Frequency Decoupling for Motion Magnification via Multi-Level Isomorphic Architecture

Fei Wang[1], Dan Guo[1,2]*, Kun Li[1], Zhun Zhong[1,3], Meng Wang[1,2]*

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

[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China

[3] School of Computer Science, University of Nottingham, NG8 1BB Nottingham, UK

Abstract

Video Motion Magnification (VMM) aims to reveal subtle and imperceptible motion information of objects in the macroscopic world. Prior methods directly model the motion field from the Eulerian perspective by Representation Learning that separates shape and texture or Multi-domain Learning from phase fluctuations. Inspired by the frequency spectrum, we observe that the low-frequency components with stable energy always possess spatial structure and less noise, making them suitable for modeling the subtle motion field. To this end, we present FD4MM, a new paradigm of Frequency Decoupling for Motion Magnification with a Multi-level Isomorphic Architecture to capture multi-level high-frequency details and a stable low-frequency structure (motion field) in video space. Since high-frequency details and subtle motions are susceptible to information degradation due to their inherent subtlety and unavoidable external interference from noise, we carefully design Sparse High/Low-pass Filters to enhance the integrity of details and motion structures, and a Sparse Frequency Mixer to promote seamless recoupling. Besides, we innovatively design a contrastive regularization for this task to strengthen the model's ability to discriminate irrelevant features, reducing undesired motion magnification. Extensive experiments on both Real-world and Synthetic Datasets show that our FD4MM outperforms SOTA methods. Meanwhile, FD4MM reduces FLOPs by 1.63× and boosts inference speed by 1.68× than the latest method. Our code is available at https://github.com/Jiafei127/FD4MM.

1. Introduction

Human eyes have a limited resolution range to perceive the subtle motion in the macroscopic world [22, 37]. **V**ideo **M**otion **M**agnification (VMM), as a "motion microscope",
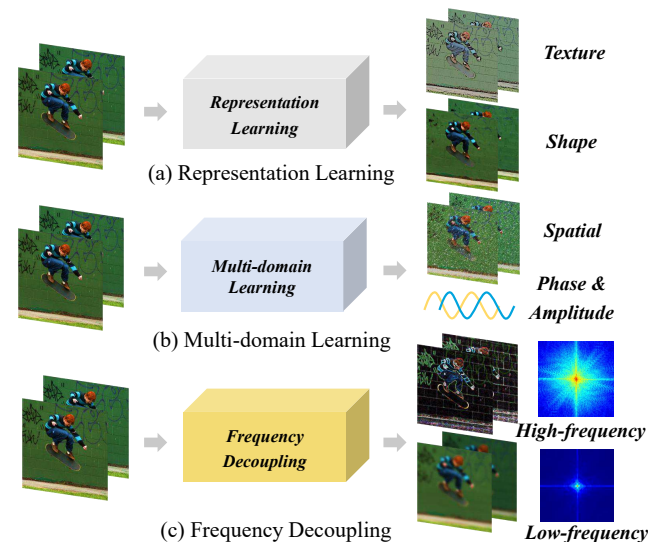

Figure 1. **Learning-based methods for motion magnification.** (a) Representation Learning methods [12, 34, 43], (b) Multi-domain Learning method [44] and (c) our Frequency Decoupling method. Inspired by the theory of frequency spectrum [41, 64], we utilise it to separate high- and low-frequency features and leverage their discriminative characteristics for motion magnification.

vividly reveals subtle variations in the macroscopic world and uncovers important invisible information [29, 52, 61]. It plays a crucial role in various downstream applications, such as micro-expression recognition [32, 62], robotic sonography [1, 18] and material property estimation [9, 10, 13, 66], *etc.* VMM is a complex task that involves generating pixel-level motion in videos. Especially when the motion occurs subtle, it is susceptible to confusion with inevitable photography noise [3, 34, 61, 71], which comprises photon noise due to the quantum properties and uncertainty of light as well as thermal noise inherent in charge-coupled devices (CCDs) of the acquisition device, resulting in amplified noise, undesired motion and distortion.

Early research drew inspiration in fluid mechanics [61], utilizing spatial decomposition [51, 61] and hand-crafted

---

*Corresponding authors.

filters [45–47, 69] from the Eulerian perspective to mitigate the negative influence of noise on motion magnification. However, these methods are prone to ringing artifacts and require the selection of different hyperparameters for optimal magnification results in specific scenarios, limiting the applicability to downstream tasks. Recent studies [4, 12, 34, 43, 44] have turned their attention to learning-based methods due to the strong representation learning capabilities, exhibiting fewer ringing artifacts and better noise characterization. (1) **Representation learning methods** [4, 12, 34] (in Fig. 1(a)). These methods rely on the representation consistency principle, forcing the encoder to disentangle texture-shape representations using data-induced color translations. However, incomplete disentanglement (*e.g.*, leaving partial texture clues in the shape representation) may inadvertently introduce undesired motions [12, 43], and the data-induced training is redundant, increasing the training difficulty and cost. (2) **Multi-domain learning method** [44] (in Fig. 1(b)). Singh *et al.* [44] consider the phase difference between two frames (called phase fluctuation) to represent the motion field. However, phase acquisition requires complex predefined traditional algorithms and overhead a more computational cost, *e.g.*, FLOPs and inference time in Table 2.

Admittedly, spatial decomposition remains a major challenge for efficiently capturing subtle motion. Although the above spatial decomposition methods provide effective solutions for motion magnification, we find that frequency decoupling based on the Fourier spectrum can serve as a new perspective for distinguishing different spatial features, as shown in Fig. 2. The high-frequency features reveal more energetic spatial details, while the low-frequency features exhibit a centrally clustered spectral distribution with more stable spatial structures. This paper applies this theory to the VMM task to achieve effective spatial decomposition for motion magnification by exploring diverse spectral energy distributions. We choose the stable low-frequency features to model the motion field, and high-frequency features with more energy to preserve spatially detailed features, such as appearance clues.

In this paper, we present FD4MM, a new paradigm of **F**requency **D**ecoupling for **M**otion **M**agnification to capture spatial high-frequency details while ensuring stable low-frequency structures for constructing the motion field. Specifically, based on the principle of spectral entropy distribution [7, 64], we design an adaptive frequency decoupling encoder to separate spatial high- and low-frequency features. Then, we propose a Multi-level Isomorphic Architecture that progressively separates high-frequency components within low-frequency features to provide a stable structure with less energy entropy for modeling motion. Given the susceptibility of high-frequency details and subtle motions to information degradation from inherent
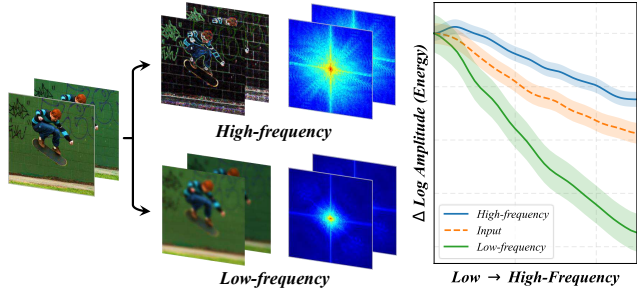

Figure 2. **Our idea of frequency decoupling for spatial decomposition.** High-frequency features reveal spatial details, enabling an expanded bright field in the spectrum, implying more energy. Low-frequency energy clusters in the central region represent stable spatial structures appropriate for modeling motion.

subtlety and unavoidable external interference from photographic noise, we should retain their maximum information integrity as noise-free before magnification.

To achieve this, we propose Sparse High-pass and Low-pass Filters within the advanced Transformer to act on the high-frequency details and the inter-frame low-frequency motion field, respectively. They are equipped with high-pass and low-pass operators and employ a sparse strategy to mitigate degradation caused by noise, enabling the model to focus on more accurate details and motion structures. Afterwards, the filtered motion field is amplified by a Pointwise Nonlinear Magnifier. Finally, we integrate the decoupled features through a Sparse Frequency Mixer, which promotes seamless recoupling of high-frequency details and magnified low-frequency features. Additionally, we develop a novel contrastive regularization loss to further reduce undesired motion magnification while enhancing the robustness of the model.

Overall, our main contributions are as follows:

- We introduce a new paradigm for VMM called FD4MM, which aims to decouple the high- and low-frequency features for motion magnification via a multi-level isomorphic frequency decoupling architecture.
- We propose Sparse High-pass and Low-pass Filters based on the Transformer framework to mitigate the degradation of details and structures caused by noise. Also, a Sparse Frequency Mixer is developed for seamless recoupling.
- We design a novel contrastive regularization to strengthen the model's ability to discriminate irrelevant features, thereby reducing undesired motion magnification.
- Extensive qualitative and quantitative experiments show that FD4MM performs favorably against SOTA methods with fewer FLOPs and faster inference speed.

2. Related Works

Hand-crafted Magnification Filters. Early methods primarily focused on Eulerian perspective [45–47, 51, 61, 69], aiming to capture the variation occurring within a fixed re-

gion without tracking each pixel's motion trajectory. Based on motion or phase fluctuations captured by Laplacian pyramid [61] or steerable pyramid [51] operators, traditional Eulerian methods gradually introduced hand-crafted filters suitable for various scenarios, such as acceleration [69], jerk [45], anisotropy [46], and bilateral filters [47]. They amplify the interested motion relying on prior motion variation but lack consideration to suppress inappropriate amplification amplitude, unavoidable occlusion, and unexpected ringing artifacts [44]. Furthermore, in these methods, many hyperparameters must be re-calibrated for different motion scenes to be applicable.

Learning-based Magnification. Recent research interests have shifted towards learning-based approaches to provide more scene-generalizable magnification [4, 12, 34, 43, 44]. Oh *et al.* [34] pioneered the disentangled texture and shape representation learning for motion magnification, achieving comparable results to hand-crafted filters in static and dynamic scenes. Despite their success, inducing representation separation through color transformation during training in these works may result in incomplete disentanglement [12] (*i.e.*, leaving partial texture clues in the shape representation), thus leading to undesired motion and flickering artifacts. Singh *et al.* [43] proposed a lightweight proxy model to mitigate issues before magnification. They [44] extended spatial decomposition to multi-domain learning, modeling motion by phase fluctuations in the frequency domain and denoising in the spatial domain. However, the complex phase mapping differences between domains often lead to information degradation and unnatural artifacts. In contrast, our method utilizes a multi-level isomorphic architecture to achieve adaptive frequency decoupling, capturing a more stable motion field for magnification.

3. Methodology

3.1. Preliminaries

Definition. Motion Magnification focuses on spatial intensity variations by establishing the motion field $\delta(x, t)$ between spatial coordinates x *vs.* pixel intensities of the 1D signal along the temporal sequence t to obtain the amplified signal $I_m(x, t)$ with a magnification factor α, which can extend to 2D space [61]. Therefore, an effective spatial decomposition strategy is crucial for a stable motion field.

Overall Pipeline. We aim to make frequency decoupling a new paradigm for learnable spatial decomposition in motion magnification. As shown in Fig. 3, our overall pipeline is designed as a Multi-level Isomorphic Architecture to capture multi-scale high-frequency details $\sum_{i \in \{s, m, d\}} H_i(x, t)$ (*s*hallow, *m*iddle and *d*eep bands) and a stable low-frequency motion field $\delta(x, t) = \Delta L_d(x, t)$ (*d*eep) of the query frame $I(x, t)$ for magnification. Then, as shown in Fig. 4, Sparse High-pass and Low-pass Fil-
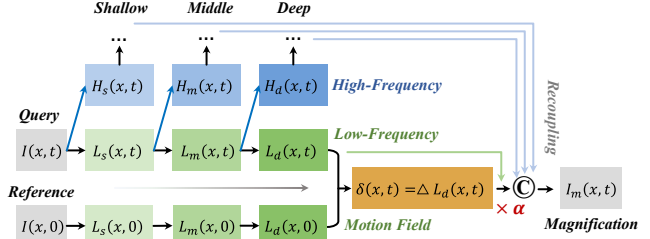

Figure 3. **Pipeline of the Multi-level Isomorphic Architecture based on Frequency Decoupling.** It aims to decouple a stable motion field and multi-level high-frequency details for magnification and recoupling with a magnification factor α, respectively.

ters are developed to minimize the degradation of high-frequency details and low-frequency structure caused by noise. Next, the filtered motion field is manipulated in a Point-wise Nonlinear Magnifier to achieve magnification. Finally, a Sparse Frequency Mixer is proposed to capture the magnified low-frequency to guide the seamless recoupling with high-frequency details, reducing the difficult-to-eliminate ringing artifacts. Overall, the magnified frame $I_m(x, t)$ can be represented as:

$$I_m(x, t) \approx \underbrace{\sum_{i \in \{s, m, d\}} H_i(x, t)}_{High-Freq.} + \alpha \underbrace{\Delta L_d(x, t)}_{Low-Freq.} \frac{\partial f(x)}{\partial x}. \quad (1)$$

3.2. Adaptive Frequency Decoupling Encoder

High-frequency signals typically reflect the detailed features of the image, while low-frequency signals reflect the global structure [7, 15, 28]. Here, we propose an Adaptive Frequency Decoupling Encoder (AFDE) to achieve spatial frequency decomposition. Specifically, given a pair of reference and query frames $\{I(x, 0), I(x, t)\} \in \mathbb{R}^{H \times W \times 3}$ in the video, we apply a 3×3 convolution layer to obtain the initial feature $\{F(x, 0), F(x, t)\} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$, respectively. Taking $F(x, t)$ as an example, a dilated convolution $\mathcal{W}_r$ with a dilation rate of $r = 2$ is applied to capture the low-frequency component $L(x, t) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$. The dilated receptive field can smooth the image details to encode approximate low-frequency energy. Instead, the high-frequency detail $H(x, t) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$ can be obtained by removing $L(x, t)$ directly from the $F(x, t)$, similar to the way of detail coefficients in the wavelet transform [2], as:

$$\begin{cases} L(x, t) = \vartheta \cdot \mathcal{W}_r F(x, t), \\ H(x, t) = \vartheta \cdot (F(x, t) - \mathcal{W}_r F(x, t)), \end{cases} \quad (2)$$

where ϑ denotes the nonlinear activation function GELU. To further obtain more stable low-frequency, we design a Multi-level Isomorphic Architecture to recursively separate frequency at different scales. The separation of the frequency bands consists of shallow (*s*), medium (*m*) and deep (*d*) levels, each of which depends on the isomorphic AFDE module and downsampling. So far, we obtain the
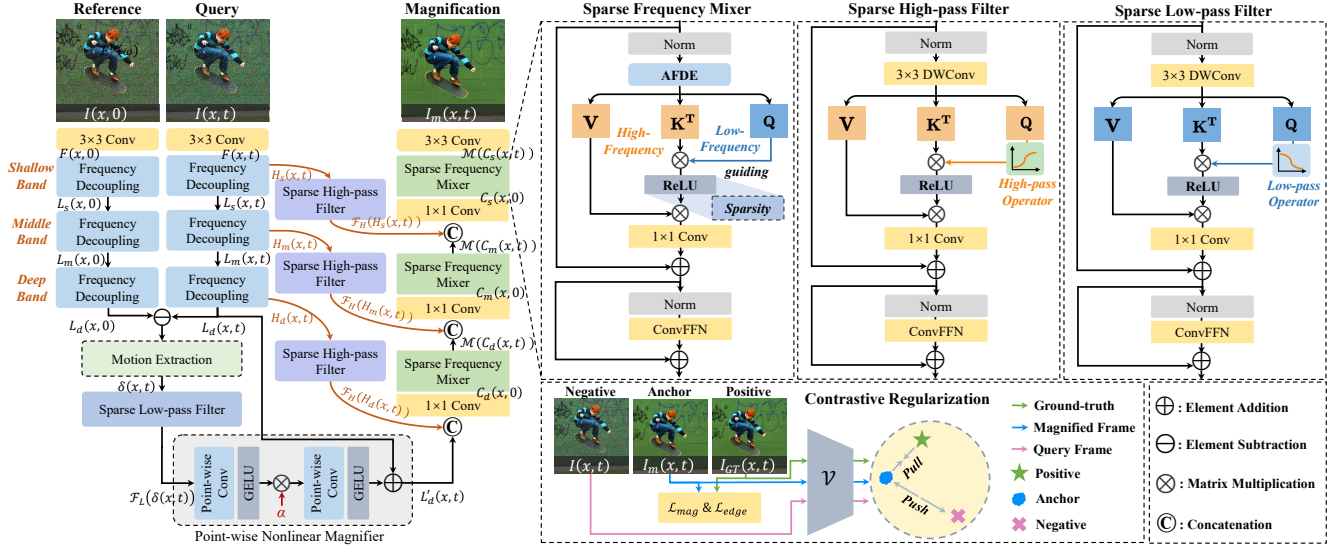
Figure 4. **Overall pipeline of the proposed FD4MM.** Based on Multi-level Isomorphic FD4MM Architecture, the multi-level high-frequency details $\{H_s(x,t), H_m(x,t), H_d(x,t)\}$ and the stable motion field ($\delta(x,t) = \Delta L_d(x,t)$), obtained from the reference $I(x,0)$ and query frames $I(x,t)$, are filtered by Sparse High- and Low-pass Filters ($\mathcal{F}_H(\cdot)$, $\mathcal{F}_L(\cdot)$) to minimize the degradation of high-frequency details and low-frequency motion structures caused by noise, respectively. Then, the motion field $\mathcal{F}_L(\delta(x,t))$ is amplified by the Point-wise Nonlinear Magnifier. Next, the Sparse Frequency Mixer $\mathcal{M}(\cdot)$ allows the magnified low-frequency to guide the high-frequency details to complete the seamless recoupling to avoid ringing artifacts, ending up a magnified frame $I_m(x,t)$. Besides, we introduce a novel Contrastive Regularization to eliminate undesired magnification results, thus enhancing the model's robustness and magnification effects.

multi-level high-frequency details $H_s(x,t) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$, $H_m(x,t) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 2C}$, $H_d(x,t) \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 4C}$, and motion field $\delta(x,t) = \Delta L_d(x,t) \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 4C}$.

3.3. Sparse High-pass and Low-pass Filters

Obtaining a stable and as noise-free as possible motion field is crucial to accurate magnification. Therefore, we revisit the advanced attention mechanism [65] and introduce Sparse High- and Low-pass Filters ($\mathcal{F}_H(\cdot)$ and $\mathcal{F}_L(\cdot)$) via the Transformer infrastructure, acting on the high-frequency details and the motion field in Fig. 4, respectively.

Considering that subtle motion is susceptible to structural degradation caused by noise and frequency attenuation, $\mathcal{F}_L(\cdot)$ emphasizes to address these issues before the magnification operation. Given the input $\delta(x,t)$, it first undergoes a 1×1 convolution and a 3×3 depth-wise convolution to encode channel-wise context. Next, to preserve overall motion structure, we capture the low-frequency of *query* using a learnable low-pass operator (AvgPool [7, 35, 41]) before reshaping to obtain each single-headed h projections $\{\mathbf{Q}_L, \mathbf{K}, \mathbf{V}\} \in \mathbb{R}^{(\frac{H}{8} \times \frac{W}{8}) \times C'}$, where $C' = \frac{4C}{h}$. Therefore, the standard attention matrix $\mathbf{A}_L^{(h)} \in \mathbb{R}^{C' \times C'}$ as:

$$\mathbf{A}_L^{(h)} = \text{Softmax}(\frac{\mathbf{Q}_L \mathbf{K}^T}{\tau}), \qquad (3)$$

where $\mathbf{K}^T$ is the transpose of $\mathbf{K}$ and τ denotes a learnable temperature parameter defined by $\tau = \sqrt{C'}$.

Previous work [6, 25, 26] pointed out that the atten-

tion $\mathbf{A}_L^{(h)}$ using all tokens is redundant, as it may involve noise interactions between unrelated features, which hinders noise suppression and results in magnified distortion and artifacts. Recent studies [26, 27, 39, 59] have found that ReLU, an activation function with gating properties, can effectively aggregate positive knowledge while removing negative features without additional operators such as dropout [24, 30] and Top-k [53]. Thus, our work replaces Softmax with ReLU as an alternative way to compute sparse attention matrix $\mathbf{SA}_L^{(h)} \in \mathbb{R}^{C' \times C'}$, and Eq. 3 is rewritten as:

$$\mathbf{SA}_L^{(h)} = \text{ReLU}(\frac{\mathbf{Q}_L \mathbf{K}^T}{\tau}). \qquad (4)$$

Thus, the output is obtained by weighted aggregation of the $\mathbf{SA}_L^{(h)}$ with $\mathbf{V}$ and concatenating the residual across all heads. Besides, we employ a simple Convolutional Feed-Forwards Network (ConvFFN) [11, 73] for a more flexible information update to obtain the noise-free and structure-complete motion field $\mathcal{F}_L(\delta(x,t)) \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 4C}$.

As for $\mathcal{F}_H(\cdot)$, we employ an overall architecture similar to $\mathcal{F}_L(\cdot)$ to enhance the desired high-frequency details and suppress noise at multi-level bands. A critical aspect lies in performing sparse attention computation using the projection *query* that undergoes a high-pass operator (Max-Pool [41]). Here, the results of the $\mathcal{F}_H(\cdot)$ for multi-level high-frequency details are $\mathcal{F}_H(H_d(x,t)) \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 4C}$, $\mathcal{F}_H(H_m(x,t)) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 2C}$, $\mathcal{F}_H(H_s(x,t)) \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$.

3.4. Point-wise Nonlinear Magnifier

This section is crucial for manipulating the motion field $\mathcal{F}_L(\delta(x,t))$ magnification in the Eulerian perspective. Unlike previous approaches [34, 43], our Point-wise Nonlinear Magnifier expects to preserve its nonlinear spatial intensity transformation during amplification while avoiding excessive checkerboard artifacts and spatial frequency collapse [33, 51]. Therefore, it abandons the local operation of a large convolution kernel in favor of a simple point-wise convolution and a more stable nonlinear activation function, GELU, to achieve the global feature intensity interactions.

$$L'_d(x,t) = L_d(x,t) + \mathcal{W}_p(\alpha \cdot \mathcal{W}_p(\mathcal{F}_L(\delta(x,t)))), \quad (5)$$

where $\mathcal{W}_p(\cdot)$ denotes a 1×1 point-wise convolution with GELU activation and $L'_d(x,t) \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 4C}$ is the magnified low-frequency component with magnification factor α.

3.5. Sparse Frequency Mixer

A challenge in VMM is the "expanded" or "retracted" of high-frequency details when superimposed on the magnified low-frequency structure [34, 61]. However, achieving a perfect high- and low-frequency alignment is often unattainable, resulting in ringing artifacts at high-frequency boundaries. To address this, we develop a Sparse Frequency Mixer $\mathcal{M}(\cdot)$ to progressively promote seamless recoupling of $\{H_{s,m,d}(x,t)\}$ and $L'_d(x,t)$ on a level-by-level basis to avoid ringing artifacts. For the deep band, we first concatenate $\{L'_d(x,t), \mathcal{F}_H(H_d(x,t))\}$ and compress their feature channels through 1×1 convolution to obtain the coupled feature $\mathcal{R}_d(x,t) \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 4C}$. Then, $\mathcal{M}(\cdot)$ integrates the frequency decoupling capabilities of the AFDE into sparse attention to further encoding the contextual high- and low-frequency vectors of the normalized $\mathcal{R}_d(x,t)$. For the deep band implementation, we utilize the low-frequency as $\mathbf{Q}_d \in \mathbb{R}^{(\frac{H}{8} \times \frac{W}{8}) \times C'}$ to guide sparse attention computation with the high-frequency $\{\mathbf{K}_d, \mathbf{V}_d\} \in \mathbb{R}^{(\frac{H}{8} \times \frac{W}{8}) \times C'}$ by:

$$\mathbf{Q}_d, \{\mathbf{K}_d, \mathbf{V}_d\} = \mathcal{E}(\mathcal{R}_d(x,t)),$$
$$\mathbf{SA}_M^{(h)} = \text{ReLU}(\frac{\mathbf{Q}_d \mathbf{K}_d^T}{\tau}), \quad (6)$$

where $\mathcal{E}(\cdot)$ is the AFDE module and $\mathbf{SA}_M^{(h)} \in \mathbb{R}^{C' \times C'}$ is the sparse attention matrix of $\mathcal{M}(\cdot)$ with ReLU. Similarly, the final output $\mathcal{M}(\mathcal{R}_d(x,t)) \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 4C}$ is obtained by weighted aggregation and concatenation of $\mathbf{SA}_M^{(h)}$ with $\mathbf{V}_d$, and updated by the ConvFFN.

Overall, the recouple process involves all the high-frequency details with deep, middle and shallow bands $\{\mathcal{F}_H(H_d(x,t)), \mathcal{F}_H(H_m(x,t)), \mathcal{F}_H(H_s(x,t))\}$ and the magnified low-frequency $L'_d(x,t)$ progressively recoupling by the above $\mathcal{M}(\cdot)$ at each level. Inter-level PixelShuffle [40, 49] is employed to transform sub-pixel spatial resolution and latent feature dimensions, resulting in the final magnified frame $I_m(x,t) \in \mathbb{R}^{H \times W \times 3}$.

3.6. Loss Function

Contrastive Regularization. Simple consistency and perceptual loss are insensitive to noise and tend to overfit the model at the wrong magnified positions, failing to approximate the correct magnification. Contrastive learning [5, 17, 72] has emerged as an effective paradigm in visual tasks as a self-supervised technique. It aims to learn a representation that pulls "positive" anchors closer together and pushes "negative" anchors away in metric space. Our work involves two considerations: (1) Construction of positive and negative pairs. To obtain accurate and clear magnified frames, positive and negative pairs are respectively constructed from ground-truth $I_{GT}(x,t)$ and query frames $I(x,t)$, while the magnified frames $I_m(x,t)$ serve as anchors. (2) Choice of metric space. We follow [19, 50, 60] to utilize the feature map extracted from the Conv3-2 layer of a pre-trained VGG-19 [42] $\Lambda_v(\cdot)$ as the perceptual metric space and minimizing the metric distance of all samples N in the batch by Charbonnier penalty term as:

$$\mathcal{L}_{CR} = \sum_{n=1}^{N} \sqrt{\frac{\|\Lambda_v(I_m^n(x,t)), \Lambda_v(I_{GT}^n(x,t))\|^2 + \varepsilon^2}{\|\Lambda_v(I_m^n(x,t)), \Lambda_v(I^n(x,t))\|^2 + \varepsilon^2}}, \quad (7)$$

where ε is a constant value, empirically set to 10^{-3}.

Optimization. The final loss function $\mathcal{L}$ of FD4MM is a weighted sum of the following three loss terms:

$$\mathcal{L} = \mathcal{L}_{mag} + \mathcal{L}_{edge} + \lambda \mathcal{L}_{CR}, \quad (8)$$

where $\mathcal{L}_{mag}$ is a basic Charbonnier penalty term for minimizing $\{I_m(x,t), I_{GT}(x,t)\}$ and we set $\lambda = 0.1$ for the contrastive loss $\mathcal{L}_{CR}$. Similar to [44] using edge loss, we find that utilizing the Laplacian of Gaussian (LoG) edge detection operator [68] $E_{LoG}(\cdot)$ with second-order gradients is more sensitive in smoothing the magnified global structure, avoiding motion blur and texture damage induced by amplifying subtle motions. Thus, we integrate it into the edge loss $\mathcal{L}_{edge}$ with Charbonnier penalty, as:

$$\mathcal{L}_{edge} = \sqrt{\|E_{LoG}(I_m(x,t)) - E_{LoG}(I_{GT}(x,t))\|^2 + \varepsilon^2}. \quad (9)$$

4. Experiments

4.1. Experiment Setup

Training Dataset. Following the protocol in [34, 43, 44], all existing learning-based methods are trained on the same training dataset proposed by [34]. In this field, performance evaluation is conducted through cross-dataset testing.

Real-world Test Dataset. We examine all available real-world video datasets released for this task, which includes two inference modes [34] and refers to twelve videos of various motion scenes [14, 34, 44–46, 51, 69] in Table 1. In static mode (inference on initial frame $I(x,0)$ and current frame $I(x,t)$), we test subtle breathing motions of the baby [34, 51, 61], rapid vibrations of the fork [14], *etc*. In dynamic mode (inference on continuous frames, $\{I(x,t-1)$ and $I(x,t)\}$, we test forward jumps of the cattoy [34, 69],
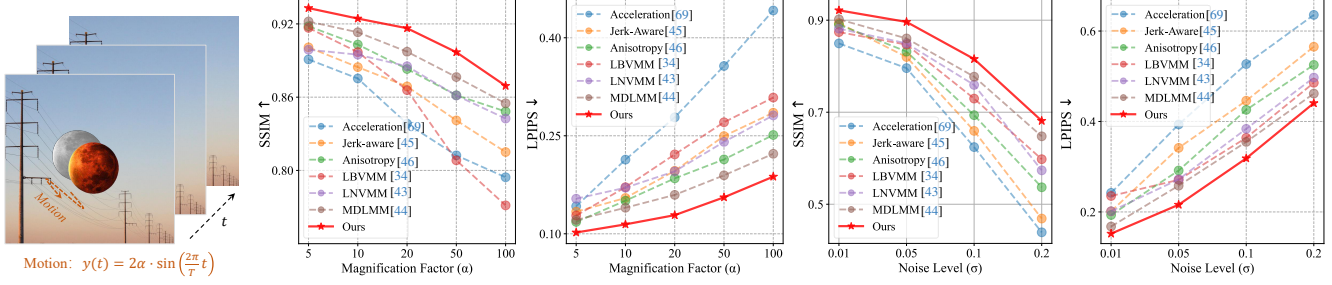
(a) A sample of Synthetic Dataset. (b) Comparison of magnification accuracy. (c) Comparison of noise robustness.

Motion: $y(t) = 2\alpha \cdot \sin\left(\frac{2\pi}{T}t\right)$

Figure 5. **Performance comparison with SOTA methods for SSIM↑ and LPIPS↓ scores on the Synthetic Dataset.** (a) Schematic of a synthetic video. (b) Comparison of magnification accuracy for different α. (c) Analysis of model robustness under different σ.

Method	Venue	Static Mode						Dynamic Mode						Avg.
		Baby	Fork	Drum	Engine	Crane	Face	Gunshot	Cattoy	Eye	Bottle	Drill	Balloon	
Acceleration [69]	CVPR'17	0.7081	0.6786	0.6432	0.6663	0.7338	0.6212	0.6049	0.6342	0.6145	0.5096	0.6592	0.6182	0.6401
Jerk-Aware [45]	CVPR'18	0.7089	0.6826	0.6739	0.6706	0.7410	0.6255	0.6176	0.6415	0.6170	0.5141	0.6768	0.6281	0.6498
Anisotropy [46]	CVPR'19	0.7124	0.6835	0.6926	0.6765	0.7423	0.6289	0.6199	0.6482	0.6188	0.5164	0.6833	0.6290	0.6535
LBVMM [34]	ECCV'18	0.7069	0.6744	0.6873	0.6744	0.7532	0.6320	0.6155	0.6415	0.6163	0.5130	0.7018	0.6283	0.6537
LNVMM [43]	WACV'23	0.6715	0.6742	0.6949	0.6436	0.7665	0.6521	0.6088	0.6407	0.6074	0.5080	0.7146	0.6205	0.6516
MDLMM [44]	CVPR'23	0.6571	0.6820	0.6946	0.6556	0.7644	0.6558	0.6087	0.6394	0.6105	0.5195	0.7183	0.6293	0.6551
Ours	-	0.7338	0.7125	0.7080	0.6938	0.7721	0.6633	0.6345	0.6946	0.6206	0.5422	0.7185	0.6417	0.6780

Table 1. **Performance comparison with SOTA methods regarding MANIQA↑ scores on Real-world Datasets.** To ensure experimental fairness, all videos have a magnification factor α of 20 in Static Mode and 10 in Dynamic Mode for inference. The results show that FD4MM achieves the best average score (Avg.) for magnification quality in all challenging scene videos.

Method	Pre-proc	Networks	FLOPs(G)	Params(M)	Times(ms)
LBVMM [34]	-	CNNs	268.6	0.98	105.7
LNVMM [43]	-	CNNs	562.0	1.81	198.5
MDLMM [44]	Fourier Transform	CNNs	65.4	**0.12**	222.1
Ours	-	Transformer	**24.9**	1.47	**82.9**

Table 2. **Fair comparison with SOTA methods for parameters, FLOPs, and inference time.**

	Networks	FLOPs(G)	Params(M)	SSIM↑	LPIPS↓
A_0	Shllow Band ($C = 24$)	11.0	0.08	0.8582	0.3132
A_1	+ Middle Band	18.3	0.37	0.8938	0.1869
A_2	**+ Deep Band**	24.9	1.47	**0.9104**	**0.1376**
A_3	+ Extra Band	33.1	5.80	0.9008	0.1449

Table 3. **Ablation studies of multi-level isomorphic architecture on the Synthetic Dataset.**

	Networks	Components	FLOPs (G)	Params (M)	SSIM↑	LPIPS↓
B_0	**Baseline**	**MIA**	6.4	0.30	0.7651	0.3750
B_1	**Main**	$+ \mathcal{F}_L(\cdot)$	9.0	0.64	0.8177	0.3005
B_2	**Components**	$+ \mathcal{F}_H(\cdot)$	16.4	1.09	0.8521	0.2197
B_3		$+ \mathcal{M}(\cdot)$	24.9	1.47	**0.9104**	**0.1376**
B_4		Swin-T [31]	33.9	1.91	0.8864	0.1654
B_5	**Transformer**	Restormer [65]	35.6	2.03	0.8973	0.1488
B_6		**Ours**	**24.9**	**1.47**	**0.9104**	**0.1376**

Table 4. **Ablation studies of main components and comparisons with mainstream Transformer blocks on the Synthetic Dataset.** MIA is the basic Multi-level Isomorphic Architecture.

	Networks	FLOPs(G)	Params(M)	SSIM↑	LPIPS↓
C_0	FD4MM w/o sparse	24.9	1.47	0.9088	0.1405
C_1	**FD4MM**	24.9	1.47	**0.9104**	**0.1376**

Table 5. **Effects of the sparse strategy on the Synthetic Dataset.**

	$\mathcal{L}_{mag}$	$\mathcal{L}_{edge}$	$\mathcal{L}_{CR}$	$\mathcal{L}_{Sobel}$	$\mathcal{L}_{perc}$	SSIM↑	LPIPS↓
D_0	✓	-	-	-	-	0.8864	0.1877
D_1	✓	-	-	✓	-	0.8901	0.1723
D_2	✓	✓	-	-	-	0.8937	0.1697
D_3	✓	✓	-	-	✓	0.9018	0.1564
D_4	✓	✓	✓	-	-	**0.9104**	**0.1376**

Table 6. **Ablation studies of loss terms on the Synthetic Dataset.**

pose a synthetic dataset with controllable magnification as a reliable criterion for quantitative assessment. The synthetic test dataset of [34] is not available, and we follow the synthesis process to produce a new one that comprises ten synthetic videos. Each video is synthesized with a foreground object from the public StickPNG library and a background image from the DIS5K [36]. Adhering to the synthetic rule of [34, 46], we move the position of the foreground object in the frame sequence to exhibit meaningful subtle motion. The video resolution is set to $640{\times}640$ pixels, and the subtle motion is defined as $y(t) = 2\alpha \cdot \sin\left(\frac{2\pi}{T}t\right)$, where t denotes the video frame index, T is the harmonic period of 60, and the magnification factor α in the original videos with 30 fps. Thus, by varying α, we can obtain the ground-truth magnified videos with different scales. Besides, following [14, 46], we add the Gaussian noise with the mean and standard deviation setting $[0, \sigma]$ onto the spatial context to assess the model's noise robustness [14, 34, 44, 46].

Evaluation Metrics. Real-world videos exhibit rich and naturally occurring variations in motion but lack accurate ground truth. Thus, we introduce an advanced no-reference image quality assessment metric called MANIQA [63] for

arm motions with gunshot recoil [44, 46], *etc.*

Synthetic Test Dataset. The magnified ground-truth videos are inaccessible for real-world videos, so we attempt to pro-
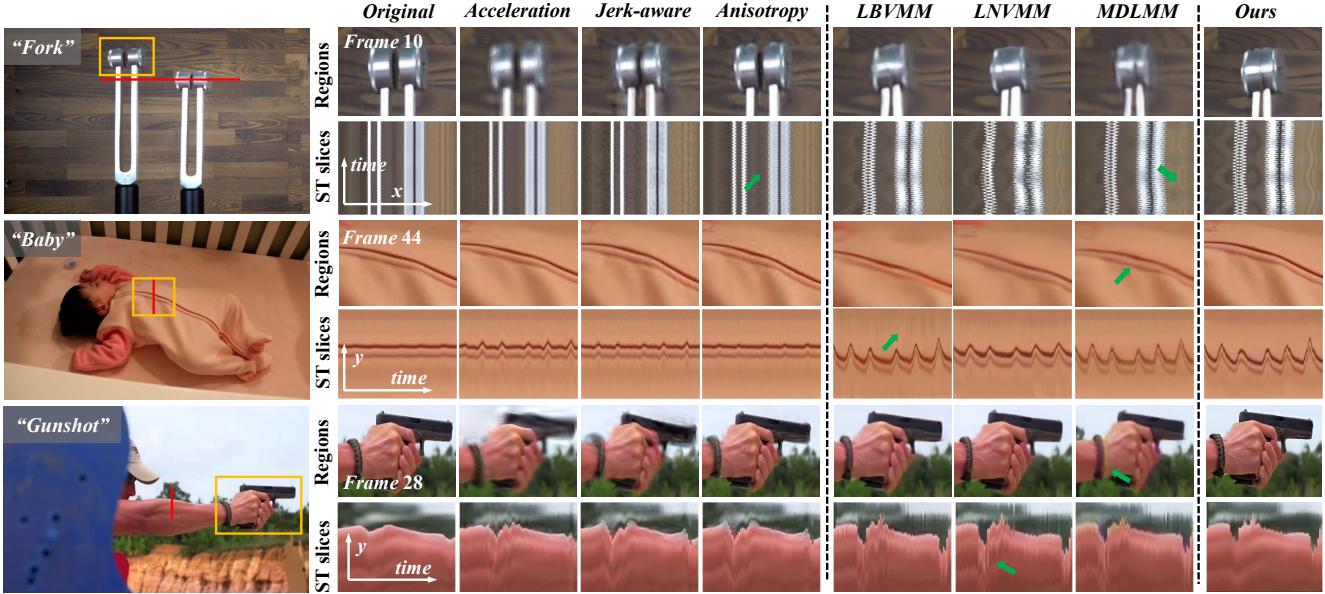
Figure 6. **Visualization examples on Real-world Datasets.** We enlarge the magnified spatial regions and display the spatiotemporal (ST) slices. It is clear that the hand-crafted filters [45, 46, 69] have much smaller magnified amplitudes and suffer from more ringing artifacts (*e.g.*, the fork's vibration). In contrast, the learning-based methods [34, 43, 44] achieve much larger magnified amplitudes and still face flickering artifacts in [34, 43] (*e.g.*, the baby's abdomen) and unnatural artifacts and deformation in [44] (*e.g.*, the arm and bracelet). Our FD4MM achieves the best appearances with satisfactory magnified amplitudes and high-quality generated images.

the Real-world test dataset. It benefits from efficiently assessing artifacts and distortion in magnified videos. For the synthetic test dataset, with accurate ground truth, we can adhere to [14, 34] by using evaluation metrics such as the structural similarity index measure SSIM [8, 16, 38, 54, 55] and LPIPS [56–58, 67] to quantitatively assess the similarity between the magnified frame and the ground truth.

Implementation Details. All methods are trained on the same training dataset of [34] to ensure the fairness of the experiments. In FD4MM, the channel size is set to $C = 24$. $\mathcal{F}_H(\cdot)$ and $\mathcal{M}(\cdot)$ from the *shallow* to *deep* bands are set to $\{2, 4, 4\}$ and $\{6, 4, 4\}$ layers respectively, and set with the consistent multi-head $h = \{4, 4, 8\}$. The layers of $\mathcal{F}_L(\cdot)$ at deep band is set to 4 and $h = 8$. We empirically set $\lambda = 0.1$ for the contrastive loss $\mathcal{L}_{CR}$. We utilize the Adam [21] optimizer with a learning rate of 1×10^{-4} for training.

4.2. Quantitative Evaluation

Comparison on Real-world Datasets. Table 1 reports Anisotropy [46] achieves higher MANIQA scores than other traditional methods but still sacrifices the motion amplitude in many scenes. The latest learning-based method MDLMM [44] gains a higher MANIQA score of 0.6551 *vs.* 0.6535 compared to Anisotropy. The proposed FD4MM is superior to MDLMM with an overall score of 0.6780 *vs.* 0.6551, indicating its superior magnification quality to respond to various motion scenes.

Comparison on Synthetic Datasets. We perform compari-

son between FD4MM and existing methods [34, 43–46, 69] on Synthetic Datasets with the magnification factor $\alpha = \{5, 10, 20, 50, 100\}$ and the noise level $\sigma = \{0.01, 0.05, 0.1, 0.2\}$. From Figs. 5 (b) and (c), the results indicate that FD4MM consistently achieves the best SSIM and LPIPS.

Model Complexity Analysis. Following [43, 44], we compare FD4MM with other learning-based methods [34, 43, 44] in terms of the model size and run-time values calculated at 720×720 resolution in Table 2. Although the latest MDLMM [44] reduces learnable parameters by adopting the traditional algorithm, our FD4MM achieves the optimal performance within acceptable parameters, considering both model complexity and inference speed.

4.3. Ablation Studies

We further perform the ablation studies in the magnification accuracy evaluation of the Synthetic Dataset.

Necessity of Multi-level Isomorphic Architecture. We evaluate the results by increasing frequency decoupling levels in Table 3. The model has a narrow bandwidth with only one or two band levels, so frequency decoupling yields no significant benefits. With a three-level isomorphic architecture, FD4MM achieves the best performance in terms of computational costs and results. However, further increasing the decoupling levels inevitably brings more computational costs and frequency attenuation. Considering the trade-offs [23, 48], we adopt the three-level isomorphic architecture to effectively cope with the task.
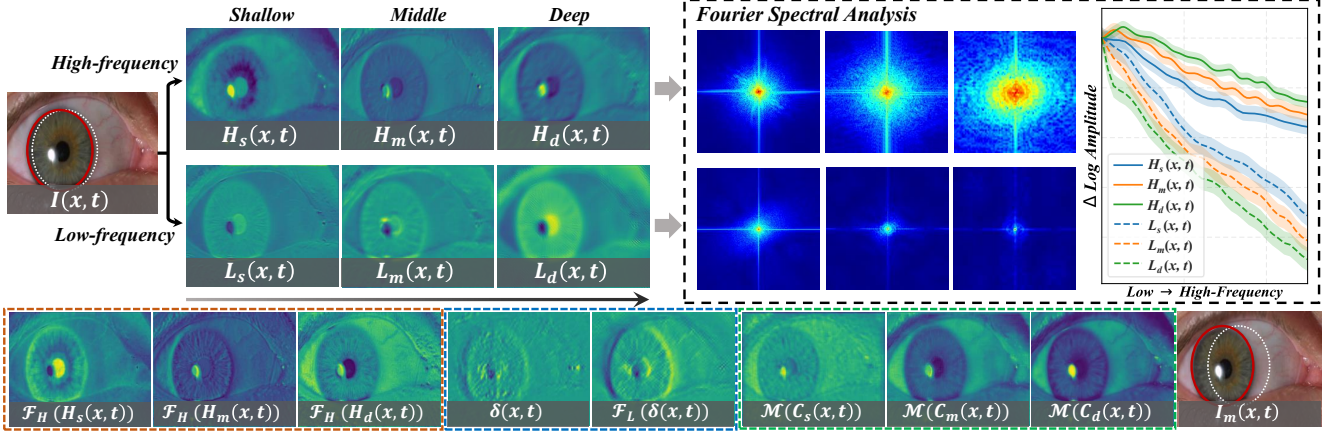
Figure 7. **Visualization of the *eye* video Sample from the Real-world Datasets.** On the left, frequency decoupling results in a significant feature difference. It is verified in the Fourier spectrum analysis on the right, where the bright field diffusing in four directions in the high-frequency spectrum indicates that it has more energy. With deeper levels, the low-frequency energy distribution is more focused in the central region, indicating greater stability. Besides, a detailed implementation process of FD4MM is provided for intuitive understanding.

Effect of Main Components and Their Sparsity. Table 4 reports the impact of each crucial component in this task. By incrementally adding each component separately, all of them contribute to improvement, especially for $\mathcal{M}(\cdot)$ (*e.g.*, SSIM from 0.8521 to 0.9104 and LPIPS from 0.2197 to 0.1376). Besides, we replace them with different mainstream Transformer blocks such as Swin-T [31] and Restormer [65] for comparison, FD4MM still exhibits the best performance. Notably, Table 5 verifies that the sparse strategy (ReLU) results in gains of 0.9104 *vs.* 0.9088 in SSIM and 0.1376 *vs.* 0.1405 in LPIPS.

Effectiveness of Loss Functions. In this study, we propose new loss functions compared to existing methods, *i.e.*, $\mathcal{L}_{edge}$ (LoG) and $\mathcal{L}_{CR}$. From Table 6, the combination of $\mathcal{L}_{edge}$ and $\mathcal{L}_{CR}$ results in a significant improvement in SSIM from 0.8864 to 0.9104. Moreover, we replace our losses with an alternative edge loss $\mathcal{L}_{Sobel}$ [70] and a perceptual loss $\mathcal{L}_{perc}$ [20] that are used in MDLMM. $\mathcal{L}_{edge}$ compared to $\mathcal{L}_{Sobel}$ yields a performance improvement of SSIM of 0.8937 *vs.* 0.8901, and $\mathcal{L}_{CR}$ has even more gains compared to $\mathcal{L}_{perc}$ for SSIM of 0.9104 *vs.* 0.9018.

4.4. Qualitative Analysis

Visualization of Motion Magnification. We perform an intuitive comparison of some challenging motion videos in Real-world Datasets. In Fig. 6, traditional hand-crafted filters [45, 46, 69] always result in small amplitudes and ringing artifacts (*e.g.*, the fork's vibration) under the same magnification factor. The learning-based methods [34, 43, 44] guarantee larger magnification amplitudes but are still accompanied by distortion, flickering, and unnatural artifacts (*e.g.*, the baby's abdomen), exhibiting disrupted spatial consistency. In contrast, FD4MM retains spatial consistency and suppresses distortion and artifacts well while enhancing magnified amplitude and overall quality.

Visualization of Frequency decoupling. In Fig. 7, high- and low-frequency features have different energy spectra. Especially in the deep band, the low-frequency feature has a more stable energy distribution and is therefore chosen for modeling the motion field $\delta(x,t)$. Next, $\mathcal{F}_H(\cdot)$ captures vital high-frequency details through its high-pass operator properties, such as the pupil and blood filaments of the eye. $\mathcal{F}_L(\cdot)$ focuses on low-frequency structures to promote well-structured motion, reducing the undesired motion magnification. Ultimately, the multi-level high-frequency details and magnified low-frequency are seamlessly recoupled by the $\mathcal{M}(\cdot)$, thereby inhibiting ringing artifacts.

5. Conclusion

We introduce FD4MM, a Multi-level Isomorphic Architecture based on frequency decoupling for motion magnification. FD4MM effectively preserves high-frequency spatial details and captures low-frequency overall structure, modeling a stable motion field. It employs Sparse High/Low-pass Filters and Sparse Frequency Mixer to handle the high-frequency details, to-be-amplified motion field, and seamless recoupling, respectively. Besides, a novel contrastive regularization is used to eliminate undesired motion and enhance model robustness. In extensive evaluations on Real-world and Synthetic test datasets, FD4MM achieves superior performance with fewer FLOPs and faster inference speed, offering a promising solution for future research.

6. Acknowledgments

References

[1] Freddy Abnousi, Guson Kang, John Giacomini, Alan Yeung, Shirin Zarafshar, Nicholas Vesom, Euan Ashley, Robert Harrington, and Celina Yong. 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, 2019.

[2] Marc Antonini, Michel Barlaud, Pierre Mathieu, and Ingrid Daubechies. Image coding using wavelet transform. *IEEE TIP*, 1:20–5, 1992.

[3] David F Barbe. Imaging devices using the charge-coupled concept. *Proceedings of the IEEE*, 63(1):38–67, 1975.

[4] Biagio Brattoli, Uta Büchler, Michael Dorkenwald, Philipp Reiser, Linard Filli, Fritjof Helmchen, Anna-Sophia Wahl, and Björn Ommer. Unsupervised behaviour analysis and magnification (ubam) using deep learning. *Nature Machine Intelligence*, 3(6):495–506, 2021.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607, 2020.

[6] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *CVPR*, pages 5896–5905, 2023.

[7] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *ICCV*, pages 3435–3444, 2019.

[8] Xiaofeng Cong, Jie Gui, Kai-Chao Miao, Jun Zhang, Bing Wang, and Peng Chen. Discrete haze level dehazing network. In *ACM MM*, pages 1828–1836, 2020.

[9] Abe Davis, Katherine L Bouman, Justin G Chen, Michael Rubinstein, Fredo Durand, and William T Freeman. Visual vibrometry: Estimating material properties from small motion in video. In *CVPR*, pages 5335–5343, 2015.

[10] Abe Davis, Katherine L Bouman, Justin G Chen, Michael Rubinstein, Oral Buyukozturk, Fredo Durand, and William T Freeman. Visual vibrometry: Estimating material properties from small motions in video. *IEEE TPAMI*, 39(4), 2017.

[11] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *CVPR*, pages 11963–11975, 2022.

[12] Michael Dorkenwald, Uta Buchler, and Bjorn Ommer. Unsupervised magnification of posture deviations across subjects. In *CVPR*, pages 8256–8266, 2020.

[13] Marc Eitner, Benjamin Miller, Jayant Sirohi, and Charles Tinney. Effect of broad-band phase-based motion magnification on modal parameter estimation. *Mechanical Systems and Signal Processing*, 146:106995, 2021.

[14] Brandon Y Feng, Hadi Alzayer, Michael Rubinstein, William T Freeman, and Jia-Bin Huang. 3d motion magnification: Visualizing subtle motions from time-varying radiance fields. In *ICCV*, pages 9837–9846, 2023.

[15] Ge Gao, Pei You, Rong Pan, Shunyuan Han, Yuanyuan Zhang, Yuchao Dai, and Hojae Lee. Neural image compression via attentional multi-scale back projection and frequency decomposition. In *ICCV*, pages 14677–14686, 2021.

[16] Dan Guo, Kun Li, Zheng-Jun Zha, and Meng Wang. Dadnet: Dilated-attention-deformable convnet for crowd counting. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1823–1832, 2019.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[18] Dianye Huang, Yuan Bi, Nassir Navab, and Zhongliang Jiang. Motion magnification in robotic sonography: Enabling pulsation-aware artery segmentation. *IROS*, 2023.

[19] Shirui Huang, Keyan Wang, Huan Liu, Jun Chen, and Yunsong Li. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In *CVPR*, pages 18145–18155, 2023.

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[22] Anh Cat Le Ngo and Raphael C-W Phan. Seeing the invisible: Survey of video motion magnification and small motion analysis. *ACM CSUR*, 52(6):1–20, 2019.

[23] Kun Li, Dan Guo, and Meng Wang. Proposal-free video grounding with contextual pyramid network. In *AAAI*, pages 1902–1910, 2021.

[24] Kun Li, Dan Guo, and Meng Wang. Vigt: proposal-free video grounding with a learnable token in the transformer. *Science China Information Sciences*, 66(10):202102, 2023.

[25] Kun Li, Jiaxiu Li, Dan Guo, Xun Yang, and Meng Wang. Transformer-based visual grounding with cross-modality interaction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–19, 2023.

[26] Xiang Li, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Dlgsanet: Lightweight dynamic local and global self-attention networks for image super-resolution. In *ICCV*, pages 12792–12801, 2023.

[27] Yufeng Li, Jiyang Lu, Hongming Chen, Xianhao Wu, and Xiang Chen. Dilated convolutional transformer for high-quality image deraining. In *CVPRW*, pages 4198–4206, 2023.

[28] Yufei Liang, Jiangning Zhang, Shiwei Zhao, Runze Wu, Yong Liu, and Shuwen Pan. Omni-frequency channel-selection representations for unsupervised anomaly detection. *IEEE TIP*, 32:4327–4340, 2023.

[29] Ce Liu, Antonio Torralba, William T Freeman, Frédo Durand, and Edward H Adelson. Motion magnification. *ACM TOG*, 24(3):519–526, 2005.

[30] Rui Liu, Young Jin Kim, Alexandre Muzio, and Hany Hassan. Gating dropout: Communication-efficient regularization for sparsely activated transformers. In *ICLR*, pages 13782–13792, 2022.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021.

[32] Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, and Khoa Luu. Micron-bert: Bert-based facial micro-expression recognition. In *CVPR*, pages 1482–1492, 2023.

[33] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

[34] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Fr'edo Durand, William T Freeman, and Wojciech Matusik. Learning-based video motion magnification. In *ECCV*, pages 633–648, 2018.

[35] Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. In *NeurIPS*, pages 14541–14554, 2022.

[36] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, pages 38–56, 2022.

[37] Michael Rubinstein, Neal Wadhwa, Fredo Durand, William T Freeman, and Hao-Yu Wu. Revealing invisible changes in the world. *Science*, 339(6119):519–519, 2013.

[38] Hao Shen, Zhong-Qiu Zhao, and Wandi Zhang. Adaptive dynamic filtering network for image denoising. In *AAAI*, pages 2227–2235, 2023.

[39] Kai Shen, Junliang Guo, Xu Tan, Siliang Tang, Rui Wang, and Jiang Bian. A study on relu and softmax in transformer. *arXiv preprint arXiv:2302.06461*, 2023.

[40] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, pages 1874–1883, 2016.

[41] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan. Inception transformer. In *NeurIPS*, pages 23495–23509, 2022.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[43] Jasdeep Singh, Subrahmanyam Murala, and G Kosuru. Lightweight network for video motion magnification. In *WACV*, pages 2041–2050, 2023.

[44] Jasdeep Singh, Subrahmanyam Murala, and G Kosuru. Multi domain learning for motion magnification. In *CVPR*, pages 13914–13923, 2023.

[45] Shoichiro Takeda, Kazuki Okami, Dan Mikami, Megumi Isogai, and Hideaki Kimata. Jerk-aware video acceleration magnification. In *CVPR*, pages 1769–1777, 2018.

[46] Shoichiro Takeda, Yasunori Akagi, Kazuki Okami, Megumi Isogai, and Hideaki Kimata. Video magnification in the wild using fractional anisotropy in temporal distribution. In *CVPR*, pages 1614–1622, 2019.

[47] Shoichiro Takeda, Kenta Niwa, Mariko Isogawa, Shinya Shimizu, Kazuki Okami, and Yushi Aono. Bilateral video magnification filter. In *CVPR*, pages 17369–17378, 2022.

[48] Shengeng Tang, Dan Guo, Richang Hong, and Meng Wang. Graph-based multimodal sequential embedding for sign language translation. *IEEE TMM*, 24:4433–4445, 2021.

[49] Shengeng Tang, Richang Hong, Dan Guo, and Meng Wang. Gloss semantic-enhanced network with online back-translation for sign language production. In *ACM MM*, pages 5630–5638, 2022.

[50] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *ECCV*, pages 146–162, 2022.

[51] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM TOG*, 32(4):1–10, 2013.

[52] Fei Wang, Dan Guo, Kun Li, and Meng Wang. Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer. *arXiv preprint arXiv:2312.04152*, 2023.

[53] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin. Kvt: k-nn attention for boosting vision transformers. In *ECCV*, pages 285–302, 2022.

[54] Yuwei Wang, Jiaxu Cai, Yuankun Liu, Xiangcheng Chen, and Yajun Wang. Motion-induced error reduction for phase-shifting profilometry with phase probability equalization. *Optics and Lasers in Engineering*, 156:107088, 2022.

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.

[56] Yanyan Wei, Zhao Zhang, Yang Wang, Mingliang Xu, Yi Yang, Shuicheng Yan, and Meng Wang. Deraincyclegan: Rain attentive cyclegan for single image deraining and rain-making. *IEEE TIP*, 30:4788–4801, 2021.

[57] Yanyan Wei, Zhao Zhang, Mingliang Xu, Richang Hong, Jicong Fan, and Shuicheng Yan. Robust attention deraining network for synchronous rain streaks and raindrops removal. In *ACM MM*, pages 6464–6472, 2022.

[58] Yanyan Wei, Zhao Zhang, Huan Zheng, Richang Hong, Yi Yang, and Meng Wang. Sginet: Toward sufficient interaction between single image deraining and semantic segmentation. In *ACM MM*, pages 6202–6210, 2022.

[59] Mitchell Wortsman, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Replacing softmax with relu in vision transformers. *arXiv preprint arXiv:2309.08586*, 2023.

[60] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *CVPR*, pages 10551–10560, 2021.

[61] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM TOG*, 31(4):1–8, 2012.

[62] Zhaoqiang Xia, Wei Peng, Huai-Qian Khor, Xiaoyi Feng, and Guoying Zhao. Revealing the invisible with model and data shrinking for composite-database micro-expression recognition. *IEEE TIP*, 29:8590–8605, 2020.

[63] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPRW*, pages 1191–1200, 2022.

[64] Guhnoo Yun, Juhan Yoo, Kijung Kim, Jeongho Lee, and Dong Hwan Kim. Spanet: Frequency-balancing token mixer

using spectral pooling aggregation modulation. In *ICCV*, pages 6113–6124, 2023.

[65] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022.

[66] Dashan Zhang, Andong Zhu, Xinlong Gong, Yuwei Wang, Jie Guo, and Xiaolong Zhang. Hybrid eulerian–lagrangian framework for structural full-field vibration quantification and modal shape visualization. *Measurement*, 219:113270, 2023.

[67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[68] Yi Zhang, Xiaoyuan Han, Han Zhang, and Liming Zhao. Edge detection algorithm of image fusion based on improved sobel operator. In *2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference*, pages 457–461, 2017.

[69] Yichao Zhang, Silvia L Pintea, and Jan C Van Gemert. Video acceleration magnification. In *CVPR*, pages 529–537, 2017.

[70] Bolun Zheng, Shanxin Yuan, Gregory Slabaugh, and Ales Leonardis. Image demoireing with learnable bandpass filters. In *CVPR*, pages 3636–3645, 2020.

[71] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *ECCV*, pages 386–403, 2022.

[72] Jinxing Zhou, Dan Guo, and Meng Wang. Contrastive positive sample propagation along the audio-visual event line. *IEEE TPAMI*, 45(6):7239–7257, 2023.

[73] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *ICCV*, pages 12780–12791, 2023.