TASK-ADAPTIVE FEATURE MATCHING LOSS FOR IMAGE DEBLURRING

Chiao-Chang Chang[1,3,⋆], *Bo-Cheng Yang*[2,3,⋆], *Yi-Ting Liu*[1,⋆], *Jun-Cheng Chen*[3], *I-Hong Jhuo*[4], *Yen-Yu Lin*[1,3]

[1]National Yang Ming Chiao Tung University, [2]National Tsing Hua University
[3]Academia Sinica, [4]Microsoft
melon88954.cs07@nycu.edu.tw, leoyang890501@gapp.nthu.edu.tw, yitingliu.cs11@nycu.edu.tw,
pullpull@citi.sinica.edu.tw, ihjhuo@gmail.com, lin@cs.nycu.edu.tw

ABSTRACT

Image deblurring is a highly challenging and ill-posed image restoration problem. Contemporary deep learning-based approaches usually tackle this problem by exploiting the encoder-decoder-based models trained by the commonly used mean squared error loss with the feature matching loss as a regularization to obtain perceptual consistent restored results as the ground truths. We argue that since the general backbone models for computing feature matching loss are usually not trained on the image deblurring task, the loss lacks specific knowledge of blur and usually leads to suboptimal performance. To address this issue, we propose a task-adaptive feature matching loss for image deblurring where we synthesize blurred images in different blur extents and employ triplet loss to finetune the backbone model for learning specific blur priors. Then, we leverage the finetuned backbone to compute feature matching loss which can greatly enhance the existing image deblurring models for better perceptual results. With extensive experiments on the GoPro and RealBlur datasets, both qualitative and quantitative results show that the SOTA deblurring models trained with the proposed loss can effectively obtain better and sharper restored images in terms of various perceptual image quality metrics than the original models while maintaining comparable PSNR and SSIM performances.

Index Terms— CLIP, feature matching loss, image deblurring

1. INTRODUCTION

Image deblurring is one of the most challenging image restoration tasks which has been studied for decades in the computer vision community. It is a highly ill-posed inverse problem with infinite potential solutions since it has to recover a sharp image from its blur version and to estimate blur kernel. Thus, it requires proper priors as the regularization to correctly recover the accurate sharp image.

Deep learning-based image deblurring approaches usually employ the encoder-decoder-based models [1] to solve this problem trained by the commonly used mean squared error (MSE) loss and the like with the feature matching loss as a regularization. The feature matching loss [2] is computed by taking the L_p-norm distance between the features of the restored and ground-truth images extracted from the intermediate layers of pretrained deep networks (e.g., AlexNet or VGGNet) on the ImageNet dataset. However, since these pretrained models for computing feature matching loss are usually trained for other image tasks instead of image deblurring, the loss lacks specific knowledge of blur and usually leads to suboptimal restored results as shown in Fig. 1.

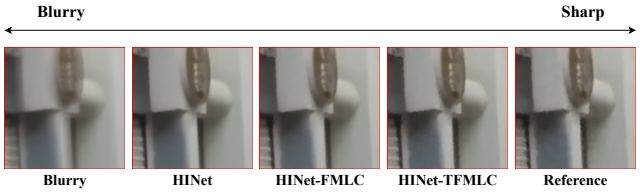

Fig. 1. Restored results of HINet [3] and our proposed task-adaptive feature matching loss based on CLIP (TFMLC). It illustrates that the restored image by the deblurring model (HINet) without employing the feature matching loss usually lacks texture details and is usually smoother than the one with (the proposed HINet-TFMLC) which is perceptually closer to the ground truth.

To address this issue, we propose a task-adaptive feature matching loss where we first adapt the pretrained deep networks in self-supervised learning paradigm for image deblurring. For this purpose, we synthesize blurred images in different blur extents from sharp moving videos (GoPro dataset) and employ triplet loss [4] to finetune the model to learn specific blur priors. Then, we can compute the feature matching loss using the finetuned model as usual to enhance the existing image deblurring models to generate the restored images which are perceptually closer to the ground truths. In our work, we employ Contrastive Language-Image Pretraining (CLIP) [5] released by OpenAI as the backbone model to compute the feature matching loss where the pretrained CLIP model is trained contrastively using 400 million image-text pairs and shows strong zero-shot capability for various vision and language tasks, including zero-shot image recognition, zero-shot reference-free quality metrics for image captioning (i.e., CLIPScore [6]), etc. We thus name our proposed loss after task-adaptive feature matching loss based on CLIP (TFMLC) and the one without adaptation after feature matching loss based on CLIP (FMLC). We apply the proposed methods with two state-of-the-art image deblurring models, MPRNet [7] and HINet [3], which are trained with their original losses along with the proposed FMLC and TFMLC and conduct careful evaluations on the challenging GoPro, RealBlur-J, and RealBlur-R datasets. Both qualitative and quantitative experimental results show that the deblurring models trained with the proposed loss are able to obtain sharper restored images with lower Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [8], Perceptual index (PI) [9], Learned Perceptual Image Patch Similarity (LPIPS) [2], and Perceptual Image-Error Assessment through Pairwise Preference (PieAPP) [10] scores which are perceptually closer to the ground truths than the original

⋆ The first three authors equally contribute to this work. The project is supported by National Science and Technology Council (NSTC), Taiwan (R.O.C) under grant numbers of 110-2221-E-001-009-MY2, 111-2221-E-001-002, 111-2634-F-002-022, 111-2628-E-A49-025-MY3, 109-2221-E-009-113-MY3.

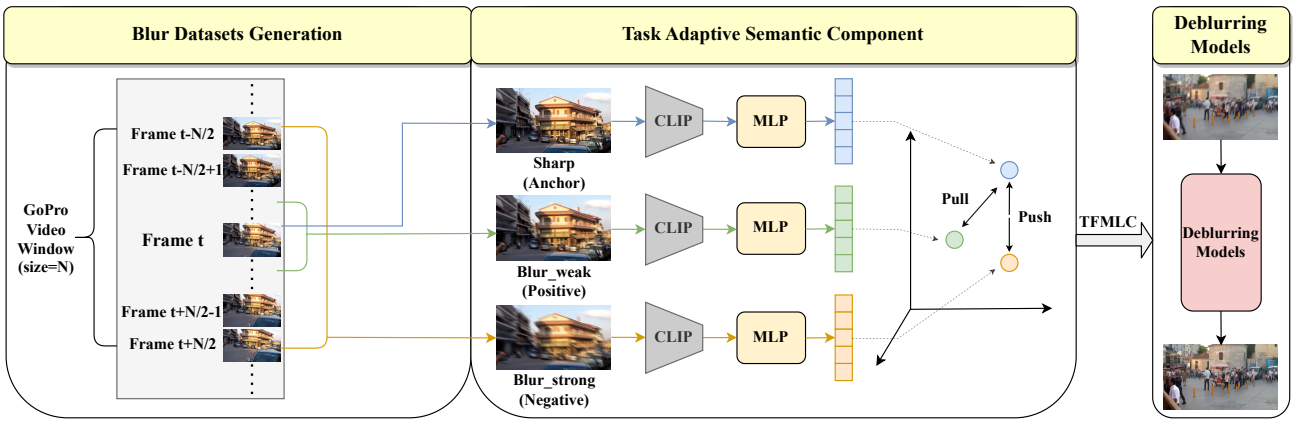

Fig. 2. The overview of the proposed task-adaptive feature matching loss based on CLIP (TFMLC). We first synthesize blurred images in different blur extents using the sharp moving videos from the GoPro dataset by taking the average frame from different sizes of video clips where longer clips usually result in more blurry average frames, and the center frame t is used as the reference sharp image. Then, we add additional MLP layers over the backbone model (CLIP) and finetune the MLPs while freezing the backbone using the triplet loss for feature adaptation. Finally, we compute the final TFMLC loss using the adapted MLP features as the regularization for image deblurring models.

models while maintaining comparable Peak Signal-to-Noise Ratio (PSNR) [11] and Structural Similarity (SSIM) [12] performances.

2. RELATED WORK

In this section, we briefly review relevant works on image quality metric and image deblurring.

Image Quality Metric: Image quality metrics are classified into two categories, subjective and objective metrics. For image deblurring, most existing models adopt objective metrics, which are divided into other two types: full-reference [11, 12, 13, 14, 15, 16, 17, 2, 18, 9, 10] and no-reference metrics [19, 20, 8, 21, 22, 23, 24]. The difference between these two methods is whether the restored images are compared with ground-truth images for evaluation. Although commonly used PSNR [11] and SSIM [12] metrics are simple to calculate and have clear physical meanings but hard to address many subtleties of human perception. On the other hand, the non-reference metric, BRISQUE [8] and PI [9], and full-reference metrics, LPIPS [2] and PieAPP [10], focus on perceptual information can better capture the perceptual similarity between images in agreement with human judgments than PSNR and SSIM. BRISQUE uses scene statistics of locally normalized luminance coefficients to quantify possible losses of "naturalness" in the image. PI combines two other non-reference image metrics and is highly correlated with the ratings of human observers. LPIPS is computed based on the intermediate features of deep networks which are pretrained on the ImageNet dataset and encode rich visual information to better capture the perceptual relation between images than other metrics. PieAPP learns to predict image perceptual error from a large-scale dataset and is well-correlated with human opinion. As a result, we pay attention to the perceptual image evaluation using the BRISQUE, PI, LPIPS, and PieAPP metrics in our paper.

Image Deblurring: With the recent breakthroughs in deep learning, people have widely started to solve the image deblurring tasks by using various deep generative models to restore the sharp images without estimating the blur kernel. Yang et al. [25] propose a two-branch deep auto-encoder framework for image deblurring to focus on high blur region with the help of the motion information from the event camera and its attention modules. Zhang et al. [26] propose two generative adversarial networks which jointly learn how to blur and how to deblur to close the gap between synthetic and real blurs. Purohit et al. [27] introduce a region adaptive dense deformable modules with a self-attentive module into a densely connected encoder-decoder design for significantly improved accuracy and speed. Liang et al. [28] propose a transformer-based method with self-attention and image warping modules to better capture the blur characteristics for improved deblurring performance on numerous challenging blur benchmarks. In this paper, we apply our proposed methods to two state-of-the-art multi-stage deblurring models MPRNet [7] and HINet [3] due to their superior performances where MPRNet adopts supervised attention module (SAM) and cross-stage feature fusion (CSFF) modules to learn the blur characteristics for effective deblurring. Similarly, HINet leverages half instance normalization block (HIN Block) along with the same SAM and CSFF modules for image deblurring.

3. THE PROPOSED APPROACH

In this section, we present the proposed task adaptive feature matching loss for image deblurring which is computed in two phases: (1) performing feature adaptation of the pretrained backbone model towards image deblurring and (2) training the image deblurring with the feature matching loss with the adapted features. Furthermore, we utilize the features extracted from the pretrained CLIP model [5] as the backbone model to compute feature matching loss based on CLIP (FMLC) and the proposed task-adaptive feature matching loss based on CLIP (TFMLC) to provide additional regularization for two state-of-the-art image deblurring models MPRNet [7] and HINet [3]. The overview of the proposed framework is illustrated in Fig. 2 and the computation details are described as follows.

Self-supervised Feature Adaptation for Image Deblurring: Since most of the backbone models for computing feature matching losses are not trained specifically for the image deblurring task, this usually leads to suboptimal performance when we train the deblurring models with them. To address this issue, we first propose a self-supervised learning strategy to synthesize blur images in different blur extents. We then append additional multilayer perceptrons (MLPs) to the backbone model followed by finetuning the MLP layers with triplet loss [4] which enforces the geometric constraint for

Table 1. The evaluation results of the proposed FMLC and TFMLC losses and compared baselines on the GoPro, RealBlur-J, and RealBlur-R datasets. Best scores are **highlighted**. Our proposed TFMLC obtains the best BRISQUE, PI, LPIPS, and PieAPP values and comparable PSNR and SSIM values simultaneously, while the proposed FMLC is the second best method.

		GoPro			RealBlur-J			RealBlur-R		
	Method	Origin	FMLC	TFMLC	Origin	FMLC	TFMLC	Origin	FMLC	TFMLC
MPRNet	PSNR↑	**31.8819**	31.8394	31.6548	26.4852	**26.4867**	26.3997	**33.9438**	33.9036	33.7888
	SSIM↑	**0.9600**	0.9599	0.9593	0.8484	**0.8488**	0.8469	**0.8083**	0.8075	0.8061
	BRISQUE↓	52.4175	51.0537	**46.3271**	48.2004	47.6759	**44.5331**	65.7721	65.5988	**63.3873**
	PI↓	5.2686	4.9309	**4.2440**	5.0767	4.8106	**4.2586**	7.2524	7.0697	**6.7775**
	LPIPS↓	0.1013	0.0844	**0.0640**	0.1612	0.1509	**0.1387**	0.0778	0.0714	**0.0668**
	PieAPP↓	0.7435	0.6963	**0.6222**	1.1182	1.0923	**1.0602**	0.6880	0.6626	**0.6285**
HINet	PSNR↑	**32.7712**	32.6660	32.4385	**26.3620**	26.3376	26.2858	**33.8045**	33.7690	33.7079
	SSIM↑	**0.9593**	0.9581	0.9553	**0.8539**	0.8519	0.8491	**0.9467**	0.9460	0.9441
	BRISQUE↓	52.0361	47.7858	**42.7883**	44.6924	44.0819	**41.8833**	66.1007	64.7499	**57.1558**
	PI↓	5.1837	4.5979	**4.1837**	5.0242	4.7747	**4.5482**	7.2053	7.0199	**6.6362**
	LPIPS↓	0.0904	0.0649	**0.0554**	0.1730	0.1685	**0.1653**	0.0775	0.0727	**0.0717**
	PieAPP↓	0.6562	0.5943	**0.5484**	1.2983	1.2821	**1.2720**	0.7259	0.7031	**0.6616**

better generalization to help the model learn the blur prior through the task of ranking the blur extents where the backbone is frozen during finetuning as shown in Fig. 2.

$$\mathcal{L}_{triplet}(a,p,n) = \max\{\mathcal{D}(a,p) - \mathcal{D}(a,n) + m, 0\}, \quad (1)$$

where a, p, n are the anchor, positive, and negative samples, respectively and $m = 1$ is the margin. For blur data synthesis, we use the GoPro dataset (i.e., GOPRO_LARGE_all) [29] which consists of 22 sharp moving videos to generate a set of triplet data where each triplet is composed of sharp (anchor) a, weak blur (positive), and strong blur (negative) samples. We take the center frame from a temporal window over the GoPro video as the sharp sample, the average frame from 0.25 to 0.75 of the window in the normalized temporal coordinate as the weak blur sample, and the average frame of the whole window as the strong blur sample. Since there are static frames, we further utilize the Lucas Kanade optical flow implementation in OpenCV to dynamically determine the window size N_w where the accumulated pixel displacement of the motion vector between the starting and ending frames are more than 25 pixels and N_w is no less than 15 frames.

Loss Function: After feature adaptation, we then can compute two feature matching losses: Feature Matching Loss based on CLIP (FMLC) and Task-Adaptive Feature Matching Loss based on CLIP (TFMLC). FMLC is a simpler structure extracting the feature of CLIP backbone model without MLP. As a result, the difference between FMLC and TFMLC are whether we use the original backbone features or adapted MLP features to compute the losses, where we denote the feature matching loss as $\mathcal{L}_C$ which represents $\mathcal{L}_{FMLC}$ or $\mathcal{L}_{TFMLC}$ in the following loss functions. In this paper, we adjust the loss functions of the original deblurring models to show that both FMLC and TFMLC are helpful for the image deblurring task. The adjusted loss function of MPRNet with three stages is:

$$\mathcal{L}_{MPRNet} = \sum_{S=1}^{3} \mathcal{L}^S_{Char} + \lambda_M \cdot \mathcal{L}^S_{Edge} + \lambda_1 \cdot \mathcal{L}^S_C, \quad (2)$$

where $\mathcal{L}_{Char}$ is the Charbonnier loss [30] and $\mathcal{L}_{Edge}$ is the edge loss, and $\lambda_M = 0.05$ and $\lambda_1 = 1$. The adjusted loss function of HINet with two HIN Blocks is:

$$\mathcal{L}_{HINet} = \sum_{S=1}^{2} -\lambda_H \cdot \mathcal{L}^S_{PSNR} + \lambda_2 \cdot \mathcal{L}^S_C, \quad (3)$$

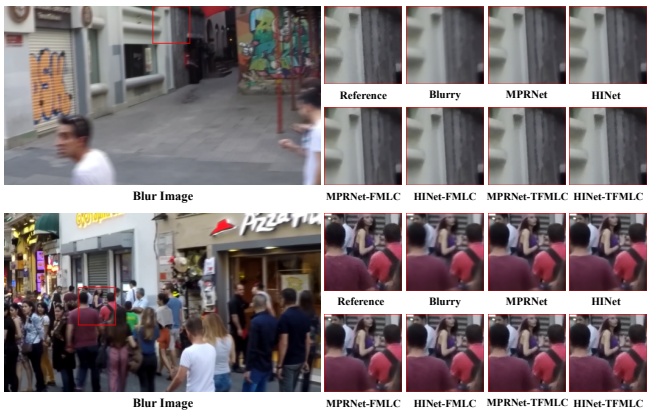

Fig. 3. Visual comparisons of the deblurring results of original MPRNet and HINet and the proposed FMLC and TFMLC enhanced models on the GoPro dataset.

where $\mathcal{L}_{PSNR}$ is the PSNR loss, and $\lambda_H = 0.5$ and $\lambda_2 = 1,000$. The parameters λ_1 and λ_2 maintain the balance between the losses of original deblurring models and our loss in similar scale.

4. EXPERIMENT

Datasets: As in [7, 3], we train the deblurring models with the proposed losses on the synthetic blur image dataset generated from the GoPro dataset [29], which contains 2,013 images for training and 1,111 images for evaluation. In addition, to verify the generalizability of the proposed method, we also evaluate the deblurring model trained using the GoPro dataset directly upon another real-world benchmark, RealBlur [31]. The RealBlur benchmark contains two datasets where the RealBlur-J dataset is generated from JPEG images processed by camera Image Signal Processor, and the RealBlur-R dataset is generated from camera raw images. The blurred images are generated with camera shakes in low-light environments where common motion blurs easily happen. Both datasets contain 980 images for evaluation.

Evaluation Metrics: Besides commonly used PSNR [11] and SSIM [12] metrics, we focus on BRISQUE, PI, LPIPS, and PieAPP

Table 2. The ablation results of image deblurring using the proposed FMLC loss computed using different backbone features of the pretrained CLIP model on the GoPro dataset. Best scores are **highlighted**. H-F means HINet trained using our proposed FMLC loss.

Method	PSNR/SSIM↑	BRISQUE/PI/LPIPS/PieAPP↓
HINet	**32.771/0.959**	52.036/5.184/0.090/0.656
H-F-RN50	32.669/0.958	47.882/4.838/0.073/0.638
H-F-RN101	32.637/0.958	44.873/4.755/0.068/0.622
H-F-ViT-B/32	32.176/0.953	43.466/4.384/0.059/**0.602**
H-F-ViT-B/16	32.047/0.952	41.124/**4.149/0.058**/0.623
H-F-ViT-L/14	31.760/0.949	**41.032**/4.204/0.060/0.649

performance to reveal the perceptual information of the restored images since they are more sensitive to perceptual information compared to traditional image metrics where lower BRISQUE, PI, LPIPS, and PieAPP indicate the image patches are more similar in reality and with higher perceptual quality.

Implementation Details: We employ the MLP features from the third residual attention block (ResBlock 3) of the ViT-B/16 backbone of the pretrained CLIP model for the proposed method. Furthermore, we add three additional MLP layers of 1,024 neurons with the ReLU and the batch normalization in-between MLPs to perform feature adaptation for image deblurring while the weights of the CLIP backbone are frozen during the process. We synthesize 1,105 triplets from 22 GoPro videos and train using the Adam optimizer [32] with a batch size of 64, initial learning rate as 1×10^{-4}, β_1 as 0.9, and β_2 as 0.999 until the loss converges. Then, we can compute the proposed TFMLC loss using the features from the last MLP layer. On the other hand, we apply the proposed loss to two deblurring models, MPRNet and HINet, along with their original losses through finetuning the models from their publicly available pretrained weights for faster training purposes. We follow most of their original training settings except the hyperparameters as follows. For **MPRNet**, we finetune the model from the officially released pretrained weights of its lite version with the initial and end learning rates as 1×10^{-6} and 1×10^{-9}, respectively. The networks are trained using cropped 256×256 patches from the image with the resolution of $1,280 \times 720$ with a batch size of 1 for 200 epochs. For **HINet**, we also finetune the model from its official pretrained weights with the learning rate as 2×10^{-5}. The networks are trained with 256×256 patches with a batch size of 4 for 2×10^4 iterations. Data augmentations include flipping and rotation operations. The image resolution of training samples is first cropped to 512×512 for data preprocessing, while that of evaluation samples is 1280×720 on the GoPro dataset and 669×760 on the RealBlur datasets.

Evaluation Results: In Table 1, we show the quantitative evaluation results of MPRNet and HINet with and without the FMLC and TFMLC losses on the GoPro, RealBur-J, and RealBlur-R datasets. The results show that the models trained with both FMLC and TFMLC can consistently achieve better perceptual image evaluation metrics than the original models across different datasets while the models employed the proposed TFMLC achieve the best BRISQUE, PI, LPIPS, and PieAPP performances, especially outperforming other approaches with a significant margin for the quality scores of BRISQUE and the PI metrics. The proportion of perceptual image evaluation metrics enhances much more than PSNR and SSIM, while PSNR and SSIM still maintain comparable performance. This trend follows the perception-distortion trade-off explained in [33].

Table 3. The ablation results of image deblurring using the proposed TFMLC loss computed by the adapted features from different layers of the pretrained ViT-B/16 CLIP model on the GoPro dataset. Best scores are **highlighted**. The H-T-RB represents applying our proposed TFMLC extracting from the ResBlock trained on HINet.

Method	PSNR/SSIM↑	BRISQUE/PI/LPIPS/PieAPP↓
HINet	**32.771/0.959**	52.036/5.184/0.090/0.656
H-T-RB 1	32.511/0.956	43.913/4.316/0.063/**0.500**
H-T-RB 3	32.439/0.955	42.788/**4.184/0.055**/0.548
H-T-RB 5	32.304/0.955	45.214/4.386/0.061/0.601
H-T-RB 7	31.948/0.951	**42.140**/4.393/0.068/0.579
H-T-RB 9	32.184/0.954	45.122/4.377/0.064/0.596
H-T-RB 11	32.211/0.954	45.139/4.417/0.064/0.619

To further verify the effectiveness of the proposed approach, we also perform the finetuning without employing the proposed losses from the pretrained weights of both deblurring models but the evaluation metrics scores are similar to the original models. In addition, we also show the visual qualitative results of different methods in Fig. 3. We find that the texture of the objects in the results from the original deblurring models is smoother, while the results of the proposed FMLC and TFMLC is closer to natural, realistic images. We also find that FMLC and TFMLC excel at addressing shadow in the images, which are obvious in Figure 1. These pieces of evidence all demonstrate the effectiveness of the proposed methods.

Ablation Studies: Since there are several different pretrained backbones available for the CLIP model, we first conduct the performance comparisons to choose the best one for image deblurring. We train the HINet for image deblurring using its original loss along with our proposed FMLC which exploits the features from the final layer of each backbone. As shown in Table 2, the backbone of ViT-B/16 based on Vision Transformer achieves more balanced results among all the perceptual metrics than others. Therefore, we adopt ViT-B/16 as the main backbone for our experiments. In addition, we further evaluate the proposed methods with the features from different ResBlocks of ViT-B/16-based image encoder for CLIP to investigate the performance influence using the features from different layers. Since the features of the middle layer strike a good balance to contain both local and semantic information, we empirically choose the middle layer as our backbone to compute the feature matching loss. As shown in Table 3, we find that the ResBlock 3 achieves a more balanced perceptual performance than others. To strike a compromise between the perceptual quality and both PSNR and SSIM scores, we mainly choose to extract image features from ResBlock 3 as our prior in our paper.

5. CONCLUSION

In this work, we present a self-supervised finetuning strategy to synthesize blurred images in different blur extents and to effectively adapt the general pretrained deep networks on the large-scale image datasets towards image deblurring to compute better feature matching loss. Both qualitative and quantitative experimental results demonstrate that two existing state-of-the-art deblurring models enhanced with our FMLC and TFMLC losses are able to obtain sharper restored images with lower BRISQUE, PI, LPIPS, and PieAPP scores than the original models while maintaining comparable PSNR and SSIM performances. This further shows the effectiveness of the proposed feature matching loss to improve the results of the image deblurring models in terms of perceptual quality.

6. REFERENCES

[1] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] L. Chen, X. Lu, J. Zhang, X. Chu, and C. Chen, "Hinet: Half instance normalization network for image restoration," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021.

[6] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Conference on Empirical Methods in Natural Language Processing*, 2021.

[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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[8] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, 2012.

[9] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 pirm challenge on perceptual image super-resolution," in *European Conference on Computer Vision (ECCV) Workshops*, 2018.

[10] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "Pieapp: Perceptual image-error assessment through pairwise preference," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[11] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *IEEE International Conference on Pattern Recognition (ICPR)*, 2010.

[12] 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*, 2004.

[13] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers*, 2003.

[14] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, 2005.

[15] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, 2006.

[16] D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, 2007.

[17] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, 2011.

[18] M. Kettunen, E. Härkönen, and J. Lehtinen, "E-LPIPS: robust perceptual image similarity via random transformation ensembles," *arXiv preprint arXiv:1906.03973*, 2019.

[19] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, 2010.

[20] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE Transactions on Image Processing*, 2012.

[21] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[22] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, 2011.

[23] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, 2013.

[24] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Processing: Image Communication*, 2014.

[25] D. Yang and M. Yamac, "Motion aware double attention network for dynamic scene deblurring," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022.

[26] K. Zhang, W. Luo, Y. Zhong, L. Ma, B. Stenger, W. Liu, and H. Li, "Deblurring by realistic blurring," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[27] K. Purohit and A. N. Rajagopalan, "Region-adaptive dense network for efficient motion deblurring," in *the AAAI Conference on Artificial Intelligence*, 2020.

[28] J. Liang, J. Cao, Y. Fan, K. Zhang, R. Ranjan, Y. Li, R. Timofte, and L. V. Gool, "Vrt: A video restoration transformer," *arXiv preprint arXiv:2201.12288*, 2022.

[29] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[30] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *International Conference on Image Processing*, 1994.

[31] J. Rim, H. Lee, J. Won, and S. Cho, "Real-world blur dataset for learning and benchmarking deblurring algorithms," in *European Conference on Computer Vision (ECCV)*, 2020.

[32] D. P Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.