Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 687074, 20 pages
http://dx.doi.org/10.1155/2015/687074

Research Article

Learning-Based Video Superresolution Reconstruction Using Spatiotemporal Nonlocal Similarity

Meiyu Liang, Junping Du, and Linghui Li

Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

Correspondence should be addressed to Junping Du; junpingdu@126.com

Received 25 May 2015; Accepted 2 August 2015

Academic Editor: William Guo

Aiming at improving the video visual resolution quality and details clarity, a novel learning-based video superresolution reconstruction algorithm using spatiotemporal nonlocal similarity is proposed in this paper. Objective high-resolution (HR) estimations of low-resolution (LR) video frames can be obtained by learning LR-HR correlation mapping and fusing spatiotemporal nonlocal similarities between video frames. With the objective of improving algorithm efficiency while guaranteeing superresolution quality, a novel visual saliency-based LR-HR correlation mapping strategy between LR and HR patches is proposed based on semicoupled dictionary learning. Moreover, aiming at improving performance and efficiency of spatiotemporal similarity matching and fusion, an improved spatiotemporal nonlocal fuzzy registration scheme is established using the similarity weighting strategy based on pseudo-Zernike moment feature similarity and structural similarity, and the self-adaptive regional correlation evaluation strategy. The proposed spatiotemporal fuzzy registration scheme does not rely on accurate estimation of subpixel motion, and therefore it can be adapted to complex motion patterns and is robust to noise and rotation. Experimental results demonstrate that the proposed algorithm achieves competitive superresolution quality compared to other state-of-the-art algorithms in terms of both subjective and objective evaluations.

1. Introduction and Motivation

Factors such as environmental changes, inaccurate focusing, optical or motion blur, subsampling, and noise disturbance can have a negative effect on video visual quality. Superresolution (SR) reconstruction technology [1–4] aims to reconstruct high-resolution (HR) video sequences from their low-resolution (LR) counterparts. With rapid and significant development of computer vision, there is a growing need for HR videos. Video visual resolution quality plays an important role in accurate moving-target tracking and recognition in intelligent video surveillance systems, which can provide more important details of moving targets. HR medical videos are also very useful for doctors to make correct diagnoses. Therefore, SR video has great research significance and application potential.

In recent years, SR reconstruction technology has been one of the most active research fields in smart image and video analytics and processing. SR techniques have been developed to solve SR problems from the frequency domain to the spatial domain. Currently relevant studies include three main categories: interpolation-based SR methods [5, 6], multiframe-based SR methods [7–9], and learning-based SR methods [10, 11]. Interpolation-based SR methods have relatively low computational cost and therefore are well suited for real-time applications. However, degradation models are not applicable to these methods if blur and noise characteristics vary for different LR video frames. Moreover, additional video details cannot be effectively recovered using these methods because some of the details of interest have usually been blurred.

Multiframe-based SR methods produce HR video sequences by fusing several LR video frames, making full use of complementary and redundant information with similar but not exactly identical details between adjacent video frames at different spatiotemporal scales. At present,

two main fields of research address this kind of method. One branch is based on accurate estimation of subpixel motion using methods such as the projections onto convex sets (POCS) method, the maximum a posteriori (MAP) estimation method, and the iterative back projection (IBP) method, which can be applied only to video sequences with relatively simple motions such as global translation. These methods cannot be adapted to more complex motion patterns such as local motion or angles of rotation. The second branch [12, 13] is based on a recently proposed novel probabilistic motion-estimation scheme based on nonlocal similarity, which does not rely on accurate estimation of subpixel motion and can be adapted to more complex motion patterns. Using this novel scheme, Protter et al. [14] proposed a nonlocal fuzzy registration scheme-based SR reconstruction framework based on a 3D nonlocal mean filter (3D NLM) [15]. Subsequently, Gao et al. [16] improved the nonlocal similarity matching method based on Zernike moment feature similarity and proposed a novel Zernike moment-based SR method which improved the noise robustness and rotation invariance of the NLM-based SR process. However, multiframe-based SR methods cannot be adapted to a larger magnification factor and usually fail when insufficient complementary and redundant information between video frames is provided.

In recent years, learning-based SR methods [17–19] have received much attention. These methods estimate the missing high-frequency details in the input LR images by learning the relationship between LR image patches and the corresponding HR patches from a training set of LR and HR image pairs. This kind of method can be adapted to larger magnification factors and can produce better superresolved results. This paper concentrates on the learning-based SR method for video SR. Until now, nearly all studies of this kind of method have focused on SR for static images. In this paper, by combining the spatiotemporal similarities between video frames, learning-based SR methods will be extended to the video SR field. In the learning-based image SR field, the representative methods are the neighbor embedding-based SR methods (NESR) and the sparse representation-based SR methods (SRR).

Motivated by locally linear embedding (LLE), Chang et al. [20] first proposed a neighbor embedding-based SR method, which reconstructed HR patches by learning a mapping from the local geometry of the LR image patch manifold to that of the HR image patch manifold. Since then, numerous other methods have been proposed and have achieved good performance. Gao et al. [21] extended this method using sparse neighbor embedding, in which the k-nearest neighbor (k-NN) of each LR patch was adaptively chosen by describing local structural information using the histograms of oriented gradients (HoG) feature. Timofte et al. [22] proposed a novel anchored neighborhood regression method for fast example-based SR, in which the nearest neighbors were computed using correlation with dictionary atoms rather than Euclidean distance. However, when dealing with a huge number of training patches, searching for the nearest neighbor can be prohibitively slow and also can require much memory. Moreover, with increasing magnification factor, the correlation between LR patches and their corresponding HR patches becomes ambiguous [23].

Recently, sparse representation and dictionary learning have been proven to be very effective for SR. In sparse representation-based SR methods, some coupled dictionary learning methods [24, 25] have been proposed for superresolution. Lin and Tang [26] proposed a novel coupled subspace learning strategy to learn mappings between different styles. They first used correlative component analysis to find the hidden spaces for each style to preserve correlative information and then learned a bidirectional transform between the two subspaces. Yang et al. [27] proposed a coupled dictionary learning model for image superresolution. They assumed that coupled HR and LR image dictionaries exist which have the same sparse representation for each pair of HR and LR patches. After learning the coupled dictionary pair, the HR patch was reconstructed on the HR dictionary with sparse coefficients coded by the LR image patch over the LR dictionary. This coupled dictionary learning-based SR method assumes that the representation coefficients of the image pair are strictly equal in the coupled subspace. However, this assumption is too strong to address the flexibility of image structures at different resolutions. To overcome this problem, in [28], a semicoupled dictionary learning-based SR method was proposed, which relaxed the above assumption and assumed that there exists a dictionary pair over which the representations of HR and LR image patches have a stable correlation mapping. He et al. [29] used a beta process for sparse coding, establishing a mapping function between HR and LR coefficients. Moreover, in the methods described in [28–30], nonlocal similarities were used to enhance SR performance.

However, these learning-based methods consider nonlocal similarities only in the spatial region of the single image. Therefore, they cannot be directly adapted to video superresolution because they do not make full use of spatiotemporal correlation between video frames, which will influence video spatiotemporal consistency to some extent. This paper aims to solve this problem by extending the concept of single frame-based nonlocal similarities to spatiotemporal nonlocal similarities. A novel learning-based video superresolution method using spatiotemporal nonlocal similarity constraint is proposed which can be adapted to larger magnification factors while effectively preserving video spatiotemporal consistency.

This paper presents a novel learning-based video superresolution reconstruction algorithm using spatiotemporal nonlocal similarity (LBST-SR). The novelty and contributions of this paper are as follows:

(1) By combining LR-HR correlation mapping learning and spatiotemporal nonlocal similarity, video SR performance is further improved via fusion of nonlocal similarity structural redundancies at different spatiotemporal scales.

(2) With the aim of improving algorithm efficiency while guaranteeing SR quality, the authors propose a novel visual saliency-based correlation mapping strategy

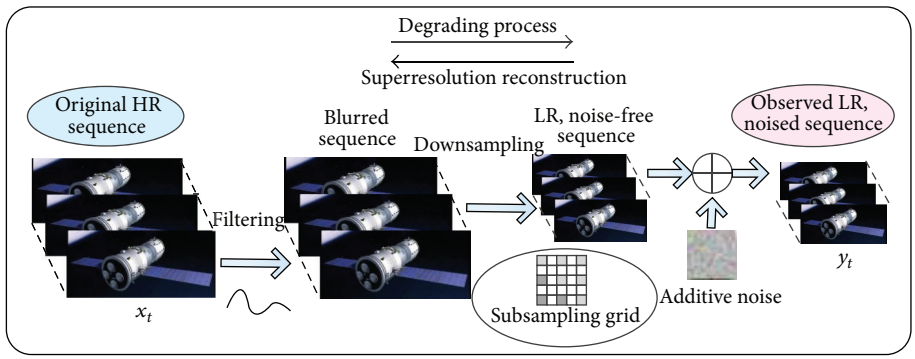


Figure 1: Observation model for video superresolution reconstruction.

between LR and HR patches based on semicoupled dictionary learning. In addition, a self-adaptive regional correlation evaluation strategy based on regional average energy and structural similarity is used in spatiotemporal similarity matching.

(3) An improved spatiotemporal nonlocal fuzzy registration scheme using pseudo-Zernike moment (PZM) and structural similarity is proposed for spatiotemporal similarity matching with the aim of further improving SR accuracy and robustness.

The remainder of the paper is organized as follows. Section 2 gives the observation model for video superresolution reconstruction. Section 3 presents the details of the proposed LBST-SR algorithm. Section 4 gives the experimental results and analysis. Conclusions are presented in Section 5.

2. Observation Model for Video Superresolution Reconstruction

The observation model for video superresolution reconstruction shown in Figure 1, which describes the relationship between HR and LR video frames for superresolution reconstruction, can be formulated as follows:

$$y_t = DB_tM_tx_t + \theta_t, \quad t = 1, 2, \ldots, T, \tag{1}$$

where x_t denotes the tth original HR video frame and y_t denotes the tth observed LR video frame, which is processed by warping M_t, blurring B_t, downsampling D, and noise disturbance θ_t. M_t describes the motions which occur during video acquisition, such as global or local translation and rotation. T denotes the frame number in the video sequence.

3. Proposed LBST-SR Algorithm

3.1. Algorithm Architecture and Mathematical Formulation. On the basis of LR-HR correlation mapping learning between LR patches and the corresponding HR patches, this paper aims to improve the performance of video superresolution reconstruction further by combining spatiotemporal domain nonlocal similarity structural redundancies at different spatiotemporal scales. Therefore, in this paper, a novel learning-based video superresolution reconstruction algorithm using spatiotemporal nonlocal similarity (LBST-SR) is proposed. Objective HR estimations of LR video frames can be obtained by learning LR-HR correlation mapping and fusing spatiotemporal nonlocal similarity information between video frames. With the aim of improving algorithm efficiency while guaranteeing superresolution quality, LR-HR correlation mapping is performed only for the visual salient object region, and then an improved nonlocal fuzzy registration scheme using pseudo-Zernike moment feature and structural similarity is proposed for spatiotemporal similarity matching and fusion. The advantages of the proposed LBST-SR algorithm mainly lie in the following three aspects: (1) it does not rely on accurate estimation of subpixel motion and therefore can be adapted to complex motion patterns (local motions, angles of rotation, etc.); (2) it has high rotation invariance effectiveness and is robust to noise and illumination; and (3) it can be adapted to larger superresolution magnification factors. The proposed algorithm architecture is shown in Figure 2. It includes the following two main processes: LR-HR correlation mapping learning and spatiotemporal nonlocal fuzzy registration and fusion.

Given an input LR video sequence $Y = \{y_m[i, j, t]\}_{t=1}^{T}$ $(m = 1, \ldots, N)$ and a set of LR and HR training pairs, the objective is to infer the corresponding HR video sequence $X = \{x_m[i, j, t]\}_{t=1}^{T}$ $(m = 1, \ldots, N)$, where $m \in [1, N]$ and N denotes the video frame number. The mathematical model of the proposed LBST-SR algorithm is formulated as minimizing the following objective energy function:

$$\widehat{X}^* = \begin{cases} \arg\min_X \left\{E_{\mathrm{SR}}^{\mathrm{CML}}(X, Y) + \lambda E_{\mathrm{SR}}^{\mathrm{STNL}}(X, Y)\right\}, & y \in R_{\mathrm{so}} \\ \arg\min_X \left\{E_{\mathrm{SR}}^{\mathrm{STNL}}(X, Y)\right\}, & y \in R_{\mathrm{nso}}, \end{cases} \tag{2}$$

where $\widehat{X}^*$ denotes the HR estimation of the video sequence. y denotes the pixel in the LR sequence Y. R_{so} denotes the salient object region in Y, and R_{nso} denotes nonsalient region

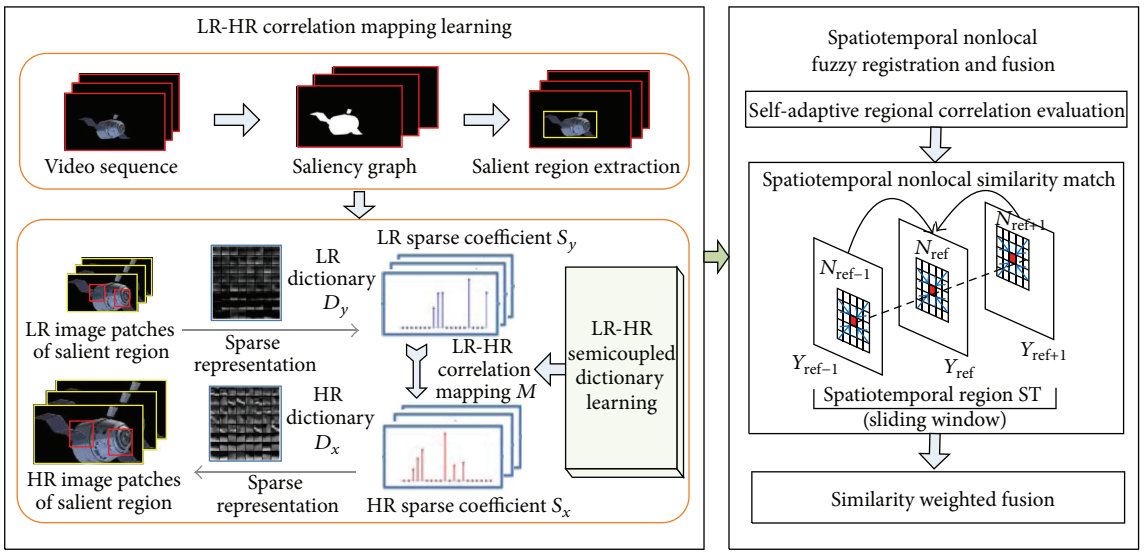


FIGURE 2: Proposed algorithm architecture.

in Y. $E_{SR}^{CML}(X,Y)$ denotes an LR-HR correlation mapping energy element, $E_{SR}^{STNL}(X,Y)$ denotes a spatiotemporal nonlocal similarity regularization constraint element, and λ is the balancing parameter between the two elements. Aiming at improving algorithm time efficiency while guaranteeing superresolution quality, the LR-HR correlation mapping is established only for the human-eye concentrated salient object region $y \in R_{so}$.

3.2. LR-HR Correlation Mapping Learning. The HR estimations of LR video frames can be obtained by learning correlation mapping between LR and HR patches. With the objective of improving algorithm efficiency while guaranteeing SR quality, the LR-HR correlation mapping is established only for the human-eye concentrated salient object region $R_{so} \in R_Y$ in the LR video frame Y. In this paper, a saliency optimization method based on robust background detection [31] is used to detect and extract the visual salient region. The learning process for LR-HR correlation mapping can be formulated as follows: given the LR patch set Y and the HR patch set X, the mapping process can be described as a process of seeking a mapping function $M = f(\cdot)$ from space Y to space X: $X = f(Y)$.

The correlation learning model based on a coupled dictionary assumes that each pair of HR and LR patches has the same sparse representation coefficients. This assumption is too strong to address the flexibility of frame structures at different resolutions, which will restrict superresolution performance. Therefore, in this research, a more flexible and stable semicoupled dictionary learning method has been used to establish correlation mapping between HR and LR patches, which assumes that there exists a stable correlation mapping between the sparse representation coefficients of HR and LR patches. In the LR-HR correlation learning process based on semicoupled dictionary learning, the LR-HR dictionary pair (D_y, D_x) and the correlation mapping matrix M can be obtained by minimizing the objective energy function given in

$$\begin{aligned} \min_{\{D_x,D_y,M\}} \quad & \|X - D_x S_x\|_F^2 + \|Y - D_y S_y\|_F^2 \\ & + \gamma \|S_x - M S_y\|_F^2 + \lambda_x \|S_x\|_1 + \lambda_y \|S_y\|_1 \\ & + \lambda_w \|M\|_F^2 \\ \text{s.t.} \quad & \|d_{x,i}\|_{l_2} \le 1, \\ & \|d_{y,i}\|_{l_2} \le 1, \\ & \forall i, \end{aligned} \tag{3}$$

where $\gamma, \lambda_x, \lambda_y$, and λ_w denote the regularization parameters needed to balance the terms in the objective function; S_y and S_x are the sparse representation coefficients of LR and HR patches, respectively; $\|X - D_x S_x\|_F^2$ and $\|Y - D_y S_y\|_F^2$ denote the reconstruction errors; $\|S_x - M S_y\|_F^2$ denotes the mapping error; and $d_{y,i}$ and $d_{x,i}$ denote the atoms of D_y and D_x, respectively.

To solve the minimization problem for the objective energy function in (3), it can be separated into three subproblems: (1) sparse coding for training samples; (2) dictionary updating; and (3) mapping updating.

Sparse Coding for Training Samples. With the initialization of M and the dictionary pair (D_y, D_x), the sparse coding coefficients S_y and S_x can be obtained by solving (4) using L_1-optimization algorithms:

$$\begin{aligned} &\min_{\{S_x\}} \|X - D_x S_x\|_F^2 + \gamma \|S_y - M_x S_x\|_F^2 + \lambda_x \|S_x\|_1, \\ &\min_{\{S_y\}} \|Y - D_y S_y\|_F^2 + \gamma \|S_x - M_y S_y\|_F^2 + \lambda_y \|S_y\|_1, \end{aligned} \tag{4}$$

where M_x denotes the mapping from S_x to S_y and M_y denotes the mapping from S_y to S_x. $\|S_y - M_x S_x\|_F^2$ denotes the mapping error generated during S_x is mapped to S_y. $\|S_x - M_y S_y\|_F^2$ denotes the mapping error generated during S_y is mapped to S_x.

Dictionary Updating. With S_y and S_x fixed, the dictionary pair (D_y, D_x) can be updated using

$$\begin{aligned} \min_{\{D_x, D_y\}} \quad & \left\|X - D_x S_x\right\|_F^2 + \left\|Y - D_y S_y\right\|_F^2 \\ \text{s.t.} \quad & \left\|d_{x,i}\right\|_{l_2} \le 1, \\ & \left\|d_{y,i}\right\|_{l_2} \le 1, \\ & \forall i. \end{aligned} \tag{5}$$

Mapping Updating. With the dictionary pair (D_y, D_x), S_y, and S_x fixed, the mapping M can be updated as follows:

$$\min_{\{M\}} \left\|S_x - M S_y\right\|_F^2 + \left(\frac{\lambda_w}{\gamma}\right) \|M\|_F^2. \tag{6}$$

By solving (6), the following expression can be derived:

$$M = S_x S_y^T \left(S_y S_y^T + \left(\frac{\lambda_w}{\gamma}\right) \cdot I\right)^{-1}, \tag{7}$$

where I is an identity matrix.

After obtaining the LR-HR correlation mapping M using the above learning process, the superresolution reconstruction is done by using it to derive the HR estimation of the salient object region in the video frame. For the salient object region $R_{\text{so}} \in R_Y$ in LR video frame Y, the following optimization problem given in (8) is solved to obtain its HR estimation:

$$\begin{aligned} \min_{\{S_{x,i}, S_{y,i}\}} & \left\|x_i - D_x S_{x,i}\right\|_F^2 + \left\|y_i - D_y S_{y,i}\right\|_F^2 \\ & + \gamma \left\|S_{x,i} - M S_{y,i}\right\|_F^2 + \lambda_x \left\|S_{x,i}\right\|_1 + \lambda_y \left\|S_{y,i}\right\|_1, \end{aligned} \tag{8}$$

where y_i is a patch of LR video frame Y and x_i is the corresponding patch in the initial estimation of HR video frame X. An initial estimation of X can be obtained using a Bicubic interpolator. Equation (8) can be solved by alternately updating $S_{x,i}$ and $S_{y,i}$. The objective HR estimation $\widehat{x}_i^{cm}$ of each patch x_i in the salient object region $R_{\text{so}} \in R_X$ of X can be derived by solving

$$\widehat{x}_i^{cm} = D_x \widehat{S}_{x,i}. \tag{9}$$

3.3. Spatiotemporal Nonlocal Fuzzy Registration and Fusion. The superresolution process based on the learned LR-HR correlation mapping uses only the spatial information in the video frame and the LR-HR mapping. Therefore, it does not make full use of the spatiotemporal relationship between video frames and therefore cannot preserve video temporal consistency. Large quantities of spatiotemporal nonlocal similarity information exist between video frames, and these nonlocal redundancies are very useful for video superresolution reconstruction. Therefore, in this research, video spatiotemporal nonlocal similarity was used to enhance further the performance of the proposed superresolution algorithm based on LR-HR correlation learning. With the objective of improving the performance and efficiency of the spatiotemporal nonlocal similarity matching, the spatiotemporal nonlocal fuzzy registration scheme was improved using the similarity weighting strategy based on PZM feature similarity and structural similarity and the self-adaptive regional correlation evaluation strategy.

3.3.1. Improved Spatiotemporal Nonlocal Fuzzy Registration Scheme Using PZM and Structural Similarity (ZSFR). Considering good rotation, translation, and scale-invariance properties and insensitivity to noise and illumination of PZM feature, the nonlocal fuzzy registration scheme could be further improved by using this feature, resulting in a more accurate and robust similarity measure between regional features in the nonlocal spatiotemporal domain for weighting calculations. In this way, the performance and robustness of SR reconstruction could be further improved. Unlike traditional methods, the improved spatiotemporal nonlocal fuzzy registration scheme does not rely on accurate estimation of subpixel motion and therefore it can be adapted to complex motion scenes and is robust to noise and rotation.

Let $\text{PZM}(k,l)$ and $\text{PZM}'(i,j)$ represent two PZM feature vectors of local regions corresponding to pixel (k,l) and pixel (i,j) in the nonlocal search region $N_{\text{nonloc}}(k,l)$ of pixel (k,l), which can be calculated as

$$\begin{aligned} \text{PZM}(k,l) &= \left(\text{PZM}_{00}, \text{PZM}_{11}, \text{PZM}_{20}, \text{PZM}_{22}, \right. \\ &\quad \left. \text{PZM}_{31}, \text{PZM}_{33}\right), \\ \text{PZM}'(i,j) &= \left(\text{PZM}'_{00}, \text{PZM}'_{11}, \text{PZM}'_{20}, \text{PZM}'_{22}, \right. \\ &\quad \left. \text{PZM}'_{31}, \text{PZM}'_{33}\right), \end{aligned} \tag{10}$$

where PZM feature with order n and repetition m ($0 \le n \le \infty$, $0 \le |m| \le n$) of video frame $f(x,y)$ is defined as

$$\begin{aligned} \text{PZM}_{nm} &= \frac{n+1}{\pi} \iint_{x^2+y^2 \le 1} f(x,y) V_{nm}^*(x,y)\, dx\, dy \\ &= \frac{n+1}{\pi} \sum_{\rho \le 1} \sum_{0 \le \theta \le 2\pi} f(\rho,\theta) V_{nm}^*(\rho,\theta)\, \rho, \\ V_{nm}(\rho,\theta) &= R_{nm}(\rho) \exp(jm\theta), \\ R_{nm}(\rho) &= \sum_{s=0}^{n-|m|} \frac{(-1)^s (2n+1-s)! \rho^{n-s}}{s!\,(n+|m|+1-s)!\,(n-|m|-s)!}, \end{aligned} \tag{11}$$

where ρ and θ are the radius and angle, respectively, of the pixels in the polar coordinate system, $\rho = \sqrt{x^2 + y^2}$, and $\theta = \tan^{-1}(y/x)$. The function $\{V_{nm}(x, y)\}$ is the basis of PZM feature, and $V^*_{nm}(x, y)$ denotes the complex conjugate of $V_{nm}(x, y)$.

The nonlocal fuzzy registration scheme based on PZM is based on a similarity match in the nonlocal spatiotemporal domain between video frames at different spatiotemporal scales, which is measured by the Euclidean distance between regional PZM feature vectors. The weight $\omega^{\mathrm{PZM}}_{\mathrm{SR}}[k, l, i, j, t]$ of each pixel in the nonlocal spatiotemporal region is calculated based on this similarity as follows:

$$\omega^{\mathrm{PZM}}_{\mathrm{SR}}[k, l, i, j, t] = \frac{1}{C(k, l)} \exp\left\{-\frac{\left\|\mathrm{PZM}(k, l) - \mathrm{PZM}'(i, j)\right\|_2^2}{\varepsilon^2}\right\}, \tag{12}$$

where ε controls the decay rate of the exponential function and the weight. $C(k, l)$ is a normalization constant, which is calculated as follows:

$$C(k, l) = \sum_{(i,j)\in N_{\mathrm{nonloc}}(k,l)} \exp\left\{-\frac{\left\|\mathrm{PZM}(k, l) - \mathrm{PZM}'(i, j)\right\|_2^2}{\varepsilon^2}\right\}. \tag{13}$$

Note that the higher the PZM order is, the more sensitive the PZM is to noise. Therefore, in the experiments performed in this study, only the first third-order moments, including PZM_{00}, PZM_{11}, PZM_{20}, PZM_{22}, PZM_{31}, and PZM_{33}, were calculated.

By analyzing the weight calculation formula for the PZM-based nonlocal fuzzy registration scheme in (12), it is clear that the time complexity is much too high and increases with the number of LR video frames and the amplification factor. To achieve further improvements in time efficiency and the edge detail-preserving ability of the superresolution algorithm, a novel spatiotemporal nonlocal fuzzy registration scheme (ZSFR) was established by improving the PZM-based spatiotemporal nonlocal fuzzy registration scheme using the similarity weighting strategy based on PZM feature similarity and structural similarity and the self-adaptive regional correlation evaluation strategy.

The improvements in the ZSFR involve two main aspects: (1) with the aim of improving algorithm efficiency, a self-adaptive regional correlation evaluation strategy based on regional average energy and regional structural similarity was constructed for nonlocal similarity matching; and (2) an improved similarity weighting strategy based on regional PZM feature similarity and regional structural similarity was proposed for spatiotemporal nonlocal similarity matching, with the aim of further improving SR performance. To describe this improved ZSFR scheme, the following three definitions are required.

Definition 1 (regional average energy). The video frame F is divided into many regions of equal size, and each region is divided into 5×5 patches. The total number of pixels in each region is Num, and the energy value of each pixel is denoted by $p_1, p_2, \ldots, p_{\mathrm{Num}}$, respectively. $\mathrm{AE}(x, y)$ is defined as the regional average energy centered on pixel (x, y) and is calculated as

$$\mathrm{AE}(x, y) = \sum_{i=1}^{\mathrm{Num}} \frac{p_i}{\mathrm{Num}}. \tag{14}$$

Definition 2 (PZM feature similarity). Given two regions centered on pixels (k, l) and (i, j), denoted by $R(k, l)$ and $R(i, j)$, respectively, the corresponding feature vectors extracted from these two regions are $\mathrm{PZM}(k, l)$ and $\mathrm{PZM}'(i, j)$. The parameter ε controls the decay rate of the exponential function. The PZM feature similarity between these two regions is defined as

$$\mathrm{RFS}(R(k, l), R(i, j)) = \exp\left\{-\frac{\left\|\mathrm{PZM}(k, l) - \mathrm{PZM}'(i, j)\right\|_2^2}{\varepsilon^2}\right\}. \tag{15}$$

Definition 3 (regional structural similarity). Given two regions centered on pixels (k, l) and (i, j), denoted by $R(k, l)$ and $R(i, j)$, respectively, $\eta_{(k,l)}$ and $\eta_{(i,j)}$ are the means of these two regions, $\sigma_{(k,l)}$ and $\sigma_{(i,j)}$ are the standard deviations of these two regions, and $\sigma_{(k,l,i,j)}$ is the covariance between the two regions. e_1 and e_2 are two constants. Then, the structural similarity $\mathrm{RSS}(R(k, l), R(i, j))$ between the two regions is defined as

$$\mathrm{RSS}(R(k, l), R(i, j)) = \frac{\left(2\eta_{(k,l)}\eta_{(i,j)} + e_1\right)\left(2\sigma_{(k,l,i,j)} + e_2\right)}{\left(\eta^2_{(k,l)} + \eta^2_{(i,j)} + e_1\right)\left(\sigma^2_{(k,l)} + \sigma^2_{(i,j)} + e_2\right)}. \tag{16}$$

In the improved spatiotemporal nonlocal fuzzy registration scheme, the regional correlation is first evaluated to divide the local regions centered on all pixels (i, j) in the nonlocal search region for pixel (k, l) into related and unrelated regions. Only related regions are used to calculate the weight, an approach which can further improve time efficiency and is beneficial for mining the most similar patterns to calculate the similarity weight. The regional correlation is calculated by combining the regional average energy and regional structural similarity. Moreover, a self-adaptive threshold δ_{adap} is introduced, which yields a self-adaptive regional correlation evaluation mechanism. If two regions are related, the criterion is defined as

$$\left|\mathrm{AE}(k, l) - \mathrm{AE}(i, j)\right| \times \left(\frac{\left(1 - \mathrm{RSS}(R(k, l), R(i, j))\right)}{2}\right) < \delta_{\mathrm{adap}}. \tag{17}$$

The self-adaptive threshold δ_{adap} is adaptively determined by the average energy $\mathrm{AE}(k, l)$ for the region centered on pixel

(k,l), which leads to a more accurate regional correlation evaluation. δ_{adap} is calculated as

$$\delta_{\text{adap}} = \lambda \text{AE}(k,l), \tag{18}$$

where λ is an adjustment factor that controls δ_{adap}. Experiments have confirmed that the best SR quality is obtained when λ is set to 0.08.

With the aim of further improving superresolution accuracy and detail-preserving ability, the similarity weight $\omega_{\text{SR}}^{\text{EPZM}}[k,l,i,j,t]$ is improved on the basis of the weighting strategy given in (12) by combining the two factors of regional PZM feature similarity and regional structural similarity. The improved similarity weight $\omega_{\text{SR}}^{\text{EPZM}}[k,l,i,j,t]$ is calculated as follows:

$$\begin{aligned}
&\omega_{\text{SR}}^{\text{EPZM}}[k,l,i,j,t] \\
&= \begin{cases} \dfrac{1}{C(k,l)} \times \text{RFS}(R(k,l), R(i,j)) \times (1 - 0.0002\text{RSS}(R(k,l), R(i,j))), & |\text{AE}(k,l) - \text{AE}(i,j)| \times \left(\dfrac{(1 - \text{RSS}(R(k,l), R(i,j)))}{2} \right) < \delta_{\text{adap}} \\ 0, & \text{otherwise}, \end{cases} \\
&= \begin{cases} \dfrac{1}{C(k,l)} \times \exp\left\{ -\dfrac{\left\| \text{PZM}(k,l) - \text{PZM}'(i,j) \right\|_2^2}{\varepsilon^2} \right\} \times (1 - 0.0002\text{RSS}(R(k,l), R(i,j))), & |\text{AE}(k,l) - \text{AE}(i,j)| \times \left(\dfrac{(1 - \text{RSS}(R(k,l), R(i,j)))}{2} \right) < \delta_{\text{adap}} \\ 0, & \text{otherwise}, \end{cases}
\end{aligned} \tag{19}$$

$$C(k,l) = \sum_{(i,j) \in N_{\text{nonloc}}(k,l)} \exp\left\{ -\frac{\left\| \text{PZM}(k,l) - \text{PZM}'(i,j) \right\|_2^2}{\varepsilon^2} \right\} \times (1 - 0.0002\,\text{RSS}(R(k,l), R(i,j))), \tag{20}$$

where (k,l) denotes the pixel to be superresolved and (i,j) denotes a pixel in the nonlocal search region $N_{\text{nonloc}}(k,l)$ centered on pixel (k,l). The parameter ε controls the decay rate of the exponential function, as well as the weight. $C(k,l)$ is a normalization constant.

3.3.2. Spatiotemporal Nonlocal Similarity Information Fusion Based on ZSFR. Spatiotemporal nonlocal similarity information fusion is based on the improved nonlocal fuzzy registration scheme using PZM feature similarity and structural similarity. By learning spatiotemporal nonlocal similarities between video frames, the similarity weight is calculated according to (19). The HR estimation of the video frame to be superresolved can then be obtained by spatiotemporal information fusion, which is implemented by a weighted average based on spatiotemporal nonlocal similarities.

Once the weight $\omega_{\text{SR}}^{\text{EPZM}}[k,l,i,j,t]$ has been determined, the HR estimation of each pixel in the video frame to be superresolved can be obtained using the weighted average of the pixels in the nonlocal spatiotemporal region. The objective superresolution energy function based on spatiotemporal nonlocal similarity can be expressed as follows:

$$\begin{aligned}
\widehat{x}_{\text{stnl}} = \arg\min_{\{x(k.l)\}} \Bigg\| & x(k,l) \\
& - \sum_{t=t_1}^{t_2} \sum_{(i,j) \in N_{\text{nonloc}}(k,l)} \omega_{\text{SR}}^{\text{EPZM}}(k,l,i,j,t)\, x(i,j) \Bigg\|_2^2,
\end{aligned} \tag{21}$$

where $[t_1, t_2]$ denotes a 3D spatiotemporal region (temporal sliding window). By minimizing the objective energy function in (21), the HR estimation $\widehat{x}_{\text{stnl}}$ of each video frame can be obtained as follows:

$$\widehat{x}_{\text{stnl}} = \frac{\sum_{(k,l) \in \Psi} \sum_{t \in [t_1,t_2]} \sum_{(i,j) \in N_{\text{nonloc}}(k,l)} \omega_{\text{SR}}^{\text{EPZM}}(k,l,i,j,t)\, x_t(i,j)}{\sum_{(k,l) \in \Psi} \sum_{t \in [t_1,t_2]} \sum_{(i,j) \in N_{\text{nonloc}}(k,l)} \omega_{\text{SR}}^{\text{EPZM}}(k,l,i,j,t)}, \tag{22}$$

where Ψ denotes the video frame to be superresolved.

Consequently, the proposed learning-based video superresolution reconstruction using spatiotemporal nonlocal similarity can be performed as follows:

$$\widehat{x}^* = \begin{cases} \arg\min_{\{x(k.l)\}} E_{\text{SR}}^{\text{CML}} + \lambda \left\| x(k,l) - \sum_{t=t_1}^{t_2} \sum_{(i,j) \in N_{\text{nonloc}}(k,l)} \omega_{\text{SR}}^{\text{EPZM}}(k,l,i,j,t)\, x(i,j) \right\|_2^2, & (k,l) \in R_{\text{so}} \\ \arg\min_{\{x(k,l)\}} \left\| x(k,l) - \sum_{t=t_1}^{t_2} \sum_{(i,j) \in N_{\text{nonloc}}(k,l)} \omega_{\text{SR}}^{\text{EPZM}}(k,l,i,j,t)\, x(i,j) \right\|_2^2, & (k,l) \in R_{\text{nso}}, \end{cases} \tag{23}$$

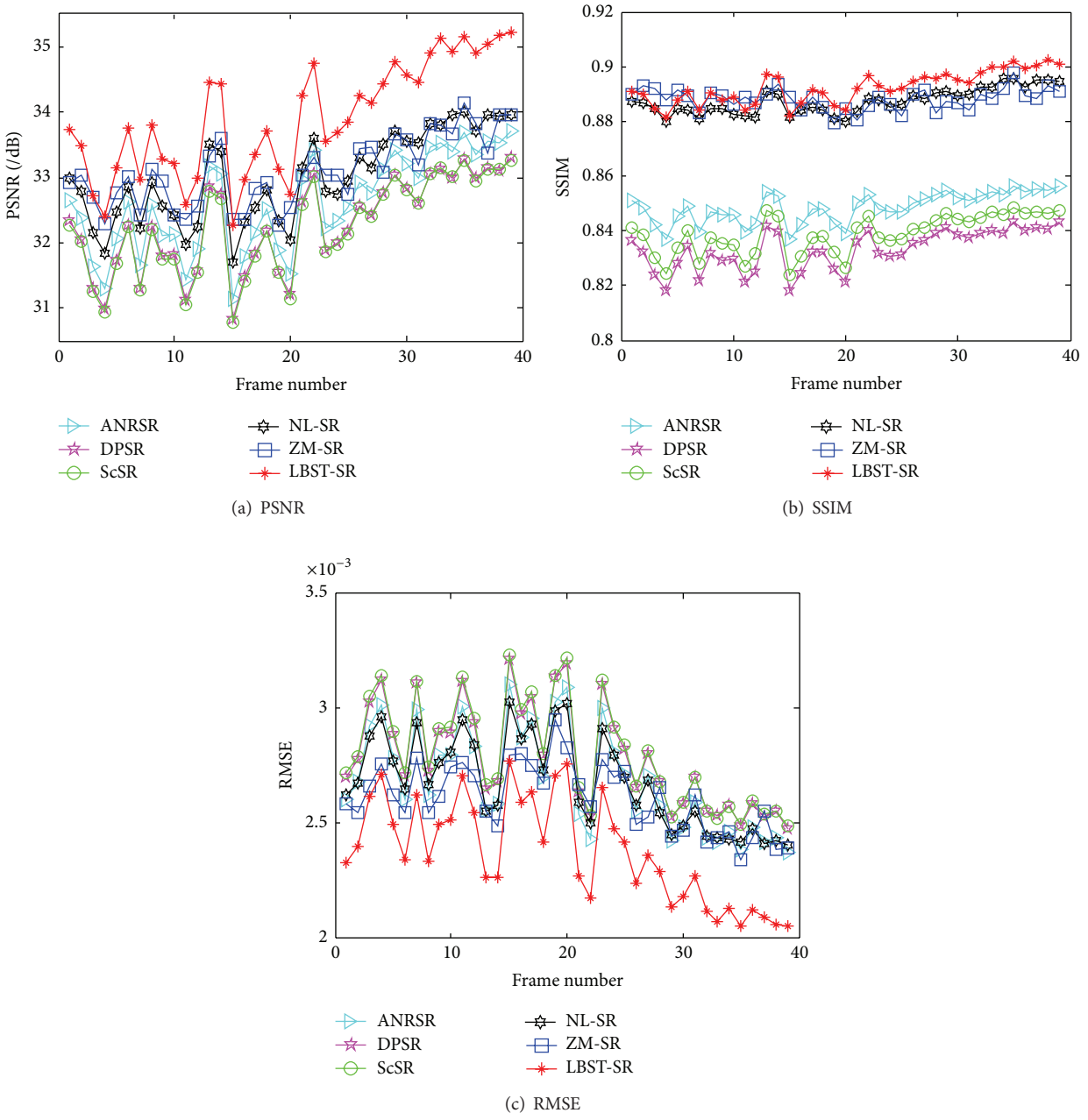


Figure 3: Objective evaluation indices of the six algorithms for the "Satellite-1" sequence.

where E_{SR}^{CML} denotes the energy function defined in (8) and λ is a balancing parameter.

3.4. Implementation Steps of the Proposed LBST-SR Algorithm. The LBST-SR algorithm implementation includes the following steps, as shown in Algorithm 4.

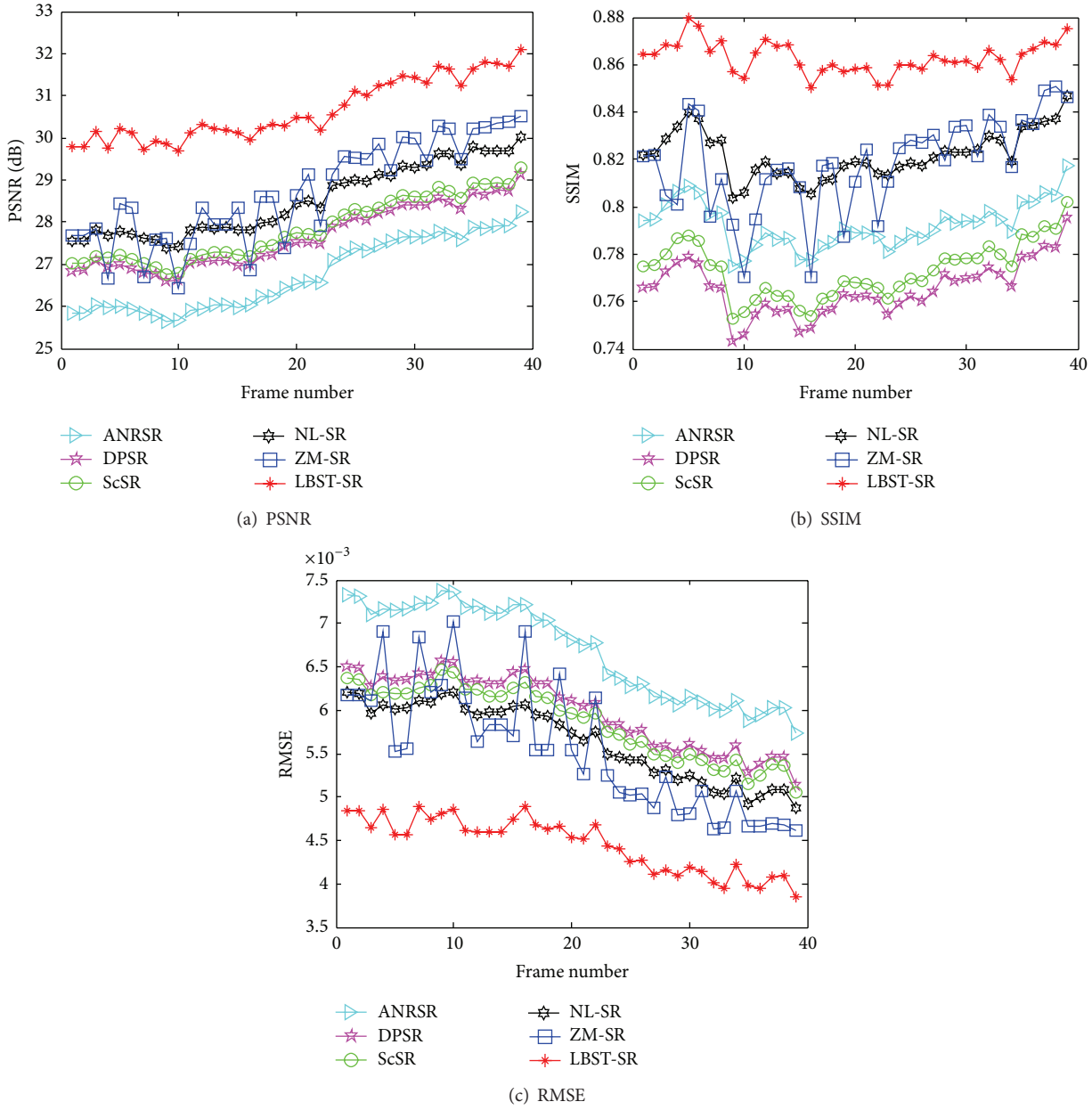
Algorithm 4. LBST-SR algorithm implementation steps are as follows:

Input. LR video sequence $\{y_m[i, j, t]\}_{t=1}^{T}$ $(m = 1, \ldots, N)$, scale amplification factor s, HR training dataset X, LR training dataset Y, nonlocal search region size $W \times W$, local region size for similarity weight calculation $B \times B$, weight-controlling filter parameter ε, and iteration scale K.

Output. The superresolved HR video sequence $\{x_k[i, j, t]\}_{t=1}^{T}$ $(k = 1, \ldots, N)$.

Training Process

Step 1. Sample LR and HR patches from LR and HR training datasets Y and X, respectively.



(a) PSNR

(b) SSIM

(c) RMSE

Figure 4: Objective evaluation indices of the six algorithms for the "Satellite-2" sequence.

Step 2. Train the LR-HR dictionary pair (D_y, D_x) and the correlation mapping matrix M by LR-HR correlation learning according to (3).

Superresolution Reconstruction Process

Step 1. Initialize LR video sequence $\{y_m[i,j,t]\}_{t=1}^{T}$ $(m = 1, \ldots, N)$ using the Bicubic interpolator with the aim of obtaining its HR initial estimation $\{Y_p[i,j,t]\}_{t=1}^{T}$ $(p = 1, \ldots, N)$.

Step 2. According to the learned dictionary pair (D_y, D_x) and the LR-HR correlation mapping M, map each LR patch of the salient region R_{so} of video frame to its HR estimation $\hat{x}$ using (8) and (9).

Step 3. Update $\hat{x}$ using the improved spatiotemporal nonlocal similarity regularization constraint in (23).

Step 4. Iteratively refine the fusion result for further optimization. Update the counter, $t = t + 1$. If $t \leq K$, return to Step 3; otherwise, end the process.

4. Experimental Results and Analysis

4.1. Experimental Dataset and Evaluation Indices. The experimental datasets in this paper consist of the benchmark video sequences taken from the http://trace.eas.asu.edu/yuv/index.html website and the spatial video sequences taken from

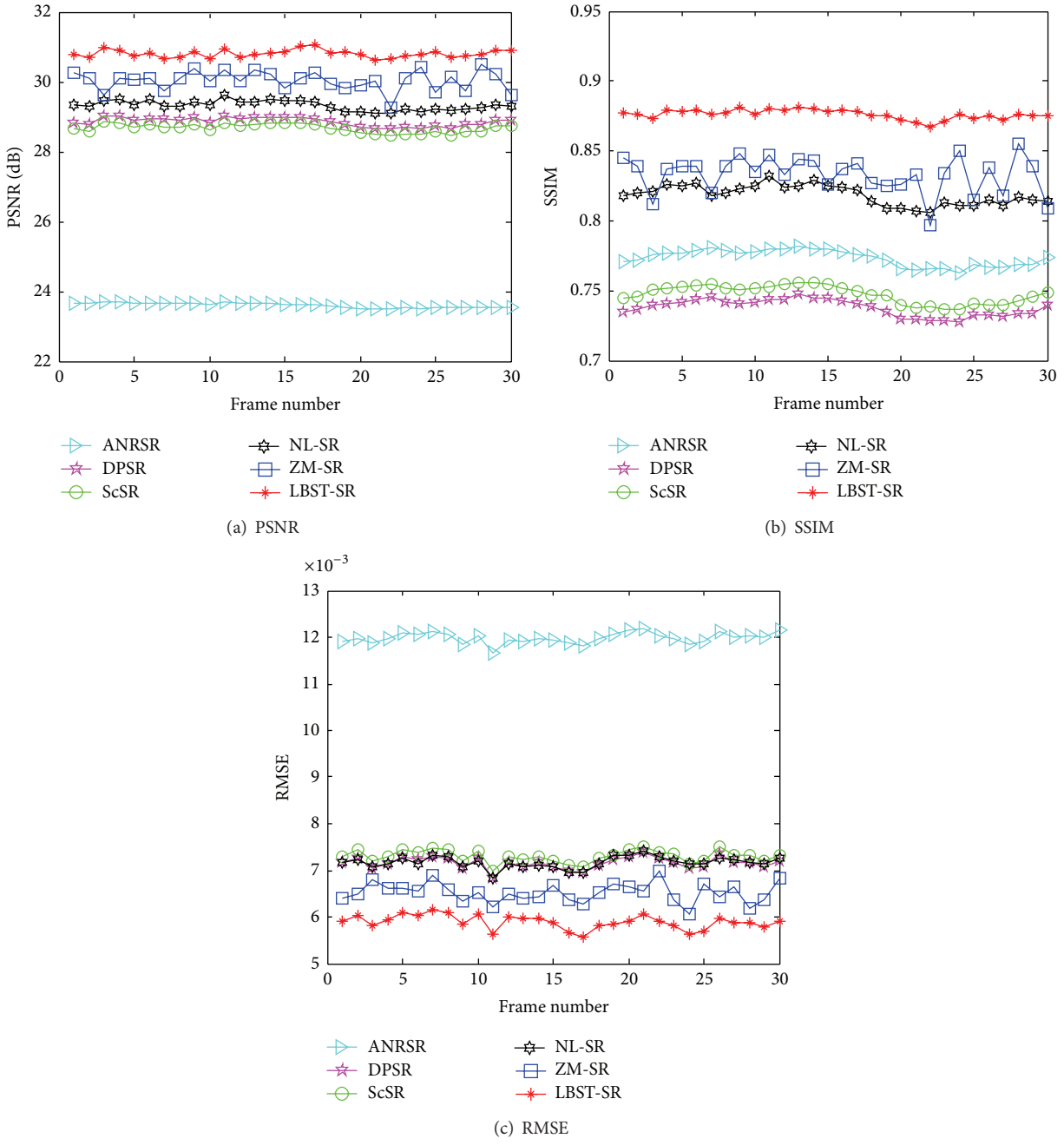


Figure 5: Objective evaluation indices of the six algorithms for the "Forman" sequence.

the YOUKU website (http://www.youku.com/). The super-resolution effects were validated in terms of both subjective visual evaluation and four objective quantitative indices: peak signal-to-noise ratio (PSNR), structural similarity (SSIM), feature similarity (FSIM), and root-mean-square error (RMSE), which were calculated as follows:

$$\mathrm{PSNR} = 10 \cdot \log_{10} \frac{255^2}{(1/(M \times N)) \sum_{i=1}^{M} \sum_{j=1}^{N} (R(i,j) - F(i,j))^2} \mathrm{dB},$$

$$\mathrm{SSIM} = \frac{(2\eta_R \eta_F + e_1)(2\sigma_{RF} + e_2)}{(\eta_R^2 + \eta_F^2 + e_1)(\sigma_R^2 + \sigma_F^2 + e_2)},$$

$$\mathrm{FSIM} = \frac{\sum_{x \in \Omega} S_L(x) \cdot [S_C(x)]^{\lambda} \cdot PC_m(x)}{\sum_{x \in \Omega} PC_m(x)},$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (R(i,j) - F(i,j))^2}, \tag{24}$$

where M and N denote the length and width of the video frame; R and F denote the reconstructed frame and

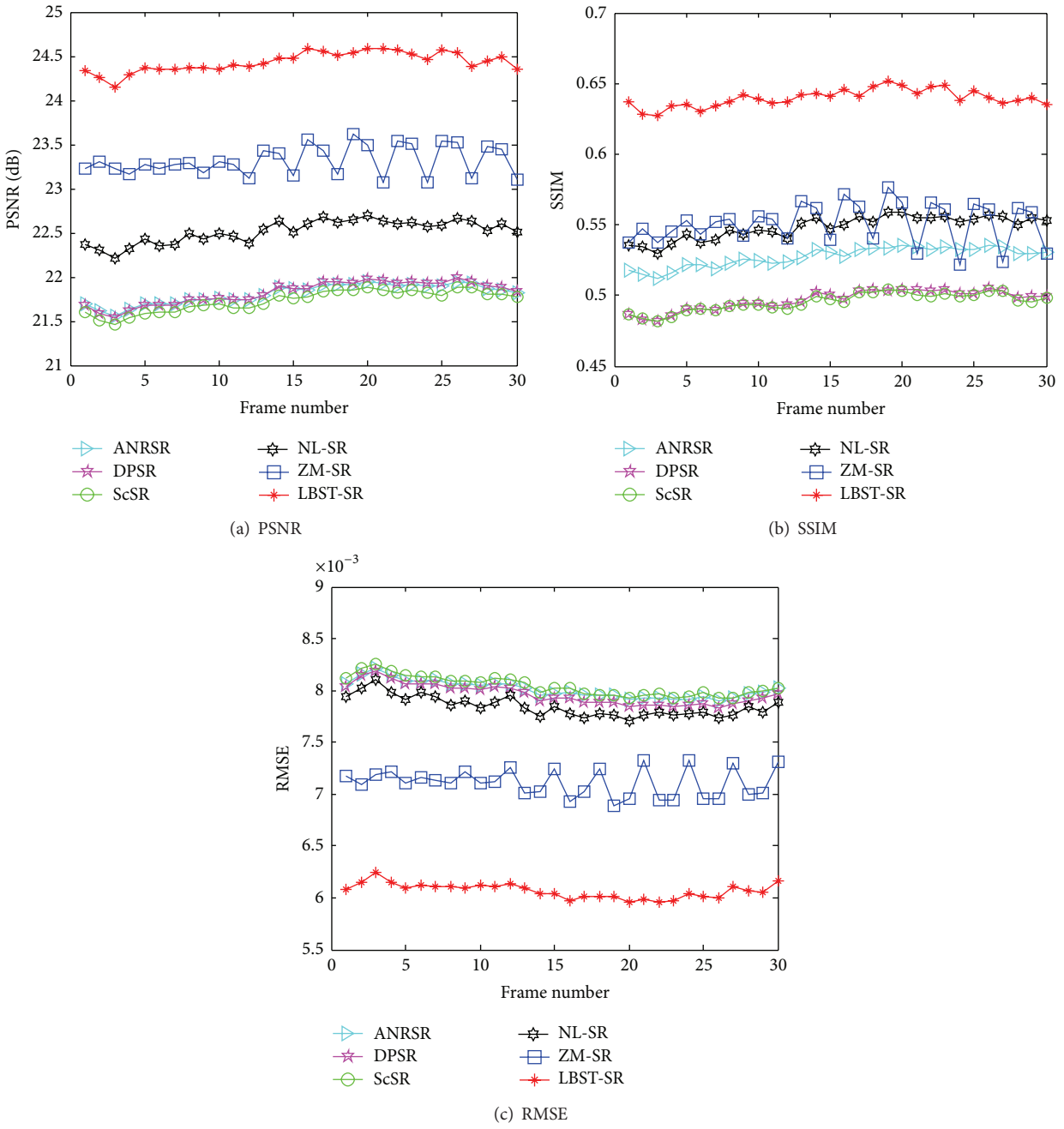


FIGURE 6: Objective evaluation indices of the six algorithms for the "Calendar" sequence.

the original frame, respectively; η_R and η_F are the means; σ_R and σ_F are the standard deviations for the original and reconstructed frames; σ_{RF} is the covariance for the original and reconstructed frames; e_1 and e_2 are constants; Ω denotes the whole spatial domain of the video frame; $S_L(x)$ is a similarity measure of the phase congruency and gradient magnitude features between R and F; $S_C(x)$ is a chrominance similarity measure between R and F; $PC_m(x)$ is used to weight the importance of $S_L(x)$ in the overall similarity between R and F, where $S_L(x)$, $S_C(x)$, and $PC_m(x)$ are calculated according to [32]. The greater the PSNR is, the closer the reconstructed frame is to the original. The closer SSIM ($0 \le \text{SSIM} \le 1$) is to 1, the greater is the similarity between the original and reconstructed frame structures. The closer FSIM ($0 \le \text{FSIM} \le 1$) is to 1, the greater is the similarity between the original and reconstructed frame features. The smaller the RMSE is, the closer the reconstructed frame is to the original.

4.2. Experimental Results and Analysis. This section describes the experiments that were carried out to evaluate the performance of the proposed LBST-SR superresolution reconstruction algorithm and a comparison of these results

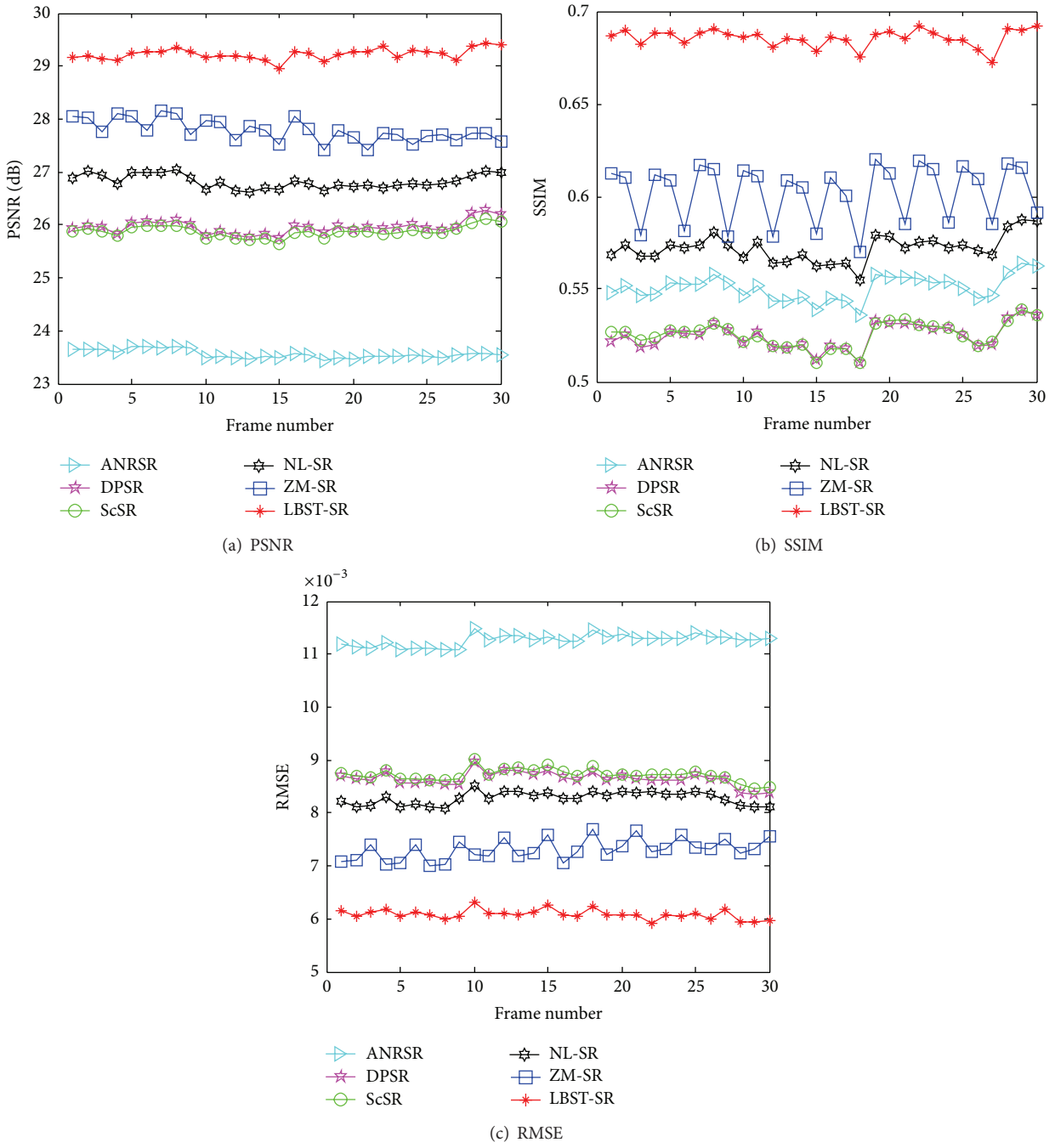
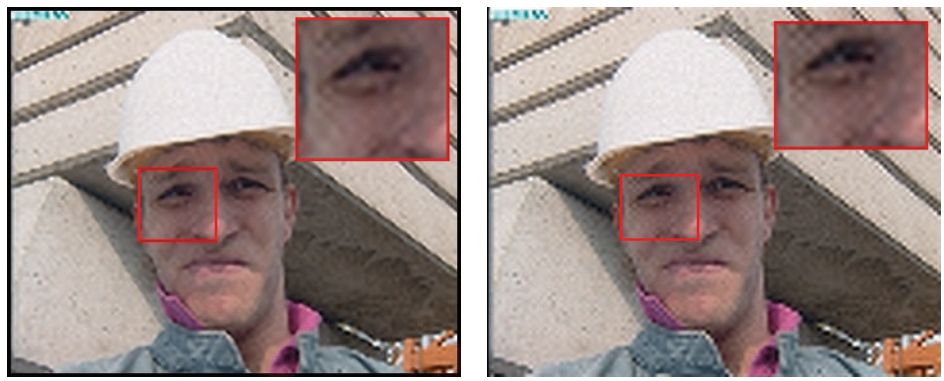


Figure 7: Objective evaluation indices of the six algorithms for the "Coastguard" sequence.

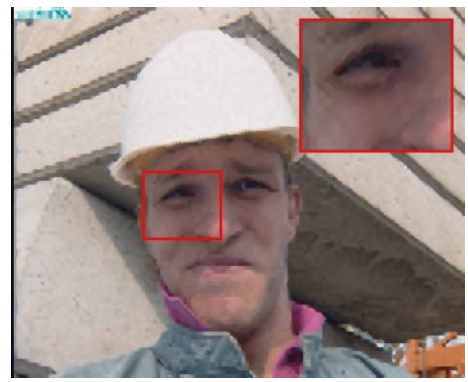
with five recently proposed representative state-of-the-art superresolution algorithms in terms of both visual quality and objective quantitative indices, including the learning-based ANRSR [22], DPSR [30], and ScSR [27] algorithms, the 3D nonlocal mean-based NL-SR [14] algorithm, and the Zernike moment-based ZM-SR [16] algorithm. In the experiments, ten benchmark and two spatial video sequences were used: "Forman," "Calendar," "Coastguard," "Suzie," "Mother_Daughter," "Miss_America," "Ice," "Football," "Carphone," "Akiyo," "Satellite-1," and "Satellite-2." Based on the motion contents, these video sequences are divided into three categories: (1) "Calendar," "Suzie," "Mother_Daughter," "Miss_America," and "Akiyo" contain small-motion objects; (2) "Forman," "Coastguard," "Carphone," "Satellite-1," and "Satellite-2" contain moderate-motion objects; and (3) "Ice" and "Football" contain fast-motion objects. Some complex motion scenes exist in these dynamic sequences, such as local motion patterns and rotations. Each video sequence was decimated by a factor of 1 : 3 and then contaminated by additive Gaussian white noise with $\sigma = 2$. In



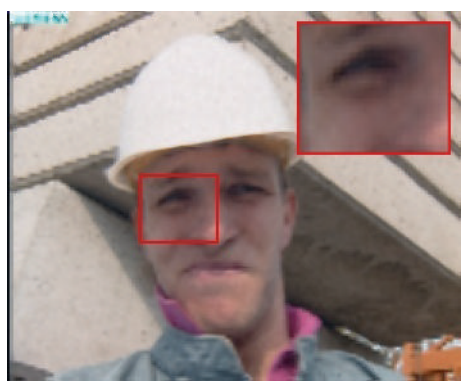
(a) ANRSR

(b) DPSR

(c) ScSR



(d) NL-SR



(e) ZM-SR

(f) LBST-SR

Figure 8: SR reconstruction visual effects for Frame 6 of the "Forman" sequence with a magnification factor of three.

the proposed LBST-SR algorithm, the spatiotemporal region used for the similarity weight calculation in the nonlocal similarity matching process was $3 \times 3 \times 6$. Superresolution with a magnification factor of three was implemented in these experiments.

4.2.1. Objective Quantitative Evaluations. The average SSIM, PSNR, FSIM, and RMSE index values of ANRSR, DPSR, ScSR, NL-SR, ZM-SR, and LBST-SR algorithms for the twelve video sequences are shown in Tables 1, 2, 3, and 4, respectively. Figures 3–7 show the PSNR, SSIM, and RMSE values of the six algorithms for the "Satellite-1," "Satellite-2," "Forman," "Calendar," and "Coastguard" sequences. The results indicate that, in most cases, the proposed LBST-SR algorithm yields better performance with higher PSNR, SSIM, and FSIM values and smaller RMSE values than the other five algorithms. In only a few cases, ZM-SR algorithm achieves slightly better effects in terms of some indices than the proposed LBST-SR algorithm. Moreover, the SSIM and FSIM index values demonstrate that the results generated by the proposed LBST-SR algorithm are much closer to the original ones than the other five algorithms in terms of structural similarity and feature similarity, because LR-HR correlation mapping learning and spatiotemporal similarity can recover high-frequency details of video frames more accurately.

Table 1: Average SSIM index values of the six algorithms.

Video sequences	ANRSR	DPSR	ScSR	NL-SR	ZM-SR	LBST-SR
Satellite-1	0.8485	0.8330	0.8388	0.8872	0.8881	**0.9118**
Satellite-2	0.7925	0.7654	0.7732	0.7732	0.8182	**0.8701**
Forman	0.7736	0.7379	0.7474	0.8184	0.8329	**0.8612**
Calendar	0.5274	0.4968	0.4954	0.5483	0.5508	**0.5979**
Coastguard	0.5507	0.5747	0.5255	0.5721	0.6022	**0.6495**
Suzie	0.7348	0.7117	0.7070	0.7456	0.7791	**0.8048**
Mother_Daughter	0.7533	0.7232	0.7239	0.7896	0.8141	**0.8186**
Miss_America	0.8043	0.7803	0.7802	0.8376	0.8664	**0.8855**
Ice	0.7596	0.7208	0.7163	0.8086	0.8061	**0.8234**
Football	0.5521	0.5210	0.5278	0.5683	0.5545	**0.5941**
Carphone	0.7211	0.6910	0.8793	0.7407	**0.9114**	0.7772
Akiyo	0.8159	0.7904	0.7926	0.8408	0.8573	**0.8598**

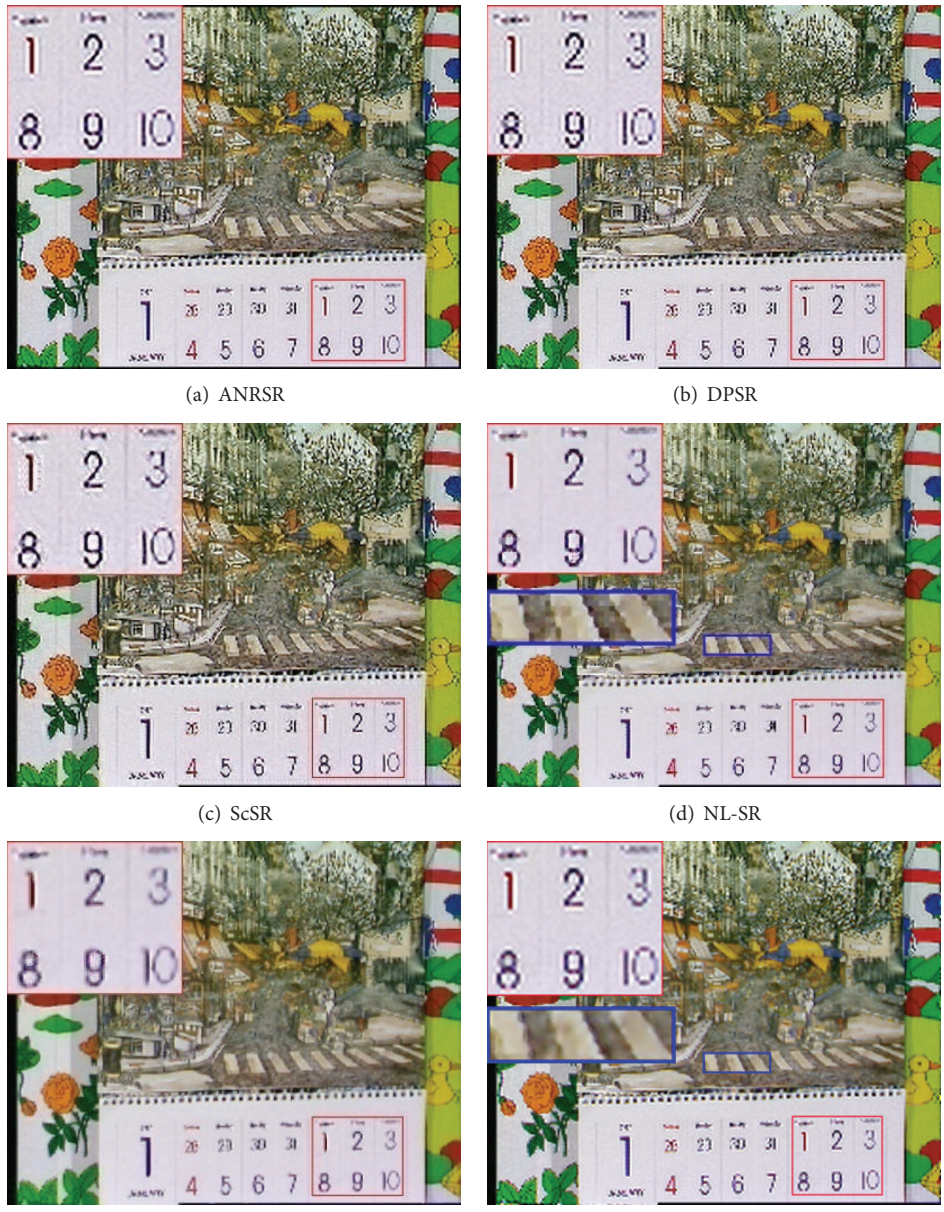
Table 2: Average PSNR index values of the six algorithms.

Video sequences	ANRSR	DPSR	ScSR	NL-SR	ZM-SR	LBST-SR
Satellite-1	32.6112	32.2610	32.2274	33.0043	33.0976	**34.3359**
Satellite-2	26.7393	27.6419	27.8432	27.8432	28.7216	**30.9721**
Forman	23.6094	28.8870	28.7007	29.3522	**30.0658**	30.0229
Calendar	21.8125	21.8218	21.7321	22.5146	23.3250	**23.3826**
Coastguard	23.5604	25.9583	25.8747	26.8223	27.7863	**28.5933**
Suzie	22.4611	27.5689	27.3860	27.5675	27.9919	**28.5858**
Mother_Daughter	23.3212	25.4622	25.4056	25.5966	**25.9141**	25.6664
Miss_America	25.4412	27.0984	26.5224	26.7219	27.0799	**27.3240**
Ice	20.0069	21.5826	21.4745	21.7167	21.8370	**21.8413**
Football	23.4603	24.3382	24.3644	25.1328	25.5998	**26.0958**
Carphone	21.6177	26.6009	26.1356	26.7163	**27.5862**	27.5311
Akiyo	26.3123	28.9223	28.9065	29.4792	**29.9924**	29.7895

Table 3: Average FSIM index values of the six algorithms.

Video sequences	ANRSR	DPSR	ScSR	NL-SR	ZM-SR	LBST-SR
Satellite-1	0.7817	0.7692	0.7781	0.7994	0.8082	**0.8139**
Satellite-2	0.8991	0.9035	0.9088	0.9088	0.9180	**0.9423**
Forman	0.8680	0.8977	0.9022	0.9214	0.9245	**0.9359**
Calendar	0.8675	0.8654	0.8654	0.8863	0.8912	**0.9027**
Coastguard	0.7852	0.8140	0.8207	0.8201	0.8399	**0.8529**
Suzie	0.8596	0.9177	0.9173	0.9291	0.9373	**0.9460**
Mother_Daughter	0.8419	0.8797	0.8830	0.8889	**0.9133**	0.9125
Miss_America	0.8819	0.9162	0.9138	0.9162	0.9307	**0.9393**
Ice	0.8330	0.8649	0.8740	0.9001	0.8978	**0.9032**
Football	0.8440	0.8405	0.8488	0.8545	0.8516	**0.8673**
Carphone	0.8368	0.8825	0.8793	0.8959	0.9114	**0.9143**
Akiyo	0.9019	0.9252	0.9264	0.9239	0.9353	**0.9360**

In terms of time efficiency of the spatiotemporal similarity matching process for ten benchmark video sequences and two spatial video sequences, the average time per video frame for the spatiotemporal nonlocal fuzzy registration scheme using PZM (ZFR) and the proposed improved nonlocal fuzzy registration scheme using PZM and structural similarity (ZSFR) is given in Table 5. Clearly, compared to ZFR scheme, the proposed PZSFR scheme improves time efficiency significantly while guaranteeing the similarity matching effect. The reason lies mainly in the use of a self-adaptive regional



(a) ANRSR (b) DPSR

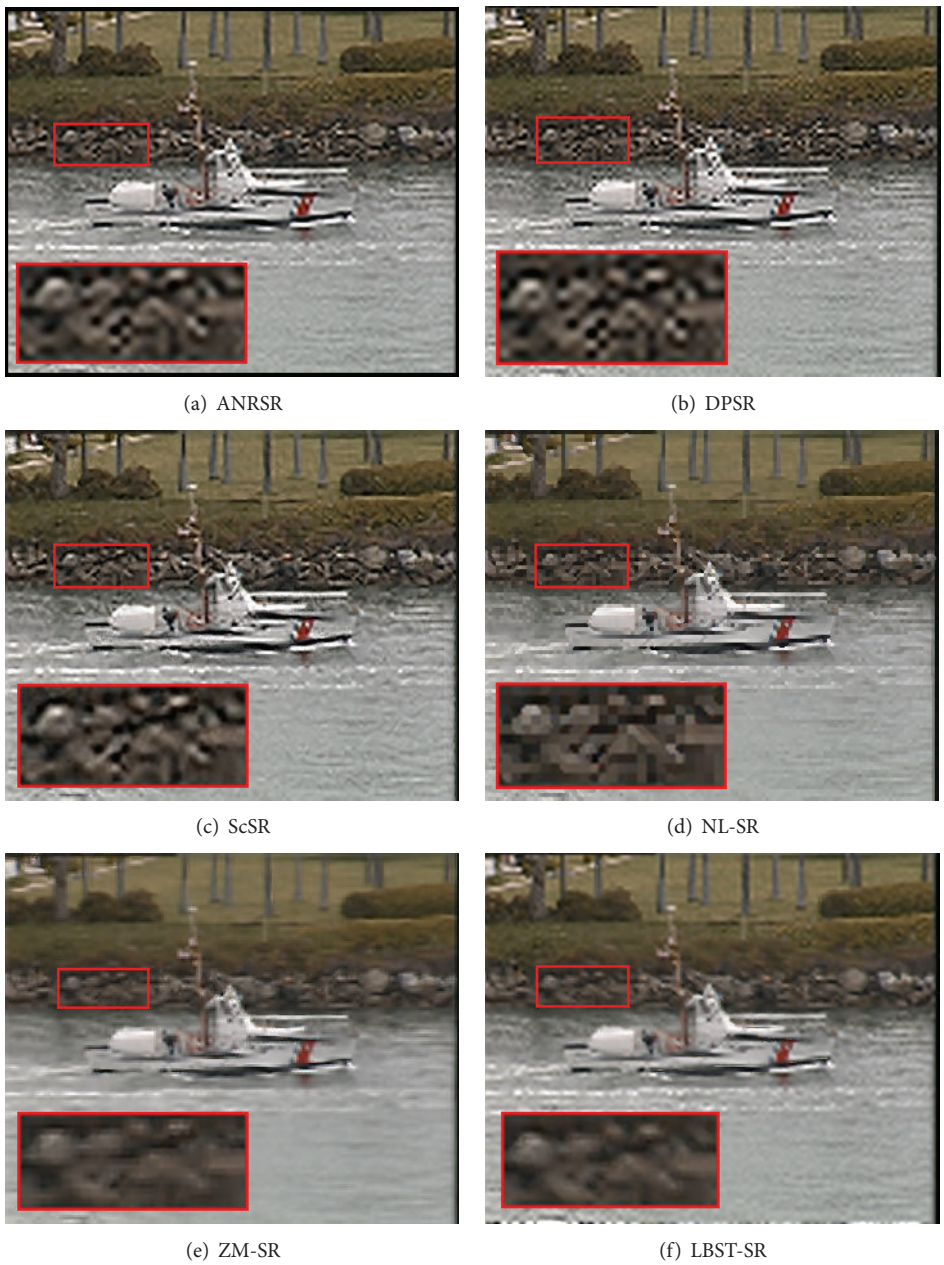
(c) ScSR (d) NL-SR

(e) ZM-SR (f) LBST-SR

Figure 9: SR reconstruction visual effects for Frame 29 of the "Calendar" sequence with a magnification factor of three.

correlation evaluation strategy based on regional average energy and regional structural similarity, which is an improvement over ZFR scheme.

4.2.2. Subjective Visual Evaluations. Figure 8 shows the SR reconstruction visual effects of the six algorithms (ANRSR, DPSR, ScSR, NL-SR, ZM-SR, and LBST-SR) for Frame 6 of the "Forman" sequence, with the magnified local textures marked by the red rectangular box. The frame contains moderate-motion objects (such as local motions of head and mouth and rotation motion of eyes) in the "Forman" sequence. By analyzing global and local detail effects (such as regions around the eyes), it is clear that the proposed LBST-SR algorithm obtains a better visual effect than the other five algorithms. The learning-based ANRSR, DPSR, and ScSR algorithms produce annoying spot artifacts and unnatural visual effects in the face regions. Edge detail blurring phenomena are produced in the ZM-SR algorithm. Some annoying block artifacts are generated in the NL-SR algorithm, which mainly occurred because local complex motions influenced the accuracy of nonlocal similarity matching and fusion between video frames. The proposed LBST-SR algorithm was able to solve this problem because the spatiotemporal similarity matching process can be adapted to complex motion patterns. In comparison, the proposed LBST-SR algorithm not only has clearer edges and contours but also produces smoother effects in the face part.



(a) ANRSR (b) DPSR

(c) ScSR (d) NL-SR

(e) ZM-SR (f) LBST-SR

Figure 10: SR reconstruction visual effects for Frame 18 of the "Coastguard" sequence with a magnification factor of three.

The superresolved results for Frame 29 of the "Calendar" sequence are shown in Figure 9, with the magnified local textures marked by red and blue rectangular boxes. The results demonstrate that the proposed algorithm generates the best visual effects and produces clearer contours and details. The "Calendar" sequence contains complex object motions, including translation motion, occluded areas, and newly appearing object areas. The proposed algorithm still performed well under such complex motion scenes, benefitting mainly from the improved spatiotemporal nonlocal fuzzy registration scheme based on PZM feature and structural similarity, which is robust to complex motion scenes. The local magnified details indicate that ANRSR, DPSR, and ScSR algorithms introduce noticeable annoying artifacts around the edges of each number. ZM-SR algorithm shows some blurring effects. In the local detail area marked by the red rectangle, the quality of the proposed algorithm is comparable to the NL-SR algorithm, but in the magnified road details area marked by the blue rectangular box, the proposed algorithm produces smoother effects, whereas discontinuous edges and annoying block artifacts are generated in the NL-SR algorithm.

Figure 10 shows the superresolved results for Frame 18 and the magnified local details of the "Coastguard" sequence. The "Coastguard" sequence contains the complex backgrounds and motions of both object and camera. Moreover,

(a) ANRSR (b) DPSR

(c) ScSR (d) NL-SR

(e) ZM-SR (f) LBST-SR

Figure 11: SR reconstruction visual effects for Frame 1 of the "Akiyo" sequence with a magnification factor of three.

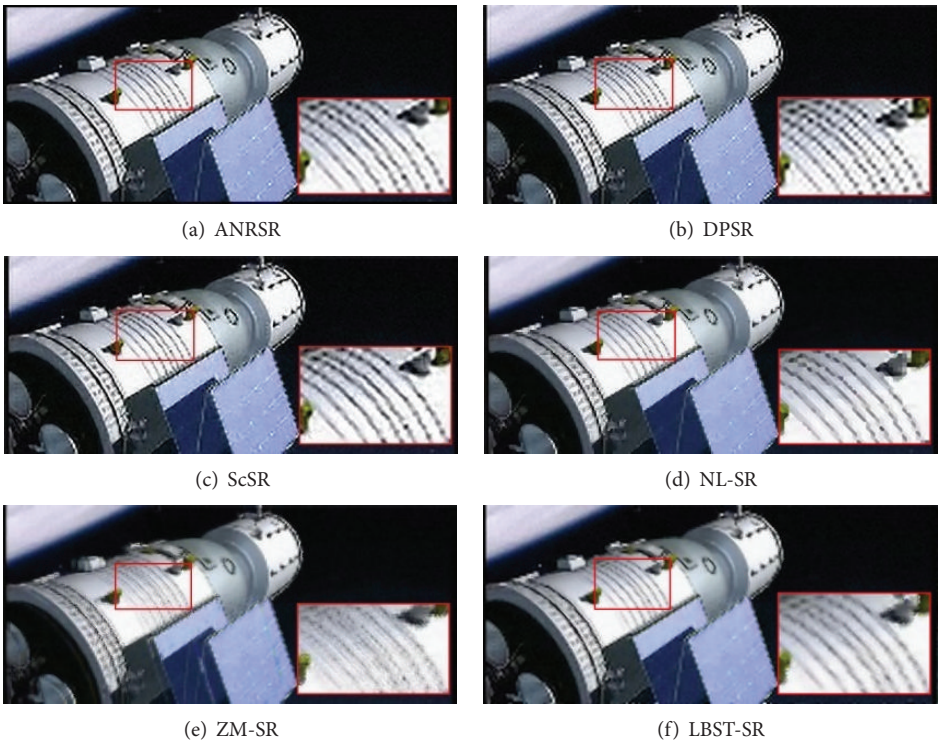
complex motions such as translation, occluded areas, and newly appearing object areas exist in this sequence. Under such complex motion scenes, the proposed LBST-SR algorithm still performed better than the other five algorithms. As can be observed from the magnified local details marked by the red rectangle and details in the background regions, annoying black spots and block artifacts are generated in the ANRSR, DPSR, and ScSR algorithms. ZM-SR algorithm produces blurred edges and details, especially in the complex stone bank background area. Annoying block artifacts and discontinuous edges are generated in the NL-SR algorithm because its nonlocal similarity matching strategy cannot be well adapted to the complex motion scenes.

The superresolved results for Frame 1 of the "Akiyo" sequence are shown in Figure 11, with local details magnified to emphasize visual quality. The magnified visual effects of the face region marked in the red rectangle demonstrate that the proposed algorithm is superior to the other five algorithms and produces a more natural and smoother visual effect. ANRSR, DPSR, and ScSR algorithms produce annoying artifacts in the face region and unnatural skin colors. NL-SR algorithm produces block effects. And some blurring phenomena are generated in the ZM-SR algorithm.

The "Satellite-2" sequence contains local motions, light variation, and more object details. Figure 12 shows the superresolved results for Frame 3 of the "Satellite-2" sequence.

Table 4: Average RMSE index values of the six algorithms.

Video sequences	ANRSR	DPSR	ScSR	NL-SR	ZM-SR	LBST-SR
Satellite-1	0.0027	0.0028	0.0028	0.0027	0.0026	**0.0023**
Satellite-2	0.0067	0.0060	0.0060	0.0059	0.0055	**0.0044**
Forman	0.0120	0.0072	0.0073	0.0072	**0.0065**	0.0067
Calendar	0.0080	0.0080	0.0080	0.0078	0.0071	**0.0070**
Coastguard	0.0113	0.0086	0.0087	0.0083	0.0073	**0.0067**
Suzie	0.0282	0.0170	0.0176	0.0176	0.0168	**0.0158**
Mother_Daughter	0.0159	0.0132	0.0132	0.0130	**0.0127**	0.0130
Miss_America	0.0269	0.0225	0.0240	0.0235	0.0225	**0.0220**
Ice	0.0216	0.0188	0.0191	0.0188	0.0186	**0.0186**
Football	0.0131	0.0116	0.0115	0.0113	0.0107	**0.0101**
Carphone	0.0300	0.0184	0.0192	0.0189	**0.0173**	0.0179
Akiyo	0.0102	0.0081	0.0081	0.0079	**0.0075**	0.0077



(a) ANRSR (b) DPSR (c) ScSR (d) NL-SR (e) ZM-SR (f) LBST-SR

Figure 12: SR reconstruction visual effects for Frame 3 of the "Satellite-2" sequence with a magnification factor of three.

The magnified details marked by the red rectangle demonstrate that the proposed LBST-SR algorithm generates the best visual effects, producing more natural visual effects and clearer details. Annoying black spot artifacts and unnatural visual effects are produced in the ANRSR, DPSR, and ScSR algorithms. Block artifacts and jagged effects are generated by the NL-SR algorithm, and ZM-SR algorithm produces blurred object details.

5. Conclusions

A novel learning-based algorithm to implement video SR reconstruction using spatiotemporal nonlocal similarity was proposed in this paper. On the basis of LR-HR correlation mapping, spatiotemporal nonlocal similarity structural redundancies were used to improve SR quality further. With the objective of improving algorithm efficiency while guaranteeing SR quality, LR-HR correlation mapping was performed only for the salient object region of the video frame, following which an improved spatiotemporal nonlocal fuzzy registration scheme was established for spatiotemporal similarity matching and fusion using the similarity weighting strategy based on pseudo-Zernike moment feature similarity and structural similarity and the self-adaptive regional correlation evaluation strategy. The proposed spatiotemporal nonlocal fuzzy registration scheme does not rely on accurate

Table 5: Comparison of time efficiency for ZFR and ZSFR schemes.

Video sequences	ZFR scheme (s)	ZSFR scheme (s)
Satellite-1	75.43	40.80
Satellite-2	52.85	25.56
Forman	41.09	18.78
Calendar	141.13	63.95
Coastguard	37.84	18.23
Suzie	8.91	4.58
Mother_Daughter	37.08	20.55
Miss_America	9.24	4.20
Ice	36.58	17.85
Football	35.65	16.70
Carphone	8.98	4.10
Akiyo	41.43	19.77

estimation of subpixel motion, and therefore it can be adapted to complex motion patterns and is robust to noise and rotation. Experimental results demonstrated that the proposed algorithm achieves competitive SR quality compared to other state-of-the-art algorithms in terms of both subjective and objective evaluations.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Basic Research Program of China (973 Program) 2012CB821200 (2012CB821206) and the National Natural Science Foundation of China (no. 61320106006).

References

[1] H. Su, L. Tang, Y. Wu, D. Tretter, and J. Zhou, "Spatially adaptive block-based super-resolution," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1031–1045, 2012.

[2] C. Liu and D. Sun, "On bayesian adaptive video super resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 346–360, 2014.

[3] L. Zhou, X. Lu, and L. Yang, "A local structure adaptive super-resolution reconstruction method based on BTV regularization," *Multimedia Tools and Applications*, vol. 71, no. 3, pp. 1879–1892, 2014.

[4] J. Chen, J. Nunez-Yanez, and A. Achim, "Video super-resolution using generalized Gaussian Markov random fields," *IEEE Signal Processing Letters*, vol. 19, no. 2, pp. 63–66, 2012.

[5] A. Giachetti and A. Nicola, "Real-time artifact-free image upscaling," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2361–2369, 2012.

[6] W. Dong, L. Zhang, R. Lukac, and G. Shi, "Sparse representation based image interpolation with nonlocal autoregressive modeling," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1382–1394, 2013.

[7] U. Mudenagudi, S. Banerjee, and P. K. Kalra, "Space-time super-resolution using graph-cut optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 995–1008, 2011.

[8] J. Chen, J. Nunez-Yanez, and A. Achim, "Video super-resolution using generalized gaussian markov random fields," *IEEE Signal Processing Letters*, vol. 19, no. 2, pp. 63–66, 2012.

[9] A. Maalouf and M.-C. Larabi, "Colour image super-resolution using geometric grouplets," *IET Image Processing*, vol. 6, no. 2, pp. 168–180, 2012.

[10] Z. Liu, Q. Huang, J. Li, and Q. Wang, "Single image super-resolution via l 0 image smoothing," *Mathematical Problems in Engineering*, vol. 2014, Article ID 974509, 8 pages, 2014.

[11] J. Yang, Z. Lin, and S. Cohen, "Fast image super-resolution based on in-place example regression," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 1059–1066, IEEE, Portland, Ore, USA, June 2013.

[12] J. Salvador, A. Kochale, and S. Schweidler, "Patch-based spatio-temporal super-resolution for video with non-rigid motion," *Signal Processing: Image Communication*, vol. 28, no. 5, pp. 483–493, 2013.

[13] H. Su, L. Tang, Y. Wu, D. Tretter, and J. Zhou, "Spatially adaptive block-based super-resolution," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1031–1045, 2012.

[14] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Transactions on Image Processing*, vol. 18, no. 1, pp. 36–51, 2009.

[15] N. Dowson and O. Salvado, "Hashed nonlocal means for rapid image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 485–499, 2011.

[16] X. Gao, Q. Wang, X. Li, D. Tao, and K. Zhang, "Zernike-moment-based image super resolution," *IEEE Transactions on Image Processing*, vol. 20, no. 10, pp. 2738–2747, 2011.

[17] M.-C. Yang and Y.-C. F. Wang, "A self-learning approach to single image super-resolution," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 498–508, 2013.

[18] F. Zhou, W. Yang, and Q. Liao, "Single image super-resolution using incoherent sub-dictionaries learning," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 891–897, 2012.

[19] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: a benchmark," in *Proceedings of the European Conference on Computer Vision (ECCV '14)*, pp. 372–386, September 2014.

[20] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, pp. 275–282, July 2004.

[21] X. Gao, K. Zhang, D. Tao, and X. Li, "Image super-resolution with sparse neighbor embedding," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3194–3205, 2012.

[22] R. Timofte, V. De, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proceedings of the 14th IEEE International Conference on Computer Vision (ICCV '13)*, pp. 1920–1927, December 2013.

[23] K. Su, Q. Tian, Q. Xue, N. Sebe, and J. Ma, "Neighborhood issue in single-frame image super-resolution," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '05)*, pp. 1–4, IEEE, July 2005.

[24] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3467–3478, 2012.

[25] J. Yang, Z. Wang, Z. Lin, X. Shu, and T. Huang, "Bilevel sparse coding for coupled feature spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2360–2367, IEEE, Providence, RI, USA, June 2012.

[26] D. Lin and X. Tang, "Coupled space learning of image style transformation," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 1699–1706, Beijing, China, October 2005.

[27] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[28] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2216–2223, June 2012.

[29] L. He, H. Qi, and R. Zaretzki, "Beta process joint dictionary learning for coupled feature spaces with application to single image super-resolution," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 345–352, IEEE, June 2013.

[30] Y. Zhu, Y. Zhang, and A. L. Yuille, "Single image super-resolution using deformable patches," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 2917–2924, IEEE, Columbus, Ohio, USA, June 2014.

[31] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 2814–2821, June 2014.

[32] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.