2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS)

A Transformer-Based Network for Multi-Stage Progressive Image Restoration

Ruyu Liu[1], Jiajia Wang[2], Haoyu Zhang[1], Jianhua Zhang[2], Xiufeng Liu[3,*]
[1]School of Information Science and Engineering, Hangzhou Normal University, Hangzhou, 311121, China
[2]School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, 300384, China
[3]Department of Technology, Management and Economics, Technical University of Denmark, Lyngby, Denmark
[*]email: xiuli@dtu.dk

2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS) | 979-8-3503-7784-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/DOCS63458.2024.10704475

***Abstract*—Image restoration is a challenging and complex problem involving recovering the original clear image from a degraded or noisy image. Existing methods for image restoration mainly use convolutional neural networks (CNNs) or Transformer models, which have different advantages and limitations in capturing spatial and channel information of the image. This paper proposes a novel Multi-Stage progressive image restoration Network based on a blend of local-global Transformers, named MSTNet. Our network consists of three stages, each using a different type of Transformer module to obtain both local and global information. The first two stages use window-based Transformer modules, which can effectively extract local spatial information within each window. The third stage uses channel-level Transformer modules to capture global channel information across the whole image. We also introduce a fusion module to combine the features from different Transformer branches and obtain a comprehensive and accurate feature representation. We conduct extensive experiments on various image restoration tasks, such as deblurring and denoising, and demonstrate the effectiveness and superiority of our network over state-of-the-art methods.**

***Index Terms*—Image restoration, Transformer, Multi-stage network, Feature fusion**

I. Introduction

In industrial activities, the assessment, restoration, and analysis of image data enable the identification of potential issues in products and production processes, leading to improved product quality and process optimization [1]. Factors such as noise, blur, rain, fog, and compression in the industrial environment can cause image degradation, reducing image quality and usability. Image restoration is recovering the original image from the degraded one, using mathematical models and computational techniques [2].

Image restoration is a challenging and complex problem, as it involves dealing with various types of degradation, which may be unknown, nonlinear, or spatially varying, and preserving the fine details and structures of the image while removing unwanted artifacts. Traditional image restoration methods rely on hand-crafted priors and models, making them computationally expensive and inconvenient for practical use [3]. Moreover, these methods often struggle to generalize effectively to diverse and realistic scenarios. With the advent of deep learning, convolutional neural networks (CNNs)-based neural networks have emerged as powerful tools for image restoration due to their ability to capture deep features and offer robust image representations [4]. However, the effectiveness of CNNs is somewhat constrained by their limited receptive field, which restricts their ability to capture long-range dependencies and global context in images. This may result in losing high-level semantic information and structural consistency in the restored images.

To overcome these limitations, Transformer models have been recently applied to image restoration, as they can exploit the self-attention mechanism to model the long-range pixel interactions and the global structure of images [5]. Transformer models can also adapt to the input content, assigning different attention weights to different pixels, depending on their relevance and importance. However, Transformer models may not be able to balance well between the spatial details and the high-level context in images, as they may focus too much on the global semantic information and neglect the local texture information [6]. Furthermore, Transformer models may not be interpretable and flexible enough for image restoration, as they may not provide clear insights into the restoration process and the degradation factors [7]. To address these challenges, we propose a novel multi-stage progressive image restoration network based on a blend of local-global Transformers, which can efficiently and effectively restore images from various types of degradation.

Our proposed method, named **M**ulti-**S**tage **T**ransformer **Net**work (MSTNet), consists of three main components: (1) A window-based Transformer, which divides the image into non-overlapping windows and applies self-attention within each window, thus capturing the local features and details of the image; (2) A channel-level Transformer, which applies self-attention across the channels of the image, thus capturing the global context and semantic information of the image; (3) A multi-stage progressive architecture, which divides the image restoration task into several subtasks, each corresponding to a different degradation factor, and uses a combination of window-based and channel-level Transformers to restore the image progressively, from coarse to fine. Moreover, we

This research was supported in part by the National Natural Science Foundation of China under Grant 62202137, 62306097, China Postdoctoral Science Foundation 2023M730599.

introduce a resolution network in the final stage, which restores the image at its original resolution, thus preserving the spatial details and avoiding the loss of information caused by downsampling and upsampling operations.

The main contributions of our paper are as follows:

- We propose a novel image restoration network based on a blend of window-based and channel-level Transformers, which can capture both the local and global information in images and balance well between the spatial details and the high-level context.
- We propose a novel multi-stage progressive image restoration network based on Transformers, which leverages the synergistic interaction of features extracted by each stage, thus enhancing the accuracy and efficiency of image restoration.
- We conduct extensive experiments on two typical image restoration tasks, such as deblurring and denoising, and demonstrate that our method achieves state-of-the-art (SoTA) results, both quantitatively and qualitatively, while having fewer parameters and lower computational complexity than existing methods.

II. Related Work

A. Image Restoration

Image restoration is the recovery of clear images from various types of images that have been degraded. Many researchers have used CNN for image restoration due to their powerful restoration capabilities [8], [9]. Among them, a U-Net network based on encoder-decoder performs well in image restoration [7], [8], [10]. The U-Net network not only acquires the multi-scale information of the image but also compresses the image into a low-dimensional representation by the encoder and then restores it to a high-quality image through the decoder. In addition, skip connections learning of residual signals is also a very effective solution, and several studies have demonstrated its effectiveness on image restoration tasks [8], [11]. Recently, several researchers have begun to use attention mechanisms to improve the quality of recovered images. Among them, both spatial attention and channel attention have been shown to be effective for image restoration [7], [12].

B. Vision Transformers

In the recent past, Transformer has shown excellent performance on advanced computer vision tasks [13]–[15]. Unlike CNN, it can obtain remote dependencies by focusing on data in the global domain. Among them, Vision Transformer [13] innovatively and successfully applies Transformer to the field of computer vision, breaking the dominance of CNN in image processing and demonstrating the strong potential and versatility of Transformer in processing non-textual data. On low-level vision tasks, it is difficult to apply directly to high-resolution images because the computational complexity of Transformer grows quadratically with the size of image resolution. For this reason, many researchers have proposed solutions. A common approach is to adopt the sliding window strategy [16], which divides the original image into multiple independent windows and then inputs the image features of each window into the Transformer model independently. This method can effectively reduce the computational complexity, but the window-based operation limits its spatial range. Based on this, [17] proposes to combine a window-based Transformer with U-Net to obtain multi-scale information and reduce computational complexity. In addition, [18] presents a channel-level Transformer to alleviate the above problem of limited spatial range. However, [18] lacks attention to spatial information. Based on the above problems, in this paper, we propose a method that jointly applies window-based Transformer and channel-level Transformer to obtain spatial information and channel information of the image. And our proposed model can improve the quality of recovered images.

C. Multi-Stage Network

Currently, many approaches [7], [19]–[22] use the multi-stage architecture to decompose a complex task into multiple easy-to-implement sub-tasks, where each stage can focus on learning features at different scales and semantic levels to achieve image restoration gradually. Among them, a common approach is to set the same model in each stage. The drawback of this design is that the model used in each stage may be specially designed for the task in that stage, while other stages may require different models. Therefore, if the same model is applied at each stage, it may lead to unsatisfactory recovery or even produce worse recovery results. In order to solve the above problem, we adopt a three-stage approach, where the first two stages combine window-based Transformer and U-Net to obtain contextual information. Because frequent downsampling operations in the first two stages will inevitably lose some detail information. Therefore, we use the channel-level Transformer in the last stage to obtain global detail information on the original resolution image.

III. Methodology

A. Model overview

As illustrated in Fig. 1, the MSTNet framework comprises three stages, with each using a different type of Transformer model to capture both spatial and channel information of the image. In the first two stages, we combine window-based Transformers with U-Net to capture contextual information. However, frequent downsampling in these stages inevitably leads to the loss of some fine details. Therefore, in the last stage, we employ channel-level Transformers to capture global detail information on the original resolution image. A fusion module is designed to combine the features from different Transformer branches. The specific design of the multi-stage network is as follows:

(1) In the first stage, the WTB-UNet network incorporates a Window-based Transformer Block (WTB) into the UNet architecture. The WTB captures spatial information by extracting the relationship between each pixel and its surrounding pixels within each window. Subsequently, a Supervised Attention Module (SAM) [7] is employed to suppress features with

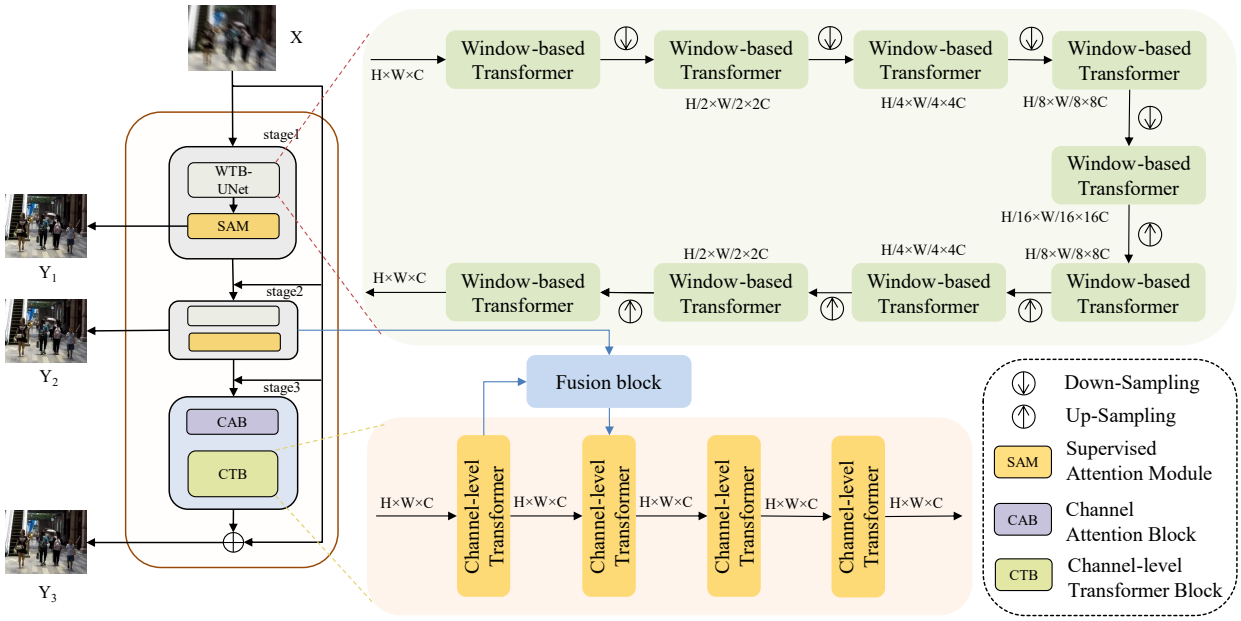


Fig. 1. The model overview of the proposed MSTNet. MSTNet consists of three stages. The first two stages mainly include a WTB-UNet block and a Supervised Attention Module (SAM) block, and the third stage includes a Channel Attention Block (CAB) and Channel-level Transformer Block (CTB).

low information content and enable only useful features to propagate to the next stage.

(2) The second stage of our network is similar to the first stage, including WTB-UNet and SAM.

(3) The last stage includes the Channel Attention Block (CAB) and Channel-level Transformer Block (CTB). The CTB consists of multiple channel-level Transformer modules, which capture global channel interaction information. After passing through the three-stage network, the original image X is restored and results in a clear, high-quality image Y_3.

To balance the trade-off between the pixel-wise accuracy and the perceptual quality of the image restoration, we use a combination of edge loss function and Charbonnier loss function [23], which can be expressed as:

$$L = \sum_{i=1}^{3} L_{char}\left(Y_i, Y\right) + \lambda L_{\text{edge}}\ \left(Y_3, Y\right)$$

$$L_{char} = \sqrt{\left\|Y_i - Y\right\|^2 + \varepsilon^2}$$

$$L_{\text{edge}}\ = \sqrt{\left\|\Delta\left(Y_i\right) - \Delta(Y)\right\|^2 + \varepsilon^2}$$

where Δ is the Laplacian operator.

B. Transformer Module

In MSTNet, we employ both window-based Transformer and channel-level Transformer to capture both spatial and channel information of the image. The core unit of Transformer for capturing image information lies in the self-attention mechanism. Fig. 2 illustrates the working principles

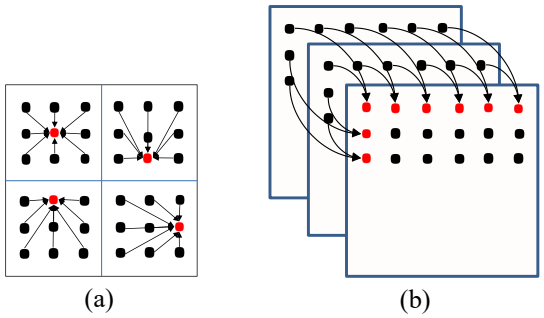


Fig. 2. (a) A window-based self-attention mechanism designed to capture the interrelations among spatial pixels; (b) A channel-level self-attention mechanism aimed at identifying the interconnections among pixels across channels.

of window-based self-attention (WSA) and channel-level self-attention (CSA). In the left figure, the red pixel represents the weighted sum of all pixels within the window, which demonstrates that WSA captures the interaction of spatial information within a window. In contrast, CSA operates at the channel level, where the red pixel in the right figure is obtained by multiplying the features of all the same positions on other channels by their weights. Therefore, CSA focuses on the interaction of channel information in the global scope. WSA can effectively obtain local spatial features, while CSA can capture global channel features. These two types of features are complementary and can improve the representation ability of the model. However, these two types of features have distinct characteristics and semantics, we further design a

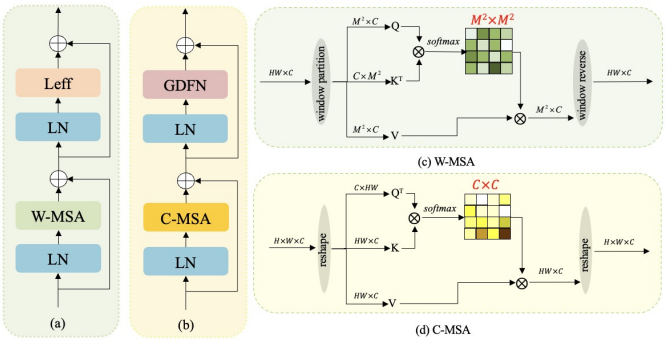


Fig. 3. (a)Window-based Transformer (b) Channel-level Transformer (c) Specific structure of W-MSA (Window-based Multi-head Self Attention) (d) Specific structure of C-MSA (Channel-wise Multi-head Self Attention)

feature fusion module to obtain a more comprehensive and accurate feature representation.

1) Window-based Transformer: The window-based Transformer can be mathematically expressed as:

$$X_l' = \text{W-MSA}\left(\text{LN}\left(X_{l-1}\right)\right) + X_{l-1}$$
$$X_l = \text{LeFF}\left(\text{LN}\left(X_l'\right)\right) + X_l$$

where X_l and X_{l-1} are the output and input features of the l-th layer, respectively, $W-MSA$ denotes the window-based multi-head self-attention mechanism, $LeFF$ denotes the feed-forward network [24], and LN denotes the layer normalization layer.

As shown in Fig. 3(c), $W-MSA$ first performs window partition operation on the feature $X \in R^{HW\times C}$, and splits the feature into multiple non-overlapping sub-features $X^i \in R^{M^2\times C}, i \in HW/M^2$. Then, $W-MSA$ uses linear projection to generate query (Q^i), key (K^{i^T}), and value (V^i) vectors and performs self-attention process:

$$\hat{X}^i = \text{Softmax}\left(Q^i K^{i^T}/\alpha + B\right) V^i$$

where $\hat{X}^i$ is the output of the self-attention process for the i-th sub-feature, α is a scaling factor and B is the relative position bias. The sub-features $\hat{X}^i$ are then concatenated by the window reverse operation to form a complete feature.

2) Channel-level Transformer: The channel-level Transformer can be mathematically expressed as:

$$X_l' = \text{C-MSA}\left(\text{LN}\left(X_{l-1}\right)\right) + X_{l-1}$$
$$X_l = \text{GDFN}\left(\text{LN}\left(X_l'\right)\right) + X_l'$$

where $C-MSA$ denotes the channel-level multi-head self-attention mechanism, and $GDFN$ denotes the gated dense fusion network. For the $C-MSA$ model.

As shown in Fig. 3(d), $C-MSA$ first performs a reshape operation on the feature $X \in R^{H\times W\times C}$, which transforms the feature into $Y \in R^{HW\times C}$. Then $C-MSA$ uses linear projection to generate query (Q), key (K) and value (V) vectors and performs self-attention process:

$$\hat{X} = V \cdot \text{Softmax}\left(S_K\left(Q^T K/\alpha\right)\right)$$

where $\hat{X}$ is the output of the self-attention process, and α is a trainable network parameter. The output $\hat{X}$ is then reshaped back to $\hat{X} \in R^{H\times W\times C}$. S_k denotes a learnable top-k selection operator, which can be expressed as:

$$[S_k(w)_{ij}] = \begin{cases} w_{ij} & w_{ij} \in \text{top}-k(\text{row}(j)) \\ 0 & \text{otherwise} \end{cases}$$

Since the channel-level self-attention mechanism is computed globally, some unimportant or even detrimental weights should be discarded. Thereforewe introduce a sparse attention method [25] to make the dense attention map into a sparse attention map, which first performs a softmax operation on each row, and then retains the top-k weights, while setting the rest of the weights to zero.

C. Fusion Module

Our model uses a fusion module (Fusion Block) to fuse the features from the window-trans and channel-trans stages. The window-trans stage extracts local spatial features within each window, while the channel-trans stage extracts global channel-wise features across the whole image. The features from the two stages have different properties and meanings, and need to be fused properly to obtain a complete and accurate feature representation for image restoration.

Existing methods for feature fusion often use simple fusion methods such as addition or multiplication, which may cause information loss or conflict when combining features from different sources. To address this issue, we propose a novel fusion module that can dynamically adjust the fusion weights based on the similarity and importance of the two stage features, using a convolutional operation. The fusion module consists of four steps, as shown in Fig. 4:

1) Reshape: We first reshape the feature from the window-trans stage $X_w \in R^{H\times W\times C}$ to match the dimension of the feature from the channel-trans stage $X_c \in R^{H\times W\times C}$ by applying a linear projection layer: $X_w = W_r X_w$, where $W_r \in R^{C\times C}$ is a learnable weight matrix.

2) Concatenate: We then concatenate the two stage features along the channel dimension, and obtain a fused feature $X_f \in R^{H\times W\times 2C}$: $X_f = [X_w, X_c]$.

3) Channel-wise Fusion: We apply a channel-wise fusion operation to the fused feature X_f, which can capture both the local spatial and global channel-wise information from the two stage features and achieve the fusion of spatial features and channel features. This operation consists of two steps: (a) We apply a 1×1 convolution layer to the fused feature X_f, followed by a ReLu activation. This step can reduce the channel dimension and increase the non-linearity of the feature. (b) We apply another 1×1 convolution layer to the output of the previous step, followed by a ReLu activation.

4) Output: The output of the fusion module, $X_o \in R^{H\times W\times C}$, is obtained by applying another linear projection layer to the fused feature X_f: $X_o = W_o X_f$, where $W_o \in R^{2C\times C}$ is a learnable weight matrix. The output feature X_o is then fed into the channel-level transformer for further processing. The fusion module can effectively

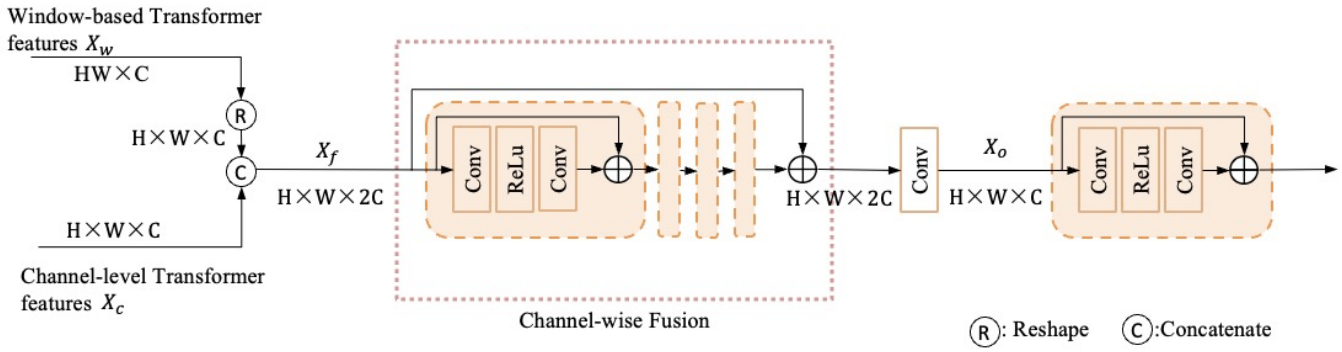


Fig. 4. Fusion module overview

integrate the features from the window-based transformer and the channel-level transformer, and produce a more informative and expressive feature for image restoration.

IV. Experiments

A. Experimental Settings

1) Experimental Configurations: Similar to the traditional Transformer model, we employed the AdamW optimizer [26] to train the MSTNet model. The momentum term was set to (0.9, 0.999), and the weight decay coefficient was set to 0.02. The initial learning rate was set to 2e-4. During the training process, we utilized a cosine decay strategy to gradually decrease the learning rate. This strategy allows the model to quickly adapt to the data in the early stages of training and gradually reduce the learning rate in the later stages to aid better convergence. The learning rate was ultimately reduced to 1e-6.

2) Parameters setting: In all Window-based Transformer blocks, we set the window size M to 8, and the initial number of channels in each stage is set to 32. In WTB-Unet, the number of window-based Transformers in each encoder and decoder is $[1, 1, 4, 4, 2, 4, 4, 1, 1]$. In CTB, the number of channel-level Transformers in each layer is $[1, 3, 3, 5]$.

3) Data Augmentation: To increase the diversity of training samples, we employ horizontal flipping, randomly flipping the image horizontally as well as rotating the image by 90, 180, or 270 degrees to generate new training samples. These augmentations serve to foster a deeper understanding of the intrinsic characteristics of images, thus bolstering accuracy and robustness.

4) Evaluation Metrics: We utilize commonly used image restoration evaluation metrics, such as PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index) [27], and model parameters, to comprehensively evaluate the model performance. These evaluation metrics provide a comprehensive quantification and comparison of the model's performance in image restoration tasks, considering factors such as the quality and structure of the restored images.

B. Datasets

We evaluate the model performance on a series of challenging datasets, with a particular focus on its ability to address image degradation issues commonly encountered in real-world scenarios, such as image blur and image noise.

1) Deblurring Datasets: (a) The **GoPro dataset** [21], a synthetic dynamic blur dataset consisting of 3,214 images of 1280 ×720 resolution, segmented into 2,103 training and 1,111 test images. Utilizing a high-speed camera to capture sharp images, these are subsequently blurred using a motion blur kernel to simulate real-world blurring effects. (b) The **HIDE dataset** [28], a real-world human-aware motion deblurring dataset, consists of 8,422 image pairs alongside 65,784 densely annotated human bounding boxes, leveraging a dual-capture system to align blurry and sharp image pairs, followed by post-processing to ensure high-quality sharp images. (c) The **RealBlur dataset** [29], consists of two part: RealBlur-R and RealBlur-J, with 4,738 pairs of real blurry and sharp images across 232 scenes, employs a similar aligned capture system and post-processing for clarity.

2) Denoising Datasets: (a) The **SIDD dataset** [30] offers 30,000 noisy images from 10 scenes, captured by five distinct smartphone cameras under varied lighting, accompanied by corresponding sharp images as ground truth. (b) The **DND dataset** [29] includes 50 real-world dynamic blurry image pairs, captured with consumer-grade cameras, illustrating the practical applicability of the proposed approach.

C. Image deblurring

We evaluate our method on the image deblurring task using the GoPro [21] dataset. We train our model on 256×256 patches cropped from the GoPro training images. To demonstrate the generalization and robustness of our model, we also apply it to the HIDE [28] and RealBlur [29] datasets. Table I shows that our method outperforms all the existing methods on all four datasets in terms of PSNR and SSIM. Specifically, our MSTNet achieves 33.40dB and 31.53dB in PSNR on the GoPro and HIDE datasets, respectively, surpassing the second-best methods by 0.34dB and 0.31dB. On the RealBlur-R and RealBlur-J [29] datasets, our method also produces

TABLE I
IMAGE DEBLURRING RESULTS. OUR MSTNET IS TRAINED ONLY ON THE GOPRO DATASET AND DIRECTLY APPLIED TO THE HIDE AND REALBLUR BENCHMARK DATASETS.

	GoPro		HIDE		RealBlur-R		RealBlur-J	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Xu et al. [3]	21.00	0.741	-	-	34.46	0.937	27.14	0.830
Nah et al. [21]	29.08	0.914	25.73	0.874	32.51	0.841	27.87	0.827
DeblurGAN [31]	28.70	0.858	24.51	0.871	33.79	0.903	27.97	0.834
SRN [22]	30.26	0.934	28.36	0.915	35.66	0.947	28.56	0.867
Zhang et al. [32]	29.19	0.931	-	-	35.48	0.947	27.80	0.847
DeblurGAN-v2 [33]	29.55	0.934	26.61	0.875	35.26	0.944	28.70	0.866
DMPHN [34]	31.20	0.940	29.09	0.924	35.70	0.948	28.42	0.860
DBGAN [35]	31.10	0.942	28.94	0.915	-	-	-	-
MTRNN [36]	31.15	0.945	29.15	0.918	35.79	0.951	28.44	0.862
SPAIR [37]	32.06	0.953	30.29	0.931	-	-	28.81	0.875
MPRNet [7]	32.66	0.959	30.96	0.939	35.99	0.952	28.70	0.873
Uformer [17]	32.97	0.967	30.83	0.952	36.22	0.957	29.06	0.884
Restormer [5]	32.92	0.961	31.22	0.942	36.19	0.957	28.96	0.879
MHNet [38]	33.06	0.969	31.14	0.950	-	-	-	-
Ours	**33.40**	**0.969**	**31.53**	**0.958**	**36.27**	**0.958**	**29.15**	**0.890**

superior results with clear and faithful image restoration. Fig. 5 illustrates the visual comparisons on the GoPro and HIDE datasets, where our method recovers more details and colors than the other methods.

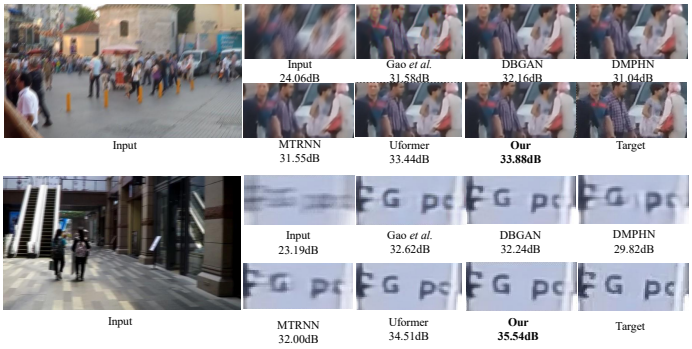


Fig. 5. Visual comparisons with SoTA methods for the image deblurring. **Top**: on the GoPro [21] dataset. **Below**: on the HIDE [28] dataset.

TABLE II
IMAGE DENOISING RESULTS. OUR MSTNET IS TRAINED ONLY ON THE SIDD IMAGES AND DIRECTLY TESTED ON DND.

	SIDD		DND	
Method	PSNR	SSIM	PSNR	SSIM
DnCNN [39]	23.66	0.583	32.43	0.790
BM3D [40]	25.65	0.685	34.51	0.851
RIDNet [41]	38.71	0.914	39.26	0.953
AINDNet [42]	38.00	0.954	39.07	0.951
VDN [43]	39.28	0.918	39.32	0.952
DANet [9]	39.47	0.909	39.59	0.955
CycleISP [44]	39.52	0.957	39.56	0.956
MIRNet [45]	39.72	0.959	39.88	0.956
MBNet [12]	39.75	0.959	39.89	0.955
MPRNet [7]	39.71	0.958	39.80	0.954
Uformer [17]	39.80	0.960	39.98	0.955
Restormer [5]	**40.02**	0.960	40.03	0.956
Ours	39.92	**0.960**	**40.06**	**0.956**

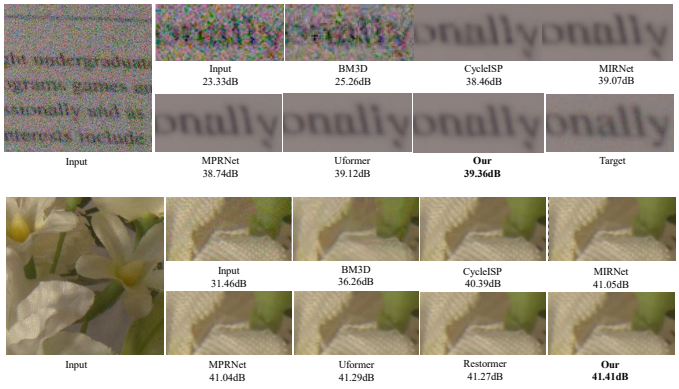


Fig. 6. Visual comparisons with SoTA methods for the image denoising. **Top:** on the SIDD [30] and **Below:** on the DND [29] datasets .

D. Image denoising

We evaluate our method on the image denoising task and compare it with several SoTA methods in two benchmark datasets SIDD [30] and DND [29]. The SIDD dataset is with synthetic noise generated by different smartphone cameras under various lighting conditions. We adopt the dataset split strategy [17] on the SIDD dataset to obtain the training and testing sets. The DND dataset contains 50 testing image pairs with real noise captured by different cameras and sensors. We train our model on the SIDD training images and test it on both datasets including DND to further validate the generalization of the model.

Table II shows the quantitative results of our method and other methods on the SIDD and DND datasets in terms of

PSNR and SSIM. Our method achieves the best performance on the DND dataset, which is a challenging real-world dataset with complex and diverse noise patterns. This demonstrates the robustness and generalization ability of our method. On the SIDD dataset, our method also performs competitively, achieving the second-best PSNR and the best SSIM among all the methods. Fig. 6 shows the visual comparisons of our method and other methods on the SIDD and DND datasets. We can see that some methods tend to produce over-smoothed or distorted results, losing some fine details and textures. Some methods can preserve some details, but also leave some noise residues or artifacts. Our method can balance the trade-off between noise removal and detail preservation, producing natural and realistic results.

E. Ablation Study

We conduct ablation experiments to validate the effectiveness of designed modules in proposed MSTNet and their impact on the performance of image restoration tasks. Considering the time cost of training large models based on transformers, we randomly cropped the images in the GoPro training dataset into 128×128 patches for training.

Transformer-based Multi-stage Network: We perform experiments to assess the influence of employing window-based Transformers and channel-level Transformers at distinct stages of the network for image restoration tasks. As outlined in Table III, the first letter W/C denotes the type of transformer utilized in the initial two stages, whereas the second letter W/C signifies the type of transformer employed in the third stage. Notably, the third row labeled W-C corresponds to our proposed method. From Table III, we can observe that our proposed method achieves the best restoration results. We can also see that using window-based Transformer in the first two stage is more effective than using channel-level Transformer, as it can capture more global information and reduce the blur level. On the other hand, using channel-level Transformer in the last stages is more beneficial than using window-based Transformer, as it can focus more on the local details and enhance the image quality. Therefore, our method combines the advantages of both types of Transformers and achieves a good balance between global and local information processing.

TABLE III
EFFECT OF THE TRANSFORMER BLOCK ON THE IMAGE DEBLURRING.

	HIDE		RealBlur-J		RealBlur-R	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
W-W	30.08	0.945	28.74	0.873	36.06	0.955
C-C	30.26	0.946	28.72	0.870	36.14	0.955
W-C	**30.52**	**0.948**	**28.93**	**0.879**	**36.19**	**0.956**

We investigate the impact of the number of stages on our model's performance by conducting experiments on the GoPro dataset. Table IV shows that our model achieves higher PSNR scores as the number of stages increases from one to three, which validates the effectiveness of our multi-stage design. Moreover, our model is also relatively computationally efficient compared with other advanced methods. Although we have a multi-stage transformer network, the total number of parameters and model size is comparable to other methods. Specifically, the main computational cost of Transformer comes from the self-attention mechanism (SA), which has quadratic time and memory complexity with respect to the spatial resolution of the input, i.e., for an image with W (width) × H (height) pixels, the time and memory complexity are $\mathrm{O}\left(\mathrm{W}^2\mathrm{H}^2\mathrm{C}\right)$, where C is the number of channels. However, window-based SA and channel-level SA have complexities of $O\left(M^2HWC\right)$ and $O\left(HWC^2\right)$, respectively, where M is the number of slices in a window.

TABLE IV
STAGE-WISE DEBLURRING PERFORMANCE OF MSTNET ON GOPRO [21].

				MSTNet		
Method	DeblurGAN-v2 [33]	DMPHN [34]	Uformer [17]	1-stage	2-stage	3-stage
PSNR	29.55	31.20	32.9	32.37	33.21	33.40
Params(M)	60.9	21.7	50.8	30.50	59.98	62.71

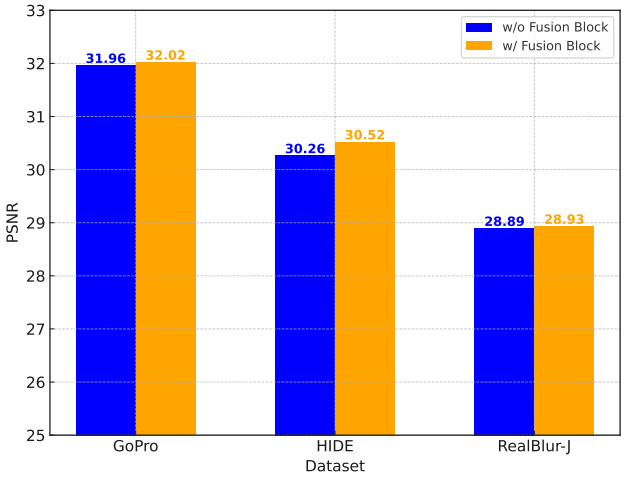
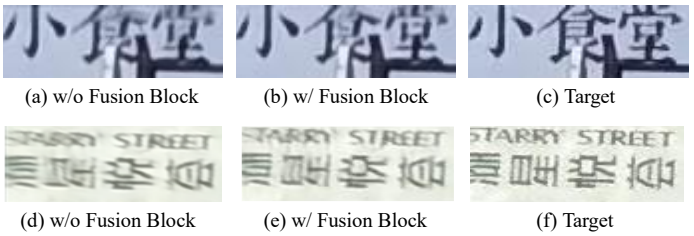


Fig. 7. Quantitative effect of the Fusion Block.



(a) w/o Fusion Block (b) w/ Fusion Block (c) Target

(d) w/o Fusion Block (e) w/ Fusion Block (f) Target

Fig. 8. Visual effect of the Fusion Block.

Fusion Module: We evaluated the fusion block on the image deblurring task using the HIDE [28] and RealBlur-J [29] datasets. As shown in Fig. 7, the fusion block increased the PSNR by 0.26dB on the HIDE dataset and by 0.04dB on the RealBlur-J dataset, demonstrating its effectiveness. Fig. 8 shows some visual comparisons with and without the fusion block. We can see that the fusion block can reduce more blur and achieve better restoration quality.

Loss function: Our loss function $L_{char} + L_{edge(3)}$ effectively performs image restoration by combining the edge loss function and the Charbonnier loss function. Specifically, we only utilize the edge loss function in the third stage to capture as much detail as possible, as shown in Figure 9. In contrast, L_{char} represents only the Charbonnier loss function is used in all three stages, while $L_{char} + L_{edge}$ indicates both the edge loss function and the Charbonnier loss function are adopted in all three stages. The experimental results clearly demonstrate the superior performance of our loss function in image restoration tasks.

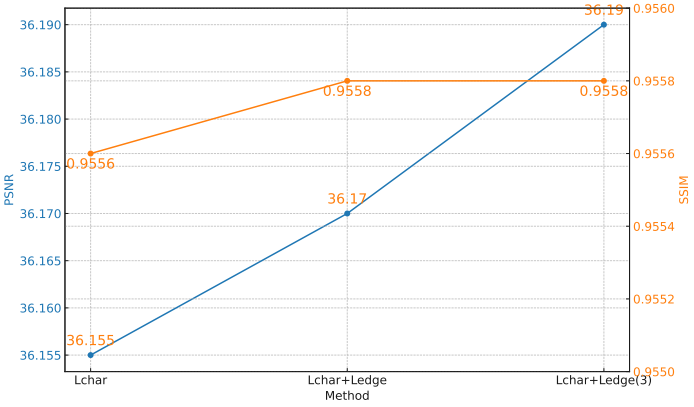


Fig. 9. Visual comparisons with SoTA methods on the SIDD [30] and DND [29] datasets for the image denoising.

V. Conclusions

This paper proposed MSTNet, a novel multi-stage progressive image restoration network based on a blend of local-global Transformers. Our multi-stage model jointly leveraged window-based Transformer and channel-level Transformer modules to capture both spatial and channel information of the image, which are complementary and enhance the representation ability of the model. In addition, the designed transformer fusion module and loss function also contribute to the improvement of image restoration performance. The experiments demonstrate that our method can significantly overcome the two typical challenges of image restoration in real-world scenarios, namely image deblurring and denoising and our model achieved superior results compared to existing methods.

References

[1] H. Yang, "Rethinking image and video restoration: An industrial perspective," *restoration*, vol. 3, p. 20, 2022.

[2] T. Liu, B. Li, X. Du, B. Jiang, L. Geng, F. Wang, and Z. Zhao, "Fair: Frequency-aware image restoration for industrial visual anomaly detection," *arXiv preprint arXiv:2309.07068*, 2023.

[3] L. Xu, S. Zheng, and J. Jia, "Unnatural l0 sparse representation for natural image deblurring," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1107–1114.

[4] C. R. Steffens, L. R. Messias, P. J. Drews-Jr, and S. S. d. C. Botelho, "Cnn based image restoration: Adjusting ill-exposed srgb images in post-processing," *Journal of Intelligent & Robotic Systems*, vol. 99, pp. 609–627, 2020.

[5] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.

[6] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

[7] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 821–14 831.

[8] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6360–6376, 2021.

[9] Z. Yue, Q. Zhao, L. Zhang, and D. Meng, "Dual adversarial network: Toward real-world noise removal and noise generation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 2020, pp. 41–58.

[10] A. Abuolaim and M. S. Brown, "Defocus deblurring using dual-pixel data," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 2020, pp. 111–126.

[11] X. Liu, M. Suganuma, Z. Sun, and T. Okatani, "Dual residual networks leveraging the potential of paired operations for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7007–7016.

[12] S. Cheng, Y. Wang, H. Huang, D. Liu, H. Fan, and S. Liu, "Nbnet: Noise basis learning for image denoising with subspace projection," in *Computer Vision and Pattern Recognition*, 2021, pp. 4896–4906.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image rethinkcognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9355–9366, 2021.

[15] Z. Huang, Y. Ben, G. Luo, P. Cheng, G. Yu, and B. Fu, "Shuffle transformer: Rethinking spatial shuffle for vision transformer," *arXiv preprint arXiv:2106.03650*, 2021.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[17] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image rethinkstoration," in *Proceedings of the IEEE/CVF conferethinknce on computer vision and pattern rethinkcognition*, 2022, pp. 17 683–17 693.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[19] X. Fu, B. Liang, Y. Huang, X. Ding, and J. Paisley, "Lightweight pyramid networks for image deraining," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 6, pp. 1794–1807, 2019.

[20] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 254–269.

[21] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3883–3891.

[22] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8174–8182.

[23] J. T. Barron, "A general and adaptive robust loss function," in *Proceedings of the IEEE/CVF Conferethnce on Computer Vision and Pattern rethinkcognition*, 2019, pp. 4331–4339.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in

Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer, 2020, pp. 213–229.

[25] X. Chen, H. Li, M. Li, and J. Pan, "Learning a sparse transformer network for effective image deraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5896–5905.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[28] Z. Shen, W. Wang, X. Lu, J. Shen, H. Ling, T. Xu, and L. Shao, "Human-aware motion deblurring," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5572–5581.

[29] J. Rim, H. Lee, J. Won, and S. Cho, "Real-world blur dataset for learning and benchmarking deblurring algorithms," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer, 2020, pp. 184–201.

[30] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1692–1700.

[31] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8183–8192.

[32] J. Zhang, J. Pan, J. Ren, Y. Song, L. Bao, R. W. Lau, and M.-H. Yang, "Dynamic scene deblurring using spatially variant recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2521–2529.

[33] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8878–8887.

[34] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5978–5986.

[35] K. Zhang, W. Luo, Y. Zhong, L. Ma, B. Stenger, W. Liu, and H. Li, "Deblurring by realistic blurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2737–2746.

[36] D. Park, D. U. Kang, J. Kim, and S. Y. Chun, "Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 327–343.

[37] K. Purohit, M. Suin, A. Rajagopalan, and V. N. Boddeti, "Spatially-adaptive image restoration using distortion-guided networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2309–2319.

[38] H. Gao and D. Dang, "Mixed hierarchy network for image restoration," *arXiv preprint arXiv:2302.09554*, 2023.

[39] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[40] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[41] S. Anwar and N. Barnes, "Real image denoising with feature attention," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3155–3164.

[42] Y. Kim, J. W. Soh, G. Y. Park, and N. I. Cho, "Transfer learning from synthetic to real-noise denoising with adaptive instance normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3482–3492.

[43] Z. Yue, H. Yong, Q. Zhao, D. Meng, and L. Zhang, "Variational denoising network: Toward blind noise modeling and removal," *Advances in neural information processing systems*, vol. 32, 2019.

[44] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Cycleisp: Real image restoration via improved data synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2696–2705.

[45] ——, "Learning enriched features for real image restoration and enhancement," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer, 2020, pp. 492–511.