Arbitrary-Scale Video Super-resolution Guided by Dynamic Context

Cong Huang[1]*, Jiahao Li[2], Lei Chu[2], Dong Liu[1], Yan Lu[2]

[1]University of Science and Technology of China,
[2] Microsoft Research Asia
hcy96@mail.ustc.edu.cn, dongeliu@ustc.edu.cn, {li.jiaha, lei.chu, yanlu}@microsoft.com

Abstract

We propose a Dynamic Context-Guided Upsampling (DCGU) module for video super-resolution (VSR) that leverages temporal context guidance to achieve efficient and effective arbitrary-scale VSR. While most VSR research focuses on backbone design, the importance of the upsampling part is often overlooked. Existing methods rely on pixelshuffle-based upsampling, which has limited capabilities in handling arbitrary upsampling scales. Recent attempts to replace pixelshuffle-based modules with implicit neural function-based and filter-based approaches suffer from slow inference speeds and limited representation capacity, respectively. To overcome these limitations, our DCGU module predicts non-local sampling locations and content-dependent filter weights, enabling efficient and effective arbitrary-scale VSR. Our proposed multi-granularity location search module efficiently identifies non-local sampling locations across the entire low-resolution grid, and the temporal bilateral filter modulation module integrates content information with the filter weight to enhance textual details. Extensive experiments demonstrate the superiority of our method in terms of performance and speed on arbitrary-scale VSR.

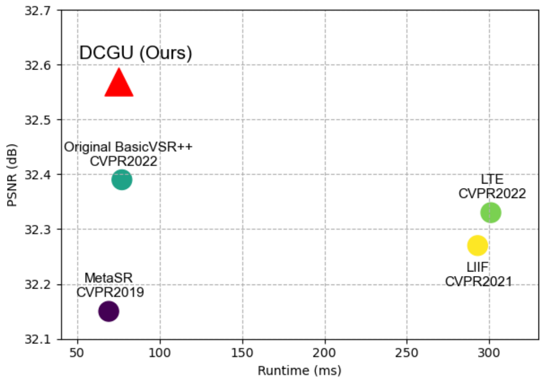


Figure 1: PSNR-Runtime comparison with different arbitrary-scale upsampling modules, including filter-based MetaSR (Hu et al. 2019), and implicit neural function-based LTE (Lee and Jin 2022) and LIIF (Chen, Liu, and Wang 2021). MetaSR, LTE, LIIF and our DCGU use BasicVSR++ (Chan et al. 2021) as the backbone.

Introduction

The task of video super-resolution (VSR) involves reconstructing high-resolution (HR) videos from low-resolution (LR) observations. While there have been numerous existing works (Wang et al. 2019; Isobe et al. 2020b, 2022; Chan et al. 2020, 2021) proposing different approaches to address the VSR problem, most of them focus on designing powerful backbones that incorporate motion alignment and feature extraction components. Interestingly, the upsampling module - a crucial yet often overlooked final step in generating HR videos (Wang, Chen, and Hoi 2020; Liu et al. 2022; Willets et al. 2017; Anwar, Khan, and Barnes 2020) - has received limited attention in these works.

Existing VSR methods typically employ the pixelshuffle-based upsampling module (Shi et al. 2016a). While this module is easy to implement, it can only support super-resolving features with fixed scale factors (e.g., x4) and cannot handle arbitrary ones. However, in real-world scenarios, scaling up LR videos with user-desired scales has more practical value. Recent work on arbitrary-scale single-image super-resolution (Hu et al. 2019; Chen, Liu, and Wang 2021; Lee and Jin 2022; Cao et al. 2022) has explored the possibility of replacing the pixelshuffle-based upsampling module and supporting arbitrary-scale super-resolution. These methods can be divided into two categories: the implicit neural function-based (Chen, Liu, and Wang 2021; Lee and Jin 2022; Cao et al. 2022) and the filter-based (Hu et al. 2019) approaches.

The implicit neural function-based methods utilize implicit neural representations to predict RGB values for specific coordinates in HR space, resulting in good visual quality. However, applying such methods directly to the VSR pipeline can severely impact inference speed due to a heavy MLP computation for each coordinate in the HR grid, making it inefficient for VSR, as shown in Fig. 1. On the other hand, the filter-based method (Hu et al. 2019) employs an MLP model to predict dynamic scale-location-dependent filter weights and then upsamples LR features at arbitrary scales. While this method has a fast inference speed as the

*This work was done when Cong Huang was an intern in Microsoft Research Asia..

filter weights only need to be predicted once for all video frames, it cannot outperform the implicit neural function-based methods in terms of result quality due to its limited representation capacity, as shown in Fig. 1. Therefore, the question arises: **Can an efficient yet effective filter-based upsampling module for arbitrary-scale VSR be designed?**

To answer this question, we revisit the filter-based method and identify two key factors that limit its representation ability: fixed local sampling locations and content-irrelevant filter weights. Specifically, the existing methods only sample local points centered at the corresponding projected coordinate in the LR grid for upsampling, which ignores long-range contextual information that can facilitate the generation of clearer structures for super-resolution. Additionally, given a scale factor and coordinate, the filter weights are the same for different contents without considering the texture similarity among them, resulting in blurry results. Solving these limitations is not straightforward and presents a significant challenge due to the absence of information in the HR grid.

Our proposed method is built upon the filter-based approach but incorporates temporal information in a video sequence to overcome the aforementioned challenges, inspired by the effectiveness of temporal information in previous work (Huang et al. 2022, 2023; Li, Li, and Lu 2022, 2023). We introduce the **D**ynamic **C**ontext-**G**uided **U**psampling (DCGU) module, which is both efficient and effective, aligning with the filter-based method's principles while utilizing natural video content coherence to address its limitations. Instead of relying solely on fixed coordinate-based sampling locations and content-irrelevant filter weight prediction, DCGU dynamically guides sampling content-dependent points and generates content-dependent weights using the temporal HR context. Specifically, we propose a confidence-guided context generation module to efficiently generate reliable HR context from temporal HR features and current LR features. Then, we propose the Multi-Granularity Locations Search (MGLS) module to identify the correlated points across the entire LR grid by leveraging feature similarities for each point in the HR grid. MGLS divides the whole-grid search into two granularity levels of search, global patch search and local pixel search, allowing it to efficiently benefit from both global and local receptive fields. Lastly, we propose the Temporal Bilateral Filter Modulation (TBFM) module to adaptively predict the filter weights using the similarity between the HR context and the LR feature points. The normalized similarity computed in MGLS can be reused in TBFM to modulate filter weights, making TBFM almost cost-free to enhance textual details.

The proposed method outperforms the filter-based method MetaSR (Hu et al. 2019) while maintaining comparable efficiency, as demonstrated in Fig.1. Moreover, the proposed method surpasses both implicit neural function-based methods LIIF (Chen, Liu, and Wang 2021) and LTE (Lee and Jin 2022) in terms of both performance and speed, highlighting its superiority.

The contributions of this paper are three-fold:

- We propose a novel dynamic context-guided upsampling module that achieves efficient and effective arbitrary-scale VSR. Unlike previous methods that rely solely on fixed coordinate-based location sampling and content-irrelevant filter weight prediction, our method leverages context guidance to learn to predict non-local sampling locations and content-dependent filter weights. This allows us to explicitly integrate content information, improving the representation ability of the filter-based method.
- We introduce the Multi-Granularity Location Search (MGLS) module that efficiently exploits long-range contextual information via a divide-and-conquer paradigm. Additionally, the Temporal Bilateral Filter Modulation (TBFM) can integrate content information into the filter weights, improving visual quality at a low cost.
- We conduct extensive experiments that demonstrate the superiority of our method in terms of performance and speed at arbitrary scales. Particularly, our method is better and several times faster than the implicit neural function-based method LTE.

Related Work

Video Super-Resolution

Existing VSR methods can be generally divided into two categories: the sliding window-based model and the recurrent model. The sliding window-based methods directly take several adjacent frames as input for each frame. Some methods (Caballero et al. 2017; Kappeler et al. 2016; Shi et al. 2016b; Tao et al. 2017) compute the optical flow to warp the adjacent frames to the center frame as the temporal context. EDVR (Wang et al. 2019) proposes cascaded deformable convolutions to align the adjacent frames in the feature space. MuCAN (Li et al. 2020) utilizes a temporal multi-correspondence aggregation strategy to boost the alignment accuracy. TGA (Isobe et al. 2020b) splits adjacent frames into groups with different temporal dilation and proposes temporal group attention.

For the recurrent model, each frame takes the feature or HR result from the previous or future frame as input. FRVSR (Sajjadi, Vemulapalli, and Brown 2018) is a pioneering work that uses the optical flow to warp the estimated HR result to the current frame. RLSP (Fuoli, Gu, and Timofte 2019) designs a fully convolutional recurrent network to propagate the hidden state that contains abstract information without explicit warping. To improve the temporal propagation efficiency, RSDN (Isobe et al. 2020a) divides the content into structure and detail components. BasicVSR (Chan et al. 2020) is a strong benchmark by redesigning the essential components therein and leveraging the bi-directional propagation. BasicVSR++ (Chan et al. 2021) further improves the performance via the second-order grid propagation and the flow-guided deformable alignment. However, above methods all use the pixelshuffle-based upsampling module and could not handle arbitrary-scale VSR.

Arbitrary-Scale Super-Resolution

Arbitrary-scale super-resolution methods aim to super-resolve LR images with arbitrary scales in a single model,

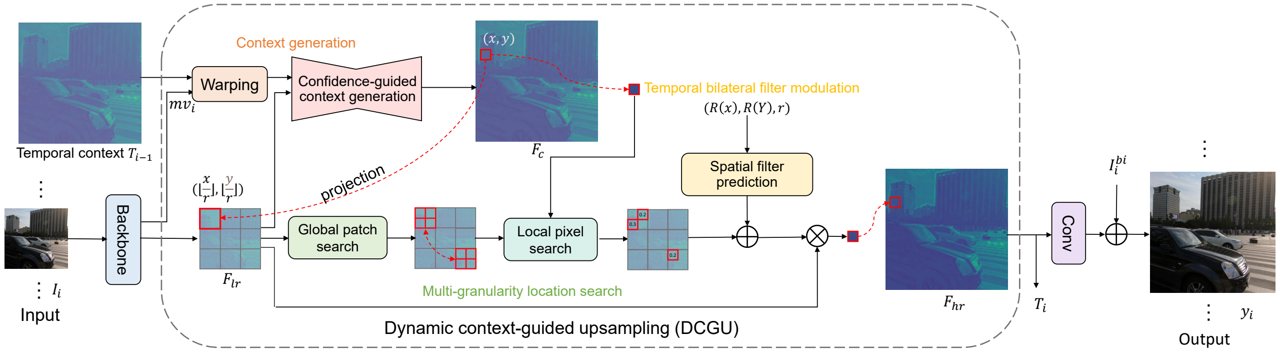

Temporal context T_{i-1}
Context generation
Warping
mv_i
Confidence-guided context generation
(x, y)
F_c
Temporal bilateral filter modulation
$(R(x), R(Y), r)$
Spatial filter prediction
Input
I_i
Backbone
$([\frac{x}{r}], [\frac{y}{r}])$
projection
F_{lr}
Global patch search
Local pixel search
0.2
0.3
0.2
Multi-granularity location search
Dynamic context-guided upsampling (DCGU)
F_{hr}
T_i
Conv
I_i^{bi}
y_i
Output

Figure 2: The overall framework of the proposed DCGU. Context generation produces the refined HR context from LR features and temporal context. With guidance of HR context, multi-granularity location search identifies sampling locations from the entire LR grid and temporal bilateral filter modulation predicts content-dependent filter weights to aggregate LR feature points.

which is convenient for practical usage. The earliest learning-based method is MetaSR (Hu et al. 2019) that uses a two-layer MLP to predict scale-dependent kernel weights to upsample the feature. The following method ArbSR (Wang et al. 2021) improves MetaSR via introducing the scale-independent attention into the backbone. Rather than predicting the scale-dependent kernel weights, the recent work LIIF (Chen, Liu, and Wang 2021) uses implicit neural representations to process the feature maps, coordinates, and scaling factor via the MLP to obtain RGB values at specific coordinates. Following LIIF, LTE (Lee and Jin 2022) proposes a dominant-frequency estimator to enable the implicit function to capture fine details. Although above methods support the arbitrary-scale super-resolution, directly applying them to VSR is sub-optimal, suffering from slow inference speeds or limited representation capacity. Recently, VideoINR (Chen et al. 2022) investigates the application of LIIF for space-time VSR while adapting it to the space VSR we focus on in this paper, also encounters limitations in terms of slow inference speed.

Method

Preliminary and Motivation

Given the LR video clip of size $h \times w \times c \times t$ and an arbitrary scale r, our target is to reconstruct an HR video clip of size $rh \times rw \times c \times t$. Specially, at time t, with the current LR features F_{lr} extracted by the backbone and the 2D coordinate $p = (x, y)$ in the HR grid, the filter-based arbitrary-scale upsampling method generates the HR feature point $F_{hr}(p)$ as:

$$F_{hr}(p) = \sum_{\Delta p \in \mathcal{R}(p,r)} f_{wp}(p, r, \Delta p) F_{lr}(\Delta p) \tag{1}$$

where $f_{wp}(p, r, \Delta p)$ is the filter weights and wp stands for weight prediction. $\mathcal{R}(p, r)$ is the fixed local sampling locations centered at the corresponding LR coordinate.

As explicated in the introduction, the effectiveness of the filter-based method is hindered by the fixed local sampling locations and content-irrelevant filter weights. To address these limitations while sustaining efficiency, our method incorporates temporal information from video sequences. We propose the dynamic context-guided upsampling (DCGU) module that introduces the context guidance to identify more informative global sampling locations, and generate content-dependent filter weights to enhance the texture detail. In specific, DCGU module could be formulated as:

$$F_{hr}(p) = \sum_{\Delta p \in \mathcal{R}(p,r,F_{lr},F_c)} f_{wp}(p, r, \Delta p, F_{lr}, F_c) F_{lr}(\Delta p) \tag{2}$$

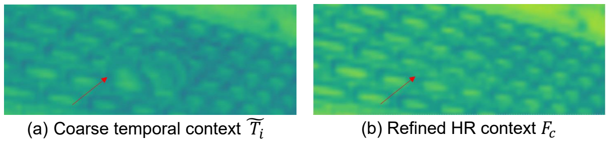
We not only integrate LR features F_{lr} in both $\mathcal{R}(\cdot)$ and $f_{wp}(\cdot)$ to enhance their content-awareness but also introduce the HR context F_c, which effectively encompasses the essential contextual information needed for upsampling, to guide the utilization of F_{lr}. In particular, we exploit the correlation between F_{lr} and F_c, enabling DCGU to aggregate more informative global sampling locations with more accurate content-dependent filter weights to yield superior performance.

As shown in Fig. 2, DCGU can be divided into three key components. Firstly, the context generation module establishes a connection between the LR grid and the HR grid by generating the HR context denoted as F_c. Secondly, the multi-granularity location search module obtains the sampling location set $R(p, r, F_{lr}, F_c)$ from the entire LR grid for each point within the HR grid. Finally, the temporal bilateral filter modulation module produces content-dependent filter weights $f_{wp}(p, r, \Delta p, F_{lr}, F_c)$.

Context Generation

The context generation module is responsible for generating the HR context that establishes a connection between the LR grid and the HR grid. However, directly deriving an accurate HR context from LR features is a challenging task that significantly increases the computational cost. To ensure both efficiency and effectiveness, we exploit the HR temporal context that is propagated from the previous time step. By adaptively fusing it with LR features, we generate F_c that captures the necessary context information for upsampling.

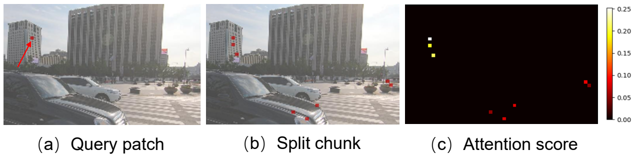
The context generation module in Fig. 2 contains two steps: warping and confidence-guided context generation.



(a) Coarse temporal context $\widetilde{T_i}$

(b) Refined HR context F_c

Figure 3: (a) The coarse temporal context suffers from misalignment. (b) The confidence-guided context module compensates for the misaligned region and generates a more accurate refined HR context.



(a) Query patch

(b) Split chunk

(c) Attention score

Figure 4: The visual analysis for global patch search.

The warping step aligns the temporal context from the previous time step to the current one. Since obtaining HR motion information is challenging, we utilize the motion information mv_i computed from LR frames in the backbone to perform motion alignment. However, as shown in Fig. 3 (a), the coarse HR temporal context $\widetilde{T_i}$ after warping may still suffer from misalignment, despite having clear texture details. To address this issue, we introduce confidence-guided context generation to generate the refined HR context F_c by adaptively fusing $\widetilde{T_i}$ with the current LR feature F_{lr}. The misaligned regions of $\widetilde{T_i}$ are compensated for by F_{lr}, as follows:

$$F_c = M \cdot \uparrow F_{lr} + (1 - M) \cdot \widetilde{T_i} \tag{3}$$

where M is the confidence map that indicates the misaligned degree and is estimated via a confidence estimation model $f_{ce}(\cdot)$ as $M = f_{ce}(\uparrow F_{lr}, \widetilde{T_i})$. $\uparrow F_{lr}$ means the bicubic-upsampled LR feature. As shown in Fig. 3 (b), the refined HR context F_c gets rid of the distorted structures and enjoys more accurate details. In the ablation study, we investigate the impact of different contexts and demonstrate that the context generated by our proposed module achieves better performance at a lower computational cost.

Multi-Granularity Location Search

The multi-granularity location search (MGLS) module in DCGU aims to efficiently identify the sampling locations from the entire LR grid, guided by the refined HR context F_c. Naive non-local search methods can be used to obtain $R(p, r, F_{lr}, F_c)$, but the computational complexity of such methods is $O(rh \times rw \times h \times w)$, which poses a significant burden on efficiency. To overcome this challenge, we leverage the inherent correlation between HR context F_c and LR features F_{lr}, specifically the strong association between point $F_c(p)$ and the points within the local patch centered at $p' = (\lfloor \frac{x}{r} \rfloor, \lfloor \frac{y}{r} \rfloor)$, where $\lfloor \rfloor$ denotes the floor function in F_{lr}. By initially calculating the non-local patch correlation in F_{lr} once, the subsequent search for point p in the HR grid can be efficiently conducted within the corresponding local patch and its globally correlated patches, thereby reducing computational redundancy.

In light of this, we propose a novel MGLS to efficiently exploit the long-range contextual information. MGLS follows the divide-and-conquer paradigm, dividing the whole-grid search between F_{lr} and F_c into two levels of granularity: global patch search (GPS) in F_{lr} and local pixel search (LPS) between F_{lr} and F_c, as illustrated in Fig. 2. GPS aims to capture the global correlations in the LR feature F_{lr}, while LPS estimates $R(p, r, F_{lr}, F_c)$ in several local windows in F_{lr} based on the result of GPS. MGLS achieves a significant reduction in computational complexity by exploiting the global receptive field efficiently, decreasing the complexity of the naive non-local search from $O(rh \times rw \times h \times w)$ to $O(rh \times rw \times s \times s)$, where s is the patch size and much smaller than h or w.

Global patch search The global patch search (GPS) module aims to identify global patch correlations throughout the entirety of F_{lr}. However, previous pair-wise non-local attention used in super-resolution method (Dai et al. 2019) introduces quadratic complexity along the input resolution, resulting in significant computational costs. Recently, sparse non-local attention-based approaches (Lee et al. 2022; Mei, Fan, and Zhou 2021) have utilized locality-sensitive hashing (Andoni et al. 2015) (LSH) to reduce these expenses. While being effective, these methods sort patches according to the hash code and split them into fixed-size chunks, potentially causing dissimilar patches to drop in the same chunk, as illustrated in Fig. 4 (b). This reliance on dissimilar patches could lead to sub-optimal performance.

To efficiently find accurate global patch correlation, we combine pair-wise non-local attention and sparse non-local attention, taking advantage of both approaches. Specifically, we first use LSH in the sparse non-local attention to obtain the initial global patch correlation. LSH works by hashing similar patches to the same bucket with high probability, thus avoiding the need to compute pair-wise attention across the entire F_{lr}. We split F_{lr} into non-overlapping patches, $F_{lr} = \{f_i | i = 0, 1, ..., N - 1\}$, and map these patches to one-dimensional vectors using a random rotation matrix R, with hash codes calculated as

$$H(f_i) = \arg\max([R \cdot \frac{f_i}{||f_i||}, -R \cdot \frac{f_i}{||f_i||}]) \tag{4}$$

where $[\cdot, \cdot]$ denotes the concatenation. Similar patches will be hashed into the same bucket with the same hash code with high probability, but the bucket may have an unbalanced amount of patches. We sort the patches according to hash codes and split them into fixed-sized chunks. Next, to filter out dissimilar patches in each chunk, we introduce pair-wise attention. The attention score for f_i in a chunk is computed as:

$$GA_{i,\cdot} = \{ \frac{\phi(f_i)\theta(f_j)}{\sum_j \phi(f_i)\theta(f_j))} | j \in N_c \} \tag{5}$$

where ϕ and θ are the projection matrix, and N_c is the set of patch indices of the chunk that f_i belongs to.

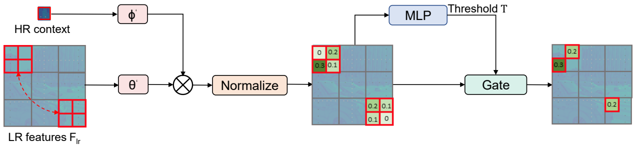
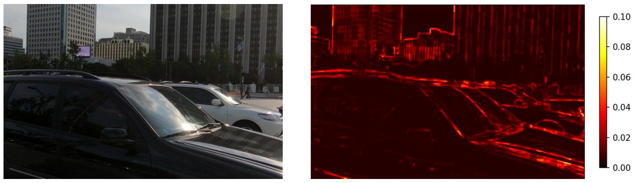


Figure 5: The paradigm overview of local pixel search.



(a) Current frame

(b) Learnt threshold for LPS

Figure 6: Visualization of the input frame and the corresponding learned threshold for LPS. The threshold changes with the frequency of image content.

Using the attention scores, we obtain globally-connected patches, $P(f_i)$, for each patch f_i by selecting the top k patches with the highest attention scores. As illustrated in Fig. 4 (c), similar patches have higher attention scores compared to their dissimilar counterparts, allowing us to filter out dissimilar patches effectively. In our experiments, setting k to 1 results in a substantial improvement.

Local pixel search After knowing the global patch correlation, the sampling location search is carried out in several local windows in F_{lr}, namely LPS. Specially, in LPS, we introduce a dynamic gating function that utilizes similarity scores to distinguish informative points. As shown in Fig. 5, given a point $F_c(p)$ in the HR grid with spatial position $p(x, y)$, we first project its position into the LR grid as $p' = (\lfloor \frac{x}{r} \rfloor, \lfloor \frac{y}{r} \rfloor)$ and locate the LR feature patch $F_{lr,p'}$ and its globally-connected patch $P(F_{lr,p'})$. We calculate the similarity score $s_{p,j}$ between $F_c(p)$ and each point LP_j in $F_{lr,p'} \cup P(F_{lr,p'})$ as :

$$s_{p,j} = \frac{\phi'(F_c(p))\theta'(LP_j)}{\sum_j \phi'(F_c(p))\theta'(LP_j)} \tag{6}$$

Instead of using a fixed threshold, the dynamic gating function adaptsively predicts the threshold using an MLP as $T_p = \mathrm{MLP}(s_{p,\cdot})$. Then, sampling locations are generated by collecting points with attention scores exceeding T_p.

We visualize the learned threshold map in Fig. 6, which varies with the frequency of image content, with higher thresholds in high-frequency regions involving rapid content change.

Discussion Our proposed method is related to the Deformable Convolutional Networks (DCN) (Dai et al. 2017) but is different in several ways. DCN uses deformable convolutional layers to predict relative offsets that are used to construct the sampling location set. While DCN can capture spatial transformations in the data, the range of offsets is usually limited to avoid model divergence, making it challenging for DCN to leverage long-range contextual information. Moreover, original DCN uses the same weights to aggregate the points for all position for all scales, making it infeasible for arbitrary-scale VSR. In contrast, our proposed method leverages context information to predict non-local sampling locations and content-dependent filter weights, enabling us to capture long-range dependencies efficiently. In the ablation study, we explore the effect of using a modified DCN in our method.

Temporal Bilateral Filter Modulation

In order to generate the final HR feature from the LR points, it is necessary to estimate filter weights. However, previous methods have not taken the content information into account when generating filter weights, resulting in blurred results. To address this issue, we propose an efficient temporal bilateral filter modulation (TBFM) that predicts content-dependent filter weights. Our approach is inspired by the traditional bilateral filter, which preserves sharp edges in images by considering range differences, and we extend this concept to arbitrary-scale VSR.

TBFM module contains two steps: predicting the spatial filter $a(\Delta p, p)$ and modulating the spatial filter with the range filter. Firstly, for spatial filter prediction, as illustrated in Fig. 2, we generate the relative position offset $O(p) = (\frac{x}{r} - \lfloor \frac{x}{r} \rfloor, \frac{y}{r} - \lfloor \frac{y}{r} \rfloor)$ and feed it, along with the scale r, to an MLP model to obtain the spatial filter. The number of locations in the sampling set is not fixed, so we introduce a additional fixed rule $s_{p,j} \in \mathrm{topn}(s_{p,\cdot})$ to ensure that the number of locations in the sampling set does not exceed the size of the spatial filter. Secondly, we modulate the spatial filter by adding the range filter, which is defined as the similarity between the HR point and the LR point. Our motivation is that the LR point, being more similar to the HR point, should have a larger weight. We generate the final filter $f_{wp}(p, r, \Delta p, F_{lr}, F_c)$ by modulating the spatial filter as:

$$f_{wp}(p, r, \Delta p, F_{lr}, F_c) = a(\Delta p, p) + \delta \cdot \mathcal{S}(F_c(p), F_{lr}(\Delta p)) \tag{7}$$

where δ is a learnable parameter and $\mathcal{S}$ is a function that calculates the similarity. Numerous options exist for $\mathcal{S}$. For instance, the function that computes the negative Euclidean distance between $F_c(p)$ and $F_{lr}(\Delta p)$ could serve as $\mathcal{S}$. However, the value ranges of $F_c(p)$ and $F_{lr}(\Delta p)$ vary, leading to unbalanced value ranges for $\mathcal{S}$ and subsequent performance degradation. We contend that a balanced value range of $\mathcal{S}$ for different content is advantageous. To achieve this, normalizing $\mathcal{S}$ among the sampling location set is an appropriate choice. Fortunately, we have already computed the normalized similarity in Eqn. 6, and we can reuse it without re-computation. Our ablation study demonstrates that our reused normalized similarity yields greater improvement than the unnormalized distance.

			BI degradation		
	Params (M)	Runtime (ms)	REDS4	Vimeo-90K-T	Vid4
BasicVSR	6.3	63	31.42/0.8909	37.18/0.9450	27.24/0.8251
BasicVSR_DCGU	6.4	61	31.57/0.8937	37.32/0.9468	27.34/0.8273
BasicVSR++	7.3	77	32.39/0.9069	37.79/0.9500	27.79/0.8400
BasicVSR++_DCGU	7.4	75	32.57/0.9082	37.96/0.9524	27.85/0.8404

Table 1: Quantitative comparison (PSNR/SSIM). All results are calculated on Y-channel except REDS4(RGB-channel).

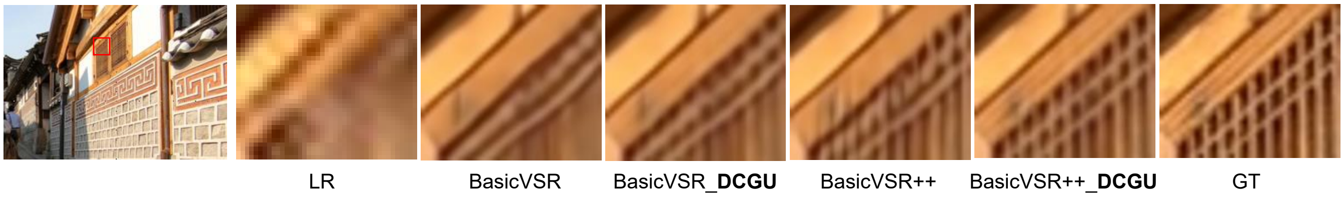


Figure 7: The visual comparison of BasicVSR, BasicVSR_DCGU, BasicVSR++ and BasicVSR++_DCGU at x4 scale.

Experiment

Dataset and Evaluation

We train our model on REDS (Nah et al. 2019) and Vimeo-90K (Xue et al. 2019). For REDS (Nah et al. 2019), we use REDS4 as testset. For Vimeo-90K (Xue et al. 2019) dataset, we use Vimeo-90K-T (Xue et al. 2019), Vid4 (Liu and Sun 2013) and UDM10 (Yi et al. 2019) as testset. We use Bicubic (BI) and Blur Downsampling (BD) to generate LR videos, separately. PSNR and SSIM are used as evaluation metrics. The model size and inference time are used to measure the efficiency. The implementation details and the result about BD degradation is in the appendix.

Performance Comparison at x4 Scale

Quantitative results As Tab. 1 shows, BasicVSR_DCGU performs better than BasicVSR (Chan et al. 2020) on all testsets. It is worth noting that although BasicVSR_DCGU has more parameters than BasicVSR, BasicVSR_DCGU enjoys a faster speed. The reason is that most parameters of DCGU come from the spatial filter prediction module that infers once for each video. Therefore, the extra parameters do not affect the speed. Moreover, the pixelshuffle-based upsampling module in the original BasicVSR performs convolution twice on HR feature, which slows down the speed. Besides, BasicVSR++_DCGU also has better performance and faster speed than BaiscVSR++. These results verify the superiority of the proposed DCGU.

Qualitative results We also present a qualitative comparison in Fig. 7. The proposed DCGU helps BasicVSR and BasicVSR++ generate finer details and sharper edges. For example, BasicVSR++_DCGU produces clearer striped details than BasicVSR++.

Performance Comparison at Arbitrary Scales

For the comparison at arbitrary scales, we construct three baselines that utilize BasicVSR++(Chan et al. 2021) as the backbone. They are created by replacing the pixelshuffle-based upsampling module with arbitrary-scale upsampling modules, including MetaSR(Hu et al. 2019), LIIF (Chen, Liu, and Wang 2021), and LTE (Lee and Jin 2022).

In-distribution evaluation Firstly, we perform an in-distribution evaluation, where all scale factors are exposed to the model during training. Tab. 2 reveals that MetaSR, LIIF, and LTE exhibit inferior performance compared to our method at all scales, demonstrating the effectiveness of our approach.

Out-of-distribution evaluation In practical applications, a wide variety of scales may be encountered, and it is not feasible to cover all scales during training. As a result, evaluating the model's performance on unseen scales, referred to out-of-distribution evaluation, is crucial for arbitrary-scale VSR. We present the results in Tab. 3. Our proposed method outperforms MetaSR, LIIF, and LTE by leveraging long-range contextual information. A visual comparison is provided in Fig. 8, which demonstrates that our method generates more detailed textures, leading to a more visually pleasing result.

Efficiency analysis We present the number of parameters for the entire model and the runtime (measured at the x4 scale) in Tab. 2. The runtime in Tab. 2 of MetaSR, LIIF, LTE and our is the runtime of the backbone of BasicVSR++ (Chan et al. 2021) plus the runtime of the upsampling module. Compared to MetaSR, LIIF and LTE exhibit slower speeds, as they require the LR feature, along with coordinate and scale, to be fed into the MLP at each time step. In contrast, our proposed method achieves a superior efficiency-performance trade-off, owing to the novel design of the dynamic context-guided upsampling module.

Ablation Study

In this section, we conduct the ablation study on other alternative designs in context generation, multi-granularity location search, and temporal bilateral filter modulation.

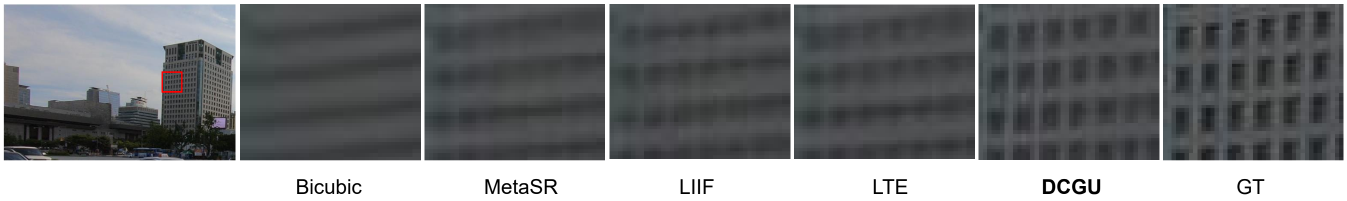


Figure 8: The visual comparison at out-of-distribution scale (x8).

	Params(M)	Runtime(ms)	x2.0	x2.5	x3.0	x3.5	x4.0
MetaSR	7.4	69	38.04/0.9626	36.60/0.9541	34.87/0.9386	33.37/0.9219	32.15/0.9040
LIIF	7.3	293	38.07/0.9627	36.58/0.9538	34.89/0.9387	33.44/0.9223	32.27/0.9054
LTE	7.4	301	38.10/0.9629	6.67/0.9543	34.92/0.9390	33.46/0.9226	32.31/0.9056
DCGU	7.4	75	38.55/0.9653	37.05/0.9569	35.19/0.9408	33.70/0.9248	32.57/0.9082

Table 2: Comparison for in-distribution scales.

	x8.0	x12.0
MetaSR	27.06/0.7544	25.11/0.6617
LIIF	27.19/0.7549	25.17/0.6629
LTE	27.20/0.7550	25.25/0.6631
DCGU	27.46/0.7573	25.46/0.6658

Table 3: Comparison for out of distribution scales.

	Ours	Bicubic	MetaSR
PSNR (dB)	32.57	32.21	32.52
Runtime (ms)	75	71	82

Table 4: Ablation study on HR context generation.

LPS	DG	GPS	DCN	PSNR (dB)
				32.27
√				32.38
√	√			32.45
√	√	√		32.57
			√	32.35

Table 5: Ablation study on multi-granularity location search.

	Base	Base + negetive L2	Ours
PSNR (dB)	32.48	32.50	32.57

Table 6: Ablation study on TBFM.

Context generation In this study, we use bicubic and MetaSR, for upsampling the LR features to generate HR contexts that guide MGLS and TBFM. As evidenced by Tab. 4, utilizing bicubic-upsampling for the HR context results in a significant performance degradation, which underscores the criticality of an accurate HR context. Although the HR context generated by MetaSR exhibits similar performance to our method, it is accompanied by a longer runtime. Consequently, these findings highlight the advantages of the proposed context generation module.

Multi-granularity location search As Tab. 5 shows, the baseline, using local information without search, suffers from a lower PSNR. Using LPS brings a 0.11 dB improvement in PSNR, while the addition of dynamic gating (DG) further enhances PSNR by 0.07 dB. These improvements emphasize the importance of adaptively preserving informative features for upsampling. Including GPS contributes to an 0.12 dB improvement, highlighting the advantages of non-local contextual information. As previously discussed, DCN (Dai et al. 2017) also could determine sampling locations by predicting the relative offset. However, when DCN is employed , PSNR degrades to 32.35 dB, thereby demonstrating the superiority of the proposed module .

Temporal bilateral filter modulation As illustrated in Tab. 6, the base model employing only the spatial filter achieves PSNR of 32.48 dB. When the negative L2 distance is utilized as the range filter to modulate the spatial filter, the improvement is only 0.02 dB. In contrast, the usage of our normalized similarity results in a more substantial performance enhancement of 0.09 dB.

Conclusion

This paper addresses the arbitrary-scale VSR problem. The dynamic context-guided upsampling (DCGU) module that are both efficient and effective is proposed. DCGU overcomes the limitations of current filter-based methods by exploiting the content coherence inherent in natural videos. DCGU introduces the confidence-guided context generation module, multi-granularity locations search module, and temporal bilateral filter modulation module to effectively and efficiently generate reliable HR context, identify correlated points across the entire LR grid and aggregate them in the content-dependent manner. The extensive experiments demonstrate the superiority of the proposed method.

References

Andoni, A.; Indyk, P.; Laarhoven, T.; Razenshteyn, I.; and Schmidt, L. 2015. Practical and optimal LSH for angular distance. *Advances in neural information processing systems*, 28.

Anwar, S.; Khan, S.; and Barnes, N. 2020. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 53(3): 1–34.

Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Real-Time Video Super-Resolution With Spatio-Temporal Networks and Motion Compensation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cao, J.; Wang, Q.; Xian, Y.; Li, Y.; Ni, B.; Pi, Z.; Zhang, K.; Zhang, Y.; Timofte, R.; and Van Gool, L. 2022. CiaoSR: Continuous Implicit Attention-in-Attention Network for Arbitrary-Scale Image Super-Resolution. *arXiv preprint arXiv:2212.04362*.

Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2020. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. *arXiv preprint arXiv:2012.02181*.

Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2021. BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment. *arXiv preprint arXiv:2104.13371*.

Chen, Y.; Liu, S.; and Wang, X. 2021. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8628–8638.

Chen, Z.; Chen, Y.; Liu, J.; Xu, X.; Goel, V.; Wang, Z.; Shi, H.; and Wang, X. 2022. Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2047–2057.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, 11065–11074.

Fuoli, D.; Gu, S.; and Timofte, R. 2019. Efficient Video Super-Resolution through Recurrent Latent Space Propagation. In *ICCV Workshops*.

Hu, X.; Mu, H.; Zhang, X.; Wang, Z.; Tan, T.; and Sun, J. 2019. Meta-SR: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1575–1584.

Huang, C.; Li, J.; Chu, L.; Liu, D.; and Lu, Y. 2023. Disentangle Propagation and Restoration for Efficient Video Recovery. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8336–8345.

Huang, C.; Li, J.; Li, B.; Liu, D.; and Lu, Y. 2022. Neural compression-based feature learning for video restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5872–5881.

Isobe, T.; Jia, X.; Gu, S.; Li, S.; Wang, S.; and Tian, Q. 2020a. Video Super-Resolution with Recurrent Structure-Detail Network. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 645–660. Cham: Springer International Publishing. ISBN 978-3-030-58610-2.

Isobe, T.; Jia, X.; Tao, X.; Li, C.; Li, R.; Shi, Y.; Mu, J.; Lu, H.; and Tai, Y.-W. 2022. Look Back and Forth: Video Super-Resolution with Explicit Temporal Difference Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17411–17420.

Isobe, T.; Li, S.; Jia, X.; Yuan, S.; Slabaugh, G.; Xu, C.; Li, Y.-L.; Wang, S.; and Tian, Q. 2020b. Video Super-Resolution With Temporal Group Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kappeler, A.; Yoo, S.; Dai, Q.; and Katsaggelos, A. 2016. Video Super-Resolution With Convolutional Neural Networks. In *IEEE Transactions on Computational Imaging*.

Lee, H.; Choi, H.; Sohn, K.; and Min, D. 2022. KNN Local Attention for Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2139–2149.

Lee, J.; and Jin, K. H. 2022. Local Texture Estimator for Implicit Representation Function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1929–1938.

Li, J.; Li, B.; and Lu, Y. 2022. Hybrid Spatial-Temporal Entropy Modelling for Neural Video Compression. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Li, J.; Li, B.; and Lu, Y. 2023. Neural Video Compression with Diverse Contexts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, Canada, June 18-22, 2023*.

Li, W.; Tao, X.; Guo, T.; Qi, L.; Lu, J.; and Jia, J. 2020. MuCAN: Multi-correspondence Aggregation Network for Video Super-Resolution. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 335–351. Cham: Springer International Publishing. ISBN 978-3-030-58607-2.

Liu, C.; and Sun, D. 2013. On Bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2): 346–360.

Liu, H.; Ruan, Z.; Zhao, P.; Dong, C.; Shang, F.; Liu, Y.; Yang, L.; and Timofte, R. 2022. Video super-resolution based on deep learning: a comprehensive survey. *Artificial Intelligence Review*, 1–55.

Mei, Y.; Fan, Y.; and Zhou, Y. 2021. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3517–3526.

Nah, S.; Baik, S.; Hong, S.; Moon, G.; Son, S.; Timofte, R.; and Lee, K. M. 2019. NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Sajjadi, M. S. M.; Vemulapalli, R.; and Brown, M. 2018. Frame-Recurrent Video Super-Resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016a. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1874–1883.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016b. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*.

Tao, X.; Gao, H.; Liao, R.; Wang, J.; and Jia, J. 2017. Detail-Revealing Deep Video Super-Resolution. In *The IEEE International Conference on Computer Vision (ICCV)*.

Wang, L.; Wang, Y.; Lin, Z.; Yang, J.; An, W.; and Guo, Y. 2021. Learning a single network for scale-arbitrary super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4801–4810.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Wang, Z.; Chen, J.; and Hoi, S. C. 2020. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3365–3387.

Willets, K. A.; Wilson, A. J.; Sundaresan, V.; and Joshi, P. B. 2017. Super-resolution imaging and plasmonics. *Chemical reviews*, 117(11): 7538–7582.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video Enhancement with Task-Oriented Flow. *International Journal of Computer Vision (IJCV)*, 127(8): 1106–1125.

Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; and Ma, J. 2019. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3106–3115.