

EulerMormer: Robust Eulerian Motion Magnification via Dynamic Filtering within Transformer

Fei Wang¹, Dan Guo^{1,2,3*}, Kun Li¹, Meng Wang^{1,2*}

¹School of Computer Science and Information Engineering, Hefei University of Technology

²Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

³Anhui Zhonghuitong Technology Co., Ltd

jiafei127@gmail.com, guodan@hfut.edu.cn, kunli.hfut@gmail.com, wangmeng@hfut.edu.cn

Abstract

Video Motion Magnification (VMM) aims to break the resolution limit of human visual perception capability and reveal the imperceptible minor motion that contains valuable information in the macroscopic domain. However, challenges arise in this task due to photon noise inevitably introduced by photographic devices and spatial inconsistency in amplification, leading to flickering artifacts in static fields and motion blur and distortion in dynamic fields in the video. Existing methods focus on explicit motion modeling without emphasizing prioritized denoising during the motion magnification process. This paper proposes a novel dynamic filtering strategy to achieve static-dynamic field adaptive denoising. Specifically, based on Eulerian theory, we separate texture and shape to extract motion representation through inter-frame shape differences, expecting to leverage these subdivided features to solve this task finely. Then, we introduce a novel dynamic filter that eliminates noise cues and preserves critical features in the motion magnification and amplification generation phases. Overall, our unified framework, EulerMormer, is a pioneering effort to first equip with Transformer in learning-based VMM. The core of the dynamic filter lies in a global dynamic sparse cross-covariance attention mechanism that explicitly removes noise while preserving vital information, coupled with a multi-scale dual-path gating mechanism that selectively regulates the dependence on different frequency features to reduce spatial attenuation and complement motion boundaries. We demonstrate extensive experiments that EulerMormer achieves more robust video motion magnification from the Eulerian perspective, significantly outperforming state-of-the-art methods. The source code is available at <https://github.com/VUT-HFUT/EulerMormer>.

Introduction

Video Motion Magnification (VMM) has garnered growing research interest due to its remarkable ability to vividly reveal subtle motions in real-world videos that are imperceptible to the human eye (Rubinstein et al. 2013; Le Ngo and Phan 2019). Existing VMM techniques behave as computer-assisted “eyes” that enable humans to see and grasp meaningful subtle motion in various challenging-to-perceive settings, such as micro-action recognition (Xia et al. 2020; Qi

*Corresponding authors.

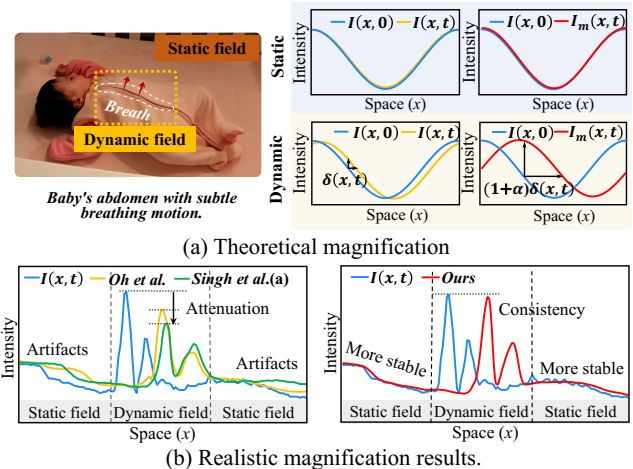


Figure 1: Theoretical basis and realistic results of VMM. Theoretically, the static field in (a) is free of position displacement, while the dynamic field should exhibit ideal position displacement to satisfy the desired motion magnification. However, in the real world, unavoidable photon noise and spatial inconsistency exist with artifacts, intensity attenuation, etc., in (b) for the magnified results (Oh et al. 2018; Singh, Murala, and Kosuru 2023a). In contrast, we achieve more robust magnification in both static and dynamic fields.

et al. 2020; Mehra et al. 2022; Nguyen et al. 2023), robotic sonography (Huang et al. 2023), clinical medicine (Abnoui et al. 2019) material properties estimation (Davis et al. 2015, 2017) and modal analysis (Eitner et al. 2021). Specifically, VMM aims to capture and amplify the imperceptible subtle motion in the video sequence while preserving fine spatial details for realistic and accurate visualization.

However, this task faces several challenges: (1) **Photon noise** (Oh et al. 2018) is inevitably present in videos due to the characteristics of charge-coupled devices (CCDs) in photographic devices and signal attenuation during transmission. The noise, indistinguishable from subtle motions, can result in flickering artifacts. (2) **Spatial inconsistency** (Wadhwa et al. 2013) measures the magnified quality, as forced motion magnification can lead to spatial frequency collapse, resulting in phenomena such as motion blur and

distortion. As the results of recent methods (Oh et al. 2018; Singh, Murala, and Kosuru 2023a) shown in Fig. 1(b), noise amplification disrupts the static field of magnified image, and spatial intensity attenuation occurs in the dynamic field.

Inspired by the theory of fluid mechanics, early research drew from the Lagrangian and Eulerian perspectives. Liu *et al.* (Liu et al. 2005) proposed the first Lagrangian-based approach, which involved tracking the motion trajectory of each pixel (optical flow) for motion magnification, but it was computationally expensive and sensitive to various noises. In contrast, Eulerian approaches (Wu et al. 2012; Wadhwa et al. 2013; Zhang, Pintea, and Van Gemert 2017; Takeda et al. 2018, 2019) relied on traditional filters to handle the motion intensity occurring in specific regions rather than tracking every pixel throughout the video. However, these Eulerian methods required fine-tuning numerous hyperparameters to adapt to different scenarios, which makes them impractical for real-world applications.

Developing effective VMM methods remains a compelling topic in the computer vision community. Recently, learning-based methods (Oh et al. 2018; Singh, Murala, and Kosuru 2023a,b) utilizing different convolutional neural networks (CNN) have attained SOTA performance. Whether they introduce proxy model regularization or frequency domain phase (Wang et al. 2022b, 2024) to optimize their models, they essentially focus on representation learning, such as motion and phase, for generating motion-magnified videos without emphasizing prioritizing denoising.

This paper focuses primarily on addressing the denoising issue in VMM. We specially design a dynamic filter module $\mathcal{F}(\cdot)$ to address the previously mentioned photon noise and spatial inconsistency in static and dynamic fields. Based on Eulerian theory, we disentangle texture and shape and further acquire the motion = Δ shape, which is expected to leverage these subdivided features to solve this task finely. Especially noteworthy is the to-be-magnified motion representation. In our framework, we utilize $\mathcal{F}(\cdot)$ to filter out noise cues from motion during the *motion magnification* phase and refine the representations of texture and magnified shape during the *amplification generation* phase. Finally, compared with the limitation of existing CNN-based methods with local receptive fields, our method is equipped with Transformer architecture in the encoder and the dynamic filter $\mathcal{F}(\cdot)$, which can ensure the contextualized global relationship learning between pixels. Overall, we provide a unified framework to filter out undesired noise cues in the representation learning of texture, shape, and motion, which results in a satisfactory magnification effect.

Our contributions can be summarized as follows:

- We introduce a novel Transformer-based EVM architecture that offers better spatial consistency and fewer artifacts and distortion in the magnified video. To our knowledge, this is a pioneering effort in learning-based VMM.
- We develop a dynamic filter implemented on a sparse attention strategy for static-dynamic field adaptive denoising and texture-shape joint refinement during the motion magnification and amplification generation phases.
- We propose a Point-wise Magnifier, which improves the

magnified representation by incorporating global nonlinear feature interactions per pixel to maintain spatial consistency and reduce flickering artifacts.

- Extensive quantitative and qualitative experiments on synthetic and real-world datasets demonstrate our favorable performances against SOTA approaches.

Related Work

Traditional Methods. Lagrangian-based approaches (Liu et al. 2005) pioneered this task by tracking the motion trajectory of each pixel for motion magnification, but dense optical flow computation is expensive and sensitive to noise. Eulerian-based methods (Wu et al. 2012; Wadhwa et al. 2013; Zhang, Pintea, and Van Gemert 2017; Takeda et al. 2018, 2019, 2022) concentrate on the specific regions where motion occurs, rather than tracking every pixel in the video. Early Eulerian-based methods (Wu et al. 2012; Wadhwa et al. 2013) altered intensities to approximate linear magnification or decomposed the motion in the frequency domain. With further research, various hand-crafted filters, such as acceleration (Zhang, Pintea, and Van Gemert 2017), jerk (Takeda et al. 2018), anisotropy (Takeda et al. 2019), and bilateral filters (Takeda et al. 2022), were explored. These works rely on the predefined bandwidth for bandpass filters to amplify specific motions, but their effectiveness requires extensive hyperparameter tuning.

Deep-Learning Methods. Learning-based approaches for the VMM have emerged but are still in their infancy, with only a handful of related works (Oh et al. 2018; Brattoli et al. 2021; Singh, Murala, and Kosuru 2023a,b). Oh *et al.* (Oh et al. 2018) proposed a CNN-based end-to-end architecture for the first attempt to learn the motion magnification representation, achieving comparable results to hand-crafted filters. Recently, Singh *et al.* (Singh, Murala, and Kosuru 2023a) proposed a lightweight CNN-based proxy model to eliminate undesired motion efficiently. Afterwards, they (Singh, Murala, and Kosuru 2023b) also utilized CNN to model the magnification signals from frequency domain phase fluctuations to avoid artifacts and blurring in the spatial domain. Unlike the above CNN methods with local receptive fields (Zheng et al. 2023; Guo et al. 2019; Zhou et al. 2021, 2022), we introduce a novel dynamic filtering strategy into Transformer-based architecture (Li et al. 2023; Li, Guo, and Wang 2021; Tang et al. 2022; Li, Guo, and Wang 2023) in this study. Intrinsically, based on the Eulerian theory, we integrate the advanced Transformer to globally model the texture, shape, and motion representations, enabling static-dynamic field adaptive denoising for motion magnification.

Preliminaries

Task Definition

Let $I(x, t)$ denote the intensity at spatial position x and time t . With $I(x, 0)=f(x)$ and $I(x, t)=f(x + \delta(x, t))$, $\delta(x, t)$ denotes a displacement function of x at time t , the goal of VMM is to synthesize $I_m(x, t)$ with respect to a magnification factor α as follows (Wu et al. 2012; Zhang et al. 2023):

$$I_m(x, t) = f(x + (1 + \alpha)\delta(x, t)). \quad (1)$$

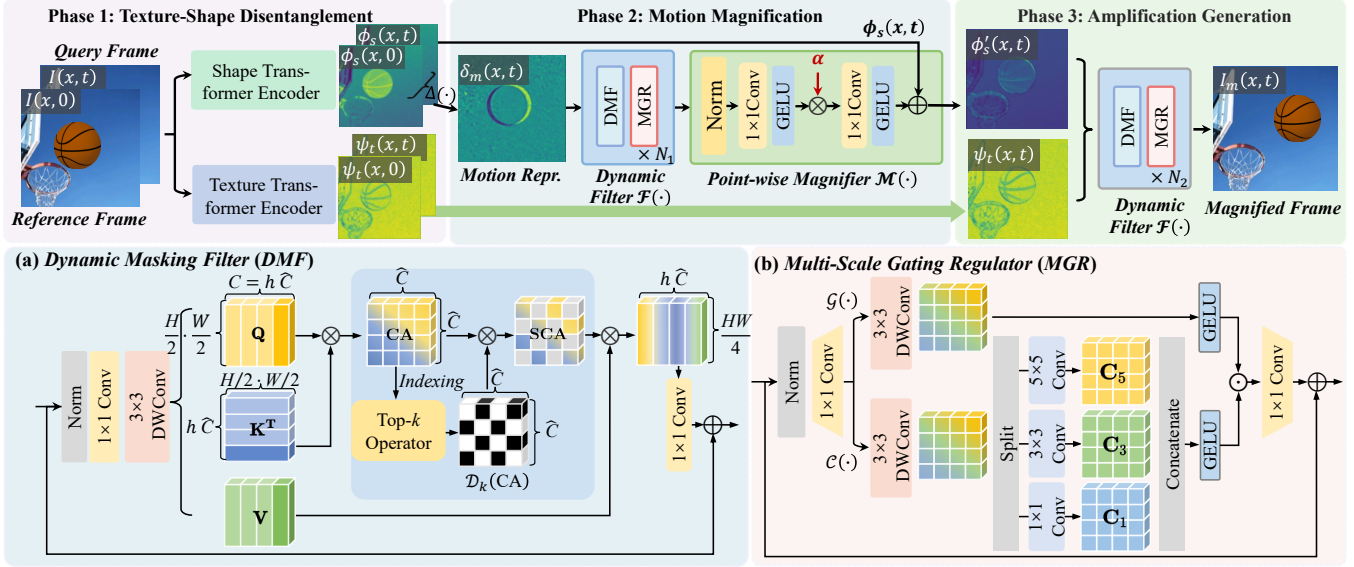


Figure 2: The overall architecture of EulerMormer for video motion magnification, which consists of three phases: (1) texture-shape disentanglement, (2) motion magnification with a dynamic filter $\mathcal{F}(\cdot)$ and a point-wise magnifier $\mathcal{M}(\cdot)$, and (3) amplification generation, which recouples and refines the original texture $\psi_t(x, t)$ and the magnified shape $\phi'_s(x, t)$ to generate high-quality magnified frames. Among them, the dynamic filter $\mathcal{F}(\cdot)$, consisting of DMF in (a) and MGR in (b), performs twice in motion magnification and amplification generation processes, which targets to achieve the static-dynamic field adaptive denoising in terms of texture, shape and motion representation learning.

We can approximate $I(x, t)$ in a first-order Taylor series expansion as:

$$I(x, t) \approx f(x) + \delta(x, t) \frac{\partial f(x)}{\partial x}, \quad (2)$$

where $\delta(x, t) \frac{\partial f(x)}{\partial x}$ is regarded as the intensity magnitude.

Combining Eqs. 1 and 2, we have the magnification:

$$I_m(x, t) \approx f(x) + (1 + \alpha) \delta(x, t) \frac{\partial f(x)}{\partial x}. \quad (3)$$

According to Eulerian learning-based VMM methods (Oh et al. 2018; Singh, Murala, and Kosuru 2023a), the motion magnification process can be disentangled into texture and shape components as shown in Eq. 4. In this work, our method belongs to this methodological scope.

$$I_m(x, t) \approx \underbrace{I(x, t)}_{Texture} + \alpha \underbrace{\delta(x, t)}_{\Delta Shape} \frac{\partial f(x)}{\partial x}. \quad (4)$$

Motivation

As described above, videos can be modeled by two independent latent variables: texture and shape. Texture representation exhibits invariance, while the motion generated by shape displacement for magnification deserves further investigation. We extract subtle motion by calculating the inter-frame shape difference between two frames, *i.e.*, motion = Δ shape. Meanwhile, the amplification of subtle motion is inevitably affected by noise, as depicted in Fig. 1, *i.e.*, photon noise in the static field and spatial inconsistency in the dynamic field. To this end, we propose a dynamic filter $\mathcal{F}(\cdot)$ in our framework designed for denoising to eliminate

artifacts and distortion caused by these noises. It is applied twice within our framework: once for denoising the motion representation and once for denoising the recoupled texture-magnified shape joint refinement, formulated as follows:

$$I_{Ours}(x, t) = \mathcal{F} \left[\underbrace{I(x, t)}_{Texture} + \alpha \mathcal{F} \left(\underbrace{\delta(x, t)}_{\Delta Shape} \frac{\partial f(x)}{\partial x} \right) \right]. \quad (5)$$

Methodology

Texture-Shape Disentanglement

Given any pair of reference and query images in a video, $[I(x, 0), I(x, t)]$, we use a 3×3 convolution layer to obtain initial feature maps $F(x, 0), F(x, t) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$, and further use a Texture Transformer Encoder $\psi_t(\cdot)$ and a Shape Transformer Encoder $\phi_s(\cdot)$ to obtain their texture and shape representations, *i.e.*, $[\psi_t(x, 0), \phi_s(x, 0)] \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$, $[\psi_t(x, t), \phi_s(x, t)] \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$, as shown in Fig. 2. Specifically, the two encoders comprise the Transformer with Multi-Dconv Head Transposed Attention (MDTA, derived from Restormer (Zamir et al. 2022)) and Multi-Scale Gated Regulator (MGR, see Sec.). MDTA replaces multi-head self-attention (MHA) in Transformer and facilitates contextualized global interaction between pixels by incorporating depth-wise convolutions and cross-covariance attention. This choice enables efficient pixel-grained representation learning, making it well-suited for this task. Our MGR utilizes the multi-scale dual-path gating mechanism to selectively integrate features at different frequencies, providing satisfactory texture and shape representations.

Motion Magnification

Obtaining a “clean” motion representation is crucial for motion magnification, as the inherent photon noise has nearly equivalent energy fields and the subtle motion change and is prone to amplify noise resulting in artifacts and distortion. We define the motion representation by implementing a simple inter-frame shape difference, *i.e.*, $\delta_m(x, t) = \Delta(\phi_s(x, t), \phi_s(x, 0)) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$. To manipulate the motion magnification, we describe two core components (DMF and MGR, see below) of the dynamic filter $\mathcal{F}(\cdot)$ and a point-wise magnifier $\mathcal{M}(\cdot)$ in detail below.

Dynamic Masking Filter (DMF). We revisit multi-head self-attention on the motion $\delta_m(x, t)$. After implementing 1×1 convolution and 3×3 depth-wise convolutions, we group $\delta_m(x, t)$ into h heads and each single-headed projection has $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{(\frac{H}{2} \times \frac{W}{2}) \times \hat{C}}$, where $\hat{C} = \frac{C}{h}$ and $h = 4$. On each head, we calculate a cross-covariance attention matrix $\mathbf{CA} \in \mathbb{R}^{\hat{C} \times \hat{C}}$ between \mathbf{K} and \mathbf{Q} . In \mathbf{CA} , a learnable temperature τ scales inner products before calculating attention weights, enhancing training stability.

$$\mathbf{CA} = \tau \mathbf{K}^T \cdot \mathbf{Q}, \quad (6)$$

In Fig. 2(a), a critical design of DMF is that we take \mathbf{CA} as a search space to perform dynamic sparse erasing. Our sparse strategy applies a dynamic filtering mechanism with the Top- k operator (Zhao et al. 2019; Wang et al. 2022a) along the channel dimension. Specifically, we adaptively select row-wise top- k contributive elements based on the channel correlation scores in \mathbf{CA} . Then, we utilize Eq. 7 to generate the corresponding binary mask for position indexing, representing the relative positions of the high-contributing elements obtained in \mathbf{CA} . Here, the dynamic mask $\mathcal{D}_k(\mathbf{CA}) \in \mathbb{R}^{\hat{C} \times \hat{C}}$ is formulated as:

$$[\mathcal{D}_k(\mathbf{CA})]_{ij} = \begin{cases} \mathbf{CA}_{ij} & \mathbf{CA}_{ij} \geq k_{ij} \\ 0 & \text{otherwise} \end{cases}, \quad (7)$$

where k_{ij} represents the k -th row-wise maximum value in \mathbf{CA}_{ij} . This allows us to dynamically degenerate the dense \mathbf{CA} into a sparse attention matrix $\mathbf{SCA} \in \mathbb{R}^{\hat{C} \times \hat{C}}$:

$$\mathbf{SCA} = \text{Softmax}(\mathcal{D}_k(\mathbf{CA})). \quad (8)$$

After the implementation of the weighted *value* \mathbf{V} sum with the sparse matrix \mathbf{SCA} , we concat all the heads’ results and output the updated motion $\delta'_m(x, t) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$. DMF is designed to explicitly remove noise from the static-dynamic fields in $\delta_m(x, t)$ and preserve the desired motion, preventing distortion and artifacts caused by amplified noise.

Multi-Scale Gating Regulator (MGR). Humans intelligently perceive visual changes across multiple scales. However, when the motion is too subtle and indistinguishable from noise, the integrity of the motion trajectory is compromised. Based on the DMF processing noise, we propose MGR that repairs the smoothness and uncertainty of the motion contours to overcome this issue. The MGR is a dual-path feedforward network consisting of multi-scale context branches $\mathcal{C}(\cdot)$ and dual-path gating $\mathcal{G}(\cdot)$, see Fig. 2(b).

We normalize and map the motion $\delta'_m(x, t)$ to a high-

dimensional space with a 1×1 convolution, where the expansion factor is $\eta = 3$. Next, after a 3×3 depth-wise convolution, the motion representation is split into dual-path gates, *i.e.*, $\mathcal{G}(\delta'_m(x, t)), \mathcal{C}(\delta'_m(x, t)) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times \frac{\eta C}{2}}$. For $\mathcal{C}(\cdot)$, we parallelly employ three depth-wise convolutions with the kernel sizes of $s \in \{1, 3, 5\}$ to capture the interactions at different frequencies $\mathbf{C}_1, \mathbf{C}_3, \mathbf{C}_5 \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times \frac{\eta C}{6}}$. Notably, high-frequency noise characterized by a small scale is effectively handled by the low-frequency characteristics of \mathbf{C}_1 . With increasing kernel sizes, \mathbf{C}_3 and \mathbf{C}_5 play a crucial role in motion contours acquisition and motion complementation. And these different frequency features are fused before passing through a layer with a nonlinear activation function of GELU. As for $\mathcal{G}(\cdot)$, a GELU activation function ensures nonlinear feature transformation. Finally, MGR regulates the output by Hadamard product \odot with $\mathcal{G}(\cdot)$ and $\mathcal{C}(\cdot)$:

$$\delta''_m(x, t) = \mathcal{G}(\delta'_m(x, t)) \odot \mathcal{C}(\delta'_m(x, t)), \quad (9)$$

where the output of updated motion $\delta''_m(x, t) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$. The combination process of DMF and MGR is defined as the dynamic filter $\mathcal{F}(\cdot)$.

Point-wise Magnifier (PWM). In this part, PWM serves as a manipulator to perform nonlinear magnification on $\delta''_m(x, t) = \mathcal{F}(\delta_m(x, t))$. It adopts a simple and efficient design with two fundamental modifications to improve magnified representation learning: (a) in order to reduce flickering artifacts, we abandon local convolutions and operate point-wise convolutions to interact with magnification across channels, thereby reducing checkerboard artifacts and being more compatible with global filtering; (b) we use the more stable GELU activation function to provide nonlinear representation learning and avoid gradient explosion. Therefore, the calculation process of PWM is as follows:

$$\phi'_s(x, t) = W_p(\alpha \cdot W_p(\delta''_m(x, t))) + \phi_s(x, t), \quad (10)$$

where $\phi'_s(x, t) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$ represents the amplified shape representation with the factor α and $W_p(\cdot)$ denotes the point-wise convolution with GELU activation layer.

Amplification Generation

We reconstruct the high-quality magnified image by recoupling the magnified shape $\phi'_s(x, t)$ with the original texture $\psi_t(x, t)$. Its challenge is avoiding high-frequency noise from $\psi_t(x, t)$ and ringing artifacts at the recoupled boundaries. For this purpose, we recouple $\phi'_s(x, t)$ and $\psi_t(x, t)$ across the feature channels and adopt the same dynamic filter $\mathcal{F}(\cdot)$ to perform the texture-magnified shape joint refinement to facilitate their fusion and boundary completeness:

$$I_m(x, t) = W_{up}(\mathcal{F}(\phi'_s(x, t), \psi_t(x, t))), \quad (11)$$

where $W_{up}(\cdot)$ denotes a layer that combines pixel shuffling operation (Shi et al. 2016) and a 3×3 convolution to perform sub-pixel level upsampling, generating the final magnified image $I_m(x, t)$. Methodologically, $\mathcal{F}(\cdot)$ in this section dynamically filters $\phi'_s(x, t)$ and $\psi_t(x, t)$ through interactive guidance along the channel to suppress noise while aiding in synthesizing smooth motion boundaries and clear details.

Loss Optimization

To optimize the proposed model, the objective function \mathcal{L} is the weighted sum of three loss terms as follows:

$$\mathcal{L} = \mathcal{L}_{mag} + \mu_1 \mathcal{L}_{dr} + \mu_2 \mathcal{L}_{edge}, \quad (12)$$

where μ_1, μ_2 are hyperparameters to balance the three loss functions. \mathcal{L}_{mag} is a basic loss term, which calculates the Charbonnier penalty (Bruhn, Weickert, and Schnörr 2005) between the magnified image I_m and ground-truth I_{GT} as:

$$\mathcal{L}_{mag} = \sqrt{\|I_m(x, t) - I_{GT}(x, t)\|^2 + \varepsilon^2}, \quad (13)$$

where ε is a constant value, being empirically set to 10^{-3} . The robust Charbonnier penalty term approximates the l_1 -loss and easily captures outliers in $I_m(x, t)$. Besides, similar to (Oh et al. 2018; Singh, Murala, and Kosuru 2023a), we use a color perturbation loss \mathcal{L}_{dr} to enforce the disentangled representation learning of shape and texture as follows:

$$\mathcal{L}_{dr} = \mathcal{L}(\phi_s(x, t), \phi_s^c(x, t)) + \mathcal{L}(\psi_t(x, t), \psi_t^c(x, t)). \quad (14)$$

where $[\phi_s(x, t), \psi_t(x, t)]$ and $[\phi_s^c(x, t), \psi_t^c(x, t)]$ are respective shape and texture representations of image $I(x, t)$ and its color perturbed image $I^c(x, t)$. Notably, in this study, we use a new loss term \mathcal{L}_{edge} , namely using a Laplacian of Gaussian (LoG) edge detector E_{LoG} (Zhang et al. 2017) with Charbonnier penalty, that is used to restrict the consistency between texture and amplified shape deformation as:

$$\mathcal{L}_{edge} = \sqrt{\|E_{LoG}(I_m(x, t)) - E_{LoG}(I_{GT}(x, t))\|^2 + \varepsilon^2}. \quad (15)$$

Experiments

Experiment Setup

Real-World Datasets. We experiment on three real-world benchmarks used in previous work: **(a) Static dataset** (Wu et al. 2012; Wadhwa et al. 2013; Oh et al. 2018) and **(b) Dynamic dataset** (Zhang, Pinteá, and Van Gemert 2017; Oh et al. 2018) contain 10 and 6 classic subtle videos in both static (slight motion, *e.g.* baby breathing) and dynamic (*e.g.*, strenuous motion and perspective shifts) scenarios. **(c) Fabric dataset** (Davis et al. 2015, 2017) contains 30 videos of subtle changes in fabric surface under wind excitation.

Synthetic Datasets. Real-world videos are rich in perceptual characteristics but lack ground truth annotations. Thus, we generate a synthetic dataset for quantitative evaluation. We select 100 objects from the public StickPNG library¹ and 600 high-resolution background images from the DIS5K dataset (Qin et al. 2022). In the data generation, we randomly place the objects onto the background images, initializing them as reference frames. Subsequently, we synthesize query frames by randomly adjusting the object direction and velocity with velocities limited to the range of (0, 2] to imitate subtle motions of objects. Therefore, we multiply the inter-frame velocities by magnification factors to synthesize the accurate ground truth for magnified motion. We create three synthetic subsets: **Synthetic-I Dataset:** Implementing random magnification factors $\alpha \in (0, 50]$; **Synthetic-II**

Dataset: Adding Poisson noise with the scale of random intensity levels $\lambda \in [3, 30]$; **Synthetic-III Dataset:** Adding Gaussian blurs with the scale of random standard deviations $\sigma \in [3, 30]$. In conclusion, the synthetic dataset contains 1,800 pairs of images and corresponding α .

Implementation Details. Following the protocol (Oh et al. 2018; Singh, Murala, and Kosuru 2023a,b), all methods are implemented with the same training data from (Oh et al. 2018) comprising 100,000 pairs of input sized 384×384 pixels. The focus in this field revolves around cross-dataset testing. We employ the Adam optimizer (Kingma and Ba 2015; Qian et al. 2023; Zhou, Guo, and Wang 2022) with the learning rate of 2×10^{-4} and the batch size of 4. For the network setting, the feature channel C is set to 48, and the numbers of the Texture Transformer Encoders and Shape Transformer Encoders are 2. The dynamic filter $\mathcal{F}(\cdot)$ is configured with $N_1 = 2$ in Phase 2 and $N_2 = 8$ in Phase 3, and the Top- k operator is set with $k = 7$. Besides, we set the loss hyperparameters as $\mu_1 = 0.1$ for \mathcal{L}_{dr} and $\mu_2 = 0.5$ for \mathcal{L}_{edge} .

Evaluation Metrics. For synthetic datasets, we employ RMSE to assess magnification error and PSNR (Shen, Zhao, and Zhang 2023; Shen et al. 2023), SSIM, and LPIPS (Zhang et al. 2018) to assess the magnification quality. For real-world datasets, we introduce an advanced no-reference image quality assessment metric, MANIQA (Yang et al. 2022). MANIQA is the NTIRE 2022 NR-IQA challenge winner achieves human-comparable quality assessment and is widely applied in image distortion and video reconstruction tasks (Wu et al. 2022; Ercan et al. 2023).

Quantitative Comparisons

Comparisons on Synthetic Datasets. We compare with existing approaches and report the experimental results in Tab. 1. On the Synthetic-I, our method performs superior to the recent best method MDLMM on magnification error and visual quality, with RMSE, PSNR, SSIM, and LPIPS values of 0.0594 *vs.* 0.0651, 25.49 dB *vs.* 24.84 dB, 0.9536 *vs.* 0.9173, and 0.0535 *vs.* 0.1228, respectively. On Synthetic-II and Synthetic-III, EulerMormer still shows significant performance gains for Poisson noise and Gaussian blur.

Comparisons on Real-World Datasets. From Tab. 2, previous works with traditional narrowband filters (Zhang, Pinteá, and Van Gemert 2017; Takeda et al. 2018, 2019) have lower MANIQA scores than ours. The MANIQA metric (Yang et al. 2022) mainly evaluates visual distortion levels. For example, compared to the previous best method Anisotropy (Takeda et al. 2019), we achieve 0.6920 *vs.* 0.6872, 0.6760 *vs.* 0.6634, and 0.7316 *vs.* 0.7288 on Static, Dynamic, and Fabric datasets, respectively.

Ablation Studies

Effectiveness of Filter $\mathcal{F}(\cdot)$. We test the dynamic filter $\mathcal{F}(\cdot)$ in Phase 2 and Phase 3 separately. Observing Tab. 3, removing $\mathcal{F}(\cdot)$ from the model significantly decreases both accuracy and quality of the magnification (*e.g.*, removing $\mathcal{F}(\cdot)$ decreases the perceptual quality LPIPS from 0.0535 to 0.1170 in Phase 2 and increases the magnification error

¹The StickPNG library is available at <https://stickpng.com/>.

Method	Venue	Synthetic-I: Magnification (α)				Synthetic-II: Poisson Noise (λ)				Synthetic-III: Gaussian Blur (σ)			
		RMSE	PSNR	SSIM	LPIPS	RMSE	PSNR	SSIM	LPIPS	RMSE	PSNR	SSIM	LPIPS
Linear	TOG'12	0.1029	20.21	0.8397	0.3247	0.1102	19.39	0.6746	0.2497	0.1347	17.21	0.5874	0.4666
Phase	TOG'13	0.0978	21.18	0.8613	0.1428	0.1053	20.30	0.6941	0.2283	0.1206	18.87	0.6109	0.4499
Acc.	CVPR'17	0.0781	22.99	0.9299	0.1346	0.0854	22.20	0.7694	0.1922	0.1011	20.62	0.6508	0.4242
Jerk	CVPR'18	0.0746	23.61	0.9333	0.1302	0.0787	23.06	0.7964	0.1844	0.0951	20.82	0.6612	0.4156
LBVMM	ECCV'18	0.0682	23.89	0.8748	0.1775	0.0700	23.65	0.8329	0.2164	0.0913	21.19	0.6645	0.4177
Aniso.	CVPR'19	0.0687	24.01	<u>0.9386</u>	0.1260	0.0745	23.72	0.8230	0.1744	0.0919	20.93	<u>0.6646</u>	<u>0.4121</u>
LNVM	WACV'23	0.0662	24.19	0.8943	0.1544	0.0681	23.92	0.8497	0.1889	0.0915	21.16	0.6581	0.4264
MDLMM	CVPR'23	<u>0.0615</u>	24.84	0.9173	<u>0.1228</u>	<u>0.0637</u>	<u>24.53</u>	<u>0.8659</u>	<u>0.1720</u>	<u>0.0896</u>	<u>21.34</u>	<u>0.6639</u>	<u>0.4205</u>
Ours	-	0.0594	25.49	0.9536	0.0535	0.0616	25.04	0.8706	0.1604	0.0867	21.89	0.6797	0.4077

Table 1: Quantitative comparison of our EulerMormer and existing methods on three subsets of the synthetic dataset: evaluating magnification accuracy, noise robustness, and blur sensitivity. Our EulerMormer achieves the best performance.

Method	Static	Dyn.	Fabric
Linear	.6288	.5169	.6597
Phase	.6696	.5861	.7120
Acc.	.6748	.6289	.7225
Jerk	.6769	.6594	.7256
LBVMM	.6830	.6409	.7234
Aniso.	<u>.6872</u>	<u>.6634</u>	<u>.7288</u>
LNVM	.6332	.6435	.7195
MDLMM	.6297	.6150	.7134
Ours	.6920	.6760	.7316

Table 2: Quantitative comparison on real-world datasets in the term of MANIQA \uparrow .

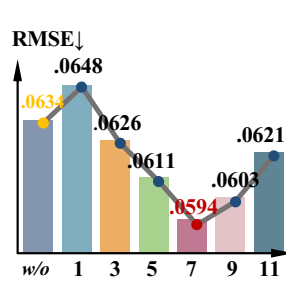


Figure 3: Ablation results of k in Top- k operator on the Synthetic-I dataset.

Phase 2		Phase 3		RMSE	PSNR	SSIM	LPIPS
DMF	MGR	DMF	MGR				
✓	✓	✗	✗	0.0747	23.04	0.8195	0.2479
✓	✓	✗	✓	0.0638	23.38	0.9389	0.0876
✓	✓	✓	✗	0.0631	24.58	0.9437	0.0691
✗	✗	✓	✓	0.0708	23.40	0.8756	0.1170
✗	✓	✓	✓	0.0622	24.72	0.9450	0.0685
✓	✗	✓	✓	<u>0.0603</u>	<u>25.34</u>	<u>0.9501</u>	<u>0.0583</u>
✓	✓	✓	✓	0.0594	25.49	0.9536	0.0535

Table 3: Ablation studies of the filter $\mathcal{F}(\cdot)$ in Phase 2 and Phase 3 on the Synthetic-I dataset.

RMSE from 0.0594 to 0.0747 in Phase 3). Moreover, we deeply discuss the two core components of $\mathcal{F}(\cdot)$, DMF and MGR. A more comprehensive analysis highlights the significant roles played by the DMF and MGR modules in denoising and artifacts-freeing (e.g., DMF improves PSNR from 24.72 to 25.49 in Phase 2, and MGR improves SSIM from 0.9437 to 0.9536 in Phases 3), thus validating the effectiveness of entire dynamic filter $\mathcal{F}(\cdot)$ in this task.

Impact of Top- k in Filter $\mathcal{F}(\cdot)$. To investigate the impact of the Top- k operator in Filter $\mathcal{F}(\cdot)$, we test $k \in \{1, 3, 5, 7, 9, 11\}$. Here, $k \in [0, \hat{C}]$ and $\hat{C} = 12$ in our experiment setup. From Fig. 3, while $k = 1$, it leads to significant sparsity of

\mathcal{L}_{mag}	\mathcal{L}_{dr}	\mathcal{L}_{edge}	\mathcal{L}_{Sobel}	RMSE	PSNR	SSIM	LPIPS
✓	✗	✗	✗	0.0678	22.05	0.9317	0.1121
✓	✓	✗	✗	0.0613	24.91	0.9405	0.0783
✓	✓	✗	✓	<u>0.0606</u>	<u>25.06</u>	<u>0.9487</u>	<u>0.0687</u>
✓	✓	✓	✗	0.0594	25.49	0.9536	0.0535

Table 4: Ablation studies of losses on the Synthetic-I dataset.

similarity-based attention matrix, resulting in a large error, i.e., RMSE of 0.0648. While $k = 11$, there is a large error too, i.e., RMSE of 0.0621. Hence, an appropriate value of k contributes to the balance of the attention sparsity calculation and magnification denoising. As a result, we set $k = 7$ with the lowest RMSE of 0.0594 as the optimal setting.

Effect of Loss Function. Tab. 4 reports the ablation studies of different loss functions. Based on the basis \mathcal{L}_{mag} , the introduction of disentangled representation loss \mathcal{L}_{dr} significantly improves the robustness of magnification, i.e., PSNR is improved from 22.05 dB to 24.91 dB. Moreover, applying the \mathcal{L}_{edge} yields gains of 0.58 dB and 0.0131 for PSNR and SSIM, respectively. Comparing it with the well-known Sobel loss \mathcal{L}_{Sobel} (Zheng et al. 2020), which focuses solely on horizontal and vertical edges, \mathcal{L}_{edge} incorporates the LoG operator for noise smoothing and edge detection, demonstrates better noise robustness, edge continuity, and effective extraction of low-contrast magnified motion boundaries.

Qualitative Analysis

(1) **Magnification visualization comparisons.** In Fig. 4, Linear (Wu et al. 2012) and Phase (Wadhwa et al. 2013) exhibit significant distortion and ringing artifacts; Acceleration, Jerk-aware and Anisotropy methods (Zhang, Pinteá, and Van Gemert 2017; Takeda et al. 2018, 2019) show insufficient amplification amplitude, and the other learning-based methods (Oh et al. 2018; Singh, Murala, and Kosuru 2023a,b) show flickering artifacts and motion distortion originating from their spatial inconsistency. In contrast, we achieve more robust results, noticeably improving artifacts and distortions while achieving satisfactory magnification amplitude. (2) **Magnification factor α .** Fig. 5 shows the magnified results of shift-up and shift-down setups of the

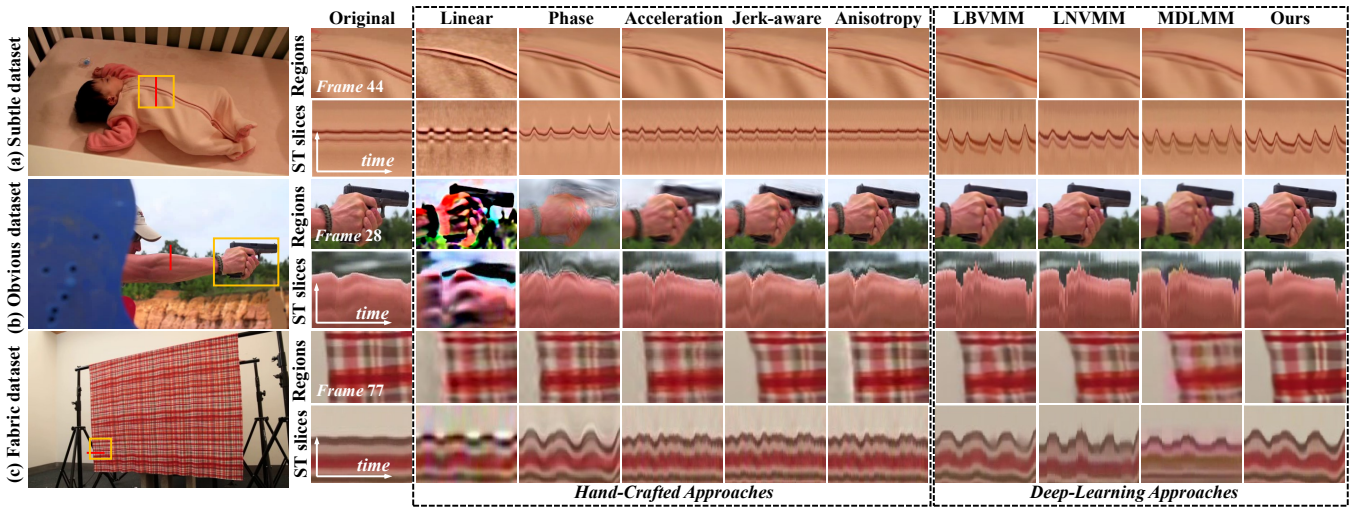


Figure 4: Qualitative results of our method with existing methods on (a) Static, (b) Dynamic and (c) Fabric datasets with magnification factors α of 20, 10, and 20, respectively. We highlight spatial regions where motion occurs and provide spatiotemporal (ST) slices of magnified motion for better comparison.

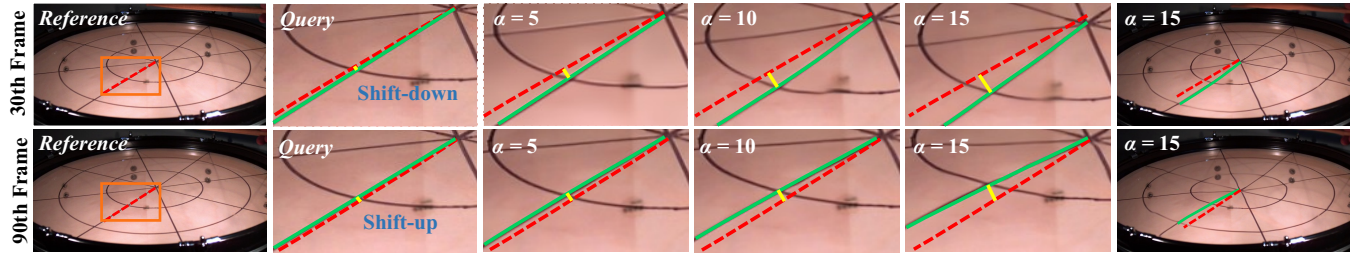


Figure 5: Magnification visualization of the *drum* video from the Static dataset. We randomly sample two frames with shift-down and shift-up motion. EulerMormer achieves reliable video motion magnification under different magnification factors α .

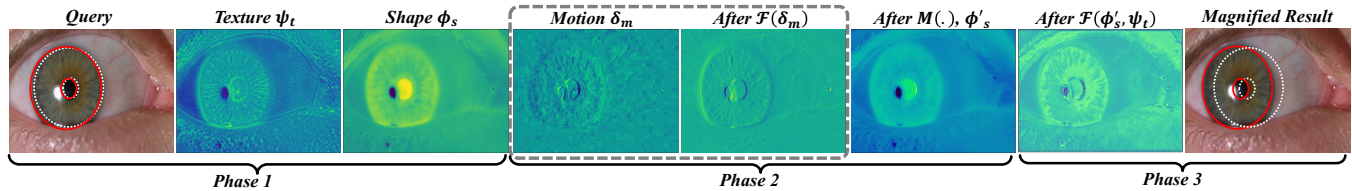


Figure 6: Dataflow of our method pipeline with the *eye* video from the Dynamic dataset. The disentangled texture, shape, and motion feature maps have distinguishable vision characteristics. The dynamic filter $\mathcal{F}(\cdot)$ effectively erases the noises in the static-dynamic field of the image and refines the texture-shape joint refinement process.

drum surface. We achieve reliable magnification at different α levels. (3) **Magnification dataflow.** Fig. 6 displays the dataflow of EulerMormer. The disentangled texture, shape, and motion feature maps have distinguishable vision characteristics. Please pay attention to the dynamic filter $\mathcal{F}(\cdot)$ in Phase 2, which effectively eliminates noise in the static field of the motion while preserving important motion information in the dynamic field. On this basis, the visualization of $\mathcal{F}(\phi'_s(x, t), \psi_t(x, t))$ also validates the ability of $\mathcal{F}(\cdot)$ to process the texture-magnified shape joint refinement.

Conclusion

In this paper, we have introduced EulerMormer, a novel Transformer-based end-to-end framework designed for VMM tasks from the Eulerian perspective, aiming to provide more robust magnification effects. The core of EulerMormer lies in embedding a dedicated dynamic filter within Transformer, enabling static-dynamic field adaptive denoising for motion and recoupling refinement. Extensive quantitative and qualitative experiments demonstrate that EulerMormer outperforms SOTA approaches.

Acknowledgments

This work was supported by the National Key R&D Program of China (2022YFB4500600), the National Natural Science Foundation of China (62272144, 72188101, 62020106007, and U20A20183), and the Major Project of Anhui Province (202203a05020011).

References

- Abnoui, F.; Kang, G.; Giacomini, J.; Yeung, A.; Zarafshar, S.; Vesom, N.; Ashley, E.; Harrington, R.; and Yong, C. 2019. A novel noninvasive method for remote heart failure monitoring: the Eulerian video Magnification apPLications In heart Failure studY (AMPLIFY). *NPJ Digital Medicine*, 2(1): 80.
- Brattoli, B.; Büchler, U.; Dorckenwald, M.; Reiser, P.; Filli, L.; Helmchen, F.; Wahl, A.-S.; and Ommer, B. 2021. Unsupervised behaviour analysis and magnification (uBAM) using deep learning. *Nature Machine Intelligence*, 3(6): 495–506.
- Bruhn, A.; Weickert, J.; and Schnörr, C. 2005. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61: 211–231.
- Davis, A.; Bouman, K. L.; Chen, J. G.; Rubinstein, M.; Buyukozturk, O.; Durand, F.; and Freeman, W. T. 2017. Visual Vibrometry: Estimating Material Properties from Small Motions in Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4).
- Davis, A.; Bouman, K. L.; Chen, J. G.; Rubinstein, M.; Durand, F.; and Freeman, W. T. 2015. Visual vibrometry: Estimating material properties from small motion in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5335–5343.
- Eitner, M.; Miller, B.; Sirohi, J.; and Tinney, C. 2021. Effect of broad-band phase-based motion magnification on modal parameter estimation. *Mechanical Systems and Signal Processing*, 146: 106995.
- Ercan, B.; Eker, O.; Erdem, A.; and Erdem, E. 2023. EVREAL: Towards a Comprehensive Benchmark and Analysis Suite for Event-based Video Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3942–3951.
- Guo, D.; Li, K.; Zha, Z.-J.; and Wang, M. 2019. Dadnet: Dilated-attention-deformable convnet for crowd counting. In *Proceedings of the 27th ACM international conference on multimedia*, 1823–1832.
- Huang, D.; Bi, Y.; Navab, N.; and Jiang, Z. 2023. Motion Magnification in Robotic Sonography: Enabling Pulsation-Aware Artery Segmentation. *arXiv preprint arXiv:2307.03698*.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.
- Le Ngo, A. C.; and Phan, R. C.-W. 2019. Seeing the invisible: Survey of video motion magnification and small motion analysis. *ACM Computing Surveys*, 52(6): 1–20.
- Li, K.; Guo, D.; and Wang, M. 2021. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1902–1910.
- Li, K.; Guo, D.; and Wang, M. 2023. ViGT: proposal-free video grounding with a learnable token in the transformer. *Science China Information Sciences*, 66(10): 202102.
- Li, K.; Li, J.; Guo, D.; Yang, X.; and Wang, M. 2023. Transformer-based Visual Grounding with Cross-modality Interaction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19: 1–19.
- Liu, C.; Torralba, A.; Freeman, W. T.; Durand, F.; and Adelson, E. H. 2005. Motion magnification. *ACM Transactions on Graphics*, 24(3): 519–526.
- Mehra, A.; Agarwal, A.; Vatsa, M.; and Singh, R. 2022. Motion Magnified 3-D Residual-in-Dense Network for Deep-Fake Detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(1): 39–52.
- Nguyen, X.-B.; Duong, C. N.; Li, X.; Gauch, S.; Seo, H.-S.; and Luu, K. 2023. Micron-BERT: BERT-Based Facial Micro-Expression Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1482–1492.
- Oh, T.-H.; Jaroensri, R.; Kim, C.; Elgharib, M.; Durand, F.; Freeman, W. T.; and Matusik, W. 2018. Learning-based video motion magnification. In *Proceedings of the European Conference on Computer Vision*, 633–648.
- Qi, H.; Guo, Q.; Juefei-Xu, F.; Xie, X.; Ma, L.; Feng, W.; Liu, Y.; and Zhao, J. 2020. Deephythm: Exposing deep-fakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4318–4327.
- Qian, W.; Guo, D.; Li, K.; Tian, X.; and Wang, M. 2023. Dual-path tokenlearner for remote photoplethysmography-based physiological measurement with facial videos. *arXiv preprint arXiv:2308.07771*.
- Qin, X.; Dai, H.; Hu, X.; Fan, D.-P.; Shao, L.; and Van Gool, L. 2022. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, 38–56.
- Rubinstein, M.; Wadhwa, N.; Durand, F.; Freeman, W. T.; and Wu, H.-Y. 2013. Revealing invisible changes in the world. *Science*, 339(6119): 519–519.
- Shen, H.; Zhao, Z.-Q.; and Zhang, W. 2023. Adaptive dynamic filtering network for image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2227–2235.
- Shen, H.; Zhao, Z.-Q.; Zhang, Y.; and Zhang, Z. 2023. Mutual Information-driven Triple Interaction Network for Efficient Image Dehazing. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7–16.
- Shi, W.; Caballero, J.; Huszar, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1874–1883.

- Singh, J.; Murala, S.; and Kosuru, G. 2023a. Lightweight Network for Video Motion Magnification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2041–2050.
- Singh, J.; Murala, S.; and Kosuru, G. 2023b. Multi Domain Learning for Motion Magnification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13914–13923.
- Takeda, S.; Akagi, Y.; Okami, K.; Isogai, M.; and Kimata, H. 2019. Video magnification in the wild using fractional anisotropy in temporal distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1614–1622.
- Takeda, S.; Niwa, K.; Isogawa, M.; Shimizu, S.; Okami, K.; and Aono, Y. 2022. Bilateral Video Magnification Filter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17369–17378.
- Takeda, S.; Okami, K.; Mikami, D.; Isogai, M.; and Kimata, H. 2018. Jerk-aware video acceleration magnification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1769–1777.
- Tang, S.; Hong, R.; Guo, D.; and Wang, M. 2022. Gloss semantic-enhanced network with online back-translation for sign language production. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5630–5638.
- Wadhwa, N.; Rubinstein, M.; Durand, F.; and Freeman, W. T. 2013. Phase-based video motion processing. *ACM Transactions on Graphics*, 32(4): 1–10.
- Wang, P.; Wang, X.; Wang, F.; Lin, M.; Chang, S.; Li, H.; and Jin, R. 2022a. Kvt: k-nn attention for boosting vision transformers. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, 285–302.
- Wang, Y.; Cai, J.; Liu, Y.; Chen, X.; and Wang, Y. 2022b. Motion-induced error reduction for phase-shifting profilometry with phase probability equalization. *Optics and Lasers in Engineering*, 156: 107088.
- Wang, Y.; Xu, H.; Zhu, H.; Rao, Y.; and Wang, Y. 2024. Nonlinear high-order harmonics correction for phase measuring profilometry. *Optics & Laser Technology*, 170: 110248.
- Wu, H.-Y.; Rubinstein, M.; Shih, E.; Guttag, J.; Durand, F.; and Freeman, W. 2012. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics*, 31(4): 1–8.
- Wu, Y.; Wang, X.; Li, G.; and Shan, Y. 2022. AnimeSR: learning real-world super-resolution models for animation videos. *Advances in Neural Information Processing Systems*, 35: 11241–11252.
- Xia, Z.; Peng, W.; Khor, H.-Q.; Feng, X.; and Zhao, G. 2020. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE Transactions on Image Processing*, 29: 8590–8605.
- Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1191–1200.
- Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5728–5739.
- Zhang, D.; Zhu, A.; Gong, X.; Wang, Y.; Guo, J.; and Zhang, X. 2023. Hybrid Eulerian–Lagrangian framework for structural full-field vibration quantification and modal shape visualization. *Measurement*, 219: 113270.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.
- Zhang, Y.; Han, X.; Zhang, H.; and Zhao, L. 2017. Edge detection algorithm of image fusion based on improved Sobel operator. In *2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference*, 457–461.
- Zhang, Y.; Pinteá, S. L.; and Van Gemert, J. C. 2017. Video acceleration magnification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 529–537.
- Zhao, G.; Lin, J.; Zhang, Z.; Ren, X.; Su, Q.; and Sun, X. 2019. Explicit sparse transformer: Concentrated attention through explicit selection. *arXiv preprint arXiv:1912.11637*.
- Zheng, B.; Yuan, S.; Slabaugh, G.; and Leonardis, A. 2020. Image demoiréing with learnable bandpass filters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3636–3645.
- Zheng, P.; Fu, H.; Fan, D.-P.; Fan, Q.; Qin, J.; Tai, Y.-W.; Tang, C.-K.; and Van Gool, L. 2023. GCoNet+: A Stronger Group Collaborative Co-Salient Object Detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhou, J.; Guo, D.; and Wang, M. 2022. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2022. Audio-visual segmentation. In *European Conference on Computer Vision*, 386–403.
- Zhou, J.; Zheng, L.; Zhong, Y.; Hao, S.; and Wang, M. 2021. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8436–8444.