2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT) | 979-8-3315-2228-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/AINIT65432.2025.11035513

# Unsupervised Image Stitching: Image Alignment Method Based on Gradual Refinement Network

Chengyu Jiao
School of Information and Automation,
Qilu University of Technology (Shandong Academy of Sciences),
Jinan, China
1098818844@qq.com

Kaiyue Bai
School of Information and Automation,
Qilu University of Technology (Shandong Academy of Sciences),
Jinan, China

Wenjing Zheng
School of Information and Automation,
Qilu University of Technology (Shandong Academy of Sciences),
Jinan, China

He Gao
Shandong Key Laboratory of Intelligent Buildings and Building energy-saving TechnologySchool of Information and Electrical Engineering, shandong Jianzhu University,
Shandong Zhengchen Technology co, LTD,
Jinan, China

Xuehan Zheng
Shandong Key Laboratory of Intelligent Buildings and Building energy-saving Technology, School of Information and Electrical Engineering, Shandong Jianzhu University,
Jinan, China

Jun Li*
School of Information and Automation,
Qilu University of Technology (Shandong Academy of Sciences),
Jinan, China
*rogerjunli@sdu.edu.cn

***Abstract*—Image alignment is a critical step in the image stitching process. Traditional image alignment methods typically use uniform grid transformations or homography transformations to achieve geometric alignment. However, these methods have limitations when dealing with complex scenes, geometric distortions, and large viewpoint changes, making it difficult to preserve image details and local consistency. To address these issues, this paper proposes an image alignment method based on a Gradual Refinement Network. Specifically, a grid offset prediction network (Gradual Refinement Network) is designed for image alignment, which enhances sensitivity to detail changes through Haar wavelet downsampling, introduces spatial and channel collaborative attention to strengthen local information extraction, and employs a coarse-to-fine strategy to achieve precise image alignment. A loss function suitable for grid offset prediction is proposed to optimize the grid offset. Experimental results show that the proposed method surpasses existing image alignment techniques in handling complex geometric deformations and large viewpoint changes.**

***Keywords—Image Alignment; Gradual Refinement Network; Deep Learning; Computer Vision***

## I. INTRODUCTION

Image stitching is a fundamental problem in the field of computer vision. Image alignment is a critical step in stitching, aiming to align images from different viewpoints through geometric transformations.

Traditional image alignment methods mainly rely on handcrafted feature extraction algorithms, such as SIFT[1] and ORB[2], which are effective in handling rotation, scaling, and lighting variations. However, they have limitations when dealing with complex geometric distortions and large viewpoint changes. To address these issues, adaptive methods based on local transformations (such as APAP[3])) have emerged. These methods enhance local alignment performance by estimating independent transformations and applying multiple affine transformations to local regions. However, these methods still suffer from alignment errors in areas with rich details or complex shapes.

In recent years, deep learning methods have shown strong robustness in complex scenes, with deep homography estimation methods gradually replacing traditional techniques, categorized into supervised and unsupervised approaches. Supervised methods rely on training data with real homography labels. Early studies used VGG-style networks to predict homography transformations, particularly excelling in alignment in low-overlap areas. However, obtaining real labels in natural scenes is challenging, and training on synthetic data may compromise generalization performance. Unsupervised methods, such as Nie[4] et al. further improved disparity tolerance using thin-plate spline transformations. Despite breakthroughs in deep image alignment methods, they still face limitations when dealing with complex scenes, geometric distortions, and large viewpoint changes, making it difficult to preserve details and local consistency.

To address these issues, this paper proposes an image alignment method based on a Gradual Refinement Network. A spatial and channel collaborative attention module (SCSA) is introduced to enhance local information extraction, and Haar wavelet downsampling is employed to improve sensitivity to detail variations. A grid constraint loss function is used for optimization, and a "coarse-to-fine" strategy is adopted to progressively refine local alignment based on initial coarse alignment, ultimately achieving precise alignment.

## II. PROPOSED METHOD

### A. Gradual Refinement Network

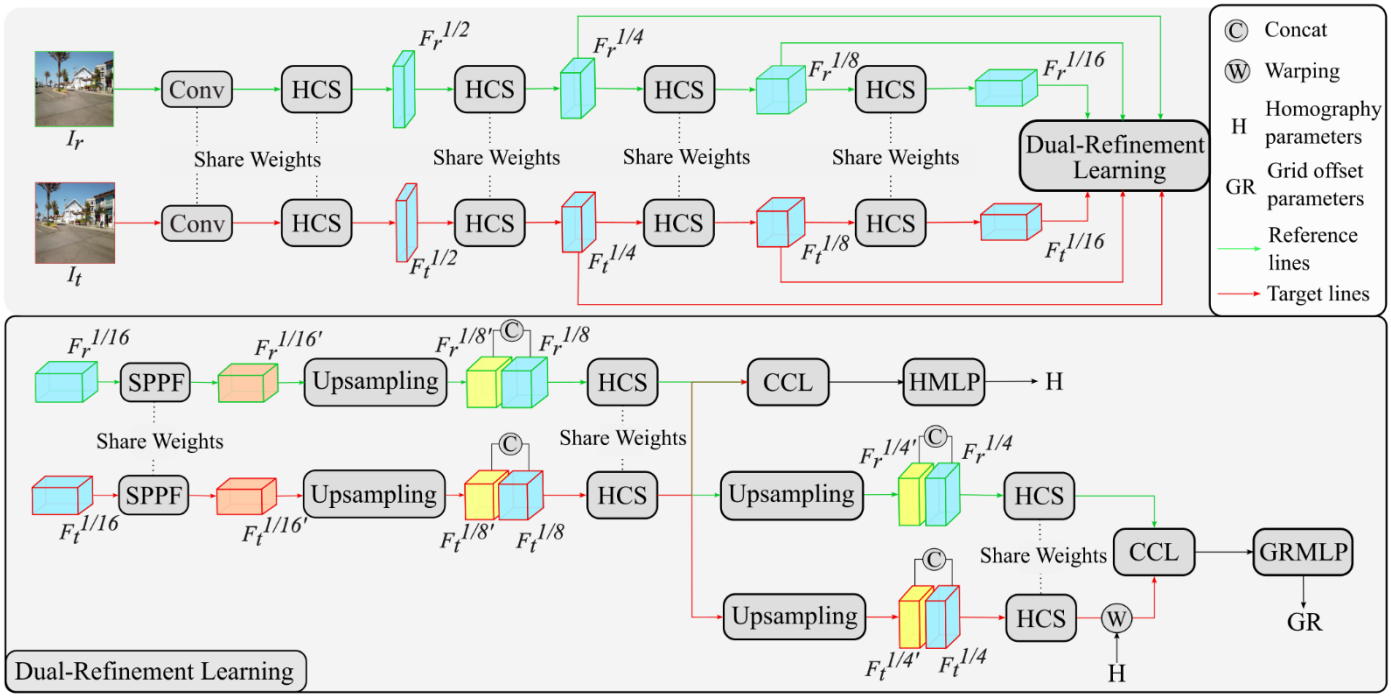


Figure 1 Gradual Refinement Network

As shown in Figure 1, given a pair of input images $(I_r, I_t) \in \mathbb{R}^{H\times W\times 3}$, the images are first downsampled via a convolutional layer to a size of $H/2\times W/2\times 64$, generating the initial feature representation F. Subsequently, four layers of the HCS module (with filter sizes of 128, 256, 512, 512) are applied to produce features at different resolutions: $F_{1/2}$, $F_{1/4}$, $F_{1/8}$, $F_{1/16}$.

In the Dual-Refinement Learning phase, to better capture detailed information, the features at different resolutions are fused. Initially, referring to equations (1) and (2), the multi-scale features obtained from the backbone network are processed through the SPPF module to enhance the representational capability of the feature maps. Subsequently, the feature maps are resized using an upsampling layer, and strengthen the fusion of features at different resolutions through the HCS module., further optimizing the feature map representation.

$$F^{"}_{1/8} = HCS(Upsample(SPPF(F_{1/16})) \oplus F_{1/8}) \quad (1)$$

$$F^{"}_{1/4} = HCS(Upsample(HCS(F^{"}_{1/8})) \oplus F_{1/4}) \quad (2)$$

Where $\oplus$ indicates feature concatenation, $F$ is the input feature flow, and $F^{"}$ is the resulting fused feature.

Subsequently, the correlation between the feature maps $F^{"}_{r1/8}$ and $F^{"}_{t1/8}$ is computed using the CCL[7] layer, which leverages a contextual association mechanism to capture global information and significantly improve alignment accuracy. A three-layer fully connected HMLP network is then employed to estimate the homography transformation parameters. The estimated transformation parameters are applied to $F^{"}_{t1/4}$ for initial coarse alignment. Finally, the transformed $F^{"}_{t1/4}$ and the aligned $F^{"}_{r1/4}$ are fed into a second CCL layer and the GRMLP layer to predict grid point offsets, achieving precise alignment:

$$Correlation_{1/8} = CCL(F^{"}_{r1/8}, F^{"}_{t1/8}) \quad (3)$$

$$H = HMLP(Correlation_{1/8}) \quad (4)$$

$$Correlation_{1/4} = CCL(F^{"}_{r1/4}, w(H, F^{"}_{t1/4})) \quad (5)$$

$$GR = GRMLP(Correlation_{1/4}) \quad (6)$$

Where *Correlation* represents the correlation volume, $H$ is the predicted homography parameter, $GR$ is the predicted grid transformation parameter, and $w(\bullet,\bullet)$ represents the homography transformation operation.

### B. HCS Module

To better capture multi-scale features and enhance feature representation, this paper proposes the HCS module. As shown in Figure 2. The HCS module processes the feature flow through an HC layer, first using Haar wavelet downsampling with a low-pass filter $H_0$ and a high-pass filter $H_1$ to extract low-frequency component $A$ and high-frequency detail components $H$, $V$, and $D$, and followed by feature extraction through a CBR layer:

$$A, H, V, D = HaarWavelet(F) \quad (7)$$

$$F_{HC} = CBR(A \oplus H \oplus V \oplus D) \quad (8)$$

Where $\oplus$ indicates the concatenation operation, and $F_{HC}$ is the output feature after processing by the HC module.

Next, the C2f module aggregates multi-scale features, employing CBS layers and multiple Bottleneck layers to extract

multi-scale information, and finally merging the features at different scales through a Concat operation.

The HCS module also introduces the SCSA attention mechanism, including the SMSA and PCSA strategies. SMSA enhances spatial features by decoupling the features through a multi-branch structure and extracting information at different scales. PCSA generates queries, keys, and values through downsampling and 1×1 convolutions, calculates an attention correlation matrix, and obtains attention weights through a Sigmoid operation. The weighted channel-enhanced feature map is then added to the original features, resulting in the enhanced output feature map:

$$F_{HCS} = SCSA(C2f(F_{HC})) \tag{9}$$

Where $F_{HCS}$ is the output feature after processing by the HCS module.

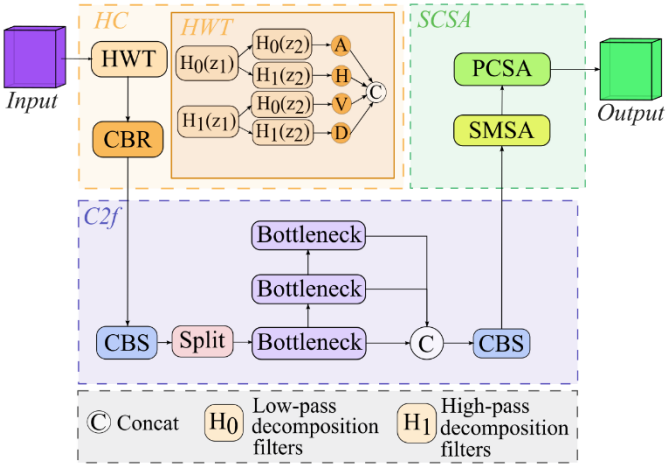


Figure 2 HCS Module

### C. Alignment Optimization

We train our network in a completely unsupervised manner. The loss function is composed of two parts: the alignment loss and the grid constraint loss.

To achieve pixel-level image alignment, we introduce the alignment loss $L_{align}$, which is computed using pixel consistency loss[4]. Let $w(\bullet,\bullet)$ represent the image distortion operation, $I_r$ be the reference image, $I_t$ be the target image, II be a matrix of ones with the same resolution as $I_r$, $H$ be the homography transformation parameters, and $GR$ be the grid offset parameters. Based on this, the alignment loss can be defined as:

$$\begin{aligned} L_{align} &= \lambda \left\| I_r \cdot w(\mathrm{II}, H) - w(I_t, H) \right\|_1 \\ &+ \lambda \left\| I_r \cdot w(\mathrm{II}, H^{-1}) - w(I_t, H^{-1}) \right\|_1 + \\ &\left\| I_r \cdot w(\mathrm{II}, GR) - w(I_t, GR) \right\|_1 \end{aligned} \tag{10}$$

To prevent image distortion caused by grid transition deformations, we design a grid constraint loss suitable for grid offsets. This loss constrains grid deformation by controlling the smoothness of the grid and the degree of deformation in non-overlapping regions.

$$L_{grid} = L_{smooth} + L_{trans} \tag{11}$$

Regarding the grid smoothness constraint, we measure the similarity between adjacent grid points along the same horizontal or vertical line by calculating the cosine similarity of their vectors. For each pair of adjacent grid points, we compute the angular difference between them and use this difference to assess the smoothness of the grid. The grid smoothness constraint can be expressed as:

$$L_{smooth} = \frac{1}{N_w}\sum_{i=1}^{N_w}(1-\cos(\theta_w(i))) + \frac{1}{N_h}\sum_{i=1}^{N_h}(1-\cos(\theta_h(i))) \tag{12}$$

Where $\cos(\theta(i))$ represents the cosine similarity between horizontally or vertically adjacent points, and $N_w$ and $N_h$ represent the number of adjacent points in the horizontal and vertical directions, respectively.

For the grid deformation constraint in non-overlapping regions $L_{trans}$, let the grid offset in the non-overlapping region be $m$, then:

$$L_{trans} = \sum_{i=1}^{N} a \sum_{j} (m_j)^2 \tag{13}$$

Where $a$ is a mask that indicates which regions belong to the non-overlapping area.

Finally, our total loss function is defined as::

$$L = L_{align} + L_{grid} \tag{14}$$

## III. Experiment

### A. Dataset and Experimental Details

Dataset: We conducted extensive experiments on the UDIS-D dataset. This dataset contains a variety of image types, including natural landscapes, architecture, and urban street scenes, with image pairs having different rotations, scaling, shifts, and viewpoint changes. The dataset aims to test the robustness and accuracy of image alignment algorithms in complex scenarios.

Experimental Details: We trained the model for 100 epochs using a single NVIDIA 2080 Ti GPU with the Adam optimizer. The initial learning rate was set to 1e-4, and a decay rate of 0.97 was applied after each iteration. The grid size was set to 13×13. All experiments were conducted using the PyTorch framework.

### B. Evaluation Metrics

We use SSIM and PSNR as metrics to assess the quality of the overlapping regions of images aligned using our method:

$$SSIM_{overlap} = SSIM(W(E) \odot I_r, W(I_t)) \tag{15}$$

$$PSNR_{overlap} = PSNR(W(E) \odot I_r, W(I_t)) \tag{16}$$

Where $W(\cdot)$ is the alignment transformation, $\odot$ denotes matrix multiplication, and $E$ is the identity matrix of the same size as $I_t$.

### C. *Quantitative Comparison*

We compared our approach with various classical image alignment methods. Specifically, we selected three traditional methods (SIFT[1], ORB[2], APAP[3]), three deep learning-based alignment algorithms (DHN[6], MDH[7], UDIS++[4]). As shown in Table 1 .The results show that our method outperforms others, with an average SSIM improvement of 1.2% and a PSNR improvement of 6%. The increase in SSIM indicates that our algorithm performs better in structural similarity, preserving more details, while the improvement in PSNR suggests better overall quality and local consistency.

Table 1: Performance Comparison on the UDIS-D Dataset

| Methods | SSIM | | | | PSNR | | | |
|---|---|---|---|---|---|---|---|---|
| | Easy | Moderate | Hard | Average | Easy | Moderate | Hard | Average |
| **I3×3** | 0.530 | 0.286 | 0.146 | 0.303 | 15.87 | 12.76 | 10.68 | 12.86 |
| **SIFT[1]+RANSAC[5]** | 0.916 | 0.833 | 0.636 | 0.779 | 28.75 | 24.08 | 18.55 | 23.27 |
| **ORB[2]+RANSAC[5]** | 0.888 | 0.772 | 0.550 | 0.718 | 27.53 | 22.85 | 17.37 | 22.06 |
| **APAP[3]** | 0.895 | 0.824 | 0.663 | 0.781 | 27.58 | 23.92 | 19.90 | 23.41 |
| **DHN[6]** | 0.930 | 0.869 | 0.714 | 0.825 | 29.70 | 26.02 | 21.44 | 25.29 |
| **MDH[7]** | 0.902 | 0.830 | 0.685 | 0.793 | 27.83 | 23.95 | 20.70 | 23.80 |
| **UDIS++[4]** | 0.933 | 0.875 | 0.739 | 0.838 | 30.19 | 25.84 | 21.57 | 25.43 |
| **Ours** | 0.947 | 0.889 | 0.742 | 0.848 | 31.53 | 27.01 | 23.47 | 26.95 |

### D. *Qualitative Comparison*

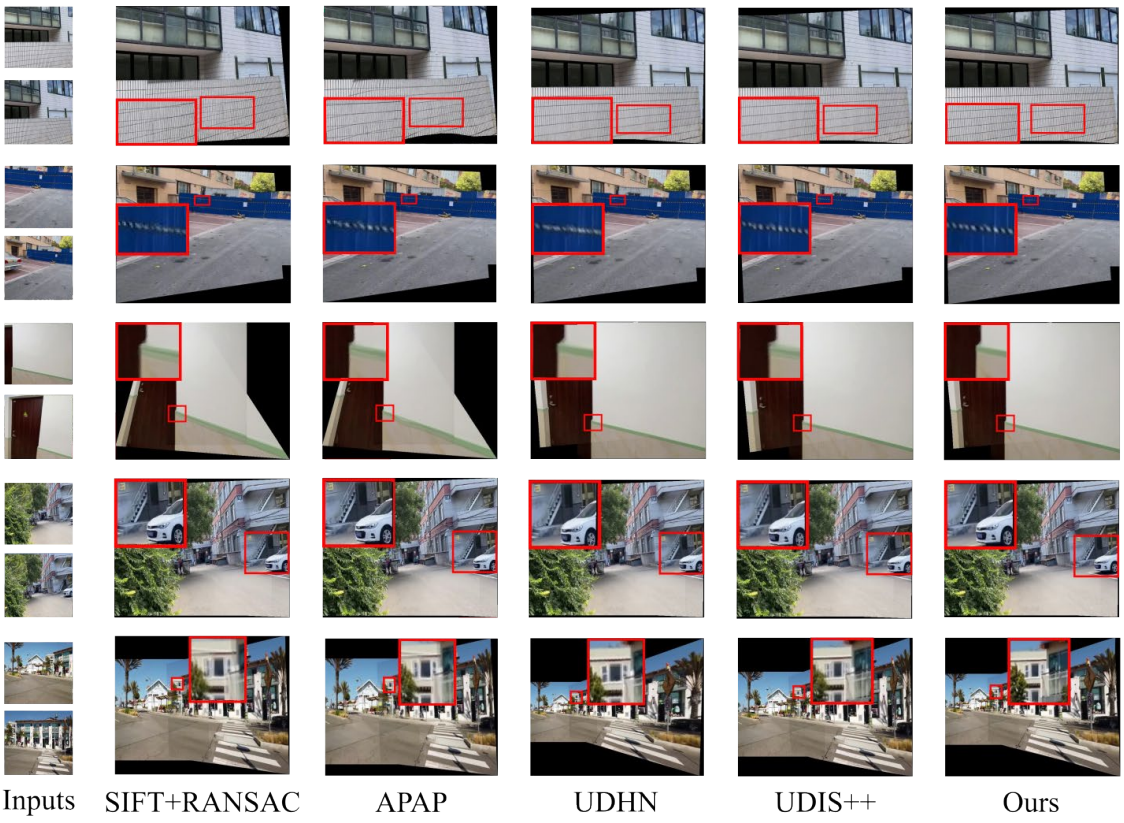


Figure 3 Image alignment results under different disparities and scenarios

In the qualitative comparison section, we visually present the performance of SIFT+RANSAC, APAP, UDHN, UDIS++, and our method in the image alignment task, particularly focusing on the alignment in overlapping regions, to intuitively assess the differences in alignment accuracy and detail recovery.

Figure 3 shows the image alignment results under four different scenarios. Traditional methods (such as SIFT+RANSAC and APAP) can achieve coarse alignment but exhibit significant errors in local details and often produce artifacts. Deep learning-based alignment methods (such as UDHN and UDIS++) improve the handling of local details but still fail to achieve accurate alignment in texture-poor or repetitive regions. In contrast, our method shows significant advantages in complex scenes, maintaining high alignment quality, reducing distortion, and producing more natural and precise alignment results.

### *E. Robustness Evaluation*

We conducted alignment experiments in a nighttime environment and compared our approach with SIFT and APAP methods. As shown in Figure 4. The experimental results show that our method successfully achieves smooth and accurate image alignment, while SIFT and APAP fail to align effectively under low-light conditions, resulting in noticeable distortion. This demonstrates that our method has stronger robustness in low-light environments and can effectively handle scarce or unclear visual information.


Figure 4 Alignment comparison in nighttime environment

### *F. Ablation Study*

To validate the effectiveness of each module, we conducted ablation experiments by removing or replacing key components and comparing the performance under different settings to gain deeper insight into the contribution of each module. The experimental results are shown in Table 2, where the symbol "√" indicates the inclusion of the corresponding module.

The results demonstrate that the HCS module and grid constraint loss function are crucial to the network model. After removing the HCS module, SSIM decreased by 16.03% and PSNR by 7.71%, leading to a significant performance drop, which confirms the importance of the HCS module in capturing both global and local information. After removing the $L_{grid}$ loss function, the network performance decreased, proving the effectiveness of this loss function in optimizing grid alignment accuracy.

Table 2: Ablation Study on the UDIS-D Dataset

| HCS | $L_{grid}$ | SSIM | PSNR |
|---|---|---|---|
| | √ | 0.712(-16.03%) | 24.87(-7.71%) |
| √ | | 0.791(-6.72%) | 25.43(-4.64%) |
| √ | √ | 0.848 | 26.95 |

## IV. Conclusion

This paper presents an image alignment method based on a Gradual Refinement Network, which utilizes deep learning techniques to estimate the homography transformation and grid offset for precise image alignment. Compared to existing methods, our approach not only improves image alignment accuracy but also enhances the visual consistency of image stitching, demonstrating strong robustness and adaptability. In the future, while further improving the model's computational efficiency and generalization ability, we will explore how to integrate more self-supervised and unsupervised learning strategies to further enhance the applicability and stability of image alignment in real-world scenarios.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China (12005108), the Shandong Provincial Natural Science Foundation Youth Project (ZR2020QF016) and Key Research and Development Plan of Shandong Province in 2022 (2022KJHZ002), Key Technologies and Applications of C-V2X Communication Safety in Vehicle Networking in Expressway Scenarios.

## References

[1] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.

[2] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: 2011 International Conference on Computer Vision, Ieee, 2011, pp. 2564–2571.

[3] J. Zaragoza, T.-J. Chin, M. S. Brown, and D. Suter, "As-projective as-possible image stitching with moving DLT," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2013, pp. 2339–2346.

[4] Nie L, Lin C, Liao K, et al. Parallax-tolerant unsupervised deep image stitching[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 7399-7408.

[5] Q. Zhou, T. Sattler, L. Leal-Taixe, Patch2pix: Epipolar-guided pixel-level corre spondences, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4669–4678.

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabi novich. Deep image homography estimation. arXiv preprint arXiv:1606.03798, 2016.

[7] Nie L, Lin C, Liao K, et al. Depth-aware multi-grid deep homography estimation with contextual correlation[J]. IEEE transactions on circuits and systems for video technology, 2021, 32(7): 4460-4472.