \bar{A}_3 - \bar{\mathcal{G}} \times_1 \bar{A}_1 \times_2 \bar{A}_2 \times_3 \bar{A}_3\|_F \\ \leq&\|\mathcal{G} \times A_2 \times A_3 \times (A_1 - \bar{A}_1)\|_F + \\ &\|\mathcal{G} \times \bar{A}_1 \times A_3 \times (A_2 - \bar{A}_2)\|_F + \\ &\|\mathcal{G} \times \bar{A}_1 \times \bar{A}_2 \times (A_3 - \bar{A}_3)\|_F + \\ &\|(\mathcal{G} - \bar{\mathcal{G}}) \times \bar{A}_1 \times \bar{A}_2 \times \bar{A}_3\|_F \\ \leq&3\alpha\gamma r\beta^2\eta^2 \cdot \frac{\varepsilon}{\lambda} + \frac{\varepsilon}{\lambda} r^{\frac{3}{2}}\beta^3\eta^3 \\ =&\varepsilon.\end{aligned} \tag{14}$$

Thus, $\bar{\boldsymbol{\mathcal{X}}}_{\text{SR}}$ is an ε-net of $\boldsymbol{\mathcal{X}}_{\text{SR}}$, and its covering number is upper bounded by

$$\mathsf{N}(\boldsymbol{\mathcal{X}}_{\text{SR}}, \varepsilon) \leq \left(\frac{3\lambda\alpha\gamma}{\varepsilon}\right)^R \prod_{n=1}^3 \left(\frac{3\sqrt{r}\lambda\beta\eta}{\varepsilon}\right)^{I_n r_n}. \tag{15}$$

Let $a = \alpha\gamma$ and $b = \sqrt{r}\beta\eta$, we have

$$\mathsf{N}(\boldsymbol{\mathcal{X}}_{\text{SR}}, \varepsilon) \leq \left(\frac{3\lambda a}{\varepsilon}\right)^R \prod_{n=1}^3 \left(\frac{3\lambda b}{\varepsilon}\right)^{I_n r_n}, \tag{16}$$

This completes the proof.

3. Proof of Lemma 2

To prove the lemma 2, we exploit Sterfling's sampling-without-replacement extension of Hoeffding's inequality [1], a common proof technique in traditional matrix completion [3]. We apply it to a tensor-based completion model with deep generative model constraints.

Lemma 4 ([1]). ***Lemma 4.*** *Let $X_1, X_2, \ldots, X_w$ be a set of samples taken without replacement from $\{x_1, x_2, \ldots, x_n\}$ of mean μ. Denote $a = \min_i x_i$ and $b = \max_i x_i$. Then*

$$\begin{aligned}&\mathsf{Pr}\left[|\frac{1}{w}\sum_{i=1}^w X_i - \mu| \geq t\right] \\ &\leq 2\exp\left(-\frac{2wt^2}{(1-(w-1)/n)(b-a)^2}\right).\end{aligned} \tag{17}$$

Consider a set of variables as follows:

$$\begin{aligned}&\mathcal{D}(i,j,k) = \\ &\left\{\left(\mathcal{Y}(i,j,k) - \mathcal{X}(i,j,k)\right)^2, \forall (i,j,k) \in \mathbb{R}^{I_1 \times I_2 \times I_3}\right\}.\end{aligned} \tag{18}$$

Then, when $(i,j,k) \in \Omega$, the sample mean of $\mathcal{D}(i,j,k)$ is the empirical loss, i.e., $\text{loss}_1(\mathcal{X})$, and when $(i,j,k) \in \mathbb{R}$, the actual mean of $\mathcal{D}(i,j,k)$ is the actual loss, i.e., $\text{loss}_2(\mathcal{X})$. It can be seen $\mathcal{D}(i,j,k) \leq \left(|\mathcal{Y}(i,j,k)| + |\mathcal{X}(i,j,k)|\right)^2 \leq (\nu + v + \alpha\gamma r^{\frac{3}{2}}\beta^3\eta^3) = \xi$, where $\nu = \max_{(i,j,k)} |\mathcal{X}_{\circledast}(i,j,k)|$ and $v = \max_{(i,j,k)} |\mathcal{N}(i,j,k)|$. Obviously, $a = \min_{(i,j,k)\in\Omega}$ with $\mathcal{D}(i,j,k) \geq 0$ and $b = \max_{(i,j,k)\in\Omega}$ with $\mathcal{D}(i,j,k) \leq \xi$. Thus, according to Lemma 4, we have

$$\begin{aligned}&\mathsf{Pr}\left[|\text{loss}_1(\mathcal{X}) - \text{loss}_2(\mathcal{X})| \geq t\right] \\ &\leq 2\exp\left(-\frac{2|\Omega|t^2}{(1-(|\Omega|-1)/N)(\xi-0)^2}\right),\end{aligned} \tag{19}$$

where $N = \prod_{n=1}^3 I_n$ is the total number of elements in the tensor. Let $\bar{\boldsymbol{\mathcal{X}}}_{SR}$ is an ε-net of $\boldsymbol{\mathcal{X}}_{SR}$, according the union bound on all $\bar{\mathcal{X}} \in \bar{\boldsymbol{\mathcal{X}}}_{SR}$, we have

$$\begin{aligned}&\mathsf{Pr}\left[\sup_{\bar{\mathcal{X}} \in \bar{\boldsymbol{\mathcal{X}}}_{SR}} |\text{loss}_1(\bar{\mathcal{X}}) - \text{loss}_2(\bar{\mathcal{X}})| \geq t\right] \\ &\leq 2|\bar{\boldsymbol{\mathcal{X}}}_{SR}| \exp\left(-\frac{2|\Omega|t^2}{(1-(|\Omega|-1)/N)\xi^2}\right).\end{aligned} \tag{20}$$

Equivalently, with probability at least $1-\delta$, the following holds:

$$\begin{aligned}&\sup_{\bar{\mathcal{X}}\in\bar{\boldsymbol{\mathcal{X}}}_{SR}}|\text{loss}_1(\bar{\mathcal{X}})-\text{loss}_2(\bar{\mathcal{X}})|\\&\leq\sqrt{\frac{(N-|\Omega|+1)\xi^2}{2N|\Omega|}\log\left(\frac{2|\bar{\boldsymbol{\mathcal{X}}}_{SR}|}{\delta}\right)}.\end{aligned}\tag{21}$$

Let $g(\Omega)\triangleq\sup_{\bar{\mathcal{X}}\in\bar{\boldsymbol{\mathcal{X}}}_{SR}}|\text{loss}_1(\bar{\mathcal{X}})-\text{loss}_2(\bar{\mathcal{X}})|$, we have $\sqrt{g(\Omega)}\geq\sup_{\bar{\mathcal{X}}\in\bar{\boldsymbol{\mathcal{X}}}_{SR}}|\sqrt{\text{loss}_1(\bar{\mathcal{X}})}-\sqrt{\text{loss}_2(\bar{\mathcal{X}})}|$, since for any non-negative a and b, $\sqrt{|a-b|}\geq|\sqrt{a}-\sqrt{b}|$ holds. Also, note that

$$\begin{aligned}|\sqrt{\text{loss}_2(\mathcal{X})}-\sqrt{\text{loss}_2(\bar{\mathcal{X}})}|&=\frac{1}{\sqrt{N}}\left|\|\mathcal{Y}-\mathcal{X}\|_F-\|\mathcal{Y}-\bar{\mathcal{X}}\|_F\right|\\&\leq\frac{1}{\sqrt{N}}\|\mathcal{Y}-\mathcal{X}-\mathcal{Y}+\bar{\mathcal{X}}\|_F\\&\leq\frac{\varepsilon}{\sqrt{N}}.\end{aligned}\tag{22}$$

Similarly, $|\sqrt{\text{loss}_1(\mathcal{X})}-\sqrt{\text{loss}_1(\bar{\mathcal{X}})}\leq\frac{\varepsilon}{\sqrt{|\Omega|}}|$ also holds. Hence, the following also holds:

$$\sup_{\mathcal{X}\in\boldsymbol{\mathcal{X}}_{SR}}|\sqrt{\text{loss}_1(\mathcal{X})}-\sqrt{\text{loss}_2(\mathcal{X})}|\leq\sqrt{g(\Omega)}+\frac{\varepsilon}{\sqrt{N}}+\frac{\varepsilon}{\sqrt{|\Omega|}}.\tag{23}$$

Therefore, with probability at least $1-\delta$, the following holds:

$$\begin{aligned}&\sup_{\mathcal{X}\in\boldsymbol{\mathcal{X}}_{SR}}|\sqrt{\text{loss}_1(\mathcal{X})}-\sqrt{\text{loss}_2(\mathcal{X})}|\\&\leq\left(\xi^2\log\left(\frac{2|\bar{\boldsymbol{\mathcal{X}}}_{SR}|}{\delta}\right)\kappa\right)^{\frac{1}{4}}+\frac{2\varepsilon}{\sqrt{|\Omega|}},\end{aligned}\tag{24}$$

where $\kappa=\frac{N-|\Omega|+1}{2N|\Omega|}$ and $|\bar{\boldsymbol{\mathcal{X}}}_{SR}|=\mathsf{N}(\boldsymbol{\mathcal{X}}_{SR},\varepsilon)$. This completes the proof.

4. Proof of Theorem 2

Let $\hat{\mathcal{X}}^*$ be the optimal solution, and its corresponding empirical loss $\text{loss}_1(\hat{\mathcal{X}}^*)$ and actual loss $\text{loss}_2(\hat{\mathcal{X}}^*)$ are as follows:

$$\begin{aligned}\sqrt{\text{loss}_1(\hat{\mathcal{X}}^*)}&=\frac{1}{\sqrt{|\Omega|}}\left\|\mathcal{M}\odot(\mathcal{Y}-\hat{\mathcal{X}}^*)\right\|_F,\\\sqrt{\text{loss}_2(\hat{\mathcal{X}}^*)}&=\frac{1}{\sqrt{N}}\left\|(\mathcal{Y}-\hat{\mathcal{X}}^*)\right\|_F.\end{aligned}\tag{25}$$

Then, we have

$$\begin{aligned}&\frac{1}{\sqrt{N}}\|\hat{\mathcal{X}}^*-\mathcal{X}_{\circledast}\|_F=\frac{1}{\sqrt{N}}\|\hat{\mathcal{X}}^*-\mathcal{Y}+\mathcal{N}\|_F\\&\leq\frac{1}{\sqrt{N}}\|\hat{\mathcal{X}}^*-\mathcal{Y}\|_F+\frac{1}{\sqrt{N}}\|\mathcal{N}\|_F.\end{aligned}\tag{26}$$

Since $\mathsf{Gap}^*(\Omega)\geq|\sqrt{\text{loss}_1(\hat{\mathcal{X}}^*)}-\sqrt{\text{loss}_2(\hat{\mathcal{X}}^*)}|$, we have

$$\begin{aligned}&\frac{1}{\sqrt{N}}\|\hat{\mathcal{X}}^*-\mathcal{X}_{\circledast}\|_F\\&\leq\frac{1}{\sqrt{|\Omega|}}\|\mathcal{M}\odot(\mathcal{Y}-\hat{\mathcal{X}}^*)\|_F+\mathsf{Gap}^*(\Omega)+\frac{1}{\sqrt{N}}\|\mathcal{N}\|_F.\end{aligned}\tag{27}$$

Considering that $\hat{\mathcal{X}}^*$ is the global optimal solution, inequality $\|\mathcal{M}\odot(\mathcal{Y}-\hat{\mathcal{X}}^*)\|_F\leq\|\mathcal{M}\odot(\mathcal{Y}-\mathcal{X})\|_F$ holds for any solution $\mathcal{X}\in\boldsymbol{\mathcal{X}}_{SR}$ including $\mathcal{X}^*$. Thus, we have

$$\begin{aligned}&\frac{1}{\sqrt{N}}\|\hat{\mathcal{X}}^*-\mathcal{X}_{\circledast}\|_F\\&\leq\frac{1}{\sqrt{|\Omega|}}\|\mathcal{M}\odot(\mathcal{Y}-\mathcal{X}^*)\|_F+\mathsf{Gap}^*(\Omega)+\frac{1}{\sqrt{N}}\|\mathcal{N}\|_F\\&\leq\frac{\|\mathcal{M}\odot(\mathcal{X}^*-\mathcal{X}_{\circledast})\|_F+\|\mathcal{M}\odot(\mathcal{X}_{\circledast}-\mathcal{Y})\|_F}{\sqrt{|\Omega|}}\\&+\mathsf{Gap}_*(\Omega)+\frac{1}{\sqrt{N}}\|\mathcal{N}\|_F\\&\leq\frac{1}{\sqrt{|\Omega|}}\|\mathcal{X}^*-\mathcal{X}_{\circledast}\|_F+\frac{1}{\sqrt{|\Omega|}}\|\mathcal{M}*\mathcal{N}\|_F\\&+\mathsf{Gap}^*(\Omega)+\frac{1}{\sqrt{N}}\|\mathcal{N}\|_F.\end{aligned}\tag{28}$$

This completes the proof.

5. More Experimental Results

In this section, we present more visual performance on various high-order data recovery tasks to further validate the effectiveness of the proposed SADTR.

Figure 1 shows the visual performance of high-order data inpainting by different methods. We observe that the proposed SADTR consistently outperforms other methods both numerically and visually, owing to its ability to establish dynamic global connections that jointly model local and non-local nonlinear dependencies, leading to a more accurate low-rank structure. Figure 2 presents the visual performance of multispectral image denoising by different methods. We observe that the proposed SADTR achieves superior numerical and visual performance, reaffirming its effectiveness in capturing accurate low-rank structures. The compact global representation and implicit sparsity constraint jointly highlight key components and suppress noise.

6. More Discussions

6.1. Ablation Study

This subsection further discusses the impact of the key components in the proposed SADTR. Specifically, we test the following variant models, including SADTR-V0 (i.e., the

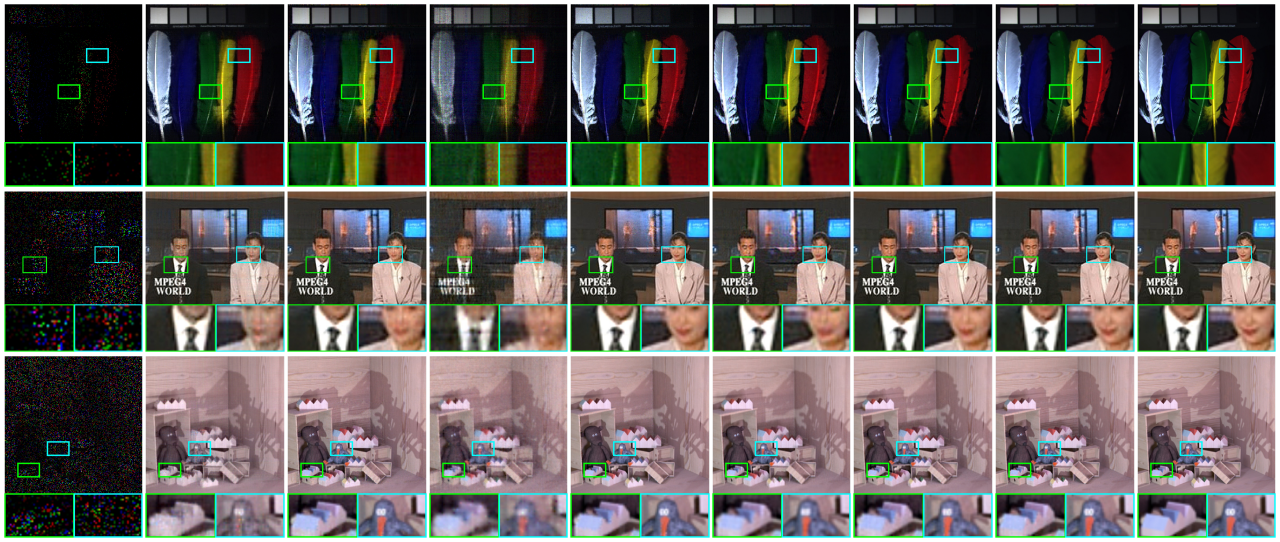

Figure 1. Visual performance of high-order data inpainting by different methods for MSI *feathers* (MR=95%), Color Videos *News* (MR=90%), and Light Field *Dino* (MR=90%).

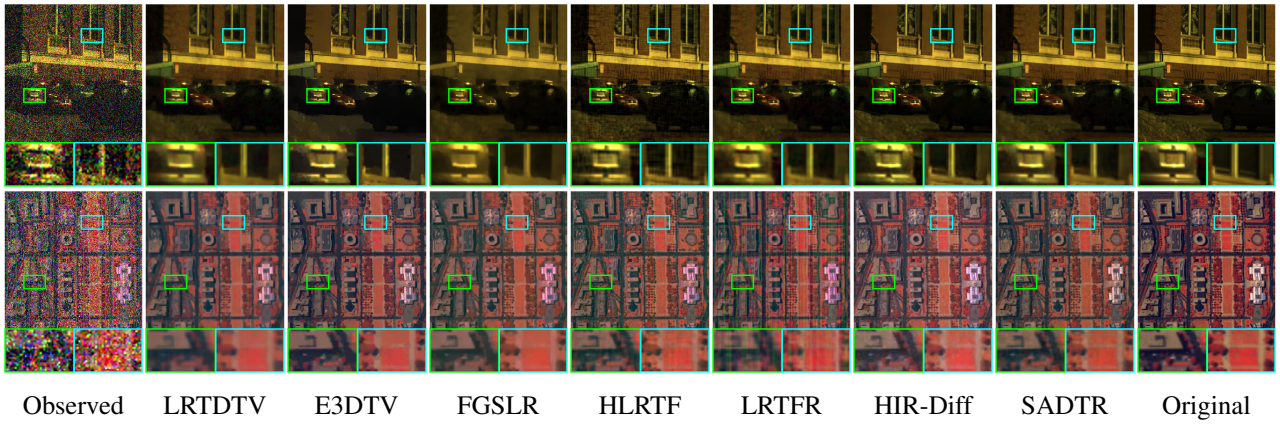

Figure 2. Visual performance of multispectral image denoising by different methods for MSI *imgb2* (Case 1) and HSI *WDC mall* (Case 3).

proposed SADTR), SADTR-V1 without FSR for all factors, SADTR-V2 without FSR for factors 2 and 3, SADTR-V3 without FSR for factors 3, and SADTR-V4 without implicit sparse representation. To test different variants of SADTR, we conduct experiments on high-order data inpainting on MSI *Toys* with MR=95% and multispectral image denoising on MSI *Beers* with Case 1.

Figure 3 illustrates the visual performance of different variants. We observe that introducing FSR and implicit sparse representation significantly improves the visual performance, demonstrating the effectiveness of these components.

6.2. Analysis of Rank

In this section, we discuss the influence of rank in the proposed SADTR model. Specifically, we take the third-order data inpainting as an example to discuss the impact of r_1, r_2, and r_3 on model performance. We conduct experiments on high-order data inpainting on MSI *Toys*, with a size of $256 \times 256 \times 31$. The candidate sets for r_1 and r_2 are $\{2, 4, 8, 16, 32, 64, 128, 256\}$, and the candidate set for r_3 is $\{2, 6, 10, 14, 18, 22, 26, 30\}$. We fix other ranks when discussing the effects of different ranks.

Figure 4 illustrates the impact of different rank values on the SADTR model's performance in high-order data inpainting tasks. We observe that the performance of SADTR gradually improves as the values of r_1, r_2, and r_3 increase. However, the values of r_1 and r_2 are more sensitive than r_3.

6.3. Analysis of Convergence

In this section, we discuss the convergence behavior of the proposed SADTR model. Here, we select data inpainting

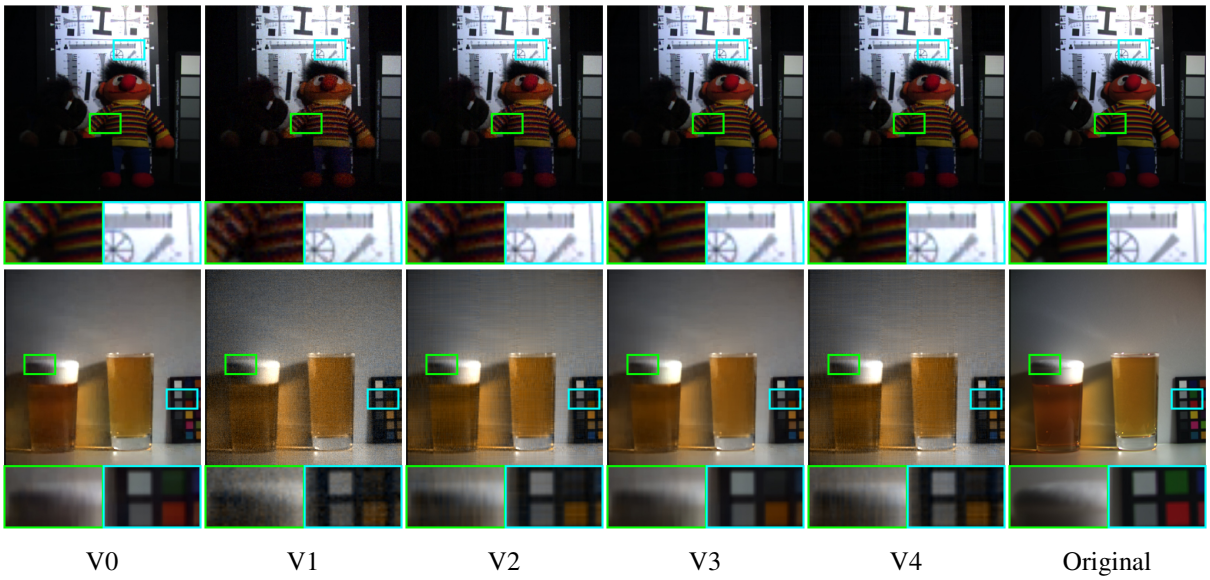

Figure 3. Different variants of SADTR exhibit different visual performance in high-order data inpainting.

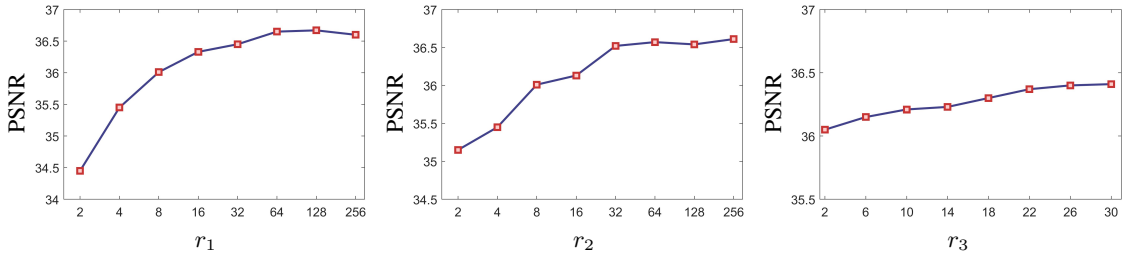

Figure 4. The impact of different rank values on the SADTR model in high-order data inpainting tasks.

task and denoising task as examples to numerically verify the convergence. Figure 5 presents the relative error in the reconstructed data at each iteration compared to their respective previous iterations. We observe that the values of the relative error achieved by the proposed method decrease and gradually tend to zero as the number of iterations increases. This justifies the numerical convergence of the proposed method.

References

[1] R. J. Serfling. Probability Inequalities for the Sum in Sampling without Replacement. *The Annals of Statistics*, 2(1):39 – 48, 1974. 2

[2] Sagar Shrestha, Xiao Fu, and Mingyi Hong. Deep spectrum cartography: Completing radio map tensors using learned neural models. *IEEE Transactions on Signal Processing*, 70: 1170–1184, 2022. 1

[3] Yu-Xiang Wang and Huan Xu. Stability of matrix factorization for collaborative filtering. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, page 163–170, 2012. 2

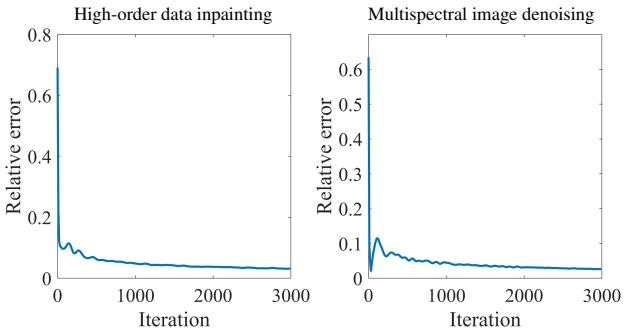

Figure 5. Relative error curves with respect to the iteration number on different tasks. Here the relative error is defined as $\|\mathcal{X} - \hat{\mathcal{X}}\|_F / \|\hat{\mathcal{X}}\|_F$, and $\mathcal{X}$ and $\hat{\mathcal{X}}$ are the results of the current iteration and its previous iteration.