remote sensing

Article

Learned Representation of Satellite Image Series for Data Compression

Liang Liao [1,2], Jing Xiao [1,*], Yating Li [3], Mi Wang [4] and Ruimin Hu [3]

[1] School of Computer Science, Wuhan University, Wuhan 430072, China; liang@nii.ac.jp
[2] National Institute of Informatics, Tokyo 101-8430, Japan
[3] National Engineering Research Center for Multimedia Software, Wuhan 430072, China; greenallee@whu.edu.cn (Y.L.); hrm@whu.edu.cn (R.H.)
[4] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan 430072, China; miwang@whu.edu.cn
* Correspondence: jing@whu.edu.cn

Received: 4 January 2020; Accepted: 31 January 2020; Published: 4 February 2020

Abstract: Real-time transmission of satellite video data is one of the fundamentals in the applications of video satellite. Making use of the historical information to eliminate the long-term background redundancy (LBR) is considered to be a crucial way to bridge the gap between the compressed data rate and the bandwidth between the satellite and the Earth. The main challenge lies in how to deal with the variant image pixel values caused by the change of shooting conditions while keeping the structure of the same landscape unchanged. In this paper, we propose a representation learning based method to model the complex evolution of the landscape appearance under different conditions by making use of the historical image series. Under this representation model, the image is disentangled into the content part and the style part. The former represents the consistent landscape structure, while the latter represents the conditional parameters of the environment. To utilize the knowledge learned from the historical image series, we generate synthetic reference frames for the compression of video frames through image translation by the representation model. The synthetic reference frames can highly boost the compression efficiency by changing the original intra-frame prediction to inter-frame prediction for the intra-coded picture (I frame). Experimental results show that the proposed representation learning-based compression method can save an average of 44.22% bits over HEVC, which is significantly higher than that using references generated under the same conditions. Bitrate savings reached 18.07% when applied to satellite video data with arbitrarily collected reference images.

Keywords: disentangled representation; image-to-image translation; time series data; high efficiency compression

1. Introduction

Currently, video satellite remote sensing has become a new trend in smart city development. Low-Earth-orbit satellites are released in an increasing number to record and monitor the landscape from space, which drives more and more extensive applications of the satellite video, such as smart transportation and disaster management. Compared to the hyperspectral image, satellite video can continuously observe a specific place, and it is helpful to detect and respond to unusual activities in time. However, owing to the gap of a large amount of data stream and the real-time demands of remote sensing data analysis, the remote surveillance applications in smart cities are greatly restricted.

Unlike the conventional city surveillance videos, satellite videos are usually taken with high resolution, for example 12,000 × 5000 for the Jilin-1 satellite videos. It can cover a large surface with great details in one frame, but it results in a large amount of video data, which is hard to transmit promptly by the current transmission channel between the satellite and Earth. Even compressing them with the latest coding standard high-efficiency video coding (HEVC) [1], it still cannot meet the needs of the real-time analysis demands [2]. More efficient and specific data compression techniques for satellite videos are in high demand.

To achieve a high compression ratio of the satellite images and videos, the dictionary-based sparse representation methods are proposed. Benefiting from the optimization algorithm, such as K-SVD [3–5] and low-rank matrix recovery theory [6–8], those methods train a dictionary based on machine learning to represent satellite data sparsely. Through constructing and optimizing the dictionary model, the remote sensing data can be efficiently represented by some atoms of the dictionary. Another type of representative work proposed to use the image prior exists in the historical images. Liu et al. [9] proposed to use a collection of historical remote sensing images as prior knowledge and introduce image feature extraction and registration to compress images. Tao et al. [10] further combined the same prior knowledge and Bayesian dictionary learning to compress remote sensing images.

Xiao et al. [2,11] discovered a new type of redundancy that exists in multi-source satellite videos, named long-term background redundancy (LBR). This type of redundancy is caused by a similar background across different images and videos captured at the same land location from different times. As the number of video clips shooting at the corresponding area increases, the redundancy across multiple video clips becomes significant. In their work, the LBR can be eliminated by creating a long-term background referencing library, containing high-definition geometrically registered images of the entire area. Then, the synthetic frames can be generated based on geometric matching, radiometric adjustment, and quality adjustment, which were further used as the reference to eliminate the LBR. The long-term background referencing library has shown its ability to represent the historical background prior to the same landscape. However, the radiometric adjustment or quality adjustment can only change the tone of the whole image, but cannot deal with the appearance change caused by seasonal variation or diurnal variation. The synthetic frame cannot achieve pixel-wise similarity to the target frame. Thus, the compression gain from the synthetic reference frame will be decreased.

In this paper, we offer a representation learning-based method to model the complex evolution of the landscape appearance under different conditions by making use of the historical image series. Each series contains multi-temporal images at the same location, namely providing different landscape appearances under various daily and seasonal conditions. The aim is to utilize the learned representation model to generate a more precise reference frame to the target frame from historical images. The challenge lies in how to deal with the variant image pixel values caused by the change of shooting conditions while keeping the structure of the same landscape unchanged. Inspired by the disentangled representation learning [12–16], we propose a disentangled representation learning of satellite image series. The description of the same landscape is determined by separating the stable structure from the changing environment. Based on that, the reference frame for the current target frame is generated by combining the stable structure extracted from the reference library and the current environmental parameters obtained from the current target frame. The synthetic reference frame can be adopted into the compression framework for satellite video compression proposed by Xiao et al. [2].

Experiments are conducted on real video clips from video satellites to evaluate the performance of the proposed method. The results reveal that the proposed method could achieve 44.22% bitrate savings on average over the main profile of HEVC compared with 30.95% bitrate savings in [2].

There are three main contributions of this work:

(1) We propose a new representation model for the remote sensing image series, which can model the consistent landscape, as well as the variational shooting conditions.

(2) We integrate the learned reorientation of historical images into the satellite video compression framework through the generation of reference frames.

(3) Our proposed video compression method outperforms the state-of-the-art compression scheme.

The remainder of this paper is organized as follows: Section 2 provides a literature review regarding related work. A detailed introduction of the proposed synthetic reference image generation method is illustrated in Section 3. Section 4 reports our experimental results, and Section 5 concludes the paper.

2. Related Work

Our work is related to the current video compression method, the representation method of historical image series, and the multi-model image-to-image translation method. Therefore, we review the related work from these three aspects.

2.1. Compression Method on Satellite Videos

In recent years, the video form of satellite data, which can capture the correlation between the images, has gained increasing popularity. To compress the satellite videos, the general video compression methods have been applied to video satellites. For example, the video compression standard H.264 [17] was equipped in the Skysat [18], which was launched by Skybox and captured 1280 × 720 (720P) of satellite videos. Satellite Jilin-1 was outfitted with the latest video compression standard HEVC. Those video compression standards aim to eliminate the spatial-temporal redundancy that exists in the intra- and inter-frames of one video, which cannot remove the long-term background redundancy from multi-source satellite video. Xiao et al. [2] and Wang et al. [11] proposed the historical knowledge library-based video compression methods. They constructed a long-term background reference library composed of high-resolution historical images. By adjusting the historical images to the target images for reference, those methods can get a tremendous increasing compression ratio, especially in terms of the intra-coded picture (I frame). However, they adopted the linear model in the pixel domain, such as radiometric adjustment and quality adjustment, to translate the historical image, which cannot deal with the complex change of the appearance of the same landscape.

2.2. Compression Method of Image Series

Similar to the characteristic of satellite videos, the image set also contains a large number of similar images with common objects or backgrounds. Multiple shooting devices usually capture them at various times from different positions. To compress this type of data, some image series representation methods were proposed [19–23]. Yue et al. [19] proposed to retrieve the most similar historical data from the cloud as a reference. Shi et al. [20] proposed to sort the images via mining the similarities between images and reorganizing the image series into the video sequence, then coding them like a video. To compress the repetitive moving object in the surveillance videos, Xiao et al. [24,25] proposed to construct one vehicle knowledge library and generate the synthetic reference object based on perspective transformation and residual compression. Chen et al. [26,27] further improved this work by lighting compensation and texture mapping. Those methods provide ways to represent relations between multiple historical image series. Still, their computational complexity is high, and the actual differences between the historical image series of satellite videos make the satellite data compression hard to find similarities at the pixel level.

2.3. Image-to-Image Translation Method

The color correction of the registered historical image can be regarded as the image-to-image translation. That is to say, images taken at any time of a location can be generated from a historical image of that location. The image structure remains unchanged while the image color is changed. Recently, Sanchez et al. [28], Gonzalez-Garcia et al. [29], Zhu et al. [30,31], and Huang et al. [12] presented work that dealt with multi-modal output. In particular, Zhu et al. [31] invented a CycleGAN model [30] to address the diversity of output. Sanchez et al. [28] proposed a cross-domain autoencoder model that integrated the variational autoencoder (VAE) and the generative adversarial network (GAN). Those methods performed well for artistic image translation. However, they often generate structural artifacts, which makes it difficult to apply them directly for generating reference images. According to the features of satellite videos, we propose a new translation model using satellite images to learn the reference frame generation task.

3. Method

3.1. Preliminary on Reference Image Generation

Preliminary studies on using the prior knowledge to improve the compression efficiency mainly focus on using images as references. In this work, the reference image can be an original historical image [32,33] or a synthesized image [24,26,34]. In our previous works [2], we used the current Google Map image as the historical reference. The reference was adjusted to generate the reference image for the compression of the target frame. Before introducing the proposed method of this paper, we first do a brief review on how to generate the reference frame of our previous work. In this paper, we make an improvement in reference frame generation to provide a more similar reference to the target frame to improve the coding efficiency.

Search for historical image: The overall framework is presented in Figure 1a. The whole process of generating the reference frame consisted of four steps: image selection, geometrical matching, radiometric adjustment, and quality adjustment. After selecting the geographically corresponding historical image from the image library, we conducted geometrical matching from the historical image to the target frame I_f, so that the reference image was geometrically fit to the target frame. Then, the radiometric adjustment was used to compensate for the color difference between the reference and the target frame. After that, the quality was adjusted to match the quality of the reference image (e.g., blurry) to the target image, which might be caused by the imaging platform. After this step, the reference image was ready for the encoding process.

The searching process is shown in Figure 2. The longitude and latitude of the center of the captured area were used to locate the current position. According to the coordinates, the lens focus of the camera, and flight altitude of the satellite, the range of the captured area could be roughly obtained, which is shown as the orange area in Figure 2. Taking the location error into account, the coverage of historical image was a little larger than that of the captured area. The final searched historical image from Google Earth is shown as the blue area in Figure 2, which is denoted as I_h.

Geometric matching: Due to the offset of shooting angle and the error of location, there will be a certain deviation of position between the historical image I_h and the captured frame I_f. Thus, feature registration was applied to I_h. Specifically, the Scale-Invariant Feature Transform (SIFT) algorithm was employed to extract feature points of I_h and I_f. Feature points were composed of feature descriptors and positions of I_h and I_f, which are denoted as f^{I_h},p^{I_h} and f^{I_f},p^{I_f}, respectively. By calculating the Euclidean distance between the feature descriptor pairs $\left\{f_i^{I_h}, f_j^{I_f}\right\}$, the feature points were matched into pairs $P_k = \left\{p_i^{I_h}, p_j^{I_f}\right\}$. The RANdom SAmple Consensus (RANSAC) matching scheme was applied for an affine transformation matrix M for I_h to I_f. Finally, the registered background reference I_g was obtained by applying M to I_h.

Radiometric adjustment: To adjust the color difference between the target frame and the reference image, a transfer model [35] was used. Different from the original $l\alpha\beta$ space, the video frames were recorded in the YUV space, in which the first channel was lightness and the other two channels were color components. We modified the transfer model into the YUV space by:

$$\begin{bmatrix} Y_c \\ U_c \\ V_c \end{bmatrix} = \begin{bmatrix} \frac{\sigma_f^Y}{\sigma_g^Y} & 0 & 0 \\ 0 & \frac{\sigma_f^U}{\sigma_g^U} & 0 \\ 0 & 0 & \frac{\sigma_f^V}{\sigma_g^V} \end{bmatrix} \left(\begin{bmatrix} Y_g \\ U_g \\ V_g \end{bmatrix} - \begin{bmatrix} \overline{Y_g} \\ \overline{U_g} \\ \overline{V_g} \end{bmatrix} \right) + \begin{bmatrix} \overline{Y_f} \\ \overline{U_f} \\ \overline{V_f} \end{bmatrix} \tag{1}$$

where $\begin{bmatrix} Y_c & U_c & V_c \end{bmatrix}^T$ and $\begin{bmatrix} Y_g & U_g & V_g \end{bmatrix}^T$ are the color values in the reference image after radiometric adjustment I_c and the previous reference image I_g, respectively. $\begin{bmatrix} \overline{Y_g} & \overline{U_g} & \overline{V_g} \end{bmatrix}^T$ and $\begin{bmatrix} \sigma_g^Y & \sigma_g^U & \sigma_g^V \end{bmatrix}^T$ are the mean and standard deviations of YUV from I_g, while $\begin{bmatrix} \overline{Y_f} & \overline{U_f} & \overline{V_f} \end{bmatrix}^T$ and $\begin{bmatrix} \sigma_f^Y & \sigma_f^U & \sigma_f^V \end{bmatrix}^T$ are the values from the target frame I_f. By using this model, the color of the reference I_r^g was adjusted according to the color statistics of the current frame.

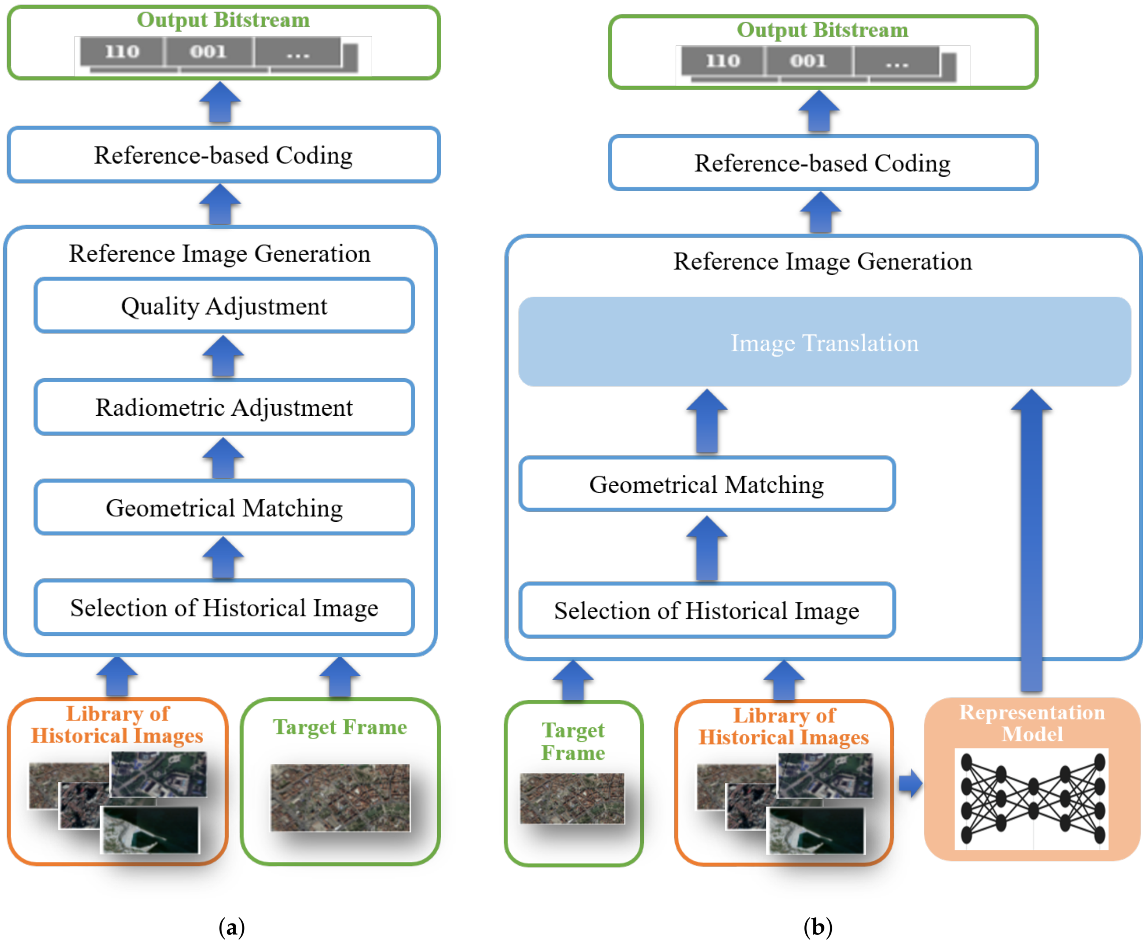


Figure 1. Comparison of the overall-frameworks. (**a**) The framework proposed in the previous method [2]; (**b**) the proposed process in this paper.

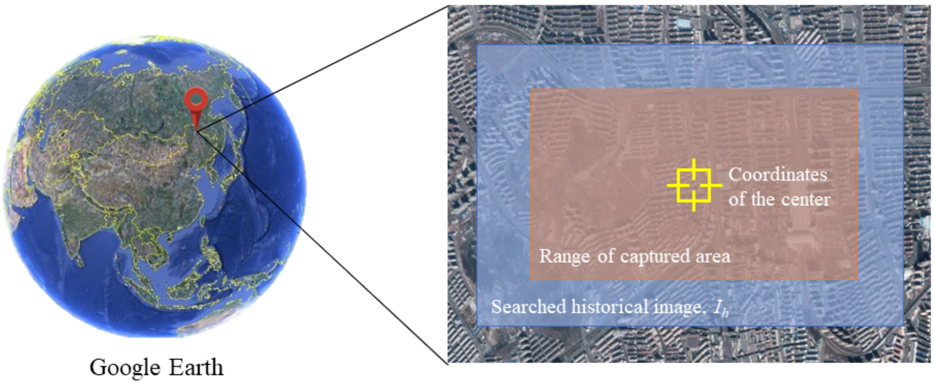


Figure 2. The process of searching for historical images.

Quality adjustment: To solve the problem that the Google reference usually has higher quality than the video frames, the quality of reference image I_c should be adjusted. The isotropic 2D Gaussian blur filter was used to simulate quality degradation of the video frame. The mean of Gaussian distribution was set to zero. The standard deviation σ was determined by minimizing the pixel value differences between I_c and I_f:

$$\arg\min_{\sigma} \sum_{k=0}^{N} (G(\sigma) * B_r^k - p_c^k) \tag{2}$$

where B_r^k is the k^{th} block from I_r^c and p_c^k is the k^{th} corresponding pixel value from the current frame. $G(\sigma)$ is the Gaussian kernel, whose values are defined by the parameter σ as follows.

After the Gaussian blur, we obtained the final reference image I_r, which will be used for the prediction in the encoding framework.

3.2. Representation Learning-Based Reference Image Generation

The inter-prediction in the existing framework of video compression highly depends on the similarity of pixels in each coding-block. However, the radiometric and quality adjustment in the previous method is a global process, which cannot provide an optimal reference for every coding-block. Therefore, instead of using only one historical image for the reference, we propose to utilize a series of historical images to improve the quality of the reference image.

Let $\boldsymbol{L} = \{L_1, L_2, ..., L_k\}$ be the image set of k locations. At each location $i (i \leq k)$, we have a series of l_i images $L_i = \left\{I_i^1, I_i^2, ..., I_i^{l_1}\right\}$. Those images from the same location share the same landscape, while the shooting conditions change. Therefore, we attempt to learn a representation of the image series. The representation consists of a content feature and a style feature for each image. The representation model learns to extract the same content feature within an image series of a location, but uses different style features to represent the individual variations. In this way, the content feature records the consistent landscape structure of a specific location, and the style feature records the variations.

Take a target frame I_f and its co-located image series $L_f = \left\{I_f^1, I_f^2, ..., I_i^{l_f}\right\}$ for example. L_f exists at both the encoding side and the decoding side of the compression. At the encoding side, the target frame I_f is input into the representation model to obtain its style feature. The style feature is then transmitted to the decoder side. Then, at the decoder side, the content code of this location is extracted from any image of L_f, which should be the same as the content code of the target frame. It combines with the transmitted style feature to reconstruct the reference image I_r.

The advantage of this representation is that the shared content feature can be treated as the redundancy among the images of the same location. Namely, it does not need to be transmitted. The style code records the individual variation of target frame, which needs to be transmitted for the reconstruction of the reference image. However, compared to the content feature, the length of the style feature is very short, which is friendly for encoding.

3.2.1. Network Architecture

In order to learn a disentangled representation of an image series, we chose to use an unsupervised disentanglement model similar to the one proposed in [12]. In the model, one image is decomposed into a content domain and a style domain. We assumed that images taken at any time in region c can be generated from the image at another time. In other words, the two images should have the same content feature, and they can convert to each other after exchanging their style features. Therefore, we propose a model to learn the mappings $M_1 : x \rightarrow y$ and $M_2 : y \rightarrow x$. As is shown in Figure 3, the content encoder learns the function $Enc_c : \mathcal{I} \rightarrow \mathcal{C}$. Taking x and y as inputs respectively, the corresponding outputs are Enc_c^x and Enc_c^y. Similarly, the style encoder learns the function $Enc_s : \mathcal{I} \rightarrow \mathcal{S}$ and extracts the corresponding features Enc_s^x, Enc_s^y from x and y. Then, Enc_c^x and Enc_s^y are combined to pass through the decoder, whose function is $Dec : \mathcal{C}, \mathcal{S} \rightarrow \mathcal{I}$. The generated image $Dec\left(Enc_c^x, Enc_s^y\right)$ is the reconstruction of image y, denoted as $\hat{y}$. The reconstruction of image x is generated in a similar process. The content encoder, style encoder, and the decoder are shared between images. In this way, the content encoder is learned to extract common information (the landscape) between remote sensing images, whilst the style encoder is learned to record the independent information (other factors).

To enforce the representation learning of the historical image series, evaluations on the common content features and the quality of reconstructed images are employed. A discriminator function $D : \mathcal{I} \rightarrow \{0, 1\}$ is introduced to tell whether the generated image is fit to the distribution learned from the historical images.

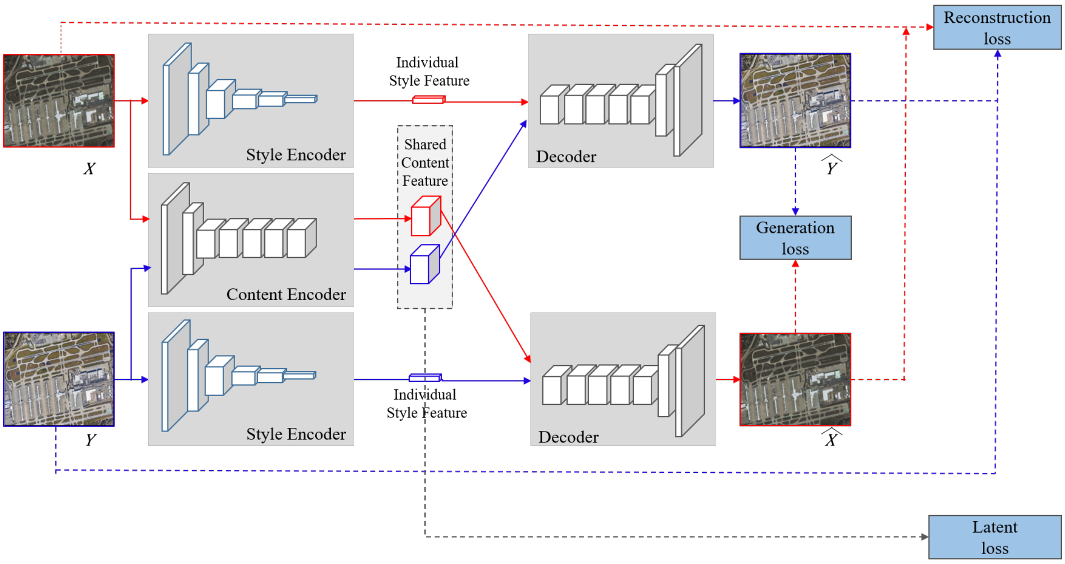


Figure 3. The overview of the proposed representation learning network.

3.2.2. Objective Function

To train the representation learning network, the loss function is composed of several terms. The shared content feature loss is used to enforce the content encoder to learn the common features. The reconstruction loss is used to apply the learning of other factors and to ensure that the encoders and decoders are inverses. Moreover, the adversarial loss is used to restrict the distribution of translated images to the image distribution of historic images.

- Shared content feature loss: As for the shared content features, Enc_c^x and Enc_c^y must be identical, i.e., $Enc_c^x = Enc_c^y$. Therefore, our goal is to minimize the L_1 distance between them. The loss function of the content encoder is defined as:

$$\mathcal{L}_c = \mathbb{E}_{x,y\sim\mathcal{I}}\left[\left|\left|Enc_c^x - Enc_c^y\right|\right|\right]. \tag{3}$$

- Reconstruction loss: The reconstruction loss contains two meanings. The first is that the disentangled image x can be reconstructed by its content feature Enc_c^x and its style feature Enc_s^x. In this case, the encoder and decoder form an auto-encoder model. The first term of reconstruction loss is the self-reconstruction of each image. Taking x for example, the self-reconstruction loss is defined as:

$$\mathcal{L}_{self}^x = \mathbb{E}_{x\sim\mathcal{I}}\left[\left|\left|Dec\left(Enc_c^x, Enc_s^x\right) - x\right|\right|\right]. \tag{4}$$

  Image y is the same.

  The second term for reconstruction loss is a cross-reconstruction loss. The exclusive feature only contains the particular information of each image. Therefore, the reconstruction of image x, $Dec\left(Enc^{y}_{c}, Enc^{x}_{s}\right)$ should be similar to x. For the sake of disentangled representation, the reconstruction loss of x is:

$$\mathcal{L}^{x}_{cross} = \mathbb{E}_{x,y\sim\mathcal{I}}\left[\left\|Dec\left(Enc^{y}_{c}, Enc^{x}_{s}\right) - x\right\|\right]. \tag{5}$$

Therefore, the overall reconstruction loss is the sum of two reconstruction loss terms:

$$\mathcal{L}_{rec} = \mathcal{L}^{x}_{self} + \mathcal{L}^{y}_{self} + \mathcal{L}^{x}_{cross} + \mathcal{L}^{y}_{cross}. \tag{6}$$

- Adversarial loss: In order to force the distribution of reconstruction to be close to the distribution of the real image, PatchGAN [36] was adopted for our model. The decoder can be regarded as the generator, and it is trained to reconstruct images, which can be classified as true by the discriminator, i.e., $Disc\left(Dec\left(Enc^{y}_{c}, Enc^{x}_{s}\right)\right) \rightarrow 1$, while the discriminator is trained to tell the reconstructed images from the real images, i.e., $Disc\left(Dec\left(Enc^{y}_{c}, Enc^{x}_{s}\right)\right) \rightarrow 0, Disc\left(x\right) \rightarrow 1$. The loss of GAN is defined as:

$$\mathcal{L}^{x}_{GAN} = \mathbb{E}_{x\sim\mathcal{I}}\left[\left(Disc(x)\right)^{2}\right] + \mathbb{E}_{x,y\sim\mathcal{I}}\left[\left(1 - Disc\left(Dec\left(Enc^{y}_{sh}, Enc^{x}_{ex}\right)\right)\right)^{2}\right]. \tag{7}$$

  The generator tries to maximize the loss while the discriminator tries to minimize it. The overall adversarial loss is the sum of both reconstructed images x and y:

$$\mathcal{L}_{GAN} = \mathcal{L}^{x}_{GAN} + \mathcal{L}^{y}_{GAN}. \tag{8}$$

- Total loss: Our model is trained jointly in an end-to-end learning manner. The objective is to minimize the following total loss defined as:

$$\mathcal{L} = \mathcal{L}_{GAN} + w_{rec}\mathcal{L}_{rec} + w_{c}\mathcal{L}_{c} \tag{9}$$

  where w_{rec} and w_{c} are the constant weights of corresponding losses.

3.2.3. Representation Learning-Based Image Translation

The procedure of image translation based on the learned representation model is shown in Figure 4. We first obtain the selected historical image after geometric matching I_g. Then, it is used to extract shared feature $Enc^{I_r}_{c}$ by the trained content encoder. The target frame I_f is used to extract style feature $Enc^{I_f}_{s}$. Then, both of them are passed through the decoder, generating the reconstruction $Dec\left(Enc^{I_g}_{c}, Enc^{I_f}_{s}\right)$. Since the reconstruction contains both the structure of I_g and the style feature of I_f, the generated image I_r is used as the reference image.

3.3. Improvement of the Compression Scheme

A video clip contains the intra-coded picture (I frame), predicted picture (P frame), and bidirectional predicted picture (B frame). Typically, the I frame is compressed by intra-frame coding, which costs a higher bitrate than inter-frame coding. The other two types of frame are usually compressed by inter-frame coding. Particularly for the satellite video, the spatial correlation is weak, while the long-term temporal correlation is strong due to the slowly moving background. Generally, the bitrate cost of the I frame is about 2–10 times over that of the P frame. Therefore, using the generated background reference as the reference frame to remove the LBR of I frames can significantly improve the coding efficiency.

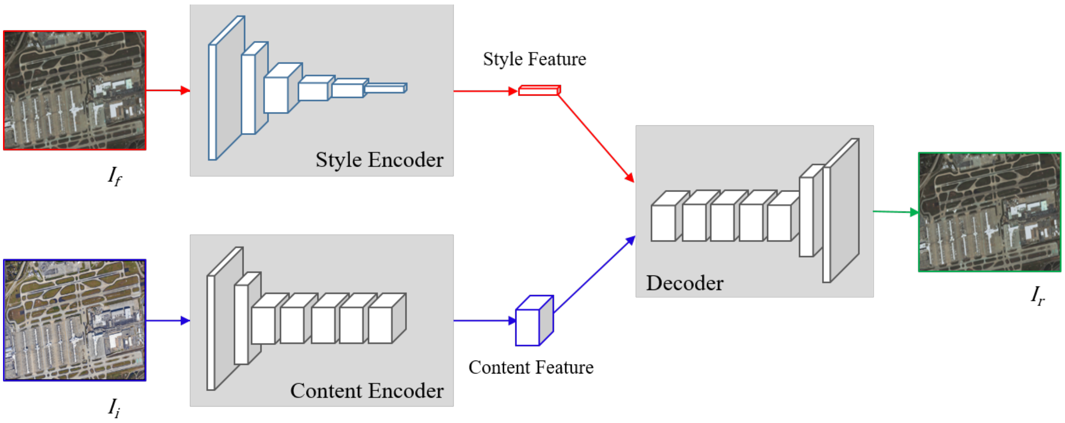


Figure 4. The flowchart of generating a synthetic reference image by using a historical image and the target frame.

The encoding procedure is sketched in Figure 5. Technically, The proposed representation learning model was trained on the server with the historical image series. The model was stored in both the encoding side (the satellite) and the decoding side (the Earth). For each target I frame, one historical image covering the shooting area was selected and was geometrically transformed to the I frame with a transformation matrix M. Then, the image passed the content encoder of the representation model to extract the content feature of this area. In the meantime, the target I frame also went through the style encoder of the representation model to obtain the style feature. Afterward, the content feature and the style feature were input into the decoder of the representation model to reconstruct the reference image. The translation procedure needed only one forward step. We adopted the method in [37] for the encoding with the reference image for the I frame. Besides the output bitstream of the frame, we also needed to transmit the transformation matrix and the style code to the decoding side, which was compressed using the Lempel–Ziv–Welch method [38].

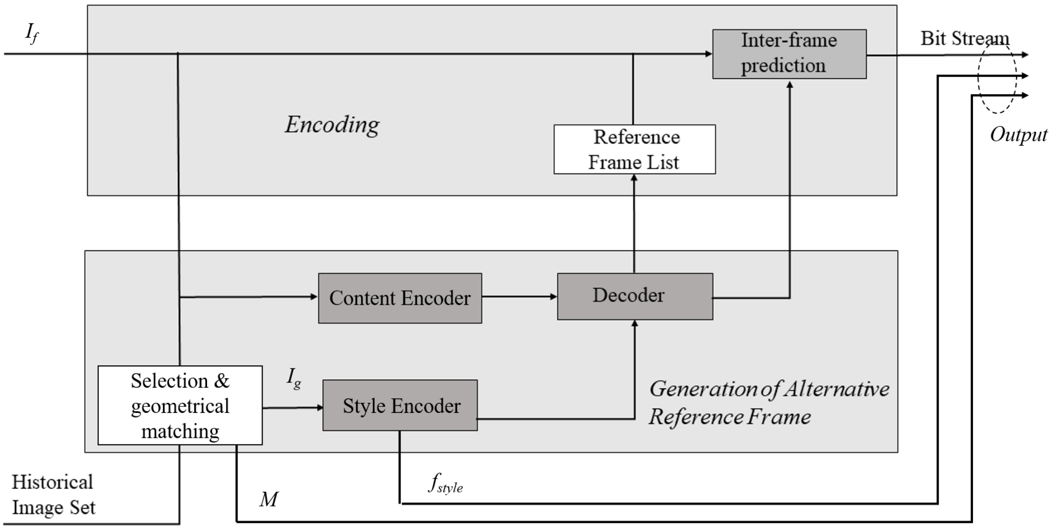


Figure 5. The framework of our proposed coding scheme. We use the learned representation model to generate reference frame. With the synthetic reference frame, we adopt inter-frame compression for both the intra-coded picture (I frame) and predicted picture (P frame).

At the decoding side, the selected image was firstly geometrically transformed using the transformation matrix. Then, its content feature was extracted by the content encoder of the representation model, which would be combined with the transmitted style feature to reconstruct the reference image. By using the reference image, the target frame could be decoded.

If the geometric structure of the current frame changed from the historical images, the current encoding process would not be affected since the wrongly generated area of the synthesized reference would not be used in the inter-frame prediction. It would result in a little decrease of the compression ratio.

In order to update the library of the structural change, we compared the content features from both the current frame and the selected historical image based on the following cross-correlation metric:

$$d(Enc_c^x, Enc_c^{ref}) = \frac{< Enc_c^x, Enc_c^{ref} >}{\|Enc_c^x\| \cdot \|Enc_c^{ref}\|}. \tag{10}$$

where Enc_c^x is the content feature from the current frame and Enc_c^{ref} is the content feature from the selected historical image.

If the difference of the content features exceeded a threshold τ, the newly reconstructed frame from the encoded bitrate would be added into the series of the historical image. The parameter τ was set based on the deviations from the content features of the image series of the same geometric structures. We used the reconstructed frame from the encoded bitrate to update the library because we could only have the reconstructed frame on the Earth for library updating.

4. Experimental Results and Discussion

4.1. Experimental Setup

4.1.1. Implementation Details

We implemented this model using the Pytorch toolbox and optimized the translation net and the discriminator using the Adam algorithm [39] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a learning rate of 0.0001 following [40]. In the training period, we use a batch size of eight, and the training is stopped after 1,000,000 iterations. The loss weight λ_{rec} and λ_c were set to 100 and 50, respectively. We chose the dimension of the style code to be eight across all datasets in [12]. We adopted the model trained on the SYNTHIA dataset [41] by Huang et al. [12] as model initialization and then fine-tuned the model on our data. Random mirroring was applied during training. For each image pair, we kept the resolution of them unchanged, but randomly cropped a patch of 256 × 256 as input to train our model. We trained our model on the server with a single Nvidia Titan X Pascal, and it took roughly five days for training. The trained representation model was about 115 megabytes. For a better understanding of the readers, we show the training process of our model in Algorithm 1.

In the test phase, the content feature and the style feature were extracted by the trained representation model from the target frame without cropping. The model rand at 0.171 s per frame (s/f) on a Nvidia Titan X Pascal for images of size 3840 × 2160. We also tested the model in a Nvidia Jetson TX2, which was selected as one part of an embedded system developed for a small satellite set to launch in the year 2020. Nvidia Jetson TX2 contains four ARM Cortex A57 cores and one GPU with 256 CUDA cores. The model ran at 0.97 s/f on a Nvidia Titan X Pascal for images of size 3840 × 2160.

In the experiment, we conducted the data compression tests based on the standard HEVC codec for different purposes (details shown in Table 1). Our method could also be integrated into other codecs since we mainly provided synthetic reference images. The implementation was based on the low-delay configuration of an HEVC test model HM16.20 [42]. This implementation was compared to the unmodified HEVC codec to test the effectiveness of the proposed method on satellite video compression. The rough reference image without translation and the synthesized reference image generated by [2] were also compared to evaluate the effectiveness of our method. The testing was implemented on a server with eight Intel i5 CPU of 2.6 GHz and an Nvidia Titan X Pascal.

Algorithm 1 Training procedure of our proposed model.

```
while iterations $t < T_{train}$ do
    Sample and crop a mini-batch of patch pairs $(x, y)$ from training data;
    Extract the content and style feature to get $(Enc_c^x, Enc_s^x) = Enc(x)$; $(Enc_c^y, Enc_s^y) = Enc(y)$
    Reconstruct image $\bar{x}$ and $\bar{y}$ by $\bar{x} = Dec(Enc_c^x, Enc_s^x)$; $\bar{y} = Dec(Enc_c^y, Enc_s^y)$
    Generate the cross domain image $\hat{x}$ and $\hat{y}$ by $\hat{x} = Dec(Enc_c^x, Enc_s^x)$; $\hat{y} = Dec(Enc_c^y, Enc_s^y)$
    Calculate $\mathcal{L}_c$ by $(Enc_c^x, Enc_c^y)$ (Equation (3))
    Calculate $\mathcal{L}_{rec}$ by $(x, \bar{x})$ and $(y, \bar{y})$ (Equation (4))
    Calculate $\mathcal{L}_{cross}$ by $(x, \hat{x})$ and $(y, \hat{y})$ (Equation (5))
    Calculate $\mathcal{L}_{GAN}$ by $(x, \bar{x})$, $(y, \bar{y})$, $(x, \hat{x})$ and $(y, \hat{y})$ (Equation (7))
    Update the encoder and decoder with $\mathcal{L}_c$, $\mathcal{L}_{rec}$, $\mathcal{L}_{cross}$ and $\mathcal{L}_{GAN}$
    Update the discriminator with $\mathcal{L}_{GAN}$
end while
```

Table 1. Experimental configuration of HEVC.

Parameter	Value	Parameter	Value
Frame structure	Low Delay IPPP	GOP size	4
QP	22, 27, 32, 37	Max partition depth	4
Fast search	Enable	Search range	64
Intra-period	20	GOP size	4
SAO	1	Rate control	-1

4.1.2. Datasets

We evaluated our proposed method on a subset of satellites video clips captured by Jilin-1. The image series from historical Google images were used for training our translation networks. The detailed introduction about these two datasets are as follows:

- Video clips from satellites: There were five video clips from the satellite Jilin-1 over the seaport of Valencia (Figure 6a), the airport of Atlanta (Figure 6b), a railway station of Munich (Figure 6c), a park of Madrid (Figure 6d) m and the urban center of Valencia (Figure 6e). Those video clips were cut out from the original 12,000 × 5000 resolution and had a unified size of 3840 × 2160 with 300 frames, 10 fps.
- Image series from historical Google images: To show the evolution of the landscape, we employed the historical images of the same landscape from Google Earth as the training data. The image series contained 5000 images, which included an average of ten images for each landscape. The example images of the test landscape from the last five years are shown in Figure 7. They showed that the structure of the same landscape was almost unchanged, but the appearance was changed due to the environment.

In subjective the experiment, evaluation Bjøntegaard metrics for the delta satellite PSNR video (BD-PSNR) data and Bjøntegaard delta rate (BD-Rate) [43] were utilized as the metrics for the objective evaluation of coding performance.

4.2. Results

4.2.1. Intermediate Results from the Background Reference Generation

We compared our representation learning-based image translation method with the following three representative methods to show its effectiveness:

- Color transform [2]: manipulating the color image by imposing the mean and standard deviation of the style image onto the content image.

- AdaIN [40]: aligning the mean and variance of the content features with those of the style features by the adaptive instance normalization (AdaIN) layer.
- PhotoWCT [43]: transferring the style of the reference image to the content image by a pair of feature transforms, whitening, and coloring.

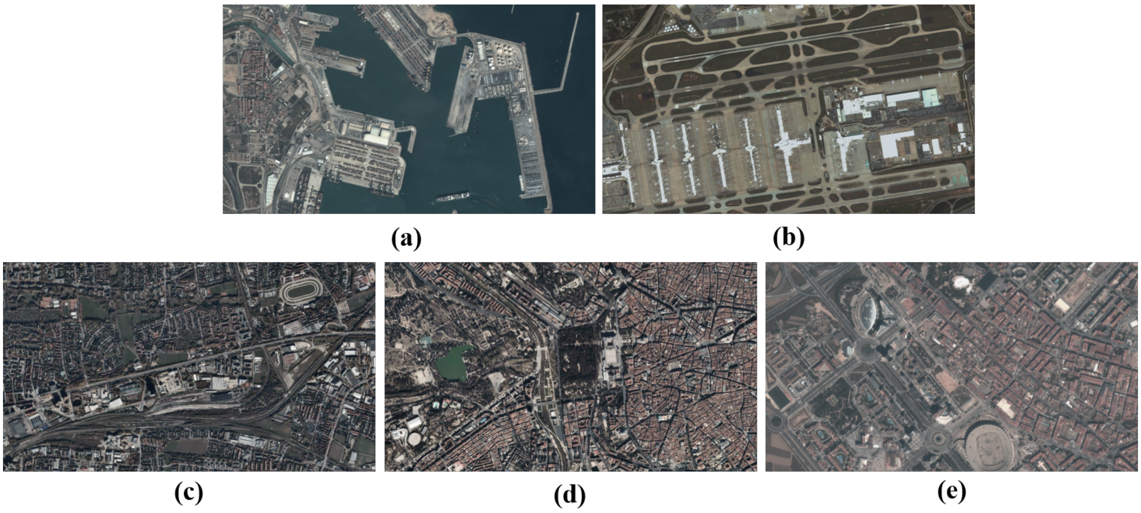


Figure 6. Typical frames of satellite video clips from experimental data. (**a**) Seaport (Valencia); (**b**) Airport (Atlanta); (**c**) Railway Station (Munich); (**d**) Park (Madrid); (**e**) Urban Center (Valencia).

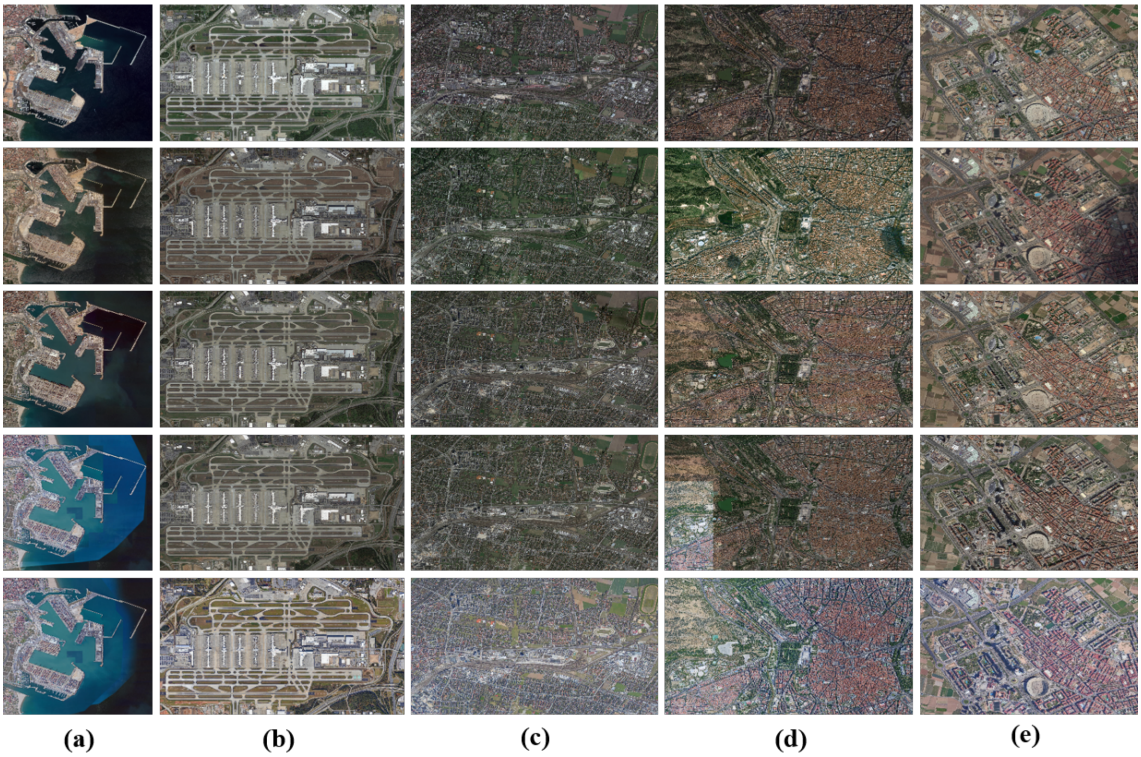


Figure 7. Image examples from historical Google images. (**a**) Seaport (Valencia); (**b**) Airport (Atlanta); (**c**) Railway Station (Munich); (**d**) Park (Madrid); (**e**) Urban Center (Valencia).

Figures 8 and 9 show the intermediate results of the park and urban center for background reference generation from the historical Google images. It can be easily noticed that the historical Google images were different from the target frames from satellite videos in terms of shooting angle, color, and texture details. With the help of geometric matching, the reference was nearly aligned with the target one. The color transform from [2] could adjust the color according to the statistics of the target frame, but it could be observed that color deviation still existed. AdaIN and PhotoWCT performed well for image color transform. However, they generated structural artifacts (e.g., distortions on object boundaries), especially in the complex scenes, such as satellite videos. We also translated the reference image to the target frame with our proposed method, which represented the consistent content feature by the high-dimensional spatial map and represented the different style by a low-dimensional vector. From the intermediate result, we could see that the proposed image translation method could successfully handle the evolution of the landscape caused by the environment changing.

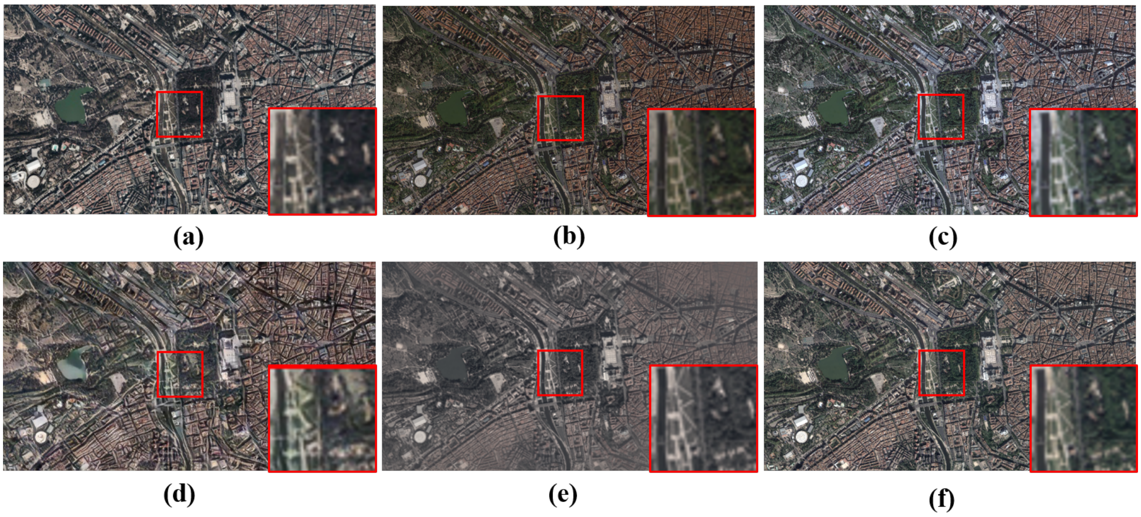


Figure 8. Reference images of the clip park. (**a**) Sample frame I_f from the satellite video clip park; (**b**) reference image I_g after geometric matching; (**c**) reference image from I_g after color transform [2]; (**d**) reference image from I_g after adaptive instance normalization (AdaIN) [40]; (**e**) reference image from I_g after PhotoWCT [43]; (**f**) reference image from I_g after image translation using the proposed method.

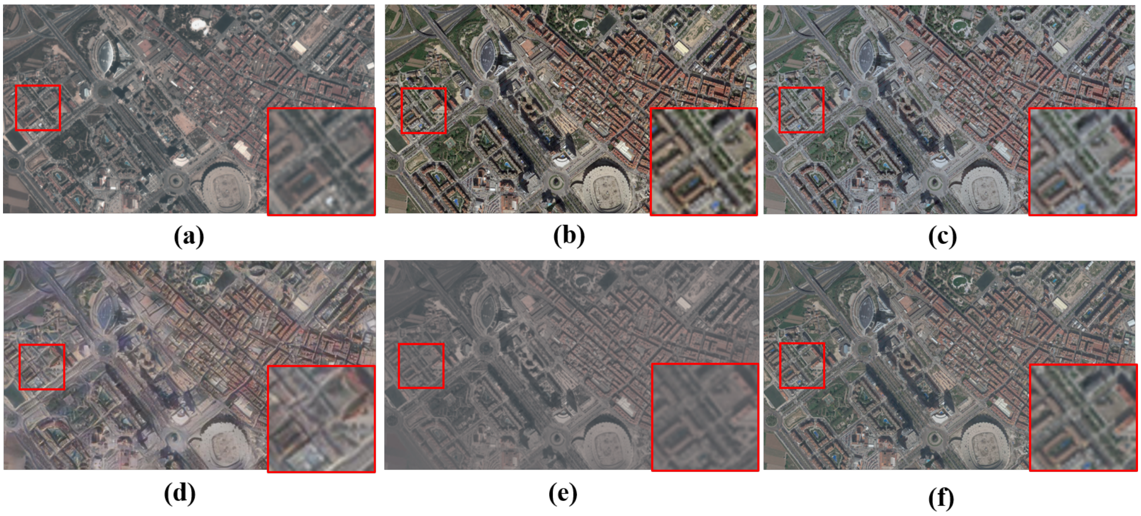


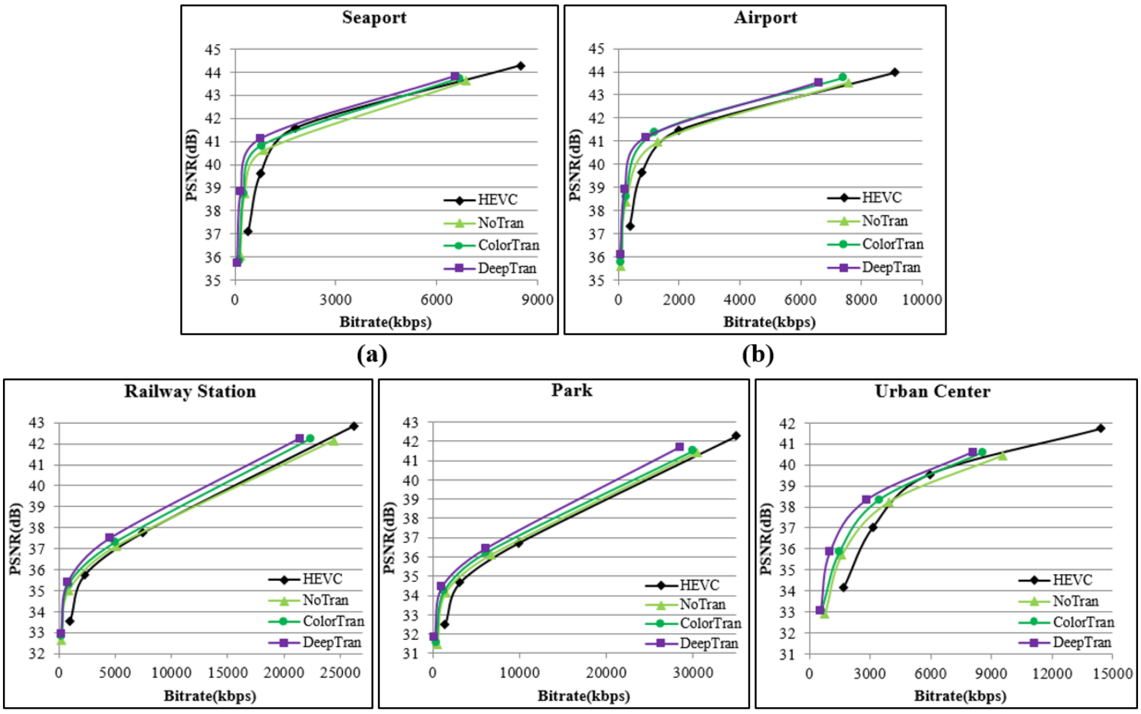
Figure 9. Reference images of the clip urban center. (**a**) Sample frame I_f from satellite video clip urban center; (**b**) reference image I_g after geometric matching; (**c**) reference image from I_g after color transform [2]; (**d**) reference image from I_g after AdaIN [40]; (**e**) reference image from I_g after PhotoWCT [43]; (**f**) reference image from I_g after image translation using the proposed method.

4.2.2. Experiments with Satellite Video Clips

The compression results of the proposed method compared with HEVC are presented in Table 2. In the experiment, we compare three methods to analyze the effectiveness of our method in generating good background references. 1) NoTran: the results from reference images generated with only geometric matching; 2) ColorTran: with the color transform method [2] 3) DeepTran: the proposed method.

In general, the average bitrate savings of our method could achieve up to 44.22%, compared to the average bitrate savings of 30.95% of ColorTran and 18.07% of NoTran. It was proven that the similarity of the background reference had a high effect on the improvement of the satellite video compression ratio. We also noticed that in different video clips with different video content, the highest bitrate reduction appeared with the seaport and airport, where there were large areas of uniform texture, such as sea and aircraft runway. The phenomenon was similar to [2] of the low-efficiency prediction at places containing a projection difference due to the projection difference in geometric matching.

The rate-distortion (RD) curves for the tested satellite video clips are shown in Figure 10, revealing similar results as we obtained from Table 2. The RD curves for the new reference-based compression method including NoTran, ColorTran, and DeepTran showed there was a significant improvement over the low bitrate situation. The improvement degree decreased over the increasing of the bitrate, especially the NoTran, which was even lower than HEVC in the videos of the railway station and urban center. The curves for the proposed method were higher than other curves in the four video clips, representing the general effectiveness of the proposed method in bitrate reduction for satellite videos.



(a)

(b)

(c)

(d)

(e)

Figure 10. RD curves of five video clips from satellite data. (**a**) Seaport (Valencia); (**b**) Airport (Atlanta); (**c**) Railway Station (Munich); (**d**) Park (Madrid); (**e**) Urban Center (Valencia).

4.3. Discussion

As shown in the results from image reconstruction, we can notice that the quality of the reconstruction is significantly improved compared to the color adjustment. This is probably owing to the high representation capability of deep networks. The common content feature from the network helps us to effectively disentangle the shared geometric structures across a series of images. In contrast, the style features record the variational factors causing the change of the images. The good reconstructed reference image further leads to the reduction of the compression bitrates as expected, since the more similar the reference to the target frame naturally leads to the higher compression ratio.

Table 2. Ablation study of the proposed method vs. HEVC with satellite data.

Method	Satellite Jilin-1	BD-PSNR (dB)	BD-Rate (%)
NoTran	Seaport	0.11	−27.73
	Airport	0.35	−25.57
	Railway Station	0.27	−10.59
	Park	0.36	−9.33
	Urban Center	0.36	−17.12
	Average	**0.29**	−**18.07**
ColorTran	Seaport	0.45	−39.28
	Airport	0.80	−42.38
	Railway Station	0.59	−22.18
	Park	0.64	−22.52
	Urban Center	0.86	−28.38
	Average	**0.67**	−**30.95**
DeepTran	Seaport	0.81	−53.33
	Airport	0.97	−52.34
	Railway Station	0.89	−36.72
	Park	0.94	−34.46
	Urban Center	1.27	−44.27
	Average	**0.98**	−**44.22**

5. Conclusions

This paper proposed a representation learning-based satellite video compression method. The key idea was to make use of the prior knowledge embedded in the historically captured images from multiple times to help the compression of the current video data since the structure of the landscape does not always change. To deal with the pixel value change caused by the daily illumination change or seasonal change, a representation model was trained to represent the distribution of the images. Taking two images from the same location, the representation model would output the same content features and different style features. The former was considered as the LBR, which did not need to be transmitted; the latter was the conditional parameters, which needed to be compressed and transmitted. By using the synthetic reference image as the reference frame, the proposed method could save an average of 44.22% bits over HEVC, which was significantly higher than that using references generated under the same conditions. Bitrate savings reached 18.07% when applied to satellite video data with arbitrarily collected reference images.

Author Contributions: Conceptualization, J.X. and R.H.; methodology, L.L.; software, L.L.; validation, L.L., Y.L., and J.X.; formal analysis, J.X.; investigation, J.X.; resources, M.W.; data curation, M.W.; writing, original draft preparation, L.L. and Y.L.; writing, review and editing, J.X; visualization, J.X.; supervision, R.H.; project administration, M.W.; funding acquisition, M.W. All authors have read and agreed to the published version of the manuscript.

Funding: National Nature Science Foundation of China: 91738302; National Nature Science Foundation of China: 61671336

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

References

1. Sjoberg, R.; Chen, Y.; Fujibayashi, A.; Hannuksela, M.M.; Samuelsson, J.; Tan, T.K.; Wang, Y.K.; Wenger, S. Overview of HEVC high-level syntax and reference picture management. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *22*, 1858–1870. [CrossRef]
2. Xiao, J.; Zhu, R.; Hu, R.; Wang, M.; Zhu, Y.; Chen, D.; Li, D. Towards Real-Time Service from Remote Sensing: Compression of Earth Observatory Video Data via Long-Term Background Referencing. *Remote. Sens.* **2018**, *10*, 876. [CrossRef]

3. Wang, L.; Lu, K.; Liu, P.; Ranjan, R.; Chen, L. IK-SVD: dictionary learning for spatial big data via incremental atom update. *Comput. Sci. Eng.* **2014**, *16*, 41–52. [CrossRef]
4. Song, W.; Deng, Z.; Wang, L.; Du, B.; Liu, P.; Lu, K. G-IK-SVD: parallel IK-SVD on GPUs for sparse representation of spatial big data. *J. Supercomput.* **2017**, *73*, 3433–3450. [CrossRef]
5. Ke, H.; Chen, D.; Shi, B.; Zhang, J.; Liu, X.; Zhang, X.; Li, X. Improving Brain E-health Services via High-Performance EEG Classification with Grouping Bayesian Optimization. *IEEE Trans. Serv. Comput.* **2019**. [CrossRef]
6. Ke, H.; Chen, D.; Shah, T.; Liu, X.; Zhang, X.; Zhang, L.; Li, X. Cloud-aided online EEG classification system for brain healthcare: A case study of depression evaluation with a lightweight CNN. *Software Pract. Exper.* **2018**. [CrossRef]
7. Jing, X.Y.; Zhu, X.; Wu, F.; You, X.; Liu, Q.; Yue, D.; Hu, R.; Xu, B. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 695–704.
8. Wu, F.; Jing, X.Y.; You, X.; Yue, D.; Hu, R.; Yang, J.Y. Multi-view low-rank dictionary learning for image classification. *Pattern Recognit.* **2016**, *50*, 143–154. [CrossRef]
9. Liu, X.; Tao, X.; Ge, N. Remote-sensing image compression using priori-information and feature registration. In Proceedings of the 2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall), Glasgow, Scotland, 11–14 May 2015; pp. 1–5.
10. Tao, X.; Li, S.; Zhang, Z.; Liu, X.; Wang, J.; Lu, J. Prior-Information-Based Remote Sensing Image Compression with Bayesian Dictionary Learning. In Proceedings of the 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), Sydney, Australia, 4–7 June 2017; pp. 1–6.
11. Wang, X.; Hu, R.; Wang, Z.; Xiao, J. Virtual background reference frame based satellite video coding. *IEEE Signal Process. Lett.* **2018**, *25*, 1445–1449. [CrossRef]
12. Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 172–189.
13. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4401–4410.
14. Lee, H.Y.; Tseng, H.Y.; Huang, J.B.; Singh, M.; Yang, M.H. Diverse image-to-image translation via disentangled representations. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 35–51.
15. Chen, D.; Tang, Y.; Zhang, H.; Wang, L. and Li, X. Incremental factorization of big time series data with blind factor approximation. *IEEE Trans. Knowl. Data Eng.* **2019**. [CrossRef]
16. Tang, Y.; Chen, D.; Wang, L.; Zomaya, A.; Chen, J. and Liu, H. Bayesian tensor factorization for multi-way analysis of multi-dimensional EEG. *Neurocomputing* **2018**, *318*, 162–174. [CrossRef]
17. Wiegand, T.; Sullivan, G.J.; Bjontegaard, G.; Luthra, A. Overview of the H. 264/AVC video coding standard. *IEEE Trans. Circuits Syst. Video Technol.* **2003**, *13*, 560–576. [CrossRef]
18. Corporation, S.I. SkySat-C Generation Satellite Sensors. Available online: https://www.satimagingcorp.com/satellite-sensors/skysat-1/ (accessed on 28 December 2019).
19. Yue, H.; Sun, X.; Yang, J.; Wu, F. Cloud-based image coding for mobile devices—Toward thousands to one compression. *IEEE Trans. Multimed.* **2013**, *15*, 845–857.
20. Shi, Z.; Sun, X.; Wu, F. Feature-based image set compression. In Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME), San Jose, CA, USA, 15–19 July 2013; pp. 1–6.
21. Wu, H.; Sun, X.; Yang, J.; Zeng, W.; Wu, F. Lossless compression of JPEG coded photo collections. *IEEE Trans. Image Process.* **2016**, *25*, 2684–2696. [CrossRef] [PubMed]
22. Wang, H.; Tian, T.; Ma, M.; Wu, J. Joint compression of near-duplicate Videos. *IEEE Trans. Multimed.* **2016**, *19*, 908–920. [CrossRef]
23. Song, X.; Peng, X.; Xu, J.; Shi, G.; Wu, F. Cloud-based distributed image coding. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 1926–1940. [CrossRef]
24. Xiao, J.; Hu, R.; Liao, L.; Chen, Y.; Wang, Z.; Xiong, Z. Knowledge-based coding of objects for multisource surveillance video data. *IEEE Trans. Multimed.* **2016**, *18*, 1691–1706. [CrossRef]

25. Xiao, J.; Liao, L.; Hu, J.; Chen, Y.; Hu, R. Exploiting global redundancy in big surveillance video data for efficient coding. *Clust. Comput.* **2015**, *18*, 531–540. [CrossRef]
26. Chen, Y.; Hu, R.; Xiao, J.; Xu, L.; Wang, Z. Multisource surveillance video data coding with hierarchical knowledge library. *Multimed. Tools Appl.* **2019**, *78*, 14705–14731. [CrossRef]
27. Chen, Y.; Hu, R.; Xiao, J.; Wang, Z. Multisource surveillance video coding with synthetic reference frame. *J. Vis. Commun. Image Represent.* **2019**, *65*, 102685. [CrossRef]
28. Sanchez, E.; Serrurier, M.; Ortner, M. Learning Disentangled Representations of Satellite Image Time Series. *arXiv Prepr.* **2019**, arXiv:1903.08863.
29. Gonzalez-Garcia, A.; van de Weijer, J.; Bengio, Y. Image-to-image translation for cross-domain disentanglement. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Montréal, QC, Canada, 2–8 December 2018; pp. 1287–1298.
30. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 27 October–3 Novemver 2017; pp. 2223–2232.
31. Zhu, J.Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A.A.; Wang, O.; Shechtman, E. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2017; pp. 465–476.
32. Auli-Llinas, F.; Marcellin, M.W.; Sanchez, V.; Serra-Sagrista, J.; Bartrina-Rapesta, J.; Blanes, I. Coding scheme for the transmission of satellite imagery. In Proceedings of the 2016 Data Compression Conference (DCC), Snowbird, UT, USA, 29 March–1 April 2016; pp. 427–436.
33. Aulí-Llinàs, F.; Marcellin, M.W.; Sanchez, V.; Bartrina-Rapesta, J.; Hernández-Cabronero, M. Dual link image coding for earth observation satellites. *IEEE Trans. Geosci. Remote. Sens.* **2018**, *56*, 5083–5096. [CrossRef]
34. Zhang, X.; Huang, T.; Tian, Y.; Gao, W. Background-modeling-based adaptive prediction for surveillance video coding. *IEEE Trans. Image Process.* **2013**, *23*, 769–784. [CrossRef] [PubMed]
35. Reinhard, E.; Adhikhmin, M.; Gooch, B.; Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **2001**, *21*, 34–41. [CrossRef]
36. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
37. Wang, X.; Hu, R.; Wang, Z.; Xiao, J.; Satoh, S. Long-Term Background Redundancy Reduction for Earth Observatory Video Coding. *IEEE Trans. Circuits Syst. Video Technol.* **2019**. [CrossRef]
38. Welch, T.A. A technique for high-performance data compression. *Computer* **1984**, *6*, 8–19. [CrossRef]
39. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv Prepr.* **2014**, arXiv:1412.6980.
40. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 27 October–3 Novemver 2017; pp. 1501–1510.
41. Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; Lopez, A.M. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3234–3243.
42. Institute, F.H.H. High Efficiency Video Coding (HEVC). Available online: https://hevc.hhi.fraunhofer.de/ (accessed on 28 December 2019).
43. Li, Y.; Liu, M.; Li, X.; Yang, M.-H. and Kautz, J. A closed-form solution to photorealistic image stylization. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 453–468.