2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Improving Video Super-Resolution with Long-Term Self-Exemplars

Guotao Meng*, Yue Wu*, and Qifeng Chen

Abstract— **Existing video super-resolution methods often utilize a few neighboring frames to generate a higher-resolution image for each frame. However, the abundant information in distant frames has not been fully exploited in these methods: corresponding patches of the same instance appear across distant frames at different scales. Based on this observation, we propose to improve the video super-resolution quality with long-term cross-scale aggregation that leverages similar patches (self-exemplars) across distant frames. Our method can be implemented as post-processing for any super-resolution methods to improve performance. Our model consists of a multi-reference alignment module to fuse the features derived from similar patches: we fuse the features of distant references to perform high-quality super-resolution. We also propose a novel and practical training strategy for reference-based super-resolution. To evaluate the performance of our proposed method, we conduct extensive experiments on our collected CarCam dataset, the Waymo Open dataset, and the REDS dataset, and the results demonstrate our method outperforms state-of-the-art methods.**

I. Introduction

Super-resolution (SR) has broad practical applications in autonomous driving and robotics since it can be used to reconstruct high-resolution (HR) visual images from low-resolution (LR) ones. While image super-resolution [1]–[14] exploits spatial information to recover missing details, video super-resolution (VSR) [15]–[27] needs to utilize additional temporal information from other frames to reconstruct a clear HR video, which can be better used for detecting small distant objects and recover critical information such as license plates.

The typical limitation of existing VSR methods is that only a few neighboring frames are utilized (usually $3 \sim 7$ frames) because, compared to utilizing the distant frames or feature maps, it is relatively easier to perform the spatial alignment on the neighboring frames or feature maps to gather information for super-resolution. However, long videos containing relevant content among long-term frames are ubiquitous, especially in driving scenarios, video surveillance, and so on. However, in previous methods, such valuable information in long-term frames is rarely exploited due to the difficulty of spatial alignment, which is caused by various reasons, including the change of camera positions, the appearance difference of the same object, and so on. To solve this problem, we propose to decompose a frame into patches and search for patches (self-exemplars) with similar content but higher resolutions from distance frames. Then the redundant information in distant frames can be fully utilized for super-resolution.

For the spatial alignment, a large portion of previous methods relies on motion compensation [17], [19], [28]–[30]. These methods first perform optical flow estimation or deformable convolution [31] to align the frames and then use aligned frames to reconstruct the image. However, estimating dense motion between two distant frames is difficult. Moreover, imperfect motion estimation often leads to unsatisfactory artifacts in the SR results of these motion-based VSR approaches. Thus, the motion-based methods are not applicable when exploiting long-term information. Another stream of VSR [16], [18], [32] is to use recurrent models to store long-term information. However, these methods usually employ a fixed-size feature vector to store previous content, and thus it is tough to memorize high-frequency long-term details. Our method fully exploits long-term self-exemplars, leading to better super-resolution performance.

According to our observation, patch recurrence exists commonly in a video. Objects may appear small and blurry in one frame but become large and clear in another frame. This is because the distance from the objects to the camera is changing. The patch of a high-resolution object in other frames can be used as a reference for super-resolution.

Thus, we introduce a long-term non-local aggregation method to leverage the most similar patches across frames, as shown in Fig. 1. Because the information contained in the whole sequence is redundant and not easy to be processed directly by a network, we propose a self-exemplar retrieval module to search self-exemplars across frames. In our method, for each frame, we first uniformly divide this frame into several patches. Then, each patch is used as a $query$ to search higher-resolution long-term self-exemplars and short-term self-exemplars with the same resolution. For the long-term self-exemplars, we propose a long-term texture aggregation module to select useful references and initially align them with $query$. Afterward, we propose a feature alignment module to align the feature maps of these self-exemplars by affine transformations. Moreover, we use a multi-reference fusion module to fuse long-term features. Then, we use a long-term and short-term feature aggregation module to fuse long-term information and short-term information to reconstruct details. Also, we propose a novel training strategy to solve the imbalance problem of data distribution.

Commonly used VSR datasets contain only several frames or have slight camera motion. To demonstrate the effectiveness of our method, we choose the driving scene as a classical application scenario as it has large camera motion

* indicates equal contribution. Guotao Meng (gmeng@connect.ust.hk) is with the Department of Electronic and Computer Engineering, HKUST. Yue Wu (ywudg@connect.ust.hk) and Qifeng Chen (cqf@ust.hk) are with the Department of Computer Science and Engineering, HKUST.

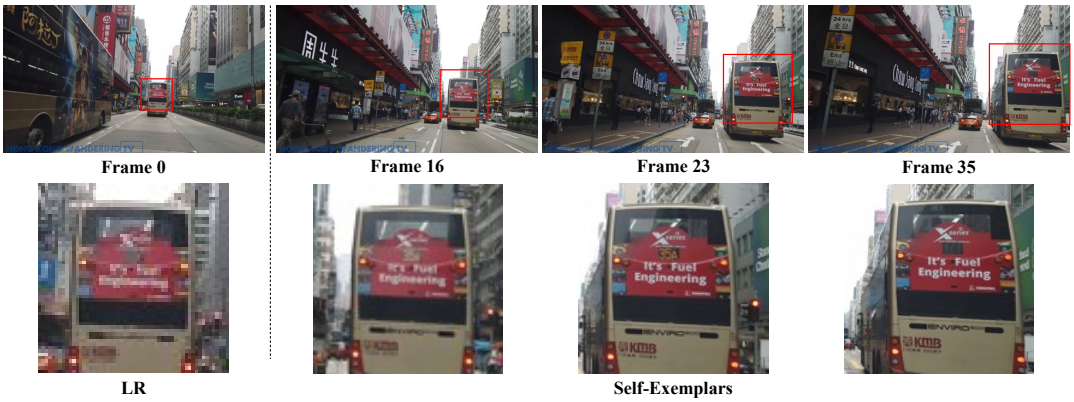

Fig. 1. It is ubiquitous that relevant content can be found in a long time span in a video. Therefore patches with similar content but higher resolution can be found in other frames. These patches with higher resolution can be used as exemplars to super-resolve the LR patches.

and evident patch recurrence. We use two datasets of driving scenarios to evaluate the performance of our method. We collected a CarCam dataset that contains 139 video sequences. We also use the public Waymo Open dataset [33]. Our method outperforms state-of-the-art (SOTA) methods on the CarCam and Waymo datasets [33]. Besides, to demonstrate the performance of our proposed method on general scenarios other than driving, we conduct experiments on the REDS dataset [34]. Our method surpasses the SOTA methods on PSNR and has a comparable performance with them on SSIM [35].

Our contributions can be summarized as follows:

- To better exploit redundant information in distant frames, we propose a novel long-term cross-scale aggregation method leveraging self-exemplars.
- We propose several novel modules to enhance the reconstruction performance: self-exemplar retrieval, multi-reference selection and pre-alignment, feature alignment, and a novel and practical training strategy.
- We collected the new CarCam dataset with 139 dashcam videos. The experiments show that our proposed method outperforms state-of-the-art VSR methods.

II. Method

Our key observation is that, for low-resolution content, it is highly likely there are similar and clearer high-resolution details in other frames, which provide valuable information for SR. Based on this observation, we propose to exploit the information from distant frames as the reference to super-resolve a video frame. For each patch in the frame to be super-resolved, we first employ *Self-exemplar Retrieval* module to search the self-exemplars with higher resolution from all the video frames as the *long-term self-exemplars* and search self-exemplars with the same resolution from neighboring frames as the *short-term self-exemplars*. Due to the limitation of model capacity, we employ a *Pre-alignment and Reference Selection* module to select the most valuable references. Afterward, we propose the *Multiple Reference Feature Alignment* module to align reference features to the LR patch feature. Finally, the features of the input patch and reference patches are fused and fed into the SR network to generate the high-resolution output.

A. Self-exemplar Retrieval

Because the time interval between distant frames can be several seconds, it is difficult to compute accurate motion compensation or align frames precisely. Furthermore, the scale and view angle of an instance may vary significantly in a video, therefore it is hard to obtain correspondence from the whole video by optical flow or object tracking. Thus, we decompose a frame into patches, and perform the patch-level self-exemplar search, without relying on motion compensation. We divide the self-exemplar into long-term and short-term based on the search range.

Long-term self-exemplars. We uniformly divide all the video frames $M = \{I_1, ..., I_T\}$ into patches with overlap. For one patch p, we search its self-exemplars from M. We first define an increasing scale sequence $C = \{c|c_1, c_2, ...\}$, where $c > 1$. Then, for each c, we apply bicubic up-sampling on p using scale c to get an up-sampled query image $p_c\uparrow$. We also sequentially apply bicubic down-sampling and up-sampling on all the frames M using scale c to obtain blurry frames $M_c\downarrow\uparrow$ to match the frequency band of $p_c\uparrow$. The patch $p_c\uparrow$ is used as a query to search larger self-exemplars in each frame from $M_c\downarrow\uparrow$ using the template matching [36]. After this operation, we obtain the patch $q_c\downarrow\uparrow$ from $M_c\downarrow\uparrow$ with the highest similarity. The cosine distance is used as the metric of the similarity. Then we extract the patch q_c from M with the same spatial location as $q_c\downarrow\uparrow$ as the self-exemplar for p of scale c. The long-term self-exemplars for p is constructed by $Q = \{q_{c_1}, q_{c_2}...\}$.

With this searching method, there may be some patches in Q that contain irrelevant content of p which are not suitable to serve as self-exemplars, especially for larger scales in C. These inaccurate exemplars are filtered out using our long-term self-exemplar selection module, which will be discussed later.

Short-term self-exemplars. Like most of the existing VSR methods, the information in the neighboring frames is also involved. For patch p, we use the patch retrieval strategy introduced above to search the most similar patches from their neighboring frames using the scale factor 1. The patch retrieval result is denoted as short-term self-exemplars Q_s.

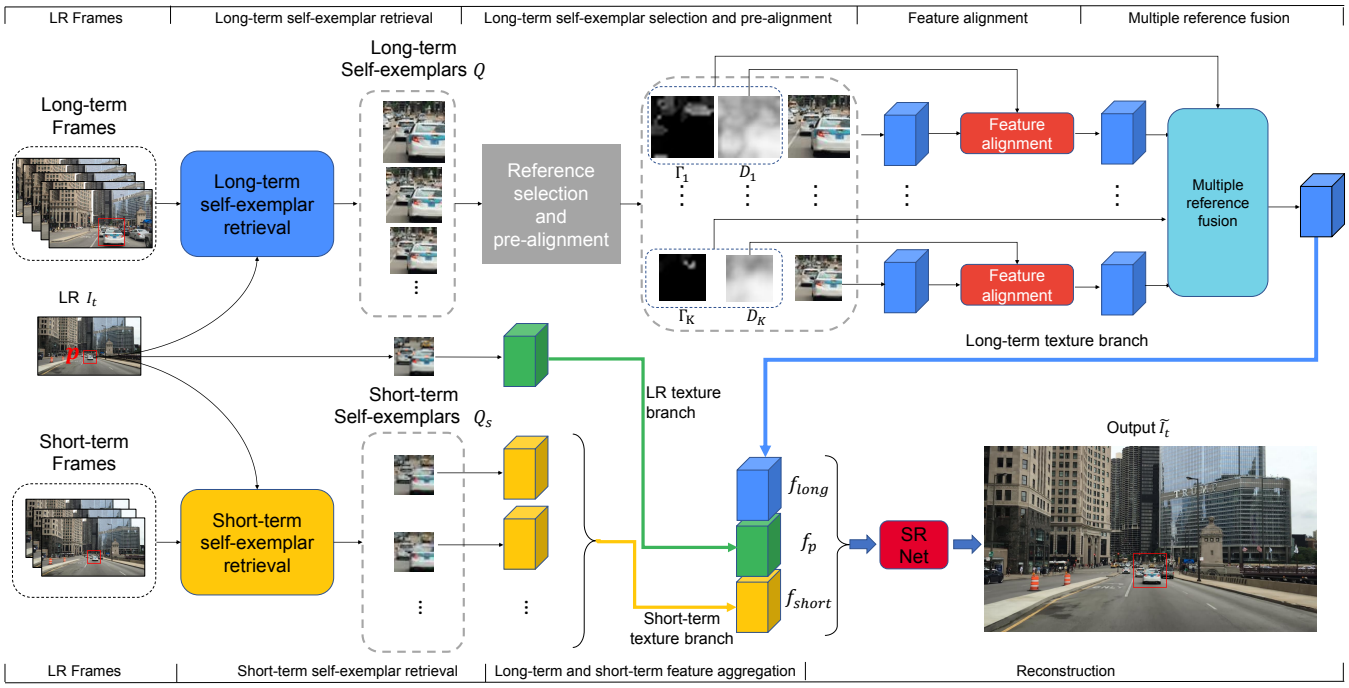

Fig. 2. The pipeline of our method. We first obtain long-term self-exemplars and short-term self-exemplars using the self-exemplar retrieval module. Then, we use a long-term self-exemplar selection and pre-alignment module to select top k references and obtain the similarity map and distance map of these self-exemplars. Then, we compute the affine transformation parameters based on distance maps. And we extract feature maps of each self-exemplar and use an affine transformation to align feature maps. We use a multi-reference fusion network to obtain the long-term feature map f_{long}. We also use feature extractors to extract features of LR patch and short-term self-exemplars, f_p and f_{short}. We concatenate these features and use an SR Net to obtain the output.

Fig. 3. Illustration of the feature swap and affine transformation alignment.

B. Long-term Texture Aggregation

Irrelevant self-exemplars will bring unwanted noise and result in performance degradation, which is discussed in detail in Sec. III-E. Thus, in this module, the feature of long-term self-exemplars is selected, aligned, and fused.

1) Pre-alignment and Long-term Self-exemplar Selection.: To select the most valuable self-exemplars, directly computing the similarity such as the cosine distance of the patches in Q with p leads to an unsatisfied result because smaller self-exemplars extracted from near frames in Q often have a higher similarity score but provide less high-frequency details. Thus, we design a novel metric to measure the quality of the patches in Q.

For each patch q in Q, we first up-sample p to $p\uparrow$ to match the spatial resolution of q. We apply down-sampling and up-sampling on q to generate $q\downarrow\uparrow$ to match the frequency band of $p\uparrow$. Then a pretrained VGG-19 [37] network ϕ is used to obtain features of $p\uparrow$ and $q\downarrow\uparrow$.

Inspired by the feature swapping operation in SRNTT [38], we unfold the features $\phi(p\uparrow)$ and $\phi(q\downarrow\uparrow)$ into N feature blocks for feature matching. The cosine distance is used to measure the similarity between the feature blocks:

$$\gamma_{g,h} = \left\langle \frac{B_g(\phi(p\uparrow))}{\|B_g(\phi(p\uparrow))\|}, \frac{B_h(\phi(q\downarrow\uparrow))}{\|B_h(\phi(q\downarrow\uparrow))\|} \right\rangle, \tag{1}$$

where $B_g(\cdot)$ denotes the g-th feature block, and $\gamma_{g,h}$ is the similarity between the g-th feature block of $\phi(p\uparrow)$ and h-th feature block of $\phi(q\downarrow\uparrow)$. Then we search over all the reference feature blocks:

$$h^g = \underset{h}{\operatorname{argmax}} \ \gamma_{g,h}, \tag{2}$$

$$\Gamma_g = \gamma_{g,h^g}, \tag{3}$$

where Γ_g is the similarity of the g-th feature block on $\phi(p\uparrow)$. Different from the feature swapping operation [38], we do not directly swap the feature blocks of $\phi(q)$ to construct the aligned feature map for p. Feature swapping takes the average of the swapped features in the overlap regions, resulting in texture corruption as shown in Fig. 3. We use a novel affine transformation-based alignment module to align the reference feature, which will be discussed in detail later.

Then we propose a score function for q to measure the distance between p and q for self-exemplars selection:

$$D_q(x_g, y_g) = \|(x_g, y_g) - (x_{h^g}, y_{h^g})\|^2, \tag{4}$$

$$score(q) = \frac{1}{N}\sum_g D_q(x_g, y_g), \tag{5}$$

where D_q is the distance map for q, N is the number of feature blocks for p. (x_g, y_g) and (x_{h^g}, y_{h^g}) are spatial

coordinates of g-th feature block of $\phi(p\uparrow)$ and h^g-th feature block of $\phi(q\downarrow\uparrow)$. If q matches p, the corresponding feature block pairs should appear at similar positions. Based on this observation, We will filter out patch q in Q when $score(q) > \delta$ because $score(q)$ represents the average distance between corresponding feature blocks. Then the largest K remaining patches in Q are selected as $Q' = \{q'_1, ...q'_K\}$. If the number of the remaining patches is smaller than K, the patches with the smallest $score(q)$ are used for Q'. δ is set as 0.1 empirically, and the selection of K is discussed in Sec. III-E.

2) Multiple Reference Feature Alignment.: After obtaining multiple references Q', we abandon the previously commonly used feature swapping align scheme because the information will be corrupted. Moreover, motion estimation can not be directly applied since the time interval between p and q may be several seconds, and the viewpoint may change largely. Thus, we propose a new affine transformation based alignment module to align the patches:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_g \left\| (x_g, y_g) - \mathcal{T}(x_{h^g}, y_{h^g}; \theta) \right\|^2, \quad (6)$$

where $\mathcal{T}$ represents the affine transformation, and θ is the parameters of $\mathcal{T}$. We use RANSAC algorithms to obtain θ^* which minimizes the sum of Euclidean distances between (x_g, y_g) and transformed (x_{h^g}, y_{h^g}) in Equation 6.

We use a sequence of residual blocks [39] to extract features f_q of each q' and f_p of p. Then we transform f_q with the affine transformation parameter θ^* to get the aligned reference feature for f_p:

$$f_q^* = \mathcal{T}(f_q; \theta^*). \quad (7)$$

Comparison to feature swapping. Feature swapping takes the average of transferred textures due to dense sampling. Therefore the original information will be corrupted. As shown in Fig. 3, the text is broken and blurry, although the reference patch contains clear and correct details. While the proposed affine transformation based alignment module keeps the structure and the details of the self-exemplars. For visualization, we swap and transform the images instead of the feature maps.

3) Multiple reference fusion.: To fuse the transformed feature maps, we use a weight network to predict the weight map of each f_q^*. This network takes the feature map f_p, f_q^*, similarity map Γ, and distance map D as input. Then we use a *softmax* function to predict a weight map w for each f_q^*:

$$f_{long} = \sum_{r=1}^{k} w_r \cdot f_{q,r}^*, \quad (8)$$

where $f_{q,r}^*$ is the aligned feature map of r-th q' in Q'.

C. Long-term and short-term feature aggregation

We use a sequence of residual blocks to extract features f_{short}, f_p of short-term self-exemplars, and p. The f_{long}, f_{short} and f_p are concatenated together and fed into an SR network to generate SR output. During training, we adopt the Charbonnier loss [40], which is defined as

$$\sqrt{\left\| \tilde{I}_t - I_t^H \right\|^2 + \epsilon^2}, \quad (9)$$

where $\tilde{I}_t$ is the predicted SR result, I_t^H is the ground truth, and ϵ is a small constant. $\tilde{I}_t$ is obtained by splicing SR results of patches.

D. Training strategy for data imbalance

The effectiveness of reference is related to the spatial locations of LR patches. The background region is less likely to have valuable references since the distance between the background and the camera changes slightly. Although the camera is moving, the size variance of the background region is small. Meanwhile, the region near the camera or self-moving objects has a larger probability of having good references.

As the background region occupies a large portion of the image, there is an imbalance in data distribution between regions with good references and regions with normal references.

Suppose we train the network using a regular training strategy, and the network will rely heavily on short-term references without utilizing long-term information.

Thus, we propose a unique training strategy to randomly replace a reference patch with a slightly adjusted ground-truth patch by an affine transformation. This strategy will solve the data imbalance by rebalancing the data distribution of the reference patches. Thus, the network can learn to draw beneficial details from long-term reference features. The probability of the replacing operation is set as 0.3 empirically.

III. Experiments

A. Datasets

Since commonly used datasets usually contain videos captured by slow cameras, thus, these datasets can not demonstrate the ability of our method in exploiting long-term content. Thus, we collect two datasets in driving scenes, including the CarCam dataset and, the Waymo Open dataset [33]. To demonstrate the generalization ability of our method on common scenarios other than driving scenes, we also evaluate our method on a commonly used SR dataset, REDS dataset.

CarCam dataset. To demonstrate the effectiveness of our method, we collect a CarCam dataset containing videos captured in different cities using different cameras. Our CarCam dataset contains 139 high-resolution video clips of driving scenarios from YouTube. The videos are captured in multiple cities, including Hong Kong, Paris, Hollywood, and Chicago. Each sequence contains 60 frames, and the shape of the frame is 1920×1080. The length of each clip is 10 seconds. To evaluate the performance quantitatively, we downsample the videos by a factor of 4 using BICUBIC to obtain LR videos. The frame rate is 6fps.

Waymo Open dataset [33]. We randomly collect 100 sequences within rich texture from the Waymo Open dataset [33]. Each sequence contains 50 frames, and the resolution of the frame is 1920×1280. We use 70 sequences as training and 30 sequences as testing. The LR videos are downsampled from the original video using BICUBIC.

REDS dataset [34]. To demonstrate the generalization ability of our method on common scenarios other than driving scenes, we use REDS to evaluate our proposed method. REDS is a commonly adopted dataset for video super-resolution evaluation. REDS consists of realistic and dynamic scenes, containing 270 sequences (266 for training and 4 for testing). Each sequence has 100 frames, and the frame rate is 24 fps.

B. Distribution of self-exemplars

To demonstrate that patches in the distant frames are used as the reference for SR, we show the distribution of the frame span between the current frame and the frames where the self-exemplars are from.

In the CarCam dataset, there are 55.99%, 38.27%, 20.45% self-exemplars from frames which have intervals larger than 5 frames, 10 frames, 20 frames. And the percentages are 68.98%, 54.16%, 35.86% for the REDS dataset and 60.17%, 41.18%, 20.99% for the Waymo Open dataset.

According to the distribution shown above, a relatively large part of the self-exemplars are from the distant frames and most of the valuable information in the whole video is utilized in our proposed method.

C. Implementation Details

In the patch retrieval process, the patch size and stride are 32×32 and 24. All the images are downsampled $2\times$ for fast searching. The scale sequence C is defined as $[1.2, 1.4, 1.7, 2.1, 2.5, 2.9, 3.5]$. After searching all the frames in the video, we select the best three long-term references due to the limitation of model capacity. The choice of this number is discussed in detail in Sec. III-E. The number of long-term references and short-term references is set as 3 and 2. The backbone of our network is a sequence of residual blocks. The network architecture and training details are provided in the supplement.

We search the self-exemplars from all the video frames. Theoretically, this is the optimal strategy in terms of super-resolution performance. Practically for a faster processing speed during inference, there can be a sliding window containing the previous and following frames to search the self-exemplars. The size of sliding windows can be adjusted based on the motion of the camera and objects.

D. Evaluation

On the CarCam and Waymo Open datasets, we evaluate our method with previous state-of-the-art methods, including PFNL [22], RBPN [18], TDAN [19], TGA [41], MuCAN [42], RSDN [32] and BasicVSR [43]. For a fair comparison, we re-train all the methods on these two datasets carefully except BasicVSR. For the methods without training code, we carefully re-implement them. For BasicVSR, we use the pre-trained model on Vimeo90K. The visual comparison is presented in Fig. 4, and the quantitative results are presented in Table I. The evaluation metrics are SSIM, PSNR, and LPIPS [45].

TABLE I

QUANTITATIVE EVALUATION OF OUR APPROACH AND STATE-OF-THE-ART VIDEO SUPER-RESOLUTION METHODS.

	CarCam			Waymo Open		
	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓
BICUBIC	0.790	25.78	0.393	0.890	31.70	0.303
PFNL [22]	0.882	28.50	0.164	0.934	34.81	0.155
RBPN [18]	0.886	28.64	0.158	0.937	35.10	0.149
TDAN [19]	0.870	28.02	0.178	0.926	34.10	0.167
TGA [41]	0.876	28.19	0.172	0.933	34.73	0.162
MUCAN [42]	0.900	29.29	0.138	0.941	35.46	0.149
RSDN [32]	0.894	29.13	0.150	0.936	35.04	0.153
BasicVSR [43]	0.891	28.48	0.187	0.939	34.46	0.164
Ours	**0.904**	**29.54**	**0.136**	**0.945**	**35.86**	**0.138**

As depicted in Table I, our method outperforms other methods by at least **0.25dB** and **0.40dB** on these two datasets. Our method also has a better perceptual quality surpassing others by at least 0.002 and 0.011 in LPIPS. The results demonstrate the effectiveness of our method. Although the LR image is blurry, our method can reconstruct high-frequency details by properly utilizing long-term self-exemplars, while previous methods can not. More visual results are provided in the supplementary material.

We also conduct experiments on REDS to demonstrate our generalization ability. The quantitative comparison is shown in Table II. It is clear that our method surpasses previous methods by at least 0.4dB in PSNR.

Our method can perform well on commonly used video super-resolution scenarios and can yield much better performance when there exists patch recurrence in distant frames.

E. Ablation study

We conduct an ablation study on several components to demonstrate the effectiveness of our proposed method.

Long-term Texture Aggregation Module. We build a baseline without utilizing information in long-term frames (WL). The long-term texture branch is removed. As shown in Table III, the baseline yields 29.129 dB in PSNR and 0.896 in SSIM. Utilizing long-term self-exemplars brings 0.41 dB improvement in PSNR, which proves the effectiveness of the long-term texture aggregation module.

Training Strategy. We build a baseline (WT) without the rebalancing training strategy. As shown in Table III, this strategy improves the result by 0.295 dB. Our proposed strategy resolves the data imbalance problem and helps promote performance when having higher-resolution self-exemplars.

Affine Transformation. As indicated in Sec. II-B.2, directly adopting commonly used feature swapping [38] instead of affine transformation (WA) will result in content corruption as shown in Fig. 3. As feature swapping unfolds a feature map into 3×3 patches and overlaps the transferred texture, the clear content will be broken. As shown in Table III, using affine transformation brings 0.417 dB improvement.

Fig. 4. Visual comparisons among different super-resolution methods on our CarCam dataset. More visual results are provided in the supplement.

TABLE II

EXPERIMENTS ON THE REDS DATASET FOR $\times 4$ SETTING. THE QUANTITATIVE RESULTS OF PRIOR WORKS ARE EXTRACTED FROM THEIR PAPERS.

Method	BICUBIC	RCAN [44]	ToFlow [30]	DUF [30]	EDVR [17]	MuCAN [42]	BasicVSR [43]	Ours
Clip_000	24.55/0.6489	26.17/0.7371	26.52/0.7540	27.30/0.7937	27.78/0.8156	27.99/0.8219	28.39/**0.8610**	**28.61**/0.8090
Clip_011	26.06/0.7261	29.34/0.8255	27.80/0.7858	28.38/0.8056	31.60/0.8779	31.84/0.8801	**32.46/0.9049**	32.10/0.8786
Clip_015	28.52/0.8034	31.85/0.8881	30.67/0.8609	31.55/0.8846	33.71/0.9161	33.90/0.9170	**34.22/0.9272**	34.01/0.9152
Clip_020	25.41/0.7386	27.74/0.8293	26.92/0.7953	27.30/0.8164	29.74/0.8809	29.78/0.8811	**30.60/0.9093**	30.41/0.8786
Average	26.14/0.7292	28.78/0.8200	27.98/0.7990	28.63/0.8251	30.71/0.8726	30.88/0.8750	**31.42/0.9007**	31.28/0.8703

TABLE III

ABLATION STUDY ON CARCAM. WL: WITHOUT LONG-TERM FRAMES; WT: WITHOUT REBALANCING TRAINING; WA: WITHOUT AFFINE TRANSFORMATION.

	WL	WT	WA	Full Model
SSIM↑	0.896	0.898	0.896	**0.904**
PSNR↑	29.129	29.247	29.125	**29.542**

TABLE IV

THE QUANTITATIVE PERFORMANCE BEFORE AND AFTER THE POST-PROCESSING WITH THE PROPOSED METHOD.

Dataset	CarCam	Waymo Open	REDS
Before	28.48/0.8884	34.46/0.9386	31.42/**0.9007**
After	**28.76/0.9064**	**36.56/0.9506**	**31.45**/0.9004

F. Efficiency analysis

The average search time of one patch for one scale is 0.24 seconds on a single CPU thread and single GPU. And the RANSAC algorithm is conducted only on the selected patches, which costs 0.0156 seconds for each patch. Since the self-exemplar search and RANSAC for each patch and each scale is independent, this processing can be significantly accelerated by multi-threading and multi-GPU parallel processing.

G. As a Post-Processing

Our method can serve as a post-processing of existing methods to further improve previous methods by exploiting long-term information. We conduct experiments on the BasicVSR [43] which is currently the state-of-the-art video super-resolution method. The output of the BasicVSR is used as the input of our method. As shown in Table IV, after our post-processing module, the performance is further increased. It is demonstrated that long-term information should be used for a better super-resolution result.

IV. CONCLUSION

The critical contribution of our proposed method is the exploitation of the long-term content in all the frames of a video, while the previous methods focus on the fusion of short-term information. In this paper, we have proposed a novel video super-resolution method with long-term cross-scale aggregation by utilizing self-exemplars across distant frames. We propose a novel long-term texture module to select, align, and fuse features derived from similar patches. We fuse the features of both long-term and short-term references and propose a novel training strategy for the data imbalance problem. Extensive experiments demonstrate the effectiveness of our proposed method. Our method has many potential applications beyond car camera scenarios, such as hand-held videos, drone videos, and surveillance.

REFERENCES

[1] H. Chen, X. He, L. Qing, and Q. Teng, "Single image super-resolution via adaptive transform-based nonlocal self-similarity modeling and learning-based gradient regularization," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1702–1717, 2017.

[2] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*, 2014.

[3] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*, 2016.

[4] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *IEEE Computer Vision and Pattern Recognition*, 2018.

[5] C.-C. Kao, Y. Wang, J. Waltman, and P. Sen, "Patch-based image hallucination for super resolution with detail reconstruction from similar sample images," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1139–1152, 2019.

[6] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *IEEE Computer Vision and Pattern Recognition*, 2016.

[7] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Computer Vision and Pattern Recognition*, 2017.

[8] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *IEEE Computer Vision and Pattern Recognition Workshops*, 2017.

[9] J. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *IEEE Computer Vision and Pattern Recognition*, 2015.

[10] Y. Liu, S. Wang, J. Zhang, S. Wang, S. Ma, and W. Gao, "Iterative network for image super-resolution," *IEEE Transactions on Multimedia*, 2021.

[11] Y. Shi, K. Wang, C. Chen, L. Xu, and L. Lin, "Structure-preserving image super-resolution via contextualized multitask learning," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2804–2815, 2017.

[12] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: enhanced super-resolution generative adversarial networks," in *European Conference on Computer Vision*, 2018.

[13] B. Yan, B. Bare, C. Ma, K. Li, and W. Tan, "Deep objective quality assessment driven single image super-resolution," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2957–2971, 2019.

[14] W. Yang, Y. Tian, F. Zhou, Q. Liao, H. Chen, and C. Zheng, "Consistent coding scheme for single-image super-resolution via independent dictionaries," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 313–325, 2016.

[15] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, and E. Wu, "Handling motion blur in multi-frame super-resolution," in *IEEE Computer Vision and Pattern Recognition*, June 2015.

[16] M. S. M. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *IEEE Computer Vision and Pattern Recognition*, 2018.

[17] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: video restoration with enhanced deformable convolutional networks," in *IEEE Computer Vision and Pattern Recognition Workshops*, 2019.

[18] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *IEEE Computer Vision and Pattern Recognition*, 2019.

[19] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: temporally-deformable alignment network for video super-resolution," in *IEEE Computer Vision and Pattern Recognition*, 2020.

[20] M. Chu, Y. Xie, J. Mayer, L. Leal-Taixe, and N. Thuerey, "Learning temporal coherence via self-supervision for gan-based video generation," in *SIGGRAPH*, 2020.

[21] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *IEEE Computer Vision and Pattern Recognition*, 2018.

[22] P. Yi, Z. Wang, K. Jiang, J. Jiang, and J. Ma, "Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations," in *IEEE International Conference on Computer Vision*, 2019.

[23] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *IEEE International Conference on Computer Vision*, 2017.

[24] L. Wang, Y. Guo, Z. Lin, X. Deng, and W. An, "Learning for video super-resolution through HR optical flow estimation," in *Asian Conference on Computer Vision*, 2018.

[25] Y. Huang, W. Wang, and L. Wang, "Bidirectional recurrent convolutional networks for multi-frame super-resolution," in *Conference on Neural Information Processing Systems*, 2015.

[26] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *IEEE Computer Vision and Pattern Recognition*, 2017.

[27] H. Lin, X. He, L. Qing, Q. Teng, and S. Yang, "Improved low-bitrate hevc video coding using deep learning based super-resolution and adaptive block patching," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3010–3023, 2019.

[28] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia, "Video super-resolution via deep draft-ensemble learning," in *IEEE International Conference on Computer Vision*, 2015.

[29] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *IEEE Computer Vision and Pattern Recognition*, 2016.

[30] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, 2019.

[31] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *IEEE International Conference on Computer Vision*, 2017.

[32] T. Isobe, X. Jia, S. Gu, S. Li, S. Wang, and Q. Tian, "Video super-resolution with recurrent structure-detail network," in *European Conference on Computer Vision*, 2020.

[33] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *IEEE Computer Vision and Pattern Recognition*, 2020.

[34] S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K. M. Lee, "Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study," in *IEEE Computer Vision and Pattern Recognition Workshops*, June 2019.

[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[36] R. Brunelli, *Template Matching Techniques in Computer Vision: Theory and Practice*. Wiley Publishing, 2009.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[38] Z. Zhang, Z. Wang, Z. L. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *IEEE Computer Vision and Pattern Recognition*, 2019.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Computer Vision and Pattern Recognition*, 2016.

[40] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep laplacian pyramid networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2599–2613, 2018.

[41] T. Isobe, S. Li, X. Jia, S. Yuan, G. G. Slabaugh, C. Xu, Y. Li, S. Wang, and Q. Tian, "Video super-resolution with temporal group attention," in *IEEE Computer Vision and Pattern Recognition*, 2020.

[42] W. Li, X. Tao, T. Guo, L. Qi, J. Lu, and J. Jia, "MuCAN: Multi-correspondence aggregation network for video super-resolution," in *European Conference on Computer Vision*, 2020.

[43] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Basicvsr: The search for essential components in video super-resolution and beyond," in *IEEE Computer Vision and Pattern Recognition*, 2021.

[44] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *European Conference on Computer Vision*, 2018.

[45] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Computer Vision and Pattern Recognition*, 2018.