Efficient Video Super-Resolution via Two-Step 2D SR and Image-to-Image Conditional Generation

Yuya Masuda*
Department of Computer Science and Communications Engineering
Waseda University
Tokyo, Japan
msd_8172@fuji.waseda.jp

Shunta Shimizu*
Department of Computer Science and Communications Engineering
Waseda University
Tokyo, Japan
shimishunta@toki.waseda.jp

Hiroshi Ishikawa
Department of Computer Science and Communications Engineering
Waseda University
Tokyo, Japan
hfs@waseda.jp

***Abstract*—Video super-resolution (VSR) has gained significant attention due to its potential to enhance video quality. However, existing VSR models are often computationally heavy, limiting their practical application. To address this, we propose a two-step approach: applying 2D super-resolution (SR) followed by image-to-image conditional generation. Initial experiments using Mean Squared Error (MSE) show significant improvements. Replacing MSE with Mean Absolute Error (MAE) and blending it with Structural Similarity Index Measure (SSIM) further enhances performance. Our method achieves superior results with reduced computational complexity, offering a practical solution for real-world VSR applications. Our code is available at https://github.com/msdsm/efficient-video-sr-2steps.**

I. Introduction

Introduction Video super-resolution (VSR) has become a crucial area of research due to its ability to enhance video quality by reconstructing high-resolution frames from low-resolution inputs. Recent models such as Implicit Alignment with Recurrent Transformers(IART) [1], Masked Intra and inter-frame Attention (MIA-VSR) [2], Recurrent Video Restoration Transformer(RVRT) [3] and Video Restoration Transformer(VRT) [4] have achieved impressive results in improving visual fidelity. Despite their success, these models often come with significant computational costs, limiting their deployment in real-time or resource-constrained environments.

The computational burden arises from the complex network architectures and extensive temporal dependencies required to process consecutive frames effectively. As video content continues to proliferate across various platforms, the need for efficient and scalable VSR solutions becomes more pressing.

To address these challenges, we propose a novel two-step approach: first, applying 2D super-resolution (SR) to each frame independently, followed by an image-to-image conditional generation process to refine the temporal coherence. This method reduces the computational overhead while maintaining or even enhancing the output quality through an optimized loss function strategy.

Our work is not only focused on improving visual quality but also on minimizing resource usage, making VSR more accessible for real-time applications and devices with limited processing power.

Our Contributions are as follows:

1) We propose a two-step VSR approach combining 2D SR and image-to-image conditional generation, reducing computational complexity.
2) We develop an optimized loss function strategy to enhance performance without increasing model size. Our method achieves competitive results compared to state-of-the-art VSR models with significantly lower resource requirements.
3) We provide an open-source implementation.

II. Related Work

A. Hybrid Attention Transformer (HAT)

The Hybrid Attention Transformer (HAT) [5] is a state-of-the-art model designed to address the limitations of existing super-resolution (SR) transformer architectures such as SwinIR [6]. HAT integrates convolutional neural networks (CNNs) and transformer-based attention mechanisms in a unified framework, achieving a balance between local feature extraction and global contextual modeling. The overall architecture of HAT is illustrated in 1. The key components of HAT are detailed below.

1) Overall Architecture: HAT adopts a three-stage architecture similar to SwinIR, consisting of:

- Shallow Feature Extraction: A single convolutional layer extracts initial low-level features from the input image.
- Deep Feature Extraction: Multiple Residual Hybrid Attention Groups (RHAGs) and a concluding convolutional layer extract high-level features, progressively enhancing details for super-resolution.
- Image Reconstruction: A sub-pixel convolution layer generates the final high-resolution image from the extracted deep features.

2) Residual Hybrid Attention Group (RHAG): The RHAG forms the core of HAT's deep feature extraction module. Each RHAG comprises:

- Hybrid Attention Blocks (HAB): These blocks combine shifted window-based multi-head self-attention ((S)W-

* Equal contribution.

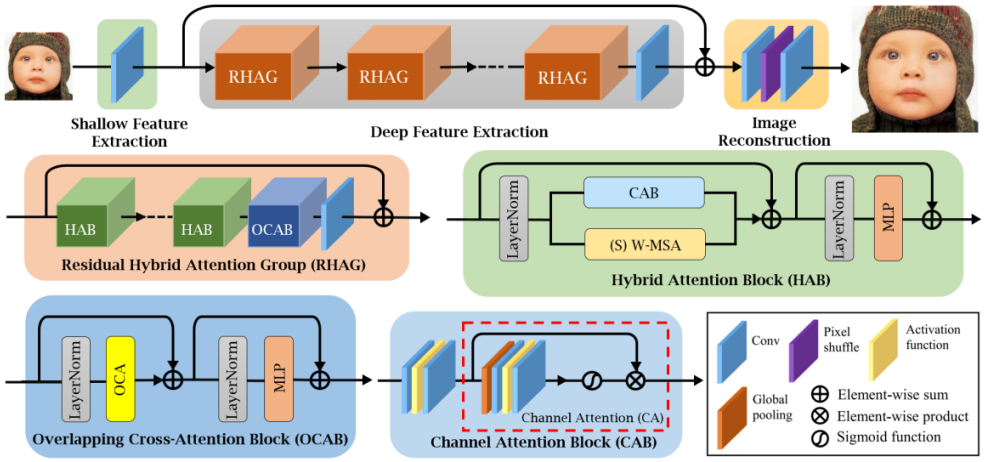

Fig. 1: Overview of HAT, reproduced from the original paper [5].

MSA) and Channel Attention Blocks (CAB) in a parallel configuration.

- Overlapping Cross Attention Block (OCAB): This component replaces the standard self-attention in transformers with overlapping cross-attention, enhancing cross-window information interactions by expanding the receptive field of key and value matrices.
- Residual Connections: Residual skip connections within and across blocks ensure efficient gradient flow and stable training.

3) Hybrid Attention Block (HAB): HAB is a novel attention mechanism combining the strengths of transformer and CNN. It computes the output as follows:

$$\begin{aligned} X_N &= \mathrm{LN}(X) \\ X_M &= (\mathrm{S})\mathrm{W\text{-}MSA}(X_N) + \alpha\,\mathrm{CAB}(X_N) + X \\ Y &= \mathrm{MLP}(\mathrm{LN}(X_M)) + X_M, \end{aligned} \tag{1}$$

where α is a learnable parameter controlling the balance between transformer-based and CNN-based features. The CAB enhances spatial feature refinement, while (S)W-MSA captures global dependencies.

4) Channel Attention Block (CAB): The CAB is a lightweight CNN module designed to focus on informative channels in the feature maps. It includes:

- Convolutional layers interleaved with activation functions (e.g., GELU).
- Global average pooling to aggregate spatial information across the feature map.
- Sigmoid gating to modulate channel-wise importance, followed by residual connections to preserve the input features.

5) Overlapping Cross Attention Block (OCAB): The OCAB modifies the self-attention mechanism by introducing overlapping regions for key and value matrices. This enhancement enables effective cross-window interaction:

- The input feature map X is split into query (Q), key (K), and value (V) components.
- K and V are expanded with overlapping regions, allowing the attention computation to capture dependencies beyond individual windows.
- The attention is computed as:

$$\mathrm{Attention}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V, \tag{2}$$

where B represents relative position encoding to incorporate spatial information.

B. U-Net

U-Net [9] is a deep learning model proposed in 2015. It features a U-shaped architecture, resembling an AutoEncoder, but it is not an AE. U-Net excels at image-to-image tasks, particularly segmentation, and has been used in models like pix2pix.

U-Net consists of two main paths:

- Contracting Path: This path extracts features through successive Convolution and Max Pooling layers, reducing spatial resolution while learning high-dimensional features.
- Expanding Path: This path reconstructs the spatial resolution using Up-convolution, integrating information from the corresponding layers in the Contracting Path via skip connections.

The skip connections preserve spatial information and significantly enhance accuracy, similar to the residual connections in ResNet [10].

III. Proposed Method

A. Overview

The overview of the proposed model is illustrated in Figure 3. The model consists of a pretrained HAT, a Conditional Left Low-Quality Image Encoder Φ_{L}, a Conditional Right Low-Quality Image Encoder Φ_{R}, and a U-Net-based Image Restoration Model Ψ. The Conditional Left/Right Low-Quality Image

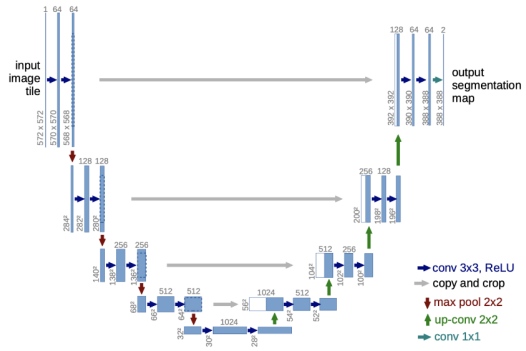

Fig. 2: U-Net Architecture, reproduced from the original paper [9]

Encoder comprises four downsampling blocks, each consisting of two convolutional layers and one pooling layer. Similarly, the Image Restoration Model consists of an encoder with five downsampling blocks, sharing a similar structure as the Conditional Left/Right Low-Quality Image Encoder, and a decoder with five upsampling blocks, each composed of one transposed convolutional layer and two convolutional layers. During training, the parameters of the HAT are fixed.

B. Conditional Low-Quality Image Encoder

The Conditional Left Low-Quality Image Encoder takes the low-resolution image $I_{\mathrm{LQ}_{t-1}} \in \mathbb{R}^{H\times W\times 3}$ at time $t-1$, as input and outputs $F_{\mathrm{LQ}_{t-1}} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 1024}$. Similarly, the Conditional Right Low-Quality Image Encoder processes the low-resolution image $I_{\mathrm{LQ}_{t+1}} \in \mathbb{R}^{H\times W\times 3}$ at time $t+1$, to obtain $F_{\mathrm{LQ}_{t+1}} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 1024}$.

$$\begin{cases} F_{\mathrm{LQ}_{t-1}} = \Phi_{\mathrm{L}}(I_{\mathrm{LQ}_{t-1}}) \\ F_{\mathrm{LQ}_{t+1}} = \Phi_{\mathrm{R}}(I_{\mathrm{LQ}_{t+1}}) \end{cases} \tag{3}$$

C. Image Restoration Model

For the low-resolution image $I_{\mathrm{LQ}_t} \in \mathbb{R}^{H\times W\times 3}$ at time t, super-resolution is performed using HAT to generate a high-resolution image $I_{\mathrm{HQ}_t} \in \mathbb{R}^{2H\times 2W\times 3}$. This high-resolution image is then fed into the encoder of the Image Restoration Model, which outputs encoded features $F_{\mathrm{HQ}_t} \in \mathbb{R}^{\frac{2H}{32}\times\frac{2W}{32}\times 1024}$.

$$F_{\mathrm{HQ}_t} = \Psi_{encoder}(I_{\mathrm{HQ}_t}) \tag{4}$$

In the intermediate layers of the model, the encoded features F_{HQ_t} are combined with the features from time $t-1$ and $t+1$ to form F_{fusion}. This conditional fusion leverages contextual features from neighboring frames to enhance the super-resolution process. Finally, the fused features F_{fusion} are decoded to generate the final output image $I_{output} \in \mathbb{R}^{2H\times 2W\times 3}$.

$$F_{\mathrm{fusion}} = F_{\mathrm{LQ}_{t-1}} + F_{\mathrm{HQ}_t} + F_{\mathrm{LQ}_{t+1}} \tag{5}$$

$$I_{output} = \Psi_{decoder}(F_{\mathrm{fusion}}) \tag{6}$$

TABLE I:
PSNR FOR DIFFERENT METHODS

	HAT	Ours (MSE)	Ours (MAE)
PSNR ↑	28.97	30.55	**30.73**

D. Loss Function

The loss function is defined to train the Conditional Left/Right Low-Quality Image Encoder and the Image Restoration Model. The Mean Absolute Error (MAE) L_{MAE} and Structural Similarity Index Measure (SSIM) Loss L_{SSIM} are computed between the output image I_{output} and the ground truth image at time t, $I_{GT} \in \mathbb{R}^{2H\times 2W\times 3}$. A weighted sum of these losses is used to train the three networks.

$$L_{\mathrm{SSIM}} = 1 - \mathrm{SSIM} \tag{7}$$

$$L_{\mathrm{total}} = \alpha L_{\mathrm{MAE}} + (1-\alpha) L_{\mathrm{SSIM}} \tag{8}$$

IV. Experiments

A. Dataset

In this study, we utilized the OpenVid-1M dataset [7]for Text-to-Video tasks and extracted 33,067 videos. For each video, 25 consecutive frames were selected, and 23 pairs of frames were created at times $t-1$, t, and $t+1$. This resulted in a dataset consisting of $33,067 \times 23$ samples. The dataset was then split into training, validation, and test sets in an 8:1:1 ratio, which were used for model training and evaluation.

B. Implementation Details

The optimization algorithm for training the Image Restoration Model and the Conditional Left/Right Image Encoder was Adam, with a learning rate of 1×10^{-4} and parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The input images corresponding to frames at times $t-1$, t, and $t+1$ were resized to 64×64 pixels. The scaling factor for HAT was set to 2x, with a total of 10 epochs and a batch size of 64.

C. Comparison of HAT and Proposed Method

The experimental results on the test set are presented in Table I. Peak Signal-to-Noise Ratio (PSNR) [8] was used as the evaluation metric. The low-resolution image at time t was input into HAT with a 2x scaling factor, and the generated image was compared to the image produced by our proposed method. The average PSNR of the generated images were calculated and compared.

From Table I, it can be observed that our method outperforms HAT in terms of PSNR. Additionally, training with Mean Absolute Error (MAE) instead of Mean Squared Error (MSE) yielded better performance, as evidenced by higher PSNR.

D. Incorporating SSIM Loss

To enhance the training process, Structural Similarity Index Measure (SSIM) Loss was integrated alongside MAE. A weighted sum of MAE and SSIM Loss was used as the loss

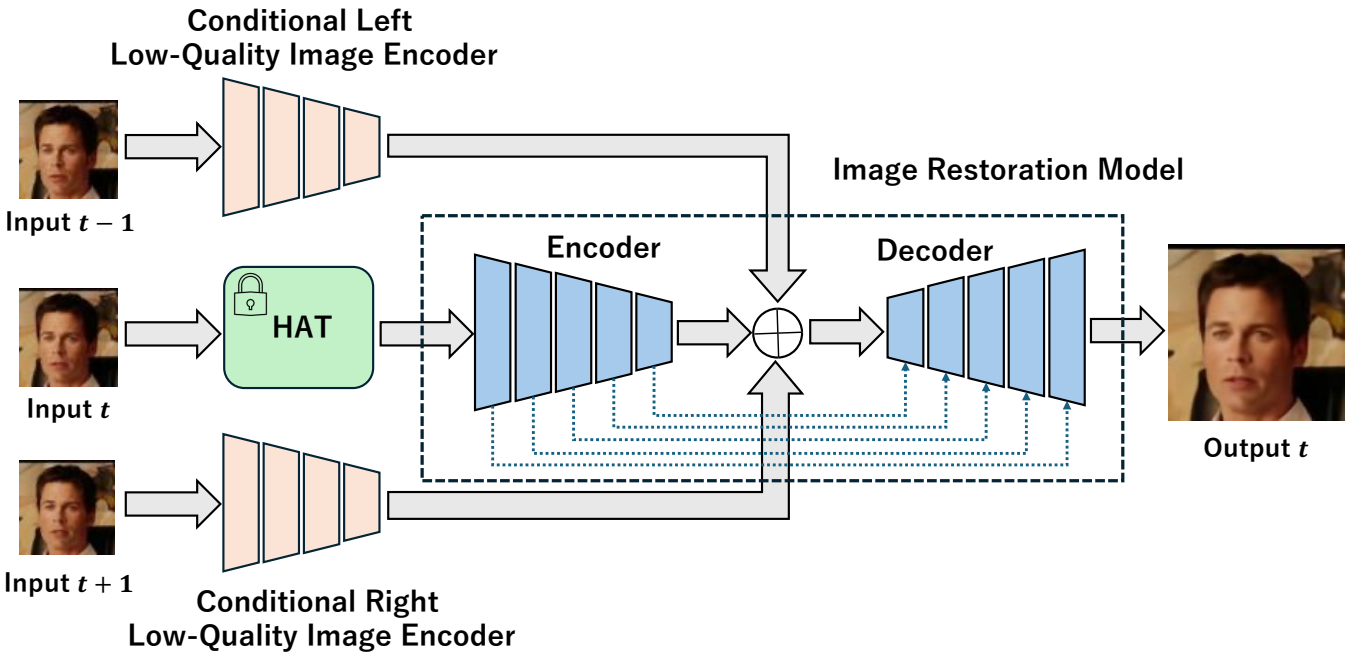

Fig. 3: Our proposed model

TABLE II:
PSNR FOR DIFFERENT α

α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
PSNR ↑	27.34	30.78	30.53	30.76	30.70	30.83	30.73	30.83	**30.86**	30.83	30.73

TABLE III:
PSNR FOR DIFFERENT α IN THE ENCODER LOSS

α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
PSNR ↑	30.55	30.61	**30.79**	30.70	30.76	30.63	30.74	30.77	30.76	30.61	30.745

function, and the hyperparameter α was adjusted to determine the optimal weight. The results are shown in Table II.

As seen in Table II, the PSNR is highest when $\alpha = 0.8$, indicating that incorporating SSIM Loss improves performance compared to using MAE alone ($\alpha = 1.0$). This demonstrates the effectiveness of SSIM Loss. Figure 4 compares the images generated by the model trained with $\alpha = 0.8$ to those generated by HAT.

E. Custom Loss for the Encoder

To enhance the encoder's performance, we propose a novel loss function that aims to align the features extracted from adjacent low-quality frames with the high-quality target frame. Specifically, given two low-quality images $I_{LQ_{t-1}}$ and $I_{LQ_{t+1}}$, the encoder outputs $F_{LQ_{t-1}}$ and $F_{LQ_{t+1}}$ from two separate encoders Φ_L and Φ_R as follows:

$$\begin{aligned} F_{LQ_{t-1}} &= \Phi_L\left(I_{LQ_{t-1}}\right), \\ F_{LQ_{t+1}} &= \Phi_R\left(I_{LQ_{t+1}}\right). \end{aligned} \tag{9}$$

The high-quality feature F_{HQ_t} is obtained by encoding the high-quality image I_{HQ_t} using a encoder $\Phi_{encoder}$:

$$F_{HQ_t} = \Phi_{encoder}\left(I_{HQ_t}\right). \tag{10}$$

The proposed encoder loss function $L_{encoder}$ is defined as:

$$\begin{aligned} L_{encoder} = {} & L_{MAE}\left(I_{output}, I_{gt}\right) \\ & + \alpha L_{MAE}\left(F_{LQ_{t+1}}, F_{HQ_t}\right) \\ & + (0.1-\alpha)\, L_{MAE}\left(F_{HQ_t}, F_{LQ_{t-1}}\right) \end{aligned} \tag{11}$$

where L_{MAE} represents the Mean Absolute Error (MAE) loss, I_{output} is the model's output, and I_{gt} is the ground truth image.

Table III summarizes the results of using this custom encoder loss. However, the performance did not surpass the previously mentioned MAE-SSIM score of 30.86.

V. Conclusion

In this work, we addressed the computational challenges of video super-resolution (VSR) by proposing a novel two-step approach that combines 2D super-resolution (SR) with subsequent image-to-image conditional generation. This method

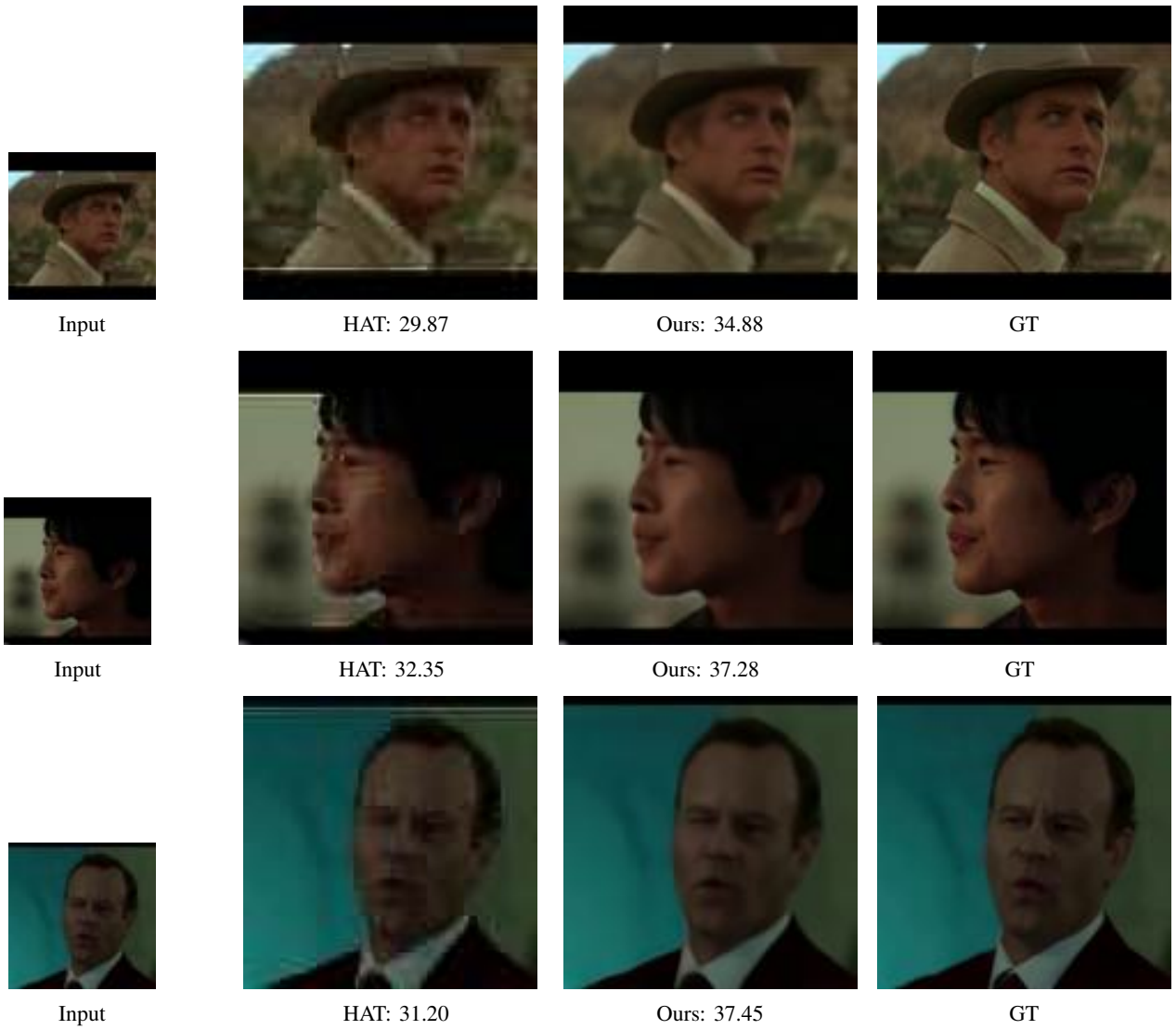

Fig. 4: Comparison of the input image (64×64) and its corresponding outputs (128×128). HAT: PSNR, Ours: PSNR are provided for clarity. All images are displayed with equal spacing and adjusted size for consistency.

effectively reduces computational complexity while maintaining high visual quality. Additionally, by optimizing the loss function using a combination of Mean Absolute Error (MAE) and Structural Similarity Index Measure (SSIM), our approach achieved superior performance compared to state-of-the-art models.

Acknowledgment. This work was partially supported by JSPS KAKENHI Grant Number JP20H00615.

REFERENCES

[1] Xu, Kai, et al. "Enhancing Video Super-Resolution via Implicit Resampling-based Alignment." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[2] Zhou, Xingyu, et al. "Video Super-Resolution Transformer with Masked Inter & Intra-Frame Attention."*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[3] Liang, Jingyun, et al. "Recurrent video restoration transformer with guided deformable attention." *Advances in Neural Information Processing Systems* 35 (2022): 378-393.

[4] Liang, Jingyun, et al. "Vrt: A video restoration transformer." *IEEE Transactions on Image Processing* (2024).

[5] Chen, Xiangyu, et al. "Activating more pixels in image super-resolution transformer." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.

[6] Liang, Jingyun, et al. "Swinir: Image restoration using swin transformer." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[7] Nan, Kepan, et al. "OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-video Generation.", *arXiv preprint arXiv:2407.02371*. 2024.

[8] Huynh-Thu, Q, et al. "Scope of validity of PSNR in image/video quality assessment." *Electronics letters*. 2018.
[9] Ronneberger, et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation.", *Medical Image Computing and Computer-Assisted Intervention.* 2015.
[10] HE, Kaiming, et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 770-778.