Semantic Lens: Instance-Centric Semantic Alignment for Video Super-Resolution

Qi Tang[1,2], Yao Zhao[1,2], Meiqin Liu[1,2*], Jian Jin[3], Chao Yao[4*]

[1]Institute of Information Science, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China
[3]Alibaba-NTU Singapore Joint Research Institute, Nanyang Technological University, Singapore
[4]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
{qitang, yzhao, mqliu}@bjtu.edu.cn, jian.jin@ntu.edu.sg, yaochao@ustb.edu.cn

Abstract

As a critical clue of video super-resolution (VSR), inter-frame alignment significantly impacts overall performance. However, accurate pixel-level alignment is a challenging task due to the intricate motion interweaving in the video. In response to this issue, we introduce a novel paradigm for VSR named **Semantic Lens**, predicated on semantic priors drawn from degraded videos. Specifically, video is modeled as instances, events, and scenes via a Semantic Extractor. Those semantics assist the Pixel Enhancer in understanding the recovered contents and generating more realistic visual results. The distilled global semantics embody the scene information of each frame, while the instance-specific semantics assemble the spatial-temporal contexts related to each instance. Furthermore, we devise a **S**emantics-**P**owered **A**ttention **C**ross-**E**mbedding (SPACE) block to bridge the pixel-level features with semantic knowledge, composed of a **G**lobal **P**erspective **S**hifter (GPS) and an **I**nstance-**S**pecific **S**emantic **E**mbedding **E**ncoder (ISEE). Concretely, the GPS module generates pairs of affine transformation parameters for pixel-level feature modulation conditioned on global semantics. After that, the ISEE module harnesses the attention mechanism to align the adjacent frames in the instance-centric semantic space. In addition, we incorporate a simple yet effective pre-alignment module to alleviate the difficulty of model training. Extensive experiments demonstrate the superiority of our model over existing state-of-the-art VSR methods.

Introduction

Owing to various constraints, videos captured in real scenes suffer from degradations such as low resolution, blur, and noise, etc. Video super-resolution (VSR) aims to recover high-resolution (HR) video frames from their low-resolution (LR) counterparts with better visual quality. With the proliferation of high-definition displays and the ever-increasing demand for superior video contents, VSR has gained widespread attention. It not only improves the viewer's visual enjoyment, but also holds potential applications in fields like film restoration (Wan et al. 2022), medical imaging (Lu et al. 2023) and satellite imagery (Deudon et al. 2020).

Compared with single image super-resolution (SISR), adjacent frames usually contain highly related contents, which
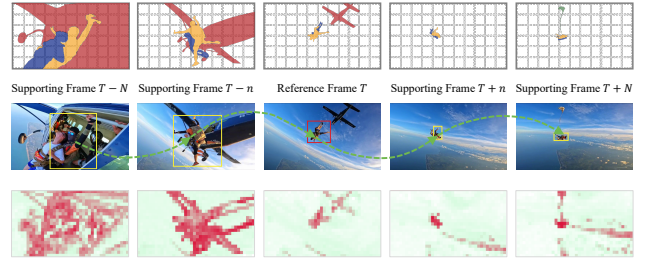
---

*Corresponding authors.

Figure 1: **Top**: Global semantics and instance-specific semantics. **Middle**: Original frames and event information (indicated by green arrows). **Bottom**: Patch PSNR heat map of five frames in a video, super-resolved by a single image super-resolution model. A clear boundary shows that PSNR is strongly related to video content.

can provide complementary information for video super-resolution. Thus, the effective utilization of temporal information across video frames can yield fine-grained visual details. In recent years, an increasing number of VSR approaches can be categorized into two paradigms, explicit and implicit inter-frame alignment. The explicit alignment attempts to align supporting frames with the reference frame by using extracted motion information, including optical flow-based approaches at the image/feature level (Chan et al. 2021), optical flow-guided deformable convolution (Chan et al. 2022), trajectory-aware attention (Liu et al. 2022a), etc. The implicit motion estimation and motion compensation seamlessly extract spatio-temporal features from multiple frames like 3D convolution (Liu et al. 2021).

However, in the temporal dimension, accurate pixel-level alignment still poses a challenging task. On one hand, different contents exhibit varying restoration difficulties (Li et al. 2023). As illustrated in Figure 1, smooth regions (indicated in light green) are easy to recover, while textured areas (marked in red) are relatively difficult. The sub-optimal restoration results of texture-rich regions will hinder the accuracy of alignment, and even for easy-to-recover contents, high-frequency artifacts may be introduced due to misalignment (Zhou et al. 2022). On the other hand, in many real scenes, the camera's viewpoint is not fixed, leading to significant changes in the background between adjacent frames.

Beyond movement from the camera, the scene itself may also be constantly changing with lighting and compositional elements. Moreover, the trajectories of instances within the video are usually inconsistent (Liu et al. 2022b, 2023b). These diverse motion cues, when overlaid, make precise inter-frame alignment difficult. For instance, when facing information loss due to occlusion, pixel-level alignment may bring holes and blur.

To address these issues, we draw inspiration from how humans perceive images, which tend to automatically understand these contents hierarchically. For example, we might intuitively regard the front-running race car on a track as the foreground, while other parts are perceived as the background. Thus, a straightforward approach is to leverage advanced video analysis techniques to model the foreground and background of videos. The intuition behind this is to establish instance-centric semantic representations, and at the same time, to attribute diverse motion information contained in the video to the corresponding instance. Each instance or background is regarded as an independent movement unit. In this way, the original complex intertwined motion is now broken down into several simple and independent movements. Besides, the change of object scale and illumination seriously affects the accuracy of alignment in pixel space, while those factors can be ignored in semantic space. Therefore, the semantic association of same object in temporal dimension can support the alignment of objects in multi-frames, thus avoiding the complex interactions of multi-instance or background noise and the challenge of accurate pixel-level alignment.

Driven by the above analysis, we introduce **Semantic Lens**, a novel paradigm for video super-resolution. It combines the magnifying power of a lens with the semantic priors, enabling not only improving resolution but also generating enjoyable visual results. Specifically, we decouple the video, originally composed of frames, into instances, events and scenes via constructing a Semantic Extractor. To make use of those semantics, we develop a Semantics-Powered Attention Cross-Embedding (SPACE) block, comprising a Global Perspective Shifter (GPS) and an Instance-Specific Semantic Embedding Encoder (ISEE). It embeds the semantics into features extracted from LR frames in a position-embedding-like manner and enables the instance-centric inter-frame alignment with the guidance of semantic priors. Experimental results demonstrate that our method consistently outperforms existing state-of-the-art methods. The main contributions are summarized as follows:

- We pioneer a novel representation for video modeled as instances, events, and scenes, providing both global semantics and instance-specific semantics to boost the performance of video super-resolution.

- We propose a Semantics-Powered Attention Cross-Embedding block to bridge semantic priors and pixel-level features, being aware of the restored contents.

- We further design Instance-Specific Semantic Embedding Encoder to perform inter-frame alignment in the instance-centric semantic space via attention mechanism.

Related Work

Video Super-Resolution Existing VSR framework generally consists of feature extraction, alignment, fusion and frame reconstruction. It is vital for them to utilize the information of inter-frame to improve their performance. For example, EDVR (Wang et al. 2019) adopts multi-scale deformable convolution (DCN) and pyramid structure to enhance the alignment. BasicVSR (Chan et al. 2021) experimentally verifies that feature-level warping can alleviate the loss of details caused by the inaccurate optical flow estimation. BasicVSR++ (Chan et al. 2022) proposes flow-guided deformable alignment to stabilize the training of DCN, which explores multiple related pixels to reduce artifacts. TCNet (Liu et al. 2023a) devises a spatio-temporal stability module to enhance the structure stability of frames and the temporal consistency of the video. RVRT (Liang et al. 2022) divides the video sequence into multiple clips and transmits information clip-by-clip, achieving the balance between performance and computational complexity. Reporting that inaccurate optical flow estimation and resampling operations will destroy the sub-pixel information in the LR video, PSRT (Shi et al. 2022) proposes patch alignment. TTVSR (Liu et al. 2022a) employs optical flow and attention mechanism to locate the most similar frame patch along the same trajectory and makes full use of the information of the whole sequence. Yet, the performance of optical flow-based methods always degrades greatly to process videos with large motion or significant lighting changes.

Visual Understanding Visual understanding at different granularity levels has been a longstanding problem in the computer vision field. Segmentation, whose essence lies in grouping pixels, encompasses various tasks based on different semantics. Semantic segmentation seeks to discern semantic categories per-pixel within an image (Long, Shelhamer, and Darrell 2015), while instance segmentation (He et al. 2017) clusters pixels with identical semantic connotation into objects. With the ability to understand at both the pixel and instance levels, it is logical to step toward the panoptic segmentation (Kirillov et al. 2019). Lately, a tendency has emerged to design a universal architecture capable of handling all segmentation tasks, like Mask2Former (Cheng et al. 2022), which excels over specialized architectures in various segmentation tasks while being straightforward to train.

In this paper, we forgo the pixel-level inter-frame alignment dominated by optical flow, favoring a content-oriented alignment based on semantic cues instead. Our method outshines state-of-the-art works in several evaluation metrics.

Method

The proposed paradigm for video super-resolution, named **Semantic Lens**, is illustrated in Figure 2, which consists of a Semantic Extractor and a Pixel Enhancer. Let $I^{LR} = \{I_t^{LR} | t \in [1, T]\}$ be a LR video sequence of height H, width W, and frame length T. The Pixel Enhancer aims to recover a high-resolution version, i.e., $I^{HR} = \{I_t^{HR} | t \in [1, T]\}$ of height sH, width sW, in which s is the up-sampling factor. It coincides with the mission of the
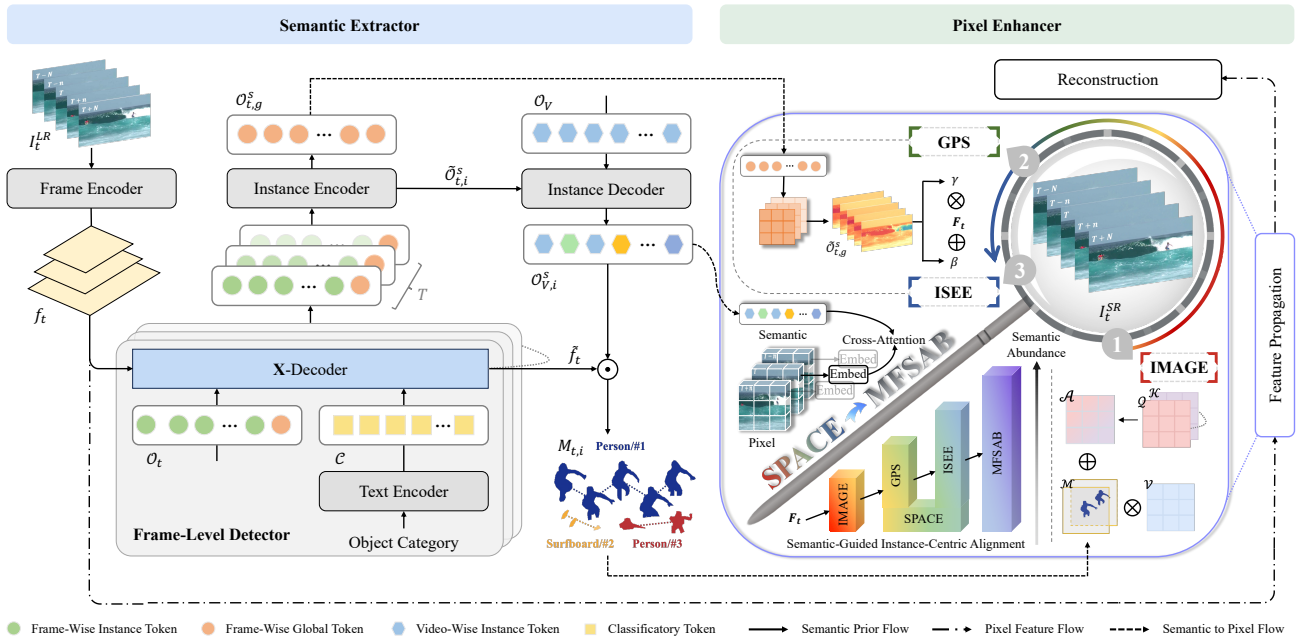
Figure 2: Overall pipeline for **Semantic Lens** consists of a Semantic Extractor and a Pixel Enhancer. The Semantic Extractor decouples low-resolution video into instances, events, and scenes, each characterized by their embodied semantics with differentiated descriptors. These semantics are employed to enhance the pixel-level features of Pixel Enhancer in a position-embedding-like manner, which yields semantic-aware features.

conventional VSR network. Differently, semantic priors derived from Semantic Extractor are embedded into the features of the Pixel Enhancer for instance-centric alignment, generating more enjoyable visual results. In this section, we will elaborate on the detailed designs of diverse semantics' utilization as well as the overall architecture.

Architecture

At each time step t, given a reference frame I_t^{LR} and $2n$ adjacent supporting frames $\{I_{t-n}^{LR}, \cdots, I_{t-1}^{LR}, I_{t+1}^{LR}, \cdots, I_{t+n}^{LR}\}$, the Semantic Lens delivers the super-resolved reference frame I_t^{SR} as output. First, the Semantic Extractor is responsible to extract diverse semantic priors from LR frames, which is mainly composed of two parts, that is, frame encoder-decoder and instance encoder-decoder. The frame encoder-decoder spatially models intra-frame correlations frame by frame, obtaining global information and instance-specific context at the frame level. While the instance encoder-decoder allows the interaction of the instance-related content along the time dimension, and establishes the video-wise instance representation under the supervision of video instance segmentation (VIS). In this way, the video content is encoded into the instance-centric semantic space. And then, those semantic information serves as the input of the subsequent stage together with the LR frames.

Pixel Enhancer is in charge of refining the original pixels and generating absent information. It adopts the bidirectional and second-order grid propagation identical to that of BasicVSR++, thereby enabling aggressive exploitation of information from the entire video. The bidirectional feature propagation branch is composed of Multi-Frame Self-Attention Blocks (MFSAB) (Shi et al. 2022). Walking past the stacked MFSABs, the refined deep features F_t of the reference frame are used to generate the high-resolution result via a reconstruction module, consisting of several convolution layers as well as pixel-shuffle layers.

With the aim of improving the inter-frame alignment, we formulate a semantic-guided instance-centric alignment schema. Specifically, features of supporting frames are first coarsely warped to reference frame via the IMAGE module. Then, the designed SPACE block is positioned before the basic unit of the feature propagation branch, and iteratively enhances the semantic abundance of original features. Based on global semantics, the GPS module generates pairs of scale and bias for spatially-adaptive feature modulation. The modulated features are further aligned with the guidance of instance-specific semantics by the ISEE module, which means that the inter-frame alignment is constrained within the same instance and its related context. Finally, the features aligned in instance-centric semantic space are fed into the subsequent MFSAB.

Semantic Extractor

The Semantic Extractor aims to encode the pixel-based representation of a video into a latent semantic space. In this semantic space, video is decoupled into instances (foreground), events (temporal), and scenes (background). Among them, instances and scenes symbolize the characters and locations recorded in the video, while events encompass the context of instances and their interactions across
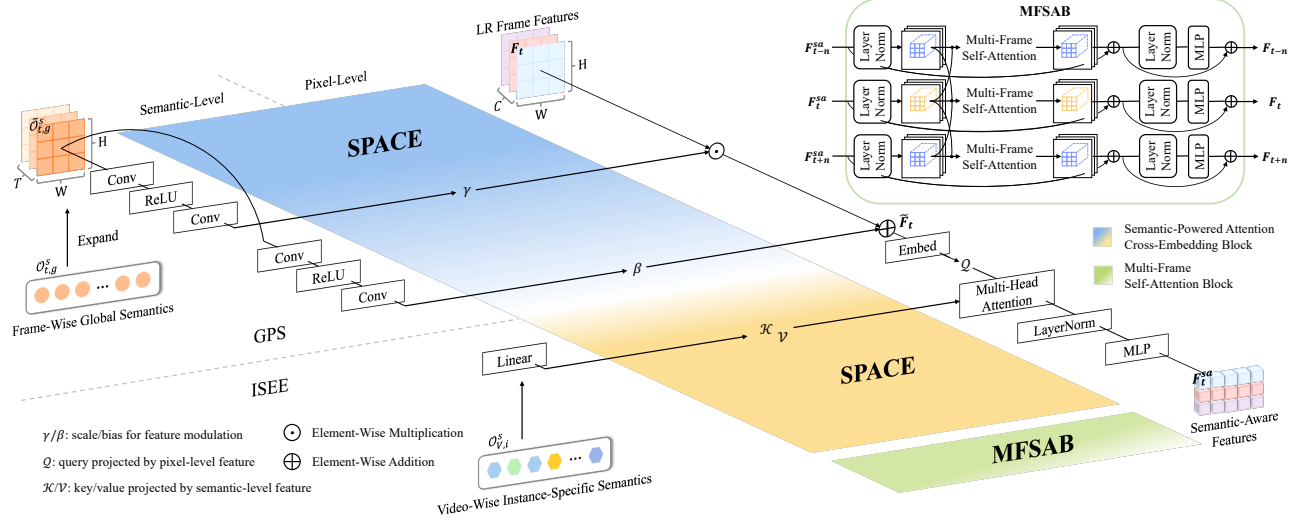
Figure 3: Illustration of Semantics-Powered Attention Cross-Embedding (SPACE) Block, composed of Global Perspective Shifter (GPS) and Instance-Specific Semantic Embedding Encoder (ISEE). It is inserted before MFSAB, the basic unit of feature propagation in Pixel Enhancer, to bridge the semantic-level priors with pixel-level features.

frames. To put it another way, the video is transformed into the instance-centric form. In this work, we employ X-Decoder (Zou et al. 2023), a frame-wise detector, to distill instance and scene information of each frame into tokens, which can be formulated as:

$$f_t = \mathbf{Enc}_{frame}(I_t^{LR}) \qquad (1)$$

$$\mathcal{O}_t^s, \tilde{f}_t = \mathbf{XDec}(f_t, \mathcal{O}_t, \mathcal{C}) \qquad (2)$$

where $\mathbf{Enc}_{frame}$ and $\mathbf{XDec}$ separately denote the frame encoder and decoder. f_t and $\tilde{f}_t$ are the pixel-level features generated by the encoder and decoder respectively. The generic non-semantic queries $\mathcal{O}_t$ and classificatory queries $\mathcal{C}$ serve as the inputs of the decoder together with f_t. $\mathcal{O}_t$ is randomly initialized, while $\mathcal{C}$ is obtained by encoding the category labels through a text encoder. $\mathcal{O}_t^s$ represents the token-level semantics, in which the last semantic token gleans the global image representation, with the rest being used for instance segmentation.

For the event, we gather instance-specific tokens from all frames to construct the video-wise instance representation. It can be used for generation of consistent instance masks given specific frame in VIS, thus it can also locate the pixels related to the same instance of multiple frames for alignment in VSR. Specifically, instance-specific tokens are divided into different local temporal windows without overlapping. In the instance encoder, window-based self-attention is employed upon instance-specific tokens which shifts along the temporal dimension, so that tokens from different frames can communicate instance-specific information. The above process can be formulated as:

$$\tilde{\mathcal{O}}_{t,i}^s = \mathbf{Enc}_{inst}(\mathcal{O}_{t,i}^s), i \in [1, N_f] \qquad (3)$$

where $\mathbf{Enc}_{inst}$ denotes the instance encoder. $\mathcal{O}_{t,i}^s, \tilde{\mathcal{O}}_{t,i}^s$ represent the i-th instance-specific token within t-th frame and those have interacted with each other along the temporal axis in instance encoder, respectively. N_f is the number of instance-specific tokens in each frame.

The construction of video-wise instance representation is to incorporate the appearance of instances across multiple frames, i.e., a more essential depiction of instance. For this reason, trainable video queries $\mathcal{O}_V$ are used to embody the video-wise instance-specific semantics and event semantics:

$$\mathcal{O}_{V,i}^s = \mathbf{Dec}_{inst}(\tilde{\mathcal{O}}_{t,i}^s, \mathcal{O}_V) \qquad (4)$$

$$M_{t,i} = \mathcal{O}_{V,i}^s \cdot \tilde{f}_t \qquad (5)$$

where $\mathbf{Dec}_{inst}$ denotes the instance decoder and $\mathcal{O}_{V,i}^s$ are the video-wise instance-specific representations. Both instance encoder and decoder consist of a stack of blocks based on the attention mechanism. $M_{t,i}$ represents the segmented masks of instances. With an eye to the limited expressiveness of the global semantic token and the intricate scenes of long video sequences, we suggest retaining the global representations of all frames to describe the scene information of the video, denoted as $\mathcal{O}_{t,g}^s$.

Semantics-Powered Attention Cross-Embedding

When aligning inter-frame features with the guidance of semantic priors, a challenge that should be considered is the discrepancy between them. To alleviate this issue, we propose the Semantics-Powered Attention Cross-Embedding (SPACE) Block, as graphically depicted in Figure 3. The SPACE block is like the bridge between Semantic Extractor and Pixel Enhancer, embedding semantic priors into pixel-level features in a position-embedding-like manner. It is composed of the Global Perspective Shifter (GPS) and the Instance-Specific Semantic Embedding Encoder (ISEE). The former generates pairs of affine transformation parameters for spatial-wise feature modulation conditioned on global semantics, while the latter utilizes the attention mechanism to enable the model to be aware of the specific content during alignment, yielding richer semantic-aware features.

Specifically, the global semantics of each frame is condensed into a single token, providing the most concise se-
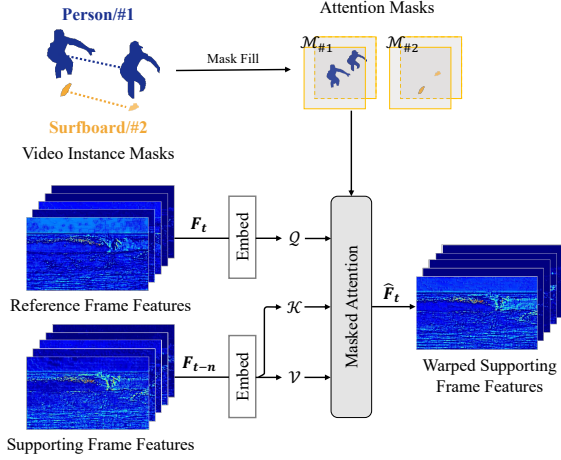
Figure 4: Illustration of Implicit Masked Attention Guided Pre-Alignment (IMAGE) Module. Attention is conducted within a local window for pre-alignment, which is modulated by the instance masks.

mantic representation of a video frame. GPS firstly extends the global semantics to match the spatial dimensions of the features from the last MFSAB using a simple dot product:

$$\tilde{\mathcal{O}}_{t,g}^s = \mathcal{O}_{t,g}^s \cdot \boldsymbol{F_t} \qquad (6)$$

where $\mathcal{O}_{t,g}^s \in \mathbb{R}^{1 \times C}$ and $\tilde{\mathcal{O}}_{t,g}^s \in \mathbb{R}^{C \times H \times W}$ are original and extended global semantics of t-th frame, respectively. C is the embedding dimension of features during propagation.

Utilizing two convolutional layers followed by a ReLU activation, it generates parameter pairs (γ, β), which are used for spatial affine transformations of the pixel-level features, representing the scale and bias, respectively:

$$(\gamma, \beta) = \mathcal{G}(\tilde{\mathcal{O}}_{t,g}^s) \qquad (7)$$

$$\tilde{\boldsymbol{F_t}} = (\boldsymbol{F_t} \odot \gamma + \beta) + \boldsymbol{F_t} \qquad (8)$$

where $\mathcal{G}$ represents the convolution layers and ReLU activation. $\tilde{\boldsymbol{F_t}}$ denotes the features embedded with global semantics, which are then fed into the ISEE. The intuition behind this methodology lies in allowing each pixel to be modulated by the global semantic prior by generating different modulation parameters, enabling adaptive feature enhancement.

ISEE is built upon the cross-attention, where the queries $\mathbf{Q}$ are derived from pixel-level features, while the keys $\mathbf{K}$ and values $\mathbf{V}$ are generated by instance-specific semantics. It bridges pixel-level features with semantic-level priors by learning the interrelationships between different pixels and instances, thereby embedding instance-specific priors into the original features via attention map $\mathbf{A}$, which is applied to values for weighted summation. The above procedures are formulated as:

$$\mathbf{Q} = \tilde{\boldsymbol{F_t}} W^Q, \mathbf{K} = \mathcal{O}_{V,i}^s W^K, \mathbf{V} = \mathcal{O}_{V,i}^s W^V \qquad (9)$$

$$\mathbf{A} = \text{SoftMax}(\mathbf{Q}\mathbf{K}^\text{T}/\sqrt{D}) \qquad (10)$$

$$\boldsymbol{F_t^{sa}} = \mathbf{AV} \qquad (11)$$

where $W^Q, W^K, W^V \in \mathbb{R}^{C \times D}$ represent the parameter matrices of projections, and D is the channel number of

projected vectors. The refined semantic-aware features $\boldsymbol{F_t^{sa}}$ will serve as the input of the subsequent MFSAB.

Aside from the SPACE, we further introduce a pre-alignment module, named Implicit Masked Attention Guided Pre-Alignment (IMAGE), as depicted in Figure 4. IMAGE utilizes the pixels from the reference frame as queries and performs a local window search in the corresponding positions of the supporting frames. By applying attention operations, the supporting frames are coarsely aligned to the reference frame, unlike previous works (Cao et al. 2022; Yang et al. 2020) which use attention to directly enhance low-resolution images with high-resolution textures. Drawing inspiration from DCN v2, we follow the idea of modulation mask to limit the magnitude of offsets, preventing the model from learning overly irrelevant content. As previously stated, the Semantic Extractor outputs both pixel-level masks and semantic-level tokens simultaneously. Instance masks, which explicitly describe the content correspondence across different frames, naturally serve as modulation masks for IMAGE and are thus used as attention masks for frame-wise pre-alignment. The benefits are two-fold. Firstly, when the pre-aligned supporting frame conducts window-based attention with the reference frame in MFSAB, it provides a broader range of information. Secondly, using instance masks as constraints can avoid performance deterioration caused by irrelevant content, meaning that coarse alignment is executed only between the same instances or backgrounds across frames. Our masked attention modulates the attention matrix via

$$\mathbf{A} = \text{SoftMax}\left(\mathcal{M} + \mathbf{Q}\mathbf{K}^\text{T}\right) \qquad (12)$$

where the attention mask $\mathcal{M}$ at feature location (x, y) is

$$\mathcal{M}(x, y) = \begin{cases} 0 & \text{if } M_{t,i}(x, y) \bigcap M_{t',i}(x, y) \neq \emptyset \\ -\infty & \text{otherwise} \end{cases} \qquad (13)$$

Here, $M_{t,i}$ and $M_{t',i}$ indicate the i-th instance masks from the reference frame and supporting frame, respectively.

Experiments

Datasets and Metrics We evaluate the performance of Semantic Lens on the benchmark datasets widely used in the field of video instance segmentation, YouTube-VIS (YTVIS) (Yang, Fan, and Xu 2019), available in three versions (2019, 2021 (Yang et al. 2021) and 2022 (Yang, Fan, and Xu 2022)). They are chosen for our research with the intent of aiding semantic priors for video super-resolution. YTVIS encapsulates 40 different predefined object categories. YTVIS-19 consists of 2,238 high-resolution video clips for training and 302 for validation, with a total of 4,883 unique video instances. The improved and extended successors, YTVIS-21 and YTVIS-22, share a training set that comprises 2,985 videos. Moreover, additional videos are included in YTVIS-21 for validation, nearly doubling the annotation quantity compared to its 2019 predecessor. On the basis of YTVIS-21, YTVIS-22 introduces 71 more complex and longer videos into the validation set. In consistency with prior studies, we evaluate our model with $4\times$ downsampling using two degradations: Bicubic downsample (BI) and downsampling employing a Gaussian filter with a standard

Method	Frames	BI degration			BD degradation		
		YTVIS-19	YTVIS-21	YTVIS-22	YTVIS-19	YTVIS-21	YTVIS-22
Bicubic	-	32.41/0.9103	31.96/0.8874	31.92/0.8834	31.89/0.8645	31.42/0.8567	31.33/0.8524
EDVR (Wang et al. 2019)	5	35.77/0.9429	35.26/0.9394	35.15/0.9358	32.28/0.8667	31.90/0.8598	31.81/0.8599
BasicVSR (Chan et al. 2021)	15	35.69/0.9446	35.35/0.9417	34.56/0.9369	32.25/0.8630	31.90/0.8562	31.81/0.8523
IconVSR (Chan et al. 2021)	15	36.25/0.9488	35.80/0.9457	35.69/0.9424	32.21/0.8625	31.83/0.8560	31.73/0.8519
BasicVSR++ (Chan et al. 2022)	30	36.33/0.9493	35.89/0.9463	35.77/0.9430	32.26/0.8645	31.88/0.8579	31.78/0.8539
TTVSR (Liu et al. 2022a)	50	36.33/0.9492	35.91/0.9465	35.81/0.9433	32.30/0.8645	31.93/0.8578	31.82/0.8534
RVRT (Liang et al. 2022)	30	36.53/0.9507	36.06/0.9481	35.93/0.9444	32.31/0.8650	31.95/0.8584	31.85/0.8544
PSRT (Shi et al. 2022)	16	<u>36.58/0.9514</u>	36.14/0.9487	36.04/0.9458	<u>34.95/0.9103</u>	<u>32.07/0.8619</u>	31.98/0.8582
IART (Xu et al. 2023)	16	36.55/0.9513	<u>36.24/0.9491</u>	<u>36.12/0.9460</u>	32.45/0.8687	32.05/<u>0.8619</u>	<u>31.97/0.8585</u>
Semantic Lens (Ours)	5	**37.19/0.9542**	**36.55/0.9499**	**36.39/0.9468**	**38.49/0.9625**	**37.93/0.9597**	**37.73/0.9558**

Table 1: Quantitative comparison (PSNR↑ and SSIM↑) on the YTVIS (2019/2021/2022) dataset for $4\times$ VSR task. The best performance is displayed in **bold**, and the second best performance is marked in <u>underline</u>.
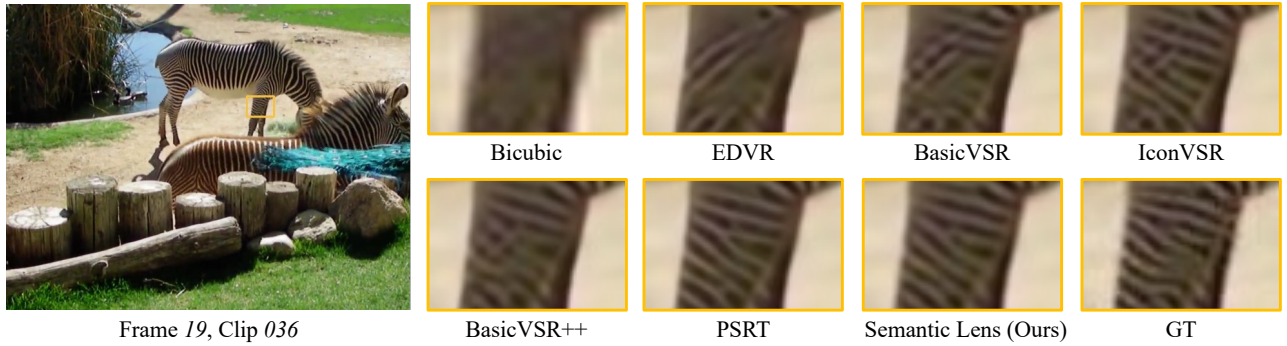

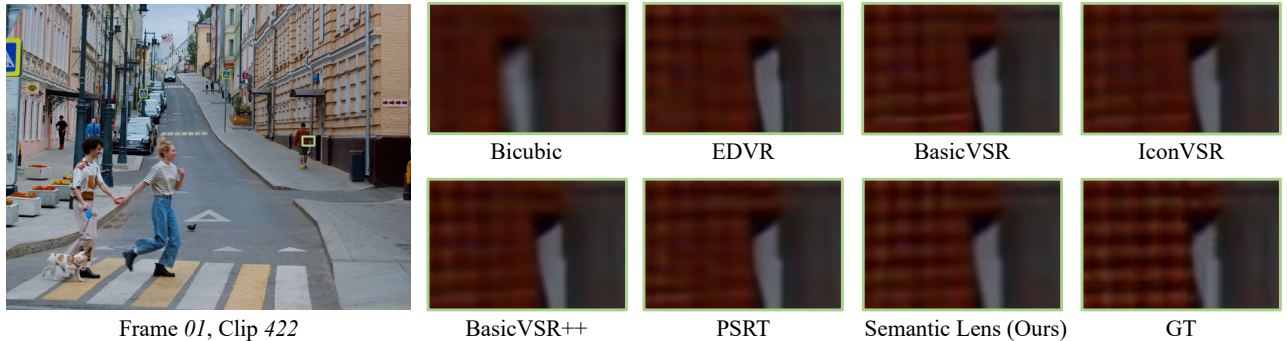Figure 5: Visual comparison of VSR ($4\times$) on YTVIS-19 dataset.

Figure 6: Visual comparison of VSR ($4\times$) on YTVIS-22 dataset.

deviation of $\sigma = 1.6$ (BD). We keep the same evaluation metrics: 1) peak signal-to-noise ratio (PSNR) and 2) structural similarity index (SSIM) as previous works.

Model Setting and Training Details For the Semantic Extractor and Pixel Enhancer, we maintain the same settings as the released code of the baseline network (Zou et al. 2023; Heo et al. 2022; Shi et al. 2022). Additionally, to align the semantic-level information with pixel-level features, we introduce a channel compression module, projecting the semantic-level features from 512 to 120. We train our model with five input frames ($T = 5$) sampled from the same video, and set the input patch size of LR frames as 64×64. For optimization, we use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.99$

and weight decay $= 10^{-4}$. The learning rate is initialized to 2×10^{-4}. The Charbonnier loss is applied on whole frames between the ground-truth I^{HR} and restored frame I^{SR}, defined as $L = \sqrt{||I^{HR} - I^{SR}||^2 + \epsilon^2}$. The whole model is trained based on X-Decoder pre-training at the image-level. Firstly, the X-Decoder with frozen weights is extended to the video-level using the video instance segmentation task. Subsequently, fine-tuning is performed on the entire Semantic Lens, where all parameters of the Semantic Extractor are kept fixed and not updated. The model is implemented with PyTorch-2.0 and trained across 4 NVIDIA 3090 GPUs, each accommodating 2 video clips.
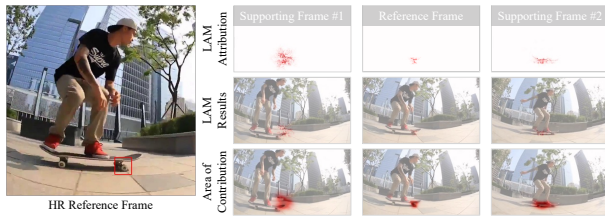
Figure 7: The attribution results of adjacent frames.

Performance Comparison

We compare Semantic Lens with several state-of-the-art methods, including 4 CNN-based VSR methods and 4 transformer-based methods. For all previous methods, we conducted evaluations using their official codes and weights. As shown in Table 1, Semantic Lens achieves state-of-the-art performance on all datasets under both degradations. Specifically, Semantic Lens provides an average improvement of 0.61 dB/0.31 dB/0.27 dB on YTVIS-19/21/22 datasets respectively, establishing the SOTA. What's more, our model achieves better performance in SSIM values.

The qualitative comparisons presented in Figure 5 and Figure 6 demonstrate the superiority of the Semantic Lens to restore finer details compared to other methods. While CNN-based methods are still struggling to restore accurate patterns, transformer-based ones, benefiting from their powerful modeling capabilities, have been dedicated to recovering more realistic textures. Compared to PSRT, the textures super-resolved by Semantic Lens is closer to the ground truth, even the distant tiny pattern can be restored intactly. This is achieved by introducing the semantic priors to enable the model to understand the restored contents.

Furthermore, we employ the interpretability tool, Local Attribution Map (LAM) (Gu and Dong 2021), to shed light on the behavior of the Semantic Lens. Given the specified target patch on the reference frame, LAM tracks the information used by the model. It highlights the pixels in multi-frames that contribute to the super-resolution most, generating the corresponding attribution maps. Representative results are shown in Figure 7. It can be observed that Semantic Lens, empowered by semantic priors, tends to concentrate more on the same instance across frames, which is particularly evident in the attribution map of supporting frame #2.

	BasicVSR	BasicVSR++	PSRT	Semantic Lens
PSNR	36.64	36.67	<u>37.15</u>	**37.19**
SSIM	0.9495	0.9436	<u>0.9538</u>	**0.9542**

Table 2: Quantitative comparison on the YTVIS-19 dataset.

To further validate the superiority of Semantic Lens, we train BasicVSR, BasicVSR++ and PSRT on the YTVIS dataset, adhering to the training strategy outlined in the original papers, with an input of 5 video frames. As reported in Table 2, we can observe that Semantic Lens achieves consistent and significant performance gain over each other method. It surpasses BasicVSR++ by a margin of up to 0.52 dB in PSNR.

Ablation Study

To verify the effectiveness of each component in the proposed Semantic Lens, we start with a baseline and gradually insert the components. Specifically, we remove the IMAGE and SPACE, retaining only MFSAB in the feature propagation branch as the baseline, while the rest of the network remains the same. From Table 3, it is apparent that each component brings considerable improvement, ranging from 0.05 dB to 0.3 dB in PSNR. Hence, it verifies that beneficial semantic priors for super-resolution are embedded into the pixel-level features.

	Baseline	GPS	ISEE	IMAGE	PSNR/SSIM
1	✓				36.75/0.9493
2	✓	✓			36.84/0.9510
3	✓	✓	✓		37.14/0.9541
4	✓	✓	✓	✓	37.19/0.9542

Table 3: Results of ablation studies on the YTVIS-19 dataset.
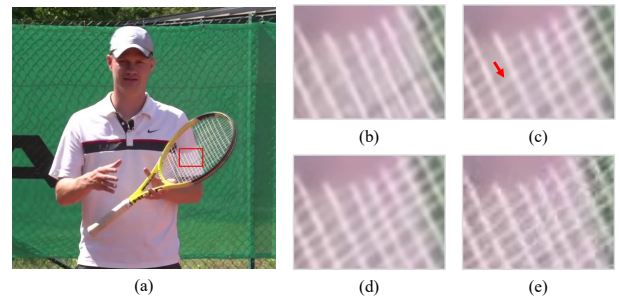

Figure 8: Visual comparisons of ablation for investigating the contribution of key modules. (a) HR Frame. (b) Baseline. (c) Baseline+SPACE. (d) Semantic Lens. (e) Patch of GT.

As shown in Table 3 and Figure 8, the proposed SPACE plays an important role in the final super-resolved result because of the adaptive embedding of diverse semantics. Compared with the baseline, the model with SPACE can achieve a more intact and natural recovery on tiny and fine textures (such as strings of the tennis racket in Figure 8 (c)). It also verifies that introducing IMAGE can indeed help improve the sharpness of the final results of videos.

Conclusion

In this paper, we introduce a video super-resolution paradigm, named Semantic Lens, that draws knowledge from semantic information. It consists of a Semantic Extractor and a Pixel Enhancer, which are responsible for extracting semantic priors and enhancing pixel-level features from low-resolution videos, respectively. To bridge the semantic and pixel-level information, we deploy a Semantics-Powered Attention Cross-Embedding Block ahead of the feature propagation basic block within the Pixel Enhancer. It iteratively embeds global semantics and instance-specific semantics into the features of VSR, aligning adjacent frames in the instance-centric semantic space.

References

Cao, J.; Liang, J.; Zhang, K.; Li, Y.; Zhang, Y.; Wang, W.; and Gool, L. V. 2022. Reference-based image super-resolution with deformable attention transformer. In *European Conference on computer vision*, 325–342.

Chan, K. C. K.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4947–4956.

Chan, K. C. K.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5962–5971.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention Mask Transformer for Universal Image Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1280–1289.

Deudon, M.; Kalaitzis, A.; Goytom, I.; Arefin, M. R.; Lin, Z.; Sankaran, K.; Michalski, V.; Kahou, S. E.; Cornebise, J.; and Bengio, Y. 2020. HighRes-Net: Recursive Fusion for Multi-Frame Super-Resolution of Satellite Imagery. arXiv:2002.06460.

Gu, J.; and Dong, C. 2021. Interpreting Super-Resolution Networks With Local Attribution Maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9199–9208.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Heo, M.; Hwang, S.; Oh, S. W.; Lee, J.; and Kim, S. J. 2022. VITA: Video Instance Segmentation via Object Token Association. In *Proceedings of Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems*, 23109–23120.

Kirillov, A.; He, K.; Girshick, R. B.; Rother, C.; and Dollár, P. 2019. Panoptic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9404–9413.

Li, G.; Ji, J.; Qin, M.; Niu, W.; Ren, B.; Afghah, F.; Guo, L.; and Ma, X. 2023. Towards High-Quality and Efficient Video Super-Resolution via Spatial-Temporal Data Overfitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10259–10269.

Liang, J.; Fan, Y.; Xiang, X.; Ranjan, R.; Ilg, E.; Green, S.; Cao, J.; Zhang, K.; Timofte, R.; and Gool, L. V. 2022.

Recurrent Video Restoration Transformer with Guided Deformable Attention. In *Proceedings of Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems*.

Liu, C.; Yang, H.; Fu, J.; and Qian, X. 2022a. Learning Trajectory-Aware Transformer for Video Super-Resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5677—-5686.

Liu, C.; Yang, H.; Fu, J.; and Qian, X. 2022b. TTVFI: Learning Trajectory-Aware Transformer for Video Frame Interpolation. arXiv:2207.09048.

Liu, H.; Zhao, P.; Ruan, Z.; Shang, F.; and Liu, Y. 2021. Large Motion Video Super-Resolution with Dual Subnet and Multi-Stage Communicated Upsampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2127–2135.

Liu, M.; Jin, S.; Yao, C.; Lin, C.; and Zhao, Y. 2023a. Temporal Consistency Learning of Inter-Frames for Video Super-Resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4): 1507–1520.

Liu, M.; Xu, C.; Yao, C.; Lin, C.; and Zhao, Y. 2023b. JNMR: Joint Non-Linear Motion Regression for Video Frame Interpolation. *IEEE Transactions on Image Processing*, 32: 5283–5295.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Lu, Z.; Liu, Y.; Jin, M.; Luo, X.; Yue, H.; Wang, Z.; Zuo, S.; Zeng, Y.; Fan, J.; Pang, Y.; et al. 2023. Virtual-Scanning Light-Field Microscopy for Robust Snapshot High-Resolution Volumetric Imaging. *Nature Methods*, 20(5): 735–746.

Shi, S.; Gu, J.; Xie, L.; Wang, X.; Yang, Y.; and Dong, C. 2022. Rethinking Alignment in Video Super-Resolution Transformers. In *Proceedings of Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems*, 36081–36093.

Wan, Z.; Zhang, B.; Chen, D.; and Liao, J. 2022. Bringing Old Films Back to Life. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 17673–17682.

Wang, X.; Chan, K. C. K.; Yu, K.; Dong, C.; and Loy, C. C. 2019. EDVR: Video Restoration With Enhanced Deformable Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1954–1963.

Xu, K.; Yu, Z.; Wang, X.; Mi, M. B.; and Yao, A. 2023. An Implicit Alignment for Video Super-Resolution. arXiv:2305.00163.

Yang, F.; Yang, H.; Fu, J.; Lu, H.; and Guo, B. 2020. Learning Texture Transformer Network for Image Super-Resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5791–5800.

Yang, L.; Fan, Y.; Fu, Y.; and Xu, N. 2021. The 3rd Large-scale Video Object Segmentation Challenge - video instance segmentation track. https://youtube-vos.org/dataset/vis/.

Yang, L.; Fan, Y.; and Xu, N. 2019. Video Instance Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 5187–5196.

Yang, L.; Fan, Y.; and Xu, N. 2022. The 4th Large-scale Video Object Segmentation Challenge - video instance segmentation track. https://youtube-vos.org/challenge/2022/.

Zhou, K.; Li, W.; Lu, L.; Han, X.; and Lu, J. 2022. Revisiting Temporal Alignment for Video Restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6043–6052.

Zou, X.; Dou, Z.-Y.; Yang, J.; Gan, Z.; Li, L.; Li, C.; Dai, X.; Behl, H.; Wang, J.; Yuan, L.; Peng, N.; Wang, L.; Lee, Y. J.; and Gao, J. 2023. Generalized Decoding for Pixel, Image, and Language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15116–15127.