Perceiving Better Moments: Cover Frame Reselection and Enhancement for Live Photos with the Live2K Dataset

Junyu Lou[1], Kai Chen[1], Weiyi You[1], Hui Zeng[2], Lei Zhang[2,3], and Shuhang Gu[1⋆]

[1] University of Electronic Science and Technology of China
[2] OPPO Research Institute
[3] The Hong Kong Polytechnic University

{junyulou.jy, shuhanggu}@gmail.com
https://github.com/CVL-UESTC/Live2K

Abstract. Modern smartphones capture Live Photos—short video bursts surrounding a still image—offering a dynamic and engaging photographic experience. However, the cover photo and video components are generated by two distinct imaging pipelines: the photo stream undergoes full computational photography processing, while the video stream is constrained by real-time efficiency and heavy compression. This intrinsic separation produces a substantial quality gap in resolution, color fidelity, and dynamic range between the cover photo and video frames. When users reselect an alternative frame from the video to replace an imperfect cover, the chosen frame often suffers from severe degradation, making direct replacement visually unsatisfactory. Restoring such frames requires simultaneous enhancement of spatial detail and color appearance, a task considerably more challenging than ordinary super-resolution or color enhancement. To address this, we define the Live Photo Cover Frame Reselection and Enhancement (LPRE) task, which leverages the intrinsic cues available within each Live Photo: the high-quality cover image as a structural and color reference, the user-reselected low-quality frame as the reconstruction target and several adjacent video frames providing temporal cues. Building upon this formulation, we construct Live2K, a real-world dataset of 2,042 Live Photos, and develop a unified one-stage baseline that integrates multi-frame fusion, guided color enhancement and super-resolution—establishing the first benchmark for Live Photo enhancement research.

Keywords: Live Photo · Image Enhancement · Image Super-Resolution

1 Introduction

Modern smartphones are typically equipped with the feature of Live Photos. When capturing a still image, the device simultaneously records a short video clip encompassing moments before and after the shutter is pressed, thereby creating a brief "living moment." During playback, this produces a more immersive and narrative experience than a conventional photo, making Live Photos widely popular among users.

⋆ Corresponding author

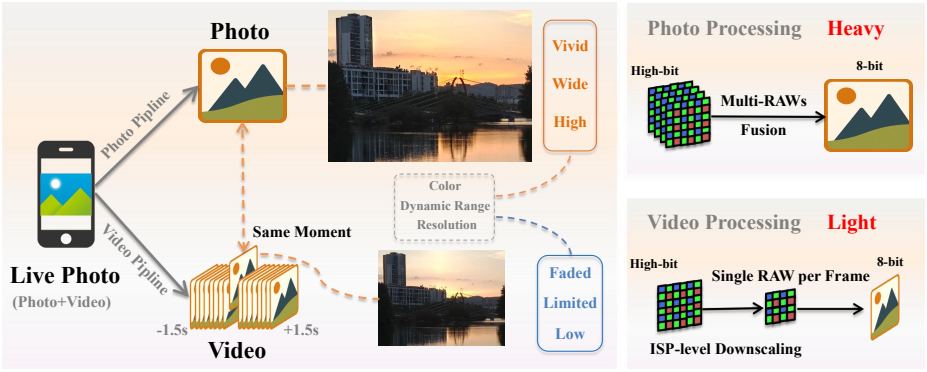

Fig. 1: Comparison between the cover frame (top) and the video frame (bottom) captured at the same moment in Live Photos. The photo pipeline processes multiple high-bit RAW data with rich system resources, resulting in a high-quality cover frame; meanwhile, the video pipeline processes each video frame with a single RAW image under a tight computational budget, resulting in low-quality video frames.

The photo and video components of a Live Photo are produced by two distinct imaging pipelines. The photo stream (used for the cover image) is generated through multi-frame fusion of several high–bit-depth RAW images captured at different exposures (long, short, and normal) to achieve a higher dynamic range and superior image quality. The video stream, in contrast, is based on a single RAW frame per time step, as is typical for real-time recording. In the Live Photo mode, the still photo and video must be captured simultaneously, which imposes strict latency and storage constraints on the video pipeline. As a result, the budget allocated to each frame is very limited. To meet these requirements, the video frames are typically downsampled at an early stage of processing and then processed with a lightweight, efficiency-oriented pipeline. As depicted in Fig. 1, this fundamental separation leads to an intrinsic quality gap between the cover image and video frames—manifested primarily in resolution, color fidelity, and dynamic range. These discrepancies stem from the imaging system itself and cannot be eliminated due to the system's latency demands. This issue often undermines user experience, as the automatically captured cover photo may correspond to an imperfect moment—such as a blink or an unideal expression—prompting users to reselect another frame from the video. Yet the reselected frame typically exhibits lower resolution, faded colors, and limited dynamic range, resulting in a visible degradation compared to the original cover, as shown in Fig. 2. Making this re-selection viable thus requires a model capable of restoring structural detail while correcting appearance inconsistencies in tone. In other words, the task demands simultaneous enhancement of both spatial resolution and color quality, which makes it more challenging than ordinary color enhancement or super-resolution.

Fortunately, every Live Photo inherently contains a high-quality still image—the original cover—that shares strong content similarity with the reselected frame. This provides a powerful cue for guided enhancement, where the cover image serves as a reference for recovering fine details and accurate color rendition. Moreover, the neigh-

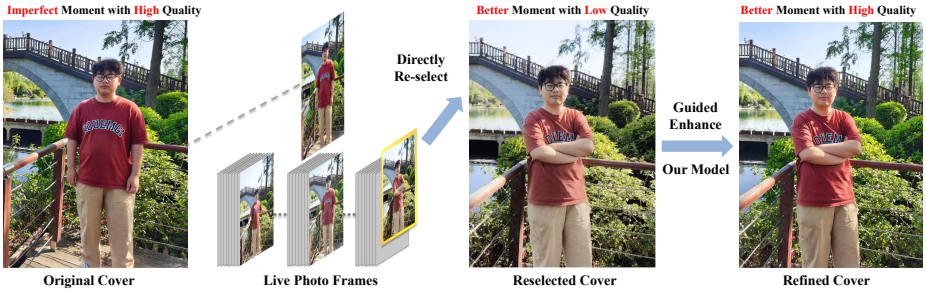

Fig. 2: When users attempt to replace an unsatisfactory cover, the newly selected cover frame is often of inferior quality. Therefore, a model is needed to perform comprehensive enhancement to restore and improve its visual quality.

boring frames in the video sequence offer complementary temporal cues that further aid the reconstruction. Together, these observations motivate a new and practical problem: the Live Photo Cover Frame Reselection and Enhancement (LPRE) task, which aims to transform a user-selected low-quality frame into a high-quality photo consistent in visual appearance with the original cover.

To enable research on this problem, we construct Live2K, a large-scale collection of 2,042 real-world Live Photos captured using two smartphone systems. Each Live Photo contains a high-quality cover image and its accompanying short video sequence. Building such a collection is non-trivial, as Live Photos are device-specific and their internal imaging pipelines are proprietary. Based on these raw Live Photos, we further construct paired data for the reselection and enhancement task. Each final pair consists of one guidance image, one ground-truth image, one corresponding low-quality frame from the video sequence, and several neighboring frames as temporal context. Importantly, the guidance image is drawn from another Live Photo captured under a visually similar scene, providing color and texture priors from a distinct yet related sample. To our knowledge, Live2K is the first dataset specifically designed for the Live Photo reselection and enhancement task. Based on Live2K, we design a unified one-stage baseline that jointly performs multi-frame fusion, guided enhancement, and super-resolution, supervised by intermediate color loss and final reconstruction loss. It is worth noting that instead of pursuing architectural novelties, our primary objective is to establish a solid and unified baseline for the newly introduced LPRE task.

Our main contributions are summarized as follows:

- We introduce the Live Photo Cover Frame Reselection and Enhancement (LPRE) task, aimed at bridging the intrinsic quality gap between cover and video frames.
- We construct Live2K, the first large-scale real-world dataset tailored for this new task.
- We present a unified baseline framework that integrates temporal fusion and guided enhancement, offering the first benchmark for Live Photo enhancement research.

2 Related Work

2.1 Image Enhancement

Traditional image enhancement approaches rely on handcrafted priors such as histogram equalization [1, 35] or Retinex-based decomposition [24, 36], which are often inadequate under complex illumination and device-dependent degradations. Deep learning has largely replaced these rule-based designs, learning content-adaptive color and tone mappings from data. Representative works include bilateral grid processing [14,30] and LUT-based models [37,40,43] as well as curve regression methods [15], achieving real-time performance while preserving perceptual quality.

Some reference-based studies [13, 20, 26, 33] explore photorealistic color transfer, where an example provides guidance for global tone alignment. However, these works primarily pursue stylistic consistency rather than faithful enhancement. In contrast, our approach leverages the high-quality cover frame in Live Photos as a reliable prior that supplies contextual and structural cues, supplementing the degraded frame to achieve more accurate and natural enhancement under cross-quality conditions.

2.2 Image Super-Resolution

Single-image super-resolution (SISR) reconstructs a high-resolution image from a single low-resolution input. Early CNN-based methods established the basic paradigm [10], followed by deeper residual [21, 28, 48], attention-based [47], and Transformer architectures [8, 27, 29, 44]. Although effective, SISR is inherently limited by the lack of complementary information within a single frame.

To overcome this, burst and multi-frame super-resolution approaches exploit short-term temporal redundancy from multiple observations. Burst SR [3, 11] focuses on handheld or smartphone bursts, aligning frames with sub-pixel precision to aggregate fine details, while video super-resolution approaches [6, 19, 39, 50] extends this idea to longer sequences with motion compensation and temporal alignment. These methods demonstrate significant improvements in reconstruction quality but remain challenged by motion variations and illumination inconsistencies.

Reference-based super-resolution approaches methods introduce an additional high-quality image to provide texture or structural priors [5, 16, 22, 25, 32, 41, 49]. By establishing feature-space correspondence, these methods can transfer fine details and recover high-frequency content when reliable guidance is available.

Our pipeline combines multi-frame inputs with cross-quality guidance in Live Photos. The short burst sequence provides temporal redundancy, while the fully processed cover frame offers reliable color and structural priors. Our method unifies these perspectives by performing correspondence-aware fusion that jointly exploits multi-frame complementarity and high-quality guidance under cross-device conditions.

3 Live2K Dataset

We construct a real-world Live Photo dataset, named **Live2K**, to support the training and evaluation of the reselected frame enhancement task.

Fig. 3: Example ground truth images in the Live2K dataset. Each column corresponds to a category: portrait, architecture, landscape, and indoor scene, from left to right.

3.1 Data Capture with Smartphones

We use an iPhone 16 Pro and an OPPO Find X8 Pro to capture 2,042 real-world Live Photos, each consisting of a cover image and its corresponding video sequence. The iPhone contributed 1,021 samples, with cover images at a resolution of 4032×3024 and video frames at 1920×1440, captured using its fusion camera system with a 24 mm f/1.78 lens. The OPPO contributed 1,021 samples, with cover images at 4096×3072 and video frames at 1728×1296, captured using a 23 mm f/1.6 lens. Since the two devices have distinct ISP pipelines and super-resolution scales, we conducted separate training and testing on the datasets from each device. The dataset covers a wide range of scenes, including portraits, architecture, natural landscapes, and indoor scenes. A small sample of images is shown in Fig. 3, which showcases the diversity of our dataset.

To provide data suitable for cross-instance guidance, the dataset was collected in a group-wise manner, where each group contains multiple Live Photos with similar content, allowing the covers to serve as mutual guidance images. Live Photos within the same group were taken from the same scene, but with variations in camera angle, distance, scene composition, and subject movement to obtain visually similar yet non-identical images. This setup allows the guidance images and video frames to closely mimic the real-world relationship between the cover and video frames in actual Live Photo captures.

3.2 Data Preprocessing

To construct data pairs, we need to find the video frame that best corresponds to the cover image at the pixel level. For each Live Photo, we first downsample the cover image to match the resolution of the video frames and convert both the cover and

video frames into the grayscale domain, which reduces color differences and allows the matching to focus on structural cues. We then use the SIFT [31] algorithm to identify the video frame most similar to the cover image, which serves as its corresponding low-quality counterpart. Since Live Photos capture video footage extending 1.5 seconds before and after the shutter press, the most similar frames are consistently identified near the median position of the sequence. Given that slight viewpoint differences may exist between the cover and the matched frame, we further warp the cover image to align with the frame's perspective using a homography estimated from SIFT feature correspondences, obtaining the ground-truth image. We also select the eight neighboring frames (four before and four after) around the matched frame, forming a 9-frame input sequence. The dataset is organized into 751 groups, in the training set, each group contains three images that can guide each other, resulting in six paired samples per group. In the test set, each group has two images, forming one paired sample. Overall, the final paired dataset includes 1,620 training images (yielding 3,240 training pairs) and 422 test images (yielding 211 test pairs).

We adopt patch-based training to reduce GPU memory usage and improve computational efficiency. For each data pair in the training set, we first perform feature matching between the ground truth (GT) and the reference (Ref) images using the SIFT [31] algorithm. Then, a sliding-window search is applied to locate, for each GT patch, the most similar patch in the Ref image. Based on the desired super-resolution scale, the corresponding region in the input frames is extracted to form a new paired sample. To ensure reliable correspondence, we apply a similarity threshold to the matched feature points, and pairs below this threshold are discarded. The thresholds are set to 160 for iPhone and 120 for OPPO. As a result, the iPhone subset contains 60,291 training pairs, where both GT and Ref images are of size 504×504 and the input images are 240×240, while the OPPO subset contains 57,755 training pairs with GT/Ref sizes of 512×512 and input size of 216×216.

3.3 Misalignment Analysis

Real-world datasets [3, 4, 42, 46] often suffer from spatial misalignment. Prior studies typically acquire inputs and ground truths (GT) using different sensors at different times, resulting in significant spatial displacement. However, leveraging the unique configuration of Live Photos, we can obtain GT and nearly simultaneous video frames captured by the same sensor. Consequently, the issue of data misalignment is substantially alleviated. We observe from the experimental results that high reconstruction quality can be achieved without any auxiliary alignment techniques.

4 Guided Enhancement Network

Given a short burst of consecutive video frames $\{\boldsymbol{X}_t\}_{t=1}^{9} \in \mathbb{R}^{3\times h\times w}$ from a Live Photo and a high-quality cover image $\boldsymbol{X}_{ref} \in \mathbb{R}^{3\times H\times W}$ as the reference, our goal is to enhance any selected video frame to a visually comparable quality to $\boldsymbol{X}_{ref}$. To this end, we propose **LPENet** (Live Photo Enhancement Network), as illustrated in Fig. 4. By integrating temporal fusion and guided enhancement into a streamlined pipeline, we aim to provide a stable foundation to facilitate future research in this domain.

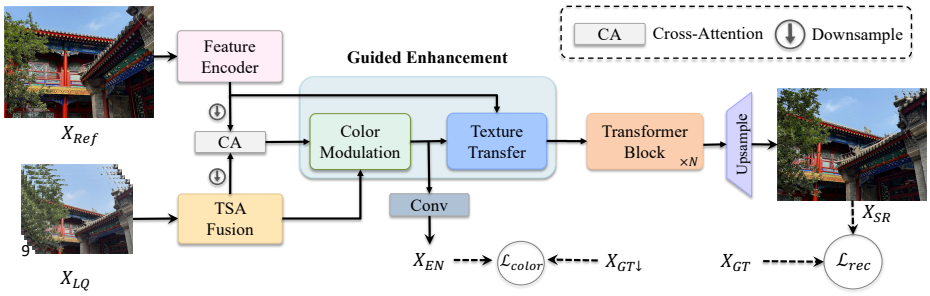

Fig. 4: The overall architecture of proposed Live Photo Enhancement Network (LPENet). The Color Modulation module consists of several Guided Color Enhancement Blocks (Fig. 5).

4.1 Feature Extraction and Multi-Frame Fusion

To reduce GPU memory consumption, each input frame is first rescaled to half the resolution of the guidance image, followed by a 2× downsampling using the PixelUnshuffle operation. While all subsequent processing is performed at this reduced resolution, performing explicit optical flow-based alignment across 9 frames remains computationally prohibitive. Therefore, instead of relying on optical flow, we employ the Temporal-Spatial Attention (TSA) module adapted from EDVR [38] to effectively exploit temporal redundancy and handle inter-frame variations. This module is tasked with implicitly aligning and aggregating features from the 9-frame input sequence. By computing attention maps across both temporal and spatial dimensions, the TSA module dynamically fuses vital contexts from neighboring frames, yielding a comprehensive fused feature representation $\boldsymbol{F}_{lq} \in \mathbb{R}^{C\times\frac{H}{4}\times\frac{W}{4}}$. Parallel to this, the Feature Encoder extracts texture information from the reference image. It comprises several convolutional layers followed by two PixelUnshuffle operations. This process maps the reference image into a deep embedding space, generating the reference feature map $\boldsymbol{F}_{Ref} \in \mathbb{R}^{C\times\frac{H}{4}\times\frac{W}{4}}$ that matches the resolution of the fused video features.

4.2 Guided Enhancement

To achieve color-consistent enhancement and spatial detail reconstruction guided by a high-quality reference, we design a two-stage Guided Enhancement module operating on the feature pair $(\boldsymbol{F}_{LQ}, \boldsymbol{F}_{Ref})$. Adopting a coarse-to-fine strategy, this module progressively refines the low-quality input by first aligning the overall color through color modulation, and subsequently injecting local details via texture transfer.

Color Modulation. In the first stage, we introduce a Color Modulation module to produce a color-corrected representation $\boldsymbol{F}_{EN}$. The module consists of multiple color enhancement blocks adapted from the NAFBlock [7] architecture, integrated with additional cross-attentive guidance and Feature-wise Linear Modulation (FiLM) [34], as illustrated in Fig. 6. Since color enhancement depends more on global context than on local texture details, we first perform a cross-attention operation on the downsampled

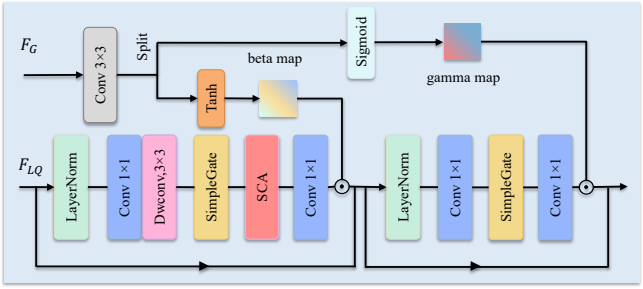

Fig. 5: The proposed Guided Color Enhancement Block.

features $(\boldsymbol{F}_{LQ\downarrow}, \boldsymbol{F}_{Ref\downarrow})$ outside the enhancement blocks to obtain a global guidance feature $\boldsymbol{F}_G$. The resulting $\boldsymbol{F}_G$ is then upsampled back to the resolution of the input to align with the enhancement backbone. This design not only reduces computational cost but also enables the network to capture global color and tone correspondences between the low-quality input and the reference. The cross-attention is formulated as:

$$\text{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\right)\boldsymbol{V}, \tag{1}$$

where d is the head dimension, $\boldsymbol{Q}$ is derived from $\boldsymbol{F}_{LQ\downarrow}$, and $\boldsymbol{K}$ and $\boldsymbol{V}$ are obtained from $\boldsymbol{F}_{Ref\downarrow}$.

The attended feature $\boldsymbol{F}_G$ serves as a shared global guidance signal for all enhancement blocks, ensuring consistent color mapping across the network. Within each block, $\boldsymbol{F}_G$ is further processed by a shallow convolutional layer to produce block-specific modulation maps:

$$[\beta, \gamma] = \text{Conv}(\boldsymbol{F}_G), \quad \beta = \tanh(\beta), \quad \gamma = \sigma(\gamma), \tag{2}$$

where $\tanh(\cdot)$ and $\sigma(\cdot)$ denote the hyperbolic tangent and sigmoid functions, respectively. The resulting modulation parameters β and γ are applied to the block's enhancement backbone in a FiLM-like manner:

$$y = x + f_1(x) \odot \beta, \quad z = y + f_2(y) \odot \gamma, \tag{3}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are two residual branches.

This design integrates global color guidance via low-resolution cross-attention and local refinement via FiLM modulation, achieving visually consistent enhancement guided by $\boldsymbol{F}_{Ref}$. To ensure that this module focuses primarily on color transformation and reduces the influence of color discrepancies during subsequent feature alignment, we append a convolutional layer at the end of this module to map $\boldsymbol{F}_{EN}$ back to the image domain. The output image $\boldsymbol{X}_{EN}$ is then compared with the downsampled ground truth $\boldsymbol{X}_{GT\downarrow} \in \mathbb{R}^{3\times\frac{H}{4}\times\frac{W}{4}}$ to compute a color-related loss.

Texture Transfer. Given the enhanced feature obtained from the previous color-guidance stage and the reference feature extracted from the high-quality cover image, our next

goal is to reconstruct high-frequency spatial details for the target frame. To align local structures, $\boldsymbol{F}_{EN}$ is divided into non-overlapping patches of size $p \times p$, and each patch $\boldsymbol{P}_{EN}^{(i)}$ searches for its k most similar patches in $\boldsymbol{F}_{Ref}$ (where $k = 3$). The top-k match indices $\mathcal{J}^*$ are determined by normalized cross-correlation:

$$\mathcal{J}^* = \text{TopK}_j \frac{\langle \boldsymbol{P}_{EN}^{(i)}, \boldsymbol{P}_{Ref}^{(j)} \rangle}{\|\boldsymbol{P}_{EN}^{(i)}\|_2 \, \|\boldsymbol{P}_{Ref}^{(j)}\|_2}.$$

To reduce computational cost, both features are downsampled by a ratio r before matching, and the resulting indices are remapped to the original coordinate grid. Since both the feature size and patch size shrink by r, the complexity decreases approximately by $1/r^4$. The corresponding top-3 matched reference patches are then gathered to form the local reference feature space. After structural alignment, we refine $\boldsymbol{F}_{EN}$ using a *Similar Window Cross Attention* (SWCA) operation. For each local query window in $\boldsymbol{F}_{EN}$, we compute multi-head cross-attention where the query tokens $\boldsymbol{Q}$ are derived from $\boldsymbol{F}_{EN}$, and the key–value tokens $\boldsymbol{K}, \boldsymbol{V}$ are extracted from the concatenated tokens of the corresponding $k = 3$ most similar reference patches. Prior to the attention calculation, $\boldsymbol{Q}$ and $\boldsymbol{K}$ are ℓ_2-normalized along the channel dimension so that the attention is computed based on cosine similarity, which improves robustness to local contrast and illumination variations. The aggregated result of cross-attention is denoted as $\boldsymbol{Y}$. To prevent over-reliance on the reference, we adopt a conservative residual fusion strategy: the input feature $\boldsymbol{F}_{EN}$ is linearly projected to $\boldsymbol{X}_{base}$, and the super-resolved feature $\boldsymbol{F}_{SR}$ is obtained by a gated interpolation between $\boldsymbol{X}_{base}$ and $\boldsymbol{Y}$:

$$\boldsymbol{F}_{SR} = (1 - \alpha)\boldsymbol{X}_{base} + \alpha \boldsymbol{Y},$$

where $\alpha \in [0, 1]$ is a learnable coefficient that adaptively balances the contribution of the reference. This mechanism preserves the structural integrity of $\boldsymbol{F}_{EN}$ while injecting high-frequency texture details from $\boldsymbol{F}_{Ref}$.

4.3 Feature Fusion and Reconstruction.

Finally, the feature $\boldsymbol{F}_{SR}$ is fused with the input feature through convolution and refined by several Swin Transformer Layers (denoted as $\text{STL}(\cdot)$):

$$\boldsymbol{F}_{out} = \text{STL}(\text{Conv}([\boldsymbol{F}_{EN}, \boldsymbol{F}_{SR}])), \tag{4}$$

An upsampling head then reconstructs the high-resolution output, producing the final $4\times$ super-resolved image:

$$\boldsymbol{X}_{SR} = \text{UpSample}_{\times 4}(\boldsymbol{F}_{out}), \tag{5}$$

4.4 Loss Function

During training, we aim for the input image to recover both the color and texture of the ground truth. To achieve this, we employ two losses — a color loss and a reconstruction loss — applied at different stages of the network. The overall loss is :

$$\mathcal{L} = \mathcal{L}_{color} + \mathcal{L}_{rec}, \tag{6}$$

Table 1: Quantitative comparison with representative methods on two settings of Live2K dataset. The best results are marked in **bold**. "/" are absent results due to the unavailable code.

Category	Methods	#Params	iPhone				OPPO			
			PSNR ↑	SSIM ↑	LPIPS ↓	ΔE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	ΔE ↓
SISR	SwinIR [27]	8.05M	26.58	0.8447	0.2353	4.21	27.38	0.8543	0.2176	4.14
BurstSR	BIPNet [11]	1.76M	25.97	0.8204	0.2618	4.78	26.93	0.8398	0.2630	4.67
	Burstormer [12]	1.56M	26.03	0.8205	0.2605	4.77	26.66	0.8311	0.2668	4.83
	BurstM [18]	13.4M	26.12	0.8262	0.2545	4.76	26.31	0.8246	0.2514	4.99
	QMambaBSR [9]	/	/	/	/	/	/	/	/	/
RefSR	ERVSR [22]	5.76M	25.40	0.8190	0.2695	4.94	26.59	0.8195	0.2708	5.06
	DATSR [5]	5.58M	26.34	0.8407	0.2492	4.50	26.64	0.8445	0.2422	4.68
	MRefSR [45]	11.8M	26.34	0.8372	0.2419	4.44	26.97	0.8475	0.2243	4.47
	RebaIR [2]	72.1M	/	/	/	/	/	/	/	/
	Ours	5.06M	**26.97**	**0.8497**	**0.2268**	**4.05**	**27.78**	**0.8657**	**0.2005**	**3.95**

Color loss. To minimize the influence of color inconsistencies before performing super-resolution, we adopt a loss function between $\boldsymbol{X}_{EN}$ and $\boldsymbol{X}_{GT\downarrow}$, which is similar to [30],

$$\mathcal{L}_{color} = \lambda_1 \mathcal{L}_2 + \lambda_2 \mathcal{L}_{ssim} + \lambda_3 \mathcal{L}_{vgg}, \tag{7}$$

where λ_1, λ_2, λ_3 are set to 10, 0.5, and 0.005.

Reconstruction loss. We adopt the commonly used $\mathcal{L}_1$ loss between $\boldsymbol{X}_{SR}$ and $\boldsymbol{X}_{GT}$ as reconstruction loss.

5 Experiments

5.1 Evaluation

During the testing phase, we perform inference on full-resolution images to address the re-selection and enhancement task of Live Photo under real-world conditions. Moreover, due to the presence of the guidance image, patch-based inference is not applicable in our setting. We evaluate our method using PSNR and SSIM metrics, both of which are computed on the Y channel (luminance) of the YCbCr color space.

5.2 Implementation Details

Since the super-resolution scale factors of both settings are non-integer, we first upsample the input image to half the spatial size of the guidance image using interpolation within the network. We then apply PixelUnshuffle to perform a 2× downsampling, followed by a 4× super-resolution reconstruction in the final stage. The detailed parameters of our network can be found in the supplementary material.

We employ the Muon optimizer [17] with auxiliary Adam [23] for bias and gain parameters. The matrix weights are updated with Muon using a $10\times$ higher learning rate, while the scalar parameters use Adam (β_1=0.9, β_2=0.99). We augment the training data with random horizontal and vertical flipping or different random rotations of 90°, 180°, and 270°. The model is trained with a batch size of 24 and an initial learning rate of 2e-4, which is cosine-annealed to 4e-6 over 200K iterations. All experiments are conducted on two NVIDIA RTX 5090 GPUs.

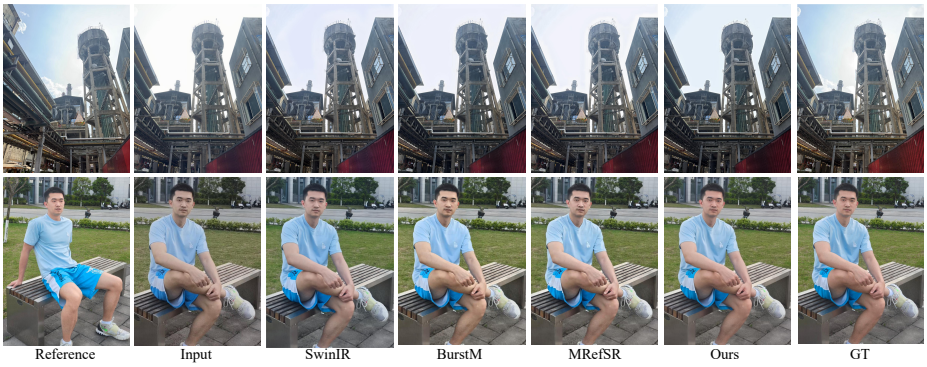

Fig. 6: Visual comparison of overall color fidelity with SOTA methods

5.3 Comparisons with Existing Methods

For comparison, we selected state-of-the-art methods from the domains of Single Image Super-Resolution (SISR), Reference-based Super-Resolution (RefSR), and Burst Image Super-Resolution (BSR). All models are retrained on our Live2K dataset for a fair comparison. We uniformly upsample the input images to half the size of the GT, then apply PixelUnshuffle for a further 2× downsampling before feeding them into the network. Crucially, performing direct inference on 4K images is computationally infeasible for almost all existing methods, as the massive memory footprint strictly exceeds standard GPU limits. To make the evaluation even possible, we were compelled to scale down these architectures. For example, we discarded the multi-scale structures in DATSR and MRefSR, retaining only operations at the lowest scale, and reduced the base channel number of BIPNet from 64 to 32. Details regarding modifications to other models are provided in the supplementary material. Other training settings remain consistent with the original implementations.

As shown in Tab. 1, our method achieves a significant improvement over previous approaches on both subcategories of the dataset. We also provide visual examples to demonstrate the comparisons in terms of overall color and local details, as presented in Fig. 6 and Fig. 7. These images clearly demonstrate our advantage in recovering textures and color from severely degraded video frames. Notably, the input image in the first row of Fig 6 is severely overexposed, resulting in an almost purely white sky. Although our method recovers some colors compared to other models, the result is still suboptimal, indicating that there is room for improvement in our utilization of the guidance image.

5.4 Ablation Study

Effects of each component. To verify the effectiveness of each component in our network, we conduct an ablation study on the OPPO subset by progressively removing modules from the full model. The last configuration represents our complete model. In the third experiment, we removed the reference image, completely discarded the detail transfer module, and employed the basic NAFBlocks in the color modulation module.

In the second experiment, we ablated the intermediate color supervision loss. In the first experiment, we discarded the multi-frame inputs, utilizing only a few convolutional layers to extract features from the input image. As illustrated in Tab. 2, introducing multi-frame information yields an improvement of 0.29 dB. Adding the color supervision loss brings a 0.06 dB gain, although marginal, it introduces almost no additional inference burden. Furthermore, incorporating the reference image results in a 0.3 dB boost.

Table 2: Ablation study on each component of our framework.

Multi-Frame	Color Loss	Reference	OPPO	
			PSNR	SSIM
			27.13	0.8516
✓			27.42	0.8596
✓	✓		27.48	0.8627
✓	✓	✓	27.78	0.8657

Table 3: Ablation study on two components within the Guided Enhancement module.

Color Modulation	Texture Transfer	OPPO	
		PSNR	SSIM
		27.48	0.8627
✓		27.66	0.8643
	✓	27.68	0.8644
✓	✓	27.78	0.8657

Effects of designs about reference. We conduct experiments to verify the effectiveness of two components in the Guided Enhancement module. As shown in Tab. 3, using either color modulation or texture transfer independently yields a PSNR improvement of approximately 0.2 dB, while combining them results in a total gain of 0.3 dB. This demonstrates that the two components are not only individually effective but also highly complementary, working synergistically to achieve better reconstruction performance.

5.5 Computational cost Comparison

We evaluate the GPU memory consumption and inference speed of models at a ground truth image resolution of 4096×3072, as listed in Tab. 4. The inference speed is reported as the average over 100 test images. The results show that our model achieves the fastest inference speed with a relatively moderate memory footprint.

Table 4: Comparison of GPU memory and runtime on 4K images.

Model	SwinIR	BIPNet	BurstM	DATSR	MRefSR	LPENet
Memory (GB)	16.25	26.93	17.47	24.65	27.58	21.36
Runtime (s)	1.419	1.036	1.80	21.05	34.86	0.865

5.6 Real-World Evaluation

The reference images and input frames in our dataset originate from distinct Live Photos, in real-world applications, the cover photo of a Live Photo naturally serves as the reference image. To evaluate our model's performance in this practical setting, we captured Live Photos with significant camera movements, utilizing the cover image as

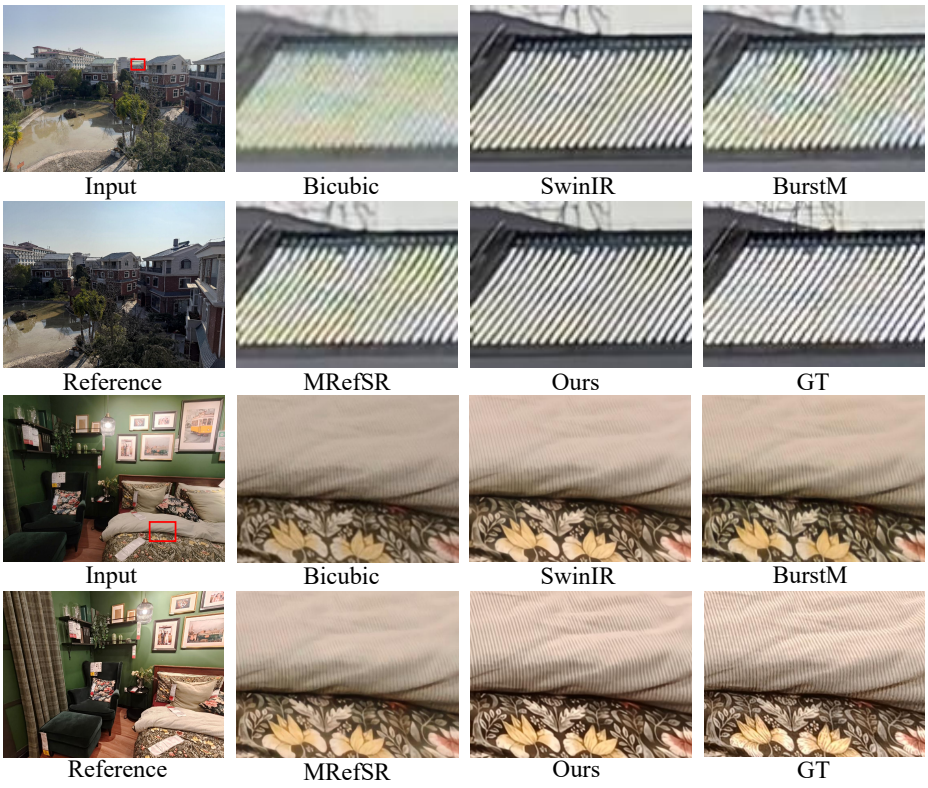

Fig. 7: Visual comparison of texture with SOTA methods.

guidance and the final 9 video frames as inputs. Ground Truth is unavailable in such real-world scenarios, we solely provide qualitative visual comparisons. As shown in Fig. 8, our approach consistently enhances the visual quality of the re-selected frames.

5.7 Cross-device Generalization

The ISP hardware pipelines and algorithms vary significantly across different manufacturers, making it impractical to directly apply a universal model to diverse devices. Consequently, device-specific training is commonly adopted in the industry. True zero-shot ISP invariance remains challenging in industrial mobile imaging systems, and our method does not fully solve this problem. Reflecting this, our sub-datasets exhibit distinct degradation profiles, such as noise, color, and scale factors, necessitating separate training. To explore the potential of cross-device generalization, we fine-tuned the model pre-trained on the OPPO dataset for 5,000 iterations using only the first 60 images from the iPhone training set. When evaluated on the iPhone test set, as shown in Tab. 5 and Fig. 9, this minimal fine-tuning yields substantial quantitative and visual improvements, demonstrating the potential of our model for cross-device transferability.

Fig. 8: Visual results of our model in real-world condition.

Table 5: Quantitative comparison on iPhone dataset.

Setting	PSNR	SSIM
w/o fine-tuned	20.63	0.6769
w/ fine-tuned	25.43	0.8132

Fig. 9: Visual performance comparison on iPhone dataset.

6 Limitation

Our dataset is mainly limited to well-lit daytime scenes, lacking night-time data. Capturing Live Photos in low-light environments naturally introduces longer exposure times, severe noise, dropped frames, and significant inter-frame motion blur. Because these degradations differ fundamentally from daytime data, our trained model may not generalize well to night scenes. Future research may focus on building a universal dataset or a domain-specific dataset tailored for extreme low-light scenarios, to effectively bridge this domain gap and ensure robust performance across all lighting conditions.

7 Conclusion

In this work, we introduced the Live Photo Cover Frame Reselection and Enhancement (LPRE) task, addressing the quality gap between the still cover and the accompanying video frames in Live Photos, with the goal of improving the visual quality of the re-selected cover frame. We analyzed the fundamental causes of this discrepancy from the perspective of the imaging pipeline and demonstrated that the issue cannot be resolved within the system itself. To facilitate research in this direction, we constructed Live2K, the first large-scale real-world dataset specifically tailored for the LPRE task. Furthermore, we proposed a unified one-stage baseline model that jointly performs multi-frame fusion, guided enhancement, and super-resolution within a single framework. Extensive experiments on Live2K, alongside evaluations in real-world scenarios, validate the effectiveness and robustness of our approach, establishing a solid benchmark for future studies.

Acknowledgment. This work was supported by National Natural Science Foundation of China (No.62476051) and Sichuan Natural Science Foundation (No.2024NSFTD0041).

References

1. Abdullah-Al-Wadud, M., Kabir, M.H., Dewan, M.A.A., Chae, O.: A dynamic histogram equalization for image contrast enhancement. IEEE transactions on consumer electronics **53**(2), 593–600 (2007)
2. Bernasconi, M., Djelouah, A., Zhang, Y., Gross, M., Schroers, C.: Rebair: Reference-based image restoration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5489–5498 (2025)
3. Bhat, G., Danelljan, M., Van Gool, L., Timofte, R.: Deep burst super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9209–9218 (2021)
4. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3086–3095 (2019)
5. Cao, J., Liang, J., Zhang, K., Li, Y., Zhang, Y., Wang, W., Gool, L.V.: Reference-based image super-resolution with deformable attention transformer. In: European conference on computer vision. pp. 325–342. Springer (2022)
6. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4947–4956 (2021)
7. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: European conference on computer vision. pp. 17–33. Springer (2022)
8. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22367–22377 (2023)
9. Di, X., Peng, L., Xia, P., Li, W., Pei, R., Cao, Y., Wang, Y., Zha, Z.J.: Qmambabsr: Burst image super-resolution with query state space model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23080–23090 (2025)
10. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015)
11. Dudhane, A., Zamir, S.W., Khan, S., Khan, F.S., Yang, M.H.: Burst image restoration and enhancement. In: Proceedings of the ieee/cvf Conference on Computer Vision and Pattern Recognition. pp. 5759–5768 (2022)
12. Dudhane, A., Zamir, S.W., Khan, S., Khan, F.S., Yang, M.H.: Burstormer: Burst image restoration and enhancement transformer. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5703–5712. IEEE (2023)
13. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
14. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Transactions on Graphics (SIGGRAPH) **36**(4), 118:1–118:12 (2017)
15. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: CVPR (2020)

16. Jiang, Y., Chan, K.C., Wang, X., Loy, C.C., Liu, Z.: Robust reference-based super-resolution via c2-matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2103–2112 (2021)
17. Jordan, K., Jin, Y., Boza, V., Jiacheng, Y., Cesista, F., Newhouse, L., Bernstein, J.: Muon: An optimizer for hidden layers in neural networks (2024), https://kellerjordan.github.io/posts/muon/
18. Kang, E., Lee, B., Im, S., Jin, K.H.: Burstm: Deep burst multi-scale sr using fourier space with optical flow. In: European Conference on Computer Vision. pp. 459–477. Springer (2024)
19. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video super-resolution with convolutional neural networks. IEEE transactions on computational imaging **2**(2), 109–122 (2016)
20. Ke, Z., Liu, Y., Zhu, L., Zhao, N., Lau, R.W.: Neural preset for color style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14173–14182 (2023)
21. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
22. Kim, Y., Lim, J., Cho, H., Lee, M., Lee, D., Yoon, K.J., Choi, H.J.: Efficient reference-based video super-resolution (ervsr): Single reference image is all you need. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1828–1837 (2023)
23. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Land, E.H., McCann, J.J.: Lightness and retinex theory. Journal of the Optical society of America **61**(1), 1–11 (1971)
25. Lee, J., Lee, M., Cho, S., Lee, S.: Reference-based video super-resolution using multi-camera video triplets. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17824–17833 (2022)
26. Li, Y., Liu, M.Y., Li, X., Yang, M.H.: A closed-form solution to photorealistic image stylization. In: ECCV (2018)
27. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
28. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
29. Long, W., Zhou, X., Zhang, L., Gu, S.: Progressive focused transformer for single image super-resolution. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2279–2288 (2025)
30. Lou, J., Zhao, X., Shi, K., Gu, S.: Learning pixel-adaptive multi-layer perceptrons for real-time image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14095–14105 (2025)
31. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the seventh IEEE international conference on computer vision. vol. 2, pp. 1150–1157. Ieee (1999)
32. Lu, L., Li, W., Tao, X., Lu, J., Jia, J.: Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6368–6377 (2021)
33. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4990–4998 (2017)

34. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
35. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing **39**(3), 355–368 (1987)
36. Rahman, Z.u., Jobson, D.J., Woodell, G.A.: Multi-scale retinex for color image enhancement. In: Proceedings of 3rd IEEE international conference on image processing. vol. 3, pp. 1003–1006. IEEE (1996)
37. Wang, T., Chen, Z., Zeng, H., Zhang, L.: Real-time image enhancer via learnable spatial-aware 3d lookup tables. In: ICCV (2021)
38. Wang, X., Chan, K.C.K., Yu, K., Dong, C., Loy, C.C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: CVPR Workshops (NTIRE) (2019)
39. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
40. Yang, C., Jin, M., Xu, Y., Zhang, R., Chen, Y., Liu, H.: Adaint: Learning adaptive intervals for 3d lookup tables on real-time image enhancement. In: CVPR (2022)
41. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5791–5800 (2020)
42. Yang, X., Xiang, W., Zeng, H., Zhang, L.: Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4781–4790 (2021)
43. Zeng, H., Cai, Y., Li, L., Cao, Z., Zhang, L.: Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. arXiv:2009.14468 (2020)
44. Zhang, L., Li, Y., Zhou, X., Zhao, X., Gu, S.: Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2856–2865 (2024)
45. Zhang, L., Li, X., He, D., Li, F., Ding, E., Zhang, Z.: Lmr: a large-scale multi-reference dataset for reference-based super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13118–13127 (2023)
46. Zhang, X., Chen, Q., Ng, R., Koltun, V.: Zoom to learn, learn to zoom. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3762–3770 (2019)
47. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018)
48. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2472–2481 (2018)
49. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image super-resolution by neural texture transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7982–7991 (2019)
50. Zhou, X., Zhang, L., Zhao, X., Wang, K., Li, L., Gu, S.: Video super-resolution transformer with masked inter&intra-frame attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25399–25408 (2024)