Dynamic Dual Trainable Bounds for Ultra-low Precision Super-Resolution Networks

Yunshan Zhong1,2, Mingbao Lin3, Xunchao Li2,K eL i3, Yunhang Shen3, Fei Chao1,2, Yongjian Wu3, and Rongrong Ji1,2(B)
1Institute of Arti■cial Intelligence, Xiamen University, Xiamen, China
zhongyunshan@stu.xmu.edu.cn, {fchao,rrji }@xmu.edu.cn
2MAC Lab, School of Informatics, Xiamen University, Xiamen, China
lixunchao@stu.xmu.edu.cn
3Tencent Youtu Lab, Shanghai, China
littlekenwu@tencent.com

Abstract. Light-weight super-resolution (SR) models have received considerable attention for their serviceability in mobile devices. Many e■orts employ network quantization to compress SR models. However, these methods su■er from severe performance degradation when quantizing the SR models to ultra-low precision (e.g., 2-bit and 3-bit) with the low-cost layer-wise quantizer. In this paper, we identify that the performance drop comes from the contradiction between the layer-wise symmetric quantizer and the highly asymmetric activation distributionin SR models. This discrepancy leads to either a waste on the quantization levels or detail loss in reconstructed images. Therefore, we propose a novel activation quantizer, referred to as Dynamic Dual TrainableBounds (DDTB), to accommodate the asymmetry of the activations.
Speci■cally, DDTB innovates in: 1) A layer-wise quantizer with trainable upper and lower bounds to tackle the highly asymmetric activations. 2) Adynamic gate controller to adaptively adjust the upper and lower bounds at runtime to overcome the drastically varying activation ranges over di■erent samples. To reduce the extra overhead, the dynamic gate con-troller is quantized to 2-bit and applied to only part of the SR networks according to the introduced dynamic intensity. Extensive experiments demonstrate that our DDTB exhibits signi■cant performance improve-ments in ultra-low precision. For example, our DDTB achieves a 0.70 dB PSNR increase on Urban100 benchmark when quantizing EDSR to 2-bit and scaling up output images to ×4. Code is at https://github.com/zysxmu/DDTB .

Keywords: Super-resolution ·Network quantization ·Dual trainable bounds ·Dynamic gate controller
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19797-0 1.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13678, pp. 1–18, 2022.https://doi.org/10.1007/978-3-031-19797-0

_
****

OSFormer: One-Stage Camou■aged Instance Segmentation with Transformers

Jialun Pei1, Tianyang Cheng2, Deng-Ping Fan3(B),H eT a n g2, Chuanbo Chen2, and Luc Van Gool3
1School of Computer Science and Technology, HUST, Wuhan, China
2School of Software Engineering, HUST, Wuhan, China
3Computer Vision Lab, ETH Zurich, Zurich, Switzerland
dengpfan@gmail.com

Abstract. We present OSFormer , the ■rst one-stage transformer framework for camou■aged instance segmentation (CIS). OSFormer isbased on two key designs. First, we design a location-sensing transformer (LST) to obtain the location label and instance-aware parameters by introducing the location-guided queries and the blend-convolutionfeed-forward network. Second, we develop a coarse-to-■ne fusion (CFF) to merge diverse context information from the LST encoder and

CNN backbone. Coupling these two components enables OSFormer toe■ciently blend l
ocal features and long-range context dependencies for
predicting camou■aged instances. Compared with two-stage frameworks,
our OSFormer reaches 41% AP and achieves good convergence e■ciencywithout requir
ing enormous training data, i.e., only 3,040 samples under
60 epochs. Code link: https://github.com/PJLallen/OSFormer .
Keywords: Camou■age
·Instance segmentation ·Transformer
1
****

Highly Accurate Dichotomous Image
Segmentation
Xuebin Qin1, Hang Dai1, Xiaobin Hu2, Deng-Ping Fan3(B),
Ling Shao4, and Luc Van Gool3
1MBZUAI, Abu Dhabi, UAE
xuebin@ualberta.ca, hang.dai@mbzuai.ac.ae
2Tencent Youtu Lab, Shanghai, China
xiaobin.hu@tum.de
3ETH Zurich, Zurich, Switzerland
dengpfan@gmail.com, vangool@vision.ee.ethz.ch
4Terminus Group, Beijing, China
ling.shao@ieee.org
Abstract. We present a systematic study on a new task called dichoto-
mous image segmentation (DIS), which aims to segment highly accurate
objects from natural images. To this end, we collected the ■rst large-scaleDIS d
ataset, called DIS5K , which contains 5,470 high-resolution (e.g.,
2K, 4K or larger) images covering camou■aged ,salient ,o r meticulous
objects in various backgrounds. DIS is annotated with extremely ■ne-
grained labels. Besides, we introduce a simple intermediate supervision
baseline (IS-Net) using both feature-level and mask-level guidance for
DIS model training. IS-Net outperforms various cutting-edge baselineson the prop
osed DIS5K, making it a general self-learned supervision net-
work that can facilitate future research in DIS. Further, we design a new
metric called human correction e■orts (HCE) which approximates the
number of mouse clicking operations required to correct the false pos-
itives and false negatives. HCE is utilized to measure the gap between
models and real-world applications and thus can complement existingmetrics. Fina
lly, we conduct the largest-scale benchmark, evaluating 16
representative segmentation models, providing a more insightful discus-
sion regarding object complexities, and showing several potential applica-tions
(e.g., background removal, art design, 3D reconstruction). Hoping
these e■orts can open up promising directions for both academic and
industries. Project page: https://xuebinqin.github.io/dis/index.html .
Keywords: Dichotomous image segmentation
·High resolution ·
Metric
We would like to thank Jiayi Zhu for his e■orts in re-organizing the dataset and
 codes.
Supplementary Information The online version contains supplementary material
available at https://doi.org/10.1007/978-3-031-19797-0 3.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13678, pp. 38–56, 2022.https://doi.org/
10.1007/978-3-031-19797-0

_
****

Boosting Supervised Dehazing Methods
via Bi-level Patch Reweighting
Xingyu Jiang1, Hongkun Dou1, Chengwei Fu1, Bingquan Dai1, Tianrun Xu2,
and Yue Deng1(B)
1School of Astronautics, Beihang University, Beijing, China

ydeng@buaa.edu.cn
2North China University of Technology, Beijing, China
Abstract. Natural images can su■er from non-uniform haze distribu-
tions in di■erent regions. However, this important fact is hardly considered
in existing supervised dehazing methods, in which all training patches areaccoun
ted for equally in the loss design. These supervised methods may
fail in making promising recoveries on some regions contaminated by heavy
hazes. Therefore, for a more reasonable dehazing losses design, the vary-ing imp
ortance of di■erent training patches should be taken into account.
Such rationale is exactly in line with the process of human learning that
di■cult concepts always require more practice in learning. To this end,we propos
e a bi-level dehazing (BILD) framework by designing an inter-
nal loop for weighted supervised dehazing and an external loop for training
patch reweighting. With simple derivations, we show the gradients of BILDexhibit
 natural connections with policy gradient and can thus explain the
BILD objective by the rewarding mechanism in reinforcement learning.
The BILD is not a new dehazing method per se, it is better recognizedas a ■exibl
e framework that can seamlessly work with general supervised
dehazing approaches for their performance boosting.
Keywords: Single image dehazing
·Bi-level optimization ·Visual
importance ·Deep learning

****

Flow-Guided Transformer for Video
Inpainting
Kaidong Zhang1, Jingjing Fu2(B), and Dong Liu1
1University of Science and Technology of China, Hefei, China
richu@mail.ustc.edu.cn ,dongeliu@ustc.edu.cn
2Microsoft Research Asia, Beijing, China
jifu@microsoft.com
Abstract. We propose a ■ow-guided transformer, which innovatively
leverage the motion discrepancy exposed by optical ■ows to instruct
the attention retrieval in transformer for high ■delity video inpainting.
More specially, we design a novel ■ow completion network to completethe corrupte
d ■ows by exploiting the relevant ■ow features in a local
temporal window. With the completed ■ows, we propagate the content
across video frames, and adopt the ■ow-guided transformer to synthe-
size the rest corrupted regions. We decouple transformers along temporal
and spatial dimension, so that we can easily integrate the locally rele-vant com
pleted ■ows to instruct spatial attention only. Furthermore, we
design a ■ow-reweight module to precisely control the impact of com-
pleted ■ows on each spatial transformer. For the sake of e■ciency, weintroduce w
indow partition strategy to both spatial and temporal trans-
formers. Especially in spatial transformer, we design a dual perspective
spatial MHSA, which integrates the global tokens to the window-basedattention. E
xtensive experiments demonstrate the e■ectiveness of the
proposed method qualitatively and quantitatively. Codes are available
athttps://github.com/hitachinsk/FGT .
Keywords: Video inpainting
·Optical ■ow ·Transformer

****

Shift-Tolerant Perceptual Similarity
Metric
Abhijay Ghildyal(B)and Feng Liu
Portland State University, Portland, OR 97201, USA
{abhijay,fliu }@pdx.edu
Abstract. Existing perceptual similarity metrics assume an image and
its reference are well aligned. As a result, these metrics are often sensitiveto

a small alignment error that is imperceptible to the human eyes. This paper studies the e■ect of small misalignment, speci■cally a small shift between the input and reference image, on existing metrics, and accord-ingly dev elops a shift-tolerant similarity metric. This paper builds upon LPIPS, a widely used learned perceptual similarity metric, and explores architectural design considerations to make it robust against impercep-tible mis alignment. Speci■cally, we study a wide spectrum of neural net-work elements, such as anti-aliasing ■ltering, pooling, striding, padding, and skip connection, and discuss their roles in making a robust met-ric. Based o n our studies, we develop a new deep neural network-based perceptual similarity metric. Our experiments show that our metric is tolerant to imperceptible shifts while being consistent with the humansimilarity judgment. Code is available at https://tinyurl.com/5n85r28r .

Keywords: Perceptual similarity metric
·Image quality assessment

1

****

Perception-Distortion Balanced ADMM
Optimization for Single-Image
Super-Resolution

Yuehan Zhang1,B oJ i1,J i aH a o2, and Angela Yao1(B)
1National University of Singapore, Singapore, Singapore
{zyuehan,jibo,ayao }@comp.nus.edu.sg
2HiSilicon Technologies, Shanghai, China
hao.jia@huawei.com

Abstract. In image super-resolution, both pixel-wise accuracy and per-ceptual ■delity are desirable. However, most deep learning methods onlyachieve h igh performance in one aspect due to the perception-distortion trade-o■, and works that successfully balance the trade-o■ rely on fus-ing results from separately trained models with ad-hoc post-processing.In this p aper, we propose a novel super-resolution model with a low-frequency constraint (LFc-SR), which balances the objective and per-ceptual quality through a single model and yields super-resolved imageswith high PSNR and perceptual scores. We further introduce an ADMM-based alternating optimization method for the non-trivial learning of the constrained model. Experiments showed that our method, without cum-bersome post-processing procedures, achieved the state-of-the-art perfor-mance. The code is available at https://github.com/Yuehan717/PDASR .

Keywords: Image super-resolution
·Perception-distortion trade-o■ ·
Constrained optimization

1

****

VQFR: Blind Face Restoration
with Vector-Quantized Dictionary
and Parallel Decoder

Yuchao Gu1,2, Xintao Wang2, Liangbin Xie2,5,C h a oD o n g4,5,G e nL i3,
Ying Shan2, and Ming-Ming Cheng1(B)
1TMCC, CS, Nankai University, Tianjin, China
cmm@nankai.edu.cn
2ARC Lab, Tencent PCG, Beijing, China
3Platform Technologies, Tencent Online Video, Beijing, China
4Shanghai AI Laboratory, Beijing, China
5Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences,
Beijing, China
https://github.com/TencentARC/VQFR/

Abstract. Although generative facial prior and geometric prior have recently demonstrated high-quality results for blind face restoration, pro-ducin g ■ne-grained facial details faithful to inputs remains a challenging problem. Motivated by the classical dictionary-based methods and the

recent vector quantization (VQ) technique, we propose a VQ-based facerestoration method – VQFR. VQFR takes advantage of high-quality low-level feature banks extracted from high-quality faces and can thus help recover realistic facial details. However, the simple application of theVQ codebook cannot achieve good results with faithful details and identity preservation. Therefore, we further introduce two special network designs. 1). We ■rst investigate the compression patch size in the VQcodebook and ■nd that the VQ codebook designed with a proper compression patch size is crucial to balance the quality and ■delity. 2). To further fuse low-level features from inputs while not “contaminating” therealistic details generated from the VQ codebook, we proposed a parallel decoder consisting of a texture decoder and a main decoder. Those two decoders then interact with a texture warping module with deformableconvolution. Equipped with the VQ codebook as a facial detail dictionary and the parallel decoder design, the proposed VQFR can largely enhance the restored quality of facial details while keeping the ■delity toprevious methods.

Keywords: Blind face restoration ·Vector quantization ·Parallel decoder

X. Wang—Project lead.Y. Gu—is an intern in ARC Lab, Tencent PCG.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19797-0 8.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13678, pp. 126–143, 2022.https://doi.org/10.1007/978-3-031-19797-0

_

****

Uncertainty Learning in Kernel Estimation for Multi-stage Blind Image Super-Resolution

Zhenxuan Fang1, Weisheng Dong1(B),X i nL i2, Jinjian Wu1, Leida Li1, and Guangming Shi1
1School of Arti■cial Intelligence, Xidian University, Xi’an, China
zxfang@stu.xidian.edu.cn, {wsdong,jinjian.wu }@mail.xidian.edu.cn, {ldli,gmshi }@xidian.edu.cn
2Lane Department of CSEE, West Virginia University, Morgantown, WV, USA
xin.li@mail.wvu.edu

Abstract. Conventional wisdom in blind super-resolution (SR) ■rst estimates the unknown degradation from the low-resolution image and then exploits the degradation information for image reconstruction. Such sequential approaches su■er from two fundamental weaknesses - i.e., the lack of robustness (the performance drops when the estimated degrada-tion is inaccurate) and the lack of transparency (network architectures are heuristic without incorporating domain knowledge). To address these issues, we propose a joint Maximum a Posteriori (MAP) approach forestimating the unknown kernel and high-resolution image simultaneously.
Our method ■rst introduces uncertainty learning in the latent space when estimating the blur kernel, aiming at improving the robustness to the esti-mation error. Then we propose a novel SR network by unfolding the joint MAP estimator with a learned Laplacian Scale Mixture (LSM) prior and the estimated kernel. We have also developed a novel approach of esti-mating both the scale prior coe■cient and the local means of the LSM model through a deep convolutional neural network (DCNN). All parameters of the MAP estimation algorithm and the DCNN parameters arejointly optimized through end-to-end training. Extensive experiments on both synthetic and real-world images show that our method achieves state-of-the-art performance for the task of blind image SR.

****

Learning Spatio-Temporal Downsampling for E■ective Video Upscaling

Xiaoyu Xiang1(B), Yapeng Tian2, Vijay Rengarajan1,
Lucas D. Young1,B oZ h u1, and Rakesh Ranjan1
1Meta Reality Labs, Menlo Park, USA
{xiangxiaoyu,apvijay,bozhufrl,rakeshr }@fb.com, lucasyoung482@gmail.com
2University of Texas at Dallas, Richardson, USA
tianyapeng92@gmail.com

Abstract. Downsampling is one of the most basic image processing operations. Improper spatio-temporal downsampling applied on videos can cause aliasing issues such as moir´ e patterns in space and the wagon-wheel e■ect in time. Consequently, the inverse task of upscaling a low-resolution, low frame-rate video in space and time becomes a challenging ill-posed problem due to information loss and aliasing artifacts. In this paper, we aim to solve the space-time aliasing problem by learn-ing a spatio-temporal downsampler. Towards this goal, we propose a neural network framework that jointly learns spatio-temporal downsampling and upsampling. It enables the downsampler to retain the keypatterns of the original video and maximizes the reconstruction performance of the upsampler. To make the downsamping results compatible with popular image and video storage formats, the downsampling resultsare encoded to uint8 with a di■erentiable quantization layer. To fully utilize the space-time correspondences, we propose two novel modules for explicit temporal propagation and space-time feature rearrangement.Experimental results show that our proposed method signi■cantly boosts the space-time reconstruction quality by preserving spatial textures and motion patterns in both downsampling and upscaling. Moreover, ourframework enables a variety of applications, including arbitrary video resampling, blurry frame reconstruction, and e■cient video storage.

Keywords: Downsampling ·Anti-aliasing ·Video upscaling

****

Learning Local Implicit Fourier Representation for Image Warping

Jaewon Lee1, Kwang Pyo Choi2, and Kyong Hwan Jin1(B)
1Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, Korea
{ljw3136,kyong.jin }@dgist.ac.kr
2Samsung Electronics, Suwon-si, Korea
kp5.choi@samsung.com

Abstract. Image warping aims to reshape images de■ned on rectangular grids into arbitrary shapes. Recently, implicit neural functions have shown remarkable performances in representing images in a continuousmanner. However, a standalone multi-layer perceptron su■ers from learning high-frequency Fourier coe■cients. In this paper, we propose a local texture estimator for image warping (LTEW) followed by an implicitneural representation to deform images into continuous shapes. Local textures estimated from a deep super-resolution (SR) backbone are multiplied by locally-varying Jacobian matrices of a coordinate transforma-tion to predict Fourier responses of a warped image. Our LTEW-based neural function outperforms existing warping methods for asymmetric-scale SR and homography transform. Furthermore, our algorithm wellgeneralizes arbitrary coordinate transformations, such as homography transform with a large magni■cation factor and equirectangular projection (ERP) perspective transform, which are not provided in training.Our source code is available at https://github.com/jaewon-lee-b/ltew .

Keywords: Image warping ·Implicit neural representation ·Fourier features ·Jacobian ·Homography transform ·Equirectangular projection (ERP)

1
****
SepLUT: Separable Image-Adaptive
Lookup Tables for Real-Time Image
Enhancement
Canqian Yang1, Meiguang Jin2,Y iX u1(B), Rui Zhang1, Ying Chen2,
and Huaida Liu2
1MoE Key Lab of Arti■cial Intelligence, AI Institute,
Shanghai Jiao Tong University, Shanghai, China
{charles.young,xuyi,zhang rui}@sjtu.edu.cn
2Alibaba Group, Hangzhou, China
{meiguang.jmg,yingchen,liuhuaida.lhd }@alibaba-inc.com
Abstract. Image-adaptive lookup tables (LUTs) have achieved great
success in real-time image enhancement tasks due to their high e■-ciency for mod
eling color transforms. However, they embed the com-
plete transform, including the color component-independent and the
component-correlated parts, into only a single type of LUTs, either 1Dor 3D, in
a coupled manner. This scheme raises a dilemma of improv-
ing model expressiveness or e■ciency due to two factors. On the one
hand, the 1D LUTs provide high computational e■ciency but lack thecritical capab
ility of color components interaction. On the other, the
3D LUTs present enhanced component-correlated transform capability
but su■er from heavy memory footprint, high training di■culty, andlimited cell u
tilization. Inspired by the conventional divide-and-conquer
practice in the image signal processor, we present SepLUT (separable
image-adaptive lookup table) to tackle the above limitations. Speci■-cally, we s
eparate a single color transform into a cascade of component-
independent and component-correlated sub-transforms instantiated as
1D and 3D LUTs, respectively. In this way, the capabilities of two sub-
transforms can facilitate each other, where the 3D LUT complements the
ability to mix up color components, and the 1D LUT redistributes theinput colors
 to increase the cell utilization of the 3D LUT and thus enable
the use of a more lightweight 3D LUT. Experiments demonstrate that
the proposed method presents enhanced performance on photo retouch-ing benchmark
 datasets than the current state-of-the-art and achieves
real-time processing on both GPUs and CPUs.
C. Yang and M. Jin—Equal contribution.
Work partially done during an internship of C. Yang at Alibaba Group.
Supplementary Information The online version contains supplementary material
available at https://doi.org/10.1007/978-3-031-19797-0 12.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13678, pp. 201–217, 2022.https://doi.or
g/10.1007/978-3-031-19797-0
_1
****
Blind Image Decomposition
Junlin Han1,2(B), Weihao Li1,P e n g f e iF a n g1,2, Chunyi Sun2,J i eH o n g1,
2,
Mohammad Ali Armin1, Lars Petersson1, and Hongdong Li2
1Data61-CSIRO, Sydney, Australia
junlin.han@data61.csiro.au
2Australian National University, Canberra, Australia
Abstract. We propose and study a novel task named Blind Image
Decomposition (BID), which requires separating a superimposed image
into constituent underlying images in a blind setting, that is, both the
source components involved in mixing as well as the mixing mecha-
nism are unknown. For example, rain may consist of multiple compo-nents, such as
 rain streaks, raindrops, snow, and haze. Rainy images
can be treated as an arbitrary combination of these components, some
of them or all of them. How to decompose superimposed images, likerainy images,

into distinct source components is a crucial step toward real-world vision systems. To facilitate research on this new task, we construct multiple benchmark datasets, including mixed image decom-position acro ss multiple domains, real-scenario deraining, and joint shadow/re■ection/watermark removal. Moreover, we propose a simple yet general Blind Image Decomposition Network (BIDeN) to serve as a strong baseline for future work. Experimental results demonstrate the tenability of our benchmarks and the e■ectiveness of BIDeN. Codes and datasets are available at GitHub .

Keywords: Image decomposition ·Low-level vision ·Rain removal

1

****

MuLUT: Cooperating Multiple Look-Up Tables for E■cient Image Super-Resolution

Jiacheng Li1, Chang Chen2, Zhen Cheng1, and Zhiwei Xiong1(B)
1University of Science and Technology of China, Hefei, China
{jclee,mywander }@mail.ustc.edu.cn, zwxiong@ustc.edu.cn
2Huawei Noah’s Ark Lab, Beijing, China
chenchang25@huawei.com

Abstract. The high-resolution screen of edge devices stimulates a strong demand for e■cient image super-resolution (SR). An emergingresearch, SR-L UT, responds to this demand by marrying the look-up table (LUT) with learning-based SR methods. However, the size of a sin-gleLUT grows exponentially with the increase of its indexing capacity. Consequently, the receptive ■eld of a single LUT is restricted, resulting in inferior performance. To address this issue, we extend SR-LUT by enabling the cooperation of Mu ltipleLUTs, termed MuLUT. Firstly, we devise two novel complementary indexing patterns and construct multi-ple LUTs in parallel. Secondly, we propose a re-indexing mechanism to enable the hierarchical indexing between multiple LUTs. In these twoways, the to tal size of MuLUT is linear to its indexing capacity, yield-ing a practical method to obtain superior performance. We examine the advantage of MuLUT on ■ve SR benchmarks. MuLUT achieves a signif-icant improvement over SR-LUT, up to 1.1 dB PSNR, while preserving its e■ciency. Moreover, we extend MuLUT to address demosaicing ofBayer-patterned images, surpassing SR-LUT on two benchmarks by a large margin.

Keywords: Image super-resolution ·Look-up table ·Image demosaicing

1

****

Learning Spatiotemporal Frequency-Transformer for Compressed Video Super-Resolution

Zhongwei Qiu1,2(B), Huan Yang3, Jianlong Fu3, and Dongmei Fu1,2
1University of Science and Technology Beijing, Beijing, China
qiuzhongwei@xs.ustb.edu.cn, fdm ustb@ustb.edu.cn
2Shunde Graduate School of University of Science and Technology Beijing, Beijing, China
3Microsoft Research, Beijing, China
{huayan,jianf }@microsoft.com

Abstract. Compressed video super-resolution (VSR) aims to restore high-resolution frames from compressed low-resolution counterparts. Most recent VSR approaches often enhance an input frame by “borrow-ing” relevant textures from neighboring video frames. Although some progress has been made, there are grand challenges to e■ectively extract and transfer high-quality textures from compressed videos where mostframes are u

sually highly degraded. In this paper, we propose a novel
Frequency-Transformer for compressed video super-resolution (FTVSR)
that conducts self-attention over a joint space-time-frequency domain.First, we
divide a video frame into patches, and transform each patch into
DCT spectral maps in which each channel represents a frequency band.
Such a design enables a ■ne-grained level self-attention on each frequencyband,
so that real visual texture can be distinguished from artifacts, and
further utilized for video frame restoration. Second, we study di■erent
self-attention schemes, and discover that a “divided attention” whichconducts a
joint space-frequency attention before applying temporal
attention on each frequency band, leads to the best video enhancement
quality. Experimental results on two widely-used video super-resolutionbenchmark
s show that FTVSR outperforms state-of-the-art approaches
on both uncompressed and compressed videos with clear visual margins.
Code are available at https://github.com/researchmm/FTVSR .
Keywords: VSR
·Transformer ·Frequency learning ·Compression
1
****

Spatial-Frequency Domain Information
Integration for Pan-Sharpening
Man Zhou1,2, Jie Huang1, Keyu Yan1,2,H uY u1, Xueyang Fu1, Aiping Liu1,
Xian Wei3, and Feng Zhao1(B)
1University of Science and Technology of China, Hefei, China
{manman,hj0117 }@mail.ustc.edu.cn, fzhao956@ustc.edu.cn
2Hefei Institute of Physical Science, Chinese Academy of Sciences, Hefei, China
3MoE Engineering Research Center of Hardware/Software Co-design Technology
and Application, East China Normal University, Shanghai, China
Abstract. Pan-sharpening aims to generate high-resolution multi-
spectral (MS) images by fusing PAN images and low-resolution MS
images. Despite its great advances, most existing pan-sharpening meth-
ods only work in the spatial domain and rarely explore the potential solu-tions
in the frequency domain. In this paper, we ■rst attempt to address
pan-sharpening in both spatial and frequency domains and propose a
Spatial-Frequency Information Integration Network, dubbed as SFIIN.
To implement SFIIN, we devise a core building module tailored with pan-
sharpening, consisting of three key components: spatial-domain infor-
mation branch, frequency-domain information branch, and dual domaininteraction.
To be speci■c, the ■rst employs the standard convolution to
integrate the local information of two modalities of PAN and MS images
in the spatial domain, while the second adopts deep Fourier transfor-mation to a
chieve the image-wide receptive ■eld for exploring global
contextual information. Followed by, the third is responsible for facilitat-
ing the information ■ow and learning the complementary representation.
We conduct extensive experiments to validate the e■ectiveness of the
proposed network and demonstrate the favorable performance againstother state-of
-the-art methods.
Keywords: Pan-sharpening
·Spatial-frequency domain
1
****

Adaptive Patch Exiting for Scalable
Single Image Super-Resolution
Shizun Wang1, Jiaming Liu2,4, Kaixin Chen1, Xiaoqi Li2,4,
Ming Lu3(B), and Yandong Guo4
1Beijing University of Posts and Telecommunications, Beijing, China
wangshizun@bupt.edu.cn
2Peking University, Beijing, China
3Intel Labs China, Beijing, China
lu199192@gmail.com

4OPPO Research Institute, Shanghai, China
Abstract. Since the future of computing is heterogeneous, scalability
is a crucial problem for single image super-resolution. Recent works try
to train one network, which can be deployed on platforms with di■erent
capacities. However, they rely on the pixel-wise sparse convolution, whichis not
hardware-friendly and achieves limited practical speedup. As image
can be divided into patches, which have various restoration di■culties,
we present a scalable method based on Adaptive Patch Exiting (APE) toachieve mor
e practical speedup. Speci■cally, we propose to train a regres-
sor to predict the incremental capacity of each layer for the patch. Once
the incremental capacity is below the threshold, the patch can exit at thespeci■
c layer. Our method can easily adjust the trade-o■ between perfor-
mance and e■ciency by changing the threshold of incremental capacity.
Furthermore, we propose a novel strategy to enable the network training ofour me
thod. We conduct extensive experiments across various backbones,
datasets and scaling factors to demonstrate the advantages of our method.
Code is available at https://github.com/littlepure2333/APE .
Keywords: Single image super-resolution
·Scalability ·E■ciency
1
****
E■cient Meta-Tuning for Content-Aware Neural Video Delivery

Xiaoqi Li1, Jiaming Liu1,2, Shizun Wang3, Cheng Lyu3,
Ming Lu4(B), Yurong Chen4,A n b a n gY a o4, Yandong Guo2,
and Shanghang Zhang1(B)
1Peking University, Beijing, China
shzhang.pku@gmail.com
2OPPO Research Institute, Beijing, China
3Beijing University of Posts and Telecommunications, Beijing, China
4Intel Labs China, Beijing, China
Abstract. Recently, Deep Neural Networks (DNNs) are utilized to
reduce the bandwidth and improve the quality of Internet video delivery.Existing
methods train corresponding content-aware super-resolution
(SR) model for each video chunk on the server, and stream low-resolution
(LR) video chunks along with SR models to the client. Although theyachieve promi
sing results, the huge computational cost of network train-
ing limits their practical applications. In this paper, we present a method
named E■cient Meta-Tuning (EMT) to reduce the computational cost.Instead of trai
ning from scratch, EMT adapts a meta-learned model to
the ■rst chunk of the input video. As for the following chunks, it ■ne-
tunes the partial parameters selected by gradient masking of previousadapted mod
el. In order to achieve further speedup for EMT, we pro-
pose a novel sampling strategy to extract the most challenging patches
from video frames. The proposed strategy is highly e■cient and bringsnegligible
additional cost. Our method signi■cantly reduces the com-
putational cost and achieves even better performance, paving the way
for applying neural video delivery techniques to practical applications.
We conduct extensive experiments based on various e■cient SR architec-
tures, including ESPCN, SRCNN, FSRCNN and EDSR-1, demonstratingthe generalizatio
n ability of our work. The code is released at https://
github.com/Neural-video-delivery/EMT-Pytorch-ECCV2022 .
Keywords: Neural video delivery
·Super-resolution ·Meta learning
X. Li, J. Liu and S. Wang—Equal contribution.
Supplementary Information The online version contains supplementary material
available at https://doi.org/10.1007/978-3-031-19797-0 18.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13678, pp. 308–324, 2022.https://doi.or
g/10.1007/978-3-031-19797-0

_1
****
Reference-Based Image Super-Resolution
with Deformable Attention Transformer
Jiezhang Cao1, Jingyun Liang1, Kai Zhang1, Yawei Li1, Yulun Zhang1(B),
Wenguan Wang1, and Luc Van Gool1,2
1Computer Vision Lab, ETH Z¨ urich, Z¨ urich, Switzerland
{jiezhang.cao,jingyun.liang,kai.zhang,yawei.li,yulun.zhang,
wenguan.wang,vangool }@vision.ee.ethz.ch
2KU Leuven, Leuven, Belgium
https://github.com/caojiezhang/DATSR
Abstract. Reference-based image super-resolution (RefSR) aims to
exploit auxiliary reference (Ref) images to super-resolve low-resolution
(LR) images. Recently, RefSR has been attracting great attention as
it provides an alternative way to surpass single image SR. However,
addressing the RefSR problem has two critical challenges: (i) It is di■cult
to match the correspondence between LR and Ref images when they aresigni■cantly
di■erent; (ii) How to transfer the relevant texture from Ref
images to compensate the details for LR images is very challenging. To
address these issues of RefSR, this paper proposes a deformable attentionTransfo
rmer, namely DATSR, with multiple scales, each of which consists
of a texture feature encoder (TFE) module, a reference-based deformable
attention (RDA) module and a residual feature aggregation (RFA) mod-ule. Speci■c
ally, TFE ■rst extracts image transformation (e.g.,b r i g h t -
ness) insensitive features for LR and Ref images, RDA then can exploit
multiple relevant textures to compensate more information for LR fea-tures, and
RFA lastly aggregates LR features and relevant textures to
get a more visually pleasant result. Extensive experiments demonstrate
that our DATSR achieves state-of-the-art performance on benchmarkdatasets quanti
tatively and qualitatively.
Keywords: Reference-based image super-resolution
·Correspondence
matching ·Texture transfer ·Deformable attention transformer
1
****
Local Color Distributions Prior for Image
Enhancement
Haoyuan Wang(B), Ke Xu, and Rynson W.H. Lau
Department of Computer Science, City University of Hong Kong, Hong Kong,
People’s Republic of China
hywang26-c@my.city.edu.hk
Abstract. Existing image enhancement methods are typically designed
to address either the over- or under-exposure problem in the inputimage. When th
e illumination of the input image contains both over-
and under-exposure problems, these existing methods may not work well.
We observe from the image statistics that the local color distributions(LCDs) of
an image su■ering from both problems tend to vary across dif-
ferent regions of the image, depending on the local illuminations. Based
on this observation, we propose in this paper to exploit these LCDs as
a prior for locating and enhancing the two types of regions (i.e., over-
/under-exposed regions). First, we leverage the LCDs to represent theseregions,
and propose a novel local color distribution embedded (LCDE)
module to formulate LCDs in multi-scales to model the correlations
across di■erent regions. Second, we propose a dual-illumination learn-ing mechan
ism to enhance the two types of regions. Third, we construct
a new dataset to facilitate the learning process, by following the camera
image signal processing (ISP) pipeline to render standard RGB imageswith both un
der-/over-exposures from raw data. Extensive experiments
demonstrate that the proposed method outperforms existing state-of-
the-art methods quantitatively and qualitatively. Codes and dataset areinhttps:/

/hywang99.github.io/lcdpnet/ .
1
****
L-CoDer: Language-Based Colorization
with Color-Object Decoupling
Transformer
Zheng Chang1, Shuchen Weng2,Y uL i3,S iL i1(B), and Boxin Shi2
1School of Arti■cial Intelligence, Beijing University of Posts and
Telecommunications, Beijing, China
{zhengchang98,lisi }@bupt.edu.cn
2NERCVT, School of Computer Science, Peking University, Beijing, China
{shuchenweng,shiboxin }@pku.edu.cn
3International Digital Economy Academy, Shenzhen, China
liyu@idea.edu.cn
Abstract. Language-based colorization requires the colorized image to
be consistent with the user-provided language caption. A most recentwork propose
s to decouple the language into color and object conditions
in solving the problem. Though decent progress has been made, its per-
formance is limited by three key issues. (i) The large gap between visionand lan
guage modalities using independent feature extractors makes it
di■cult to fully understand the language. (ii) The inaccurate language
features are never re■ned by the image features such that the languagemay fail t
o colorize the image precisely. (iii) The local region does not per-
ceive the whole image, producing global inconsistent colors. In this work,
we introduce transformer into language-based colorization to tackle theaforement
ioned issues while keeping the language decoupling property.
Our method uni■es the modalities of image and language, and further
performs color conditions evolving with image features in a coarse-to-■nemanner.
In addition, thanks to the global receptive ■eld, our method is
robust to the strong local variation. Extensive experiments demonstrate
our method is able to produce realistic colorization and outperforms prior
arts in terms of consistency with the caption.
1
****
From Face to Natural Image: Learning
Real Degradation for Blind Image
Super-Resolution
Xiaoming Li1,5, Chaofeng Chen2, Xianhui Lin3, Wangmeng Zuo1,4(B),
and Lei Zhang5
1Faculty of Computing, Harbin Institute of Technology, Harbin, China
wmzuo@hit.edu.cn
2S-Lab, Nanyang Technological University, Singapore, Singapore
3DAMO Academy, Alibaba Group, Shenzhen, China
4Peng Cheng Lab, Shenzhen, China
5Department of Computing, The Hong Kong Polytechnic University,
Hung Hom, Hong Kong
cslzhang@comp.polyu.edu.hk
Abstract. How to design proper training pairs is critical for super-
resolving real-world low-quality (LQ) images, which su■ers from the di■-culties
in either acquiring paired ground-truth high-quality (HQ) images
or synthesizing photo-realistic degraded LQ observations. Recent works
mainly focus on modeling the degradation with handcrafted or estimateddegradatio
n parameters, which are however incapable to model compli-
cated real-world degradation types, resulting in limited quality improve-
ment. Notably, LQ face images, which may have the same degradationprocess as nat
ural images, can be robustly restored with photo-realistic
textures by exploiting their strong structural priors. This motivates us
to use the real-world LQ face images and their restored HQ counterpartsto model
the complex real-world degradation (namely ReDegNet), and
then transfer it to HQ natural images to synthesize their realistic LQ

counterparts. By taking these paired HQ-LQ face images as inputs toexplicitly pr
edict the degradation-aware and content-independent repre-
sentations, we could control the degraded image generation, and subse-
quently transfer these degradation representations from face to naturalimages to
 synthesize the degraded LQ natural images. Experiments show
that our ReDegNet can well learn the real degradation process from face
images. The restoration network trained with our synthetic pairs per-forms favor
ably against SOTAs. More importantly, our method provides
a new way to handle the real-world complex scenarios by learning their
degradation representations from the facial portions, which can be usedto signi■
cantly improve the quality of non-facial areas. The source code
is available at https://github.com/csxmli2016/ReDegNet .
Keywords: Real world degradation
·Blind image super-resolution
Supplementary Information The online version contains supplementary material
available at https://doi.org/10.1007/978-3-031-19797-0 22.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13678, pp. 376–392, 2022.https://doi.or
g/10.1007/978-3-031-19797-0
_2
****

Towards Interpretable Video Super-Resolution via Alternating Optimization

Jiezhang Cao1, Jingyun Liang1, Kai Zhang1(B), Wenguan Wang1,Q i nW a n g1,
Yulun Zhang1,H a oT a n g1, and Luc Van Gool1,2
1Computer Vision Lab, ETH Z¨ urich, Z¨ urich, Switzerland
{jiezhang.cao,jingyun.liang,kai.zhang,wenguan.wang,qin.wang,
yulun.zhang,hao.tang,vangool }@vision.ee.ethz.ch
2KU Leuven, Leuven, Belgium
https://github.com/caojiezhang/DAVSR
Abstract. In this paper, we study a practical space-time video super-
resolution (STVSR) problem which aims at generating a high-frameratehigh-resolut
ion sharp video from a low-framerate low-resolution blurry
video. Suchproblemoftenoccurswhenrecordingafastdynamic eventwith
a low-framerate and low-resolution camera, and the captured video wouldsu■er fro
m three typical issues: i) motion blur occurs due to object/camera
motions during exposure time; ii) motion aliasing is unavoidable when
the event temporal frequency exceeds the Nyquist limit of temporal sam-pling; ii
i) high-frequency details are lost because of the low spatial sam-
pling rate. These issues can be alleviated by a cascade of three sepa-
rate sub-tasks, including video deblurring, frame interpolation, and super-resol
ution, which, however, would fail to capture the spatial and temporal
correlations among video sequences. To address this, we propose an inter-
pretable STVSR framework by leveraging both model-based and learning-based metho
ds. Speci■cally, we formulate STVSR as a joint video deblur-
ring, frame interpolation, and super-resolution problem, and solve it as
two sub-problems in an alternate way. For the ■rst sub-problem, we derivean inte
rpretable analytical solution and use it as a Fourier data transform
layer. Then, we propose a recurrent video enhancement layer for the second
sub-problem to further recover high-frequency details. Extensive experi-
ments demonstrate the superiority of our method in terms of quantitative
metrics and visual quality.
Keywords: Video super-resolution
·Motion blur ·Motion aliasing
1
****

Event-Based Fusion for Motion Deblurring with Cross-modal Attention

Lei Sun1,2, Christos Sakaridis2, Jingyun Liang2,Q iJ i a n g1, Kailun Yang3,

Peng Sun1, Yaozu Ye1, Kaiwei Wang1(B), and Luc Van Gool2,4
1Zhejiang University, Hangzhou, China
wangkaiwei@zju.edu.cn
2ETH Z¨ urich, Z¨ urich, Switzerland
3KIT, Karlsruhe, Germany
4KU Leuven, Leuven, Belgium
Abstract. Traditional frame-based cameras inevitably su■er from
motion blur due to long exposure times. As a kind of bio-inspired cam-
era, the event camera records the intensity changes in an asynchronousway with h
igh temporal resolution, providing valid image degradation
information within the exposure time. In this paper, we rethink the
event-based image deblurring problem and unfold it into an end-to-endtwo-stage i
mage restoration network. To e■ectively fuse event and image
features, we design an event-image cross-modal attention module applied
at multiple levels of our network, which allows to focus on relevant fea-tures f
rom the event branch and ■lter out noise. We also introduce a
novel symmetric cumulative event representation speci■cally for image
deblurring as well as an event mask gated connection between the twostages of ou
r network which helps avoid information loss. At the dataset
level, to foster event-based motion deblurring and to facilitate evalua-
tion on challenging real-world images, we introduce the Real Event Blur
(REBlur) dataset, captured with an event camera in an illumination-
controlled optical laboratory. Our Event Fusion Network (EFNet) setsthe new stat
e of the art in motion deblurring, surpassing both the prior
best-performing image-based method and all event-based methods with
public implementations on the GoPro dataset (by up to 2.47 dB) and onour REBlur
dataset, even in extreme blurry conditions. The code and
our REBlur dataset are available at https://ahupujr.github.io/EFNet/ .

****

Fast and High Quality Image Denoising
via Malleable Convolution

Yifan Jiang1(B), Bartlomiej Wronski2, Ben Mildenhall2, Jonathan T. Barron2,
Zhangyang Wang1, and Tianfan Xue2
1University of Texas at Austin, Austin, USA
yifanjiang97@utexas.edu
2Google Research, San Francisco, USA
Abstract. Most image denoising networks apply a single set of static
convolutional kernels across the entire input image. This is sub-optimal
for natural images, as they often consist of heterogeneous visual pat-
terns. Dynamic convolution tries to address this issue by using per-pixelconvolu
tion kernels, but this greatly increases computational cost. In
t h i sw o r k ,w ep r e s e n t Malle ableConv olution (MalleConv), which per
-
forms spatial-varying processing with minimal computational overhead.
MalleConv uses a smaller set of spatially-varying convolution kernels,
a compromise between static and per-pixel convolution kernels. Thesespatially-va
rying kernels are produced by an e■cient predictor network
running on a downsampled input, making them much more e■cient to
compute than per-pixel kernels produced by a full-resolution image, andalso enla
rging the network’s receptive ■eld compared with static kernels.
These kernels are then jointly upsampled and applied to a full-resolution
feature map through an e■cient on-the-■y slicing operator with mini-mum memory o
verhead. To demonstrate the e■ectiveness of MalleConv,
we use it to build an e■cient denoising network we call MalleNet .
MalleNet achieves high-quality results without very deep architectures,making it
 8.9 ×faster than the best performing denoising algorithms
while achieving similar visual quality. We also show that a single Malle-
Conv layer added to a standard convolution-based backbone can signi■-cantly redu
ce the computational cost or boost image quality at a similar

cost. More information are on our project page: https://yifanjiang.net/MalleConv.html .
Keywords: Image denoising
·Dynamic kernel ·E■ciency
1
****
TAPE: Task-Agnostic Prior Embedding
for Image Restoration
Lin Liu1, Lingxi Xie3, Xiaopeng Zhang3, Shanxin Yuan4, Xiangyu Chen5,6,
Wengang Zhou1,2, Houqiang Li1,2, and Qi Tian3(B)
1CAS Key Laboratory of Technology in GIPAS, EEIS Department,
University of Science and Technology of China, Hefei, China
2Institute of Arti■cial Intelligence, Hefei Comprehensive National Science Center,
Hefei, China
3Huawei Cloud BU, Shenzhen, China
tian.qi1@huawei.com
4Huawei Noah’s Ark Lab, London, UK
5University of Macau, Zhuhai, China
6Shenzhen Institutes of Advanced Technology, CAS, Shenzhen, China
Abstract. Learning a generalized prior for natural image restoration is
an important yet challenging task. Early methods mostly involved hand-
crafted priors including normalized sparsity, ■0gradients, dark channel
priors, etc. Recently, deep neural networks have been used to learn var-ious image priors but do not guarantee to generalize. In this paper,
we propose a novel approach that embeds a task-agnostic prior into
a transformer. Our approach, named Task-Agnostic Prior Embedding(TAPE), consists of two stages, namely, task-agnostic pre-training and
task-speci■c ■ne-tuning, where the ■rst stage embeds prior knowledge
about natural images into the transformer and the second stage extractsthe knowledge to assist downstream image restoration. Experiments on
various types of degradation validate the e■ectiveness of TAPE. The
image restoration performance in terms of PSNR is improved by as muchas 1.45 dB and even outperforms task-speci■c algorithms. More impor-
tantly, TAPE shows the ability of disentangling generalized image priors
from degraded images, which enjoys favorable transfer ability to unknowndownstream tasks.
1
****
Uncertainty Inspired Underwater Image
Enhancement
Zhenqi Fu1,W uW a n g1, Yue Huang1, Xinghao Ding1(B), and Kai-Kuang Ma2
1Xiamen University, Fujian 361005, China
{fuzhenqi,23320170155546 }@stu.xmu.edu.cn, {yhuang2010,dxh }@xmu.edu.cn
2Nanyang Technological University, Singapore 639798, Singapore
ekkma@ntu.edu.sg
Abstract. A main challenge faced in the deep learning-based Under-
water Image Enhancement (UIE) is that the ground truth high-quality
image is unavailable. Most of the existing methods ■rst generate approx-imate reference maps and then train an enhancement network with cer-
tainty. This kind of method fails to handle the ambiguity of the reference
map. In this paper, we resolve UIE into distribution estimation and con-sensus process. We present a novel probabilistic network to learn the
enhancement distribution of degraded underwater images. Speci■cally,
we combine conditional variational autoencoder with adaptive instance
normalization to construct the enhancement distribution. After that, we
adopt a consensus process to predict a deterministic result based on aset of samples from the distribution. By learning the enhancement dis-
tribution, our method can cope with the bias introduced in the reference
map labeling to some extent. Additionally, the consensus process is use-ful to c

apture a robust and stable result. We examined the proposed
method on two widely used real-world underwater image enhancement
datasets. Experimental results demonstrate that our approach enablessampling pos
sible enhancement predictions. Meanwhile, the consensus
estimate yields competitive performance compared with state-of-the-
art UIE methods. Code available at https://github.com/zhenqifu/PUIE-
Net.
Keywords: Underwater image enhancement
·Deep learning ·
Probabilistic network ·Adaptive instance normalization ·Conditional
variational autoencoder
1
****

Hourglass Attention Network for Image
Inpainting
Ye Deng1,S i q iH u i1, Rongye Meng1, Sanping Zhou1,2, and Jinjun Wang1(B)
1Xi’an Jiaotong University, Xi’an, China
{dengye,huisiqi }@stu.xjtu.edu.cn, spzhou@xjtu.edu.cn,
jinjun@mail.xjtu.edu.cn
2Shunan Academy of Arti■cial Intelligence, Ningbo, China
Abstract. Bene■ting from the powerful ability of convolutional neural
networks (CNNs) to learn semantic information and texture patterns of
images, learning-based image inpainting methods have made noticeablebreakthrough
s over the years. However, certain inherent defects (e.g.
local prior, spatially sharing parameters) of CNNs limit their perfor-
mance when encountering broken images mixed with invalid informa-tion. Compared
to convolution, attention has a lower inductive bias, and
the output is highly correlated with the input, making it more suit-
able for processing images with various breakage. Inspired by this, inthis paper
 we propose a novel attention-based network (transformer),
called hourglass attention network (HAN) for image inpainting, which
builds an hourglass-shaped attention structure to generate appropri-ate features
 for complemented images. In addition, we design a novel
attention called Laplace attention, which introduces a Laplace distance
prior for the vanilla multi-head attention, allowing the feature matchingprocess
 to consider not only the similarity of features themselves, but
also distance between features. With the synergy of hourglass atten-
tion structure and Laplace attention, our HAN is able to make fulluse of hierarc
hical features to mine e■ective information for broken
images. Experiments on several benchmark datasets demonstrate supe-
rior performance by our proposed approach. The code can be found atgithub.com/de
ngyecode/hourglassattention .
Keywords: Image inpainting
·Attention ·Transformer
1
****

Unfolded Deep Kernel Estimation
for Blind Image Super-Resolution
Hongyi Zheng1, Hongwei Yong1, and Lei Zhang1,2(B)
1The Hong Kong Polytechnic University, Hong Kong, People’s Republic of China
{cshzheng,cshyong,cslzhang }@comp.polyu.edu.hk
2OPPO Research, Shenzhen, People’s Republic of China
Abstract. Blind image super-resolution (BISR) aims to reconstruct a
high-resolution image from its low-resolution counterpart degraded byunknown blu
r kernel and noise. Many deep neural network based meth-
ods have been proposed to tackle this challenging problem without con-
sidering the image degradation model. However, they largely rely on thetraining
sets and often fail to handle images with unseen blur kernels
during inference. Deep unfolding methods have also been proposed to
perform BISR by utilizing the degradation model. Nonetheless, the exist-ing deep

unfolding methods cannot explicitly solve the data term of the unfolding objective function, limiting their capability in blur kernel estimation. In this work, we propose a novel unfolded deep kernel estimation(UDKE) method, which, for the ■rst time to our best knowledge, explicitly solves the data term with high e■ciency. The UDKE based BISR method can jointly learn image and kernel priors in an end-to-end man-ner, and it can e■ectively exploit the information in both training data and image degradation model. Experiments on benchmark datasets and real-world data demonstrate that the proposed UDKE method could wellpredict complex unseen non-Gaussian blur kernels in inference, achieving signi■cantly better BISR performance than state-of-the-art. The source code of UDKE is available at https://github.com/natezhenghy/UDKE .

Keywords: Blind image super-resolution ·Blur kernel estimation · Unfolding method

****

Event-guided Deblurring of Unknown Exposure Time Videos

Taewoo Kim1, Jeongmin Lee1,L i nW a n g2, and Kuk-Jin Yoon1(B)
1Korea Advanced Institute of Science and Technology, Daejeon, South Korea
{intelpro,jeanmichel,kjyoon }@kaist.ac.kr
2AI Thrust, HKUST Guangzhou and Department of CSE, HKUST, Hong Kong, China
linwang@ust.hk

Abstract. Motion deblurring is a highly ill-posed problem due to the loss of motion information in the blur degradation process. Since event cameras can capture apparent motion with a high temporal resolution,several attempts have explored the potential of events for guiding deblurring. These methods generally assume that the exposure time is the same as the reciprocal of the video frame rate. However, this is not true inreal situations, and the exposure time might be unknown and dynamically varies depending on the video shooting environment (e.g., illumination condition). In this paper, we address the event-guided motiondeblurring assuming dynamically variable unknown exposure time of the frame-based camera. To this end, we ■rst derive a new formulation for event-guided motion deblurring by considering the exposureand readout time in the video frame acquisition process. We then propose a novel end-to-end learning framework for event-guided motion deblurring. In particular, we design a novel Exposure Time-based EventSelection (ETES) module to selectively use event features by estimating the cross-modal correlation between the features from blurred frames and the events. Moreover, we propose a feature fusion module to fusethe selected features from events and blur frames e■ectively. We conduct extensive experiments on various datasets and demonstrate that our method achieves state-of-the-art performance. Our project code and dataset are available at: https://intelpro.github.io/UEVD/

****

ReCoNet: Recurrent Correction Network for Fast and E■cient Multi-modality Image Fusion

Zhanbo Huang1, Jinyuan Liu2,X i nF a n1(B), Risheng Liu1,3, Wei Zhong1, and Zhongxuan Luo1
1DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Dalian, China
{xin.fan,rsliu,zhongwei,zxluo }@dlut.edu.cn
2School of Software Technology, Dalian University of Technology, Dalian, China
3Peng Cheng Laboratory, Shenzhen, China

Abstract. Recent advances in deep networks have gained great atten-

tion in infrared and visible image fusion (IVIF). Nevertheless, most existing methods are incapable of dealing with slight misalignment on source images and su■er from high computational and spatial expenses. Thispaper tackles these two critical issues rarely touched in the community by developing a recurrent correction network for robust and e■cient fusion, namely ReCoNet. Concretely, we design a deformation module toexplicitly compensate geometrical distortions and an attention mechanism to mitigate ghosting-like artifacts, respectively. Meanwhile, the network consists of a parallel dilated convolutional layer and runs in a recur-rent fashion, signi■cantly reducing both spatial and computational complexities. ReCoNet can e■ectively and e■ciently alleviates both structural distortions and textural artifacts brought by slight misalignment.Extensiv e experiments on two public datasets demonstrate the superior accuracy and e■cacy of our ReCoNet against the state-of-the-art IVIF methods. Consequently, we obtain a 16% relative improvement of CC ondatasets wit h misalignment and boost the e■ciency by 86%. The source code is available at https://github.com/dlut-dimt/reconet .

Keywords: Deep learning
·Multi-modality image fusion

1

****

Content Adaptive Latents and Decoder
for Neural Image Compression

Guanbo Pan1,G u oL u2, Zhihao Hu1, and Dong Xu3(B)
1School of Software, Beihang University, Beijing, China
2School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
3Department of Computer Science, The University of Hong Kong, Hong Kong, China
dongxu@cs.hku.hk

Abstract. In recent years, neural image compression (NIC) algorithms have shown powerful coding performance. However, most of them are notadaptive to the image content. Although several content adaptive methods have been proposed by updating the encoder-side components, the adaptability of both latents and the decoder is not well exploited. In thiswork, we propose a new NIC framework that improves the content adaptability on both latents and the decoder. Speci■cally, to remove redundancy in the latents, our content adaptive channel dropping (CACD)method automat ically selects the optimal quality levels for the latents spatially and drops the redundant channels. Additionally, we propose the content adaptive feature transformation (CAFT) method to improvedecoder-side content adaptability by extracting the characteristic information of the image content, which is then used to transform the features in the decoder side. Experimental results demonstrate that our proposedmethods w ith the encoder-side updating algorithm achieve the state-of-the-art performance.

Keywords: Neural image compression
·Content adaptive coding

1

****

E■cient and Degradation-Adaptive
Network for Real-World Image
Super-Resolution

Jie Liang1,2, Hui Zeng2, and Lei Zhang1,2(B)
1The HongKong Polytechnic University, Hung Hom, Hong Kong
2OPPO Research, Shenzhen, China
cslzhang@comp.polyu.edu.hk

Abstract. E■cient and e■ective real-world image super-resolution (Real-ISR) is a challenging task due to the unknown complex degradation of real-world images and the limited computation resources in practicalapplicati ons.RecentresearchonReal-ISRhasachievedsigni■cantprogress

by modeling the image degradation space; however, these methods largely rely on heavy backbone networks and they are in■exible to handle images of di■erent degradation levels. In this paper, we propose an e■cient and e■ective degradation-adaptive super-resolution (DASR) network, whoseparameters a re adaptively speci■ed by estimating the degradation of each input image. Speci■cally, a tiny regression network is employed to pre- dict the degradation parameters of the input image, while several convolu-tional experts with the same topology are jointly optimized to specify the network parameters via a non-linear mixture of experts. The joint opti- mization of multiple experts and the degradation-adaptive pipeline signif-icantl y extend the model capacity to handle degradations of various lev- els, while the inference remains e■cient since only one adaptively speci■ed network is used for super-resolving the input image. Our extensive experi-ments demonstrate that DASR is not only much more e■ective than exist- ing methods on handling real-world images with di■erent degradation lev- els but also e■cient for easy deployment. Codes, models and datasets areavailabl e at https://github.com/csjliang/DASR .

Keywords: Real-world image super-resolution
.
Degradation-adaptive ·E■cient super-resolution

1

****

Unidirectional Video Denoising by
Mimicking Backward Recurrent Modules
with Look-Ahead Forward Ones

Junyi Li1, Xiaohe Wu1(B), Zhenxing Niu2, and Wangmeng Zuo1
1Harbin Institute of Technology, Harbin, China
csxhwu@gmail.com ,wmzuo@hit.edu.cn
2Xidian University, Xi’an, China

Abstract. While signi■cant progress has been made in deep video denoising, it remains very challenging for exploiting historical and futureframe s. Bidirectional recurrent networks (BiRNN) have exhibited appeal- ing performance in several video restoration tasks. However, BiRNN is intrinsically o■ine because it uses backward recurrent modules to propa-gate fro m the last to current frames, which causes high latency and large memory consumption. To address the o■ine issue of BiRNN, we present a novel recurrent network consisting of forward and look-ahead recurrent modules for unidirectional video denoising. Particularly, look-ahead mod- ule is an elaborate forward module for leveraging information from near-future f rames. When denoising the current frame, the hidden features by forward and look-ahead recurrent modules are combined, thereby making it feasible to exploit both historical and near-future frames. Due to thescene m otion between non-neighboring frames, border pixels missing may occur when warping look-ahead feature from near-future frame to current frame, which can be largely alleviated by incorporating forward warpingand propo sed border enlargement. Experiments show that our method achieves state-of-the-art performance with constant latency and mem- ory consumption. Code is avaliable at https://github.com/nagejacob/ FloRNN .

Keywords: Video denoising
·Recurrent neural networks ·Temporal
alignment

1

****

Self-supervised Learning for Real-World
Super-Resolution from Dual Zoomed
Observations

Zhilu Zhang1, Ruohao Wang1, Hongzhi Zhang1(B), Yunjin Chen1,2,
and Wangmeng Zuo1,2
1Harbin Institute of Technology, Harbin, China

zhanghz0451@gmail.com, wmzuo@hit.edu.cn
2Peng Cheng Laboratory, Shenzhen, China
Abstract. In this paper, we consider two challenging issues in reference-based super-resolution (RefSR), (i) how to choose a proper reference image, and (ii) how to learn real-world RefSR in a self-supervised manner. Particularly, we present a novel self-supervised learning approach for real-worldimage SR from observations at dual camera zooms (SelfDZSR). Considering the popularity of multiple cameras in modern smartphones, the more zoomed (telephoto) image can be naturally leveraged as the reference toguide the SR of the lesser zoomed (short-focus) image. Furthermore, Self-DZSR learns a deep network to obtain the SR result of short-focus image to have the same resolution as the telephoto image. For this purpose, we takethe telephoto image instead of an additional high-resolution image as the supervision information and select a center patch from it as the reference to super-resolve the corresponding short-focus image patch. To mitigate thee■ect of the misalignment between short-focus low-resolution (LR) image and telephoto ground-truth (GT) image, we design an auxiliary-LR generator and map the GT to an auxiliary-LR while keeping the spatial positionunchanged. Then the auxiliary-LR can be utilized to deform the LR features by the proposed adaptive spatial transformer networks (AdaSTN), and match the Ref features to GT. During testing, SelfDZSR can bedirectly deployed to super-solve the whole short-focus image with the reference of telephoto image. Experiments show that our method achieves better quantitative and qualitative performance against state-of-the-arts.Codes are available at https://github.com/cszhilu1998/SelfDZSR .
Keywords: Reference-based super-resolution ·Self-supervised learning ·Real world

****

Secrets of Event-Based Optical Flow
Shintaro Shiba1,2(B), Yoshimitsu Aoki1, and Guillermo Gallego2,3
1Department of Electronics and Electrical Engineering, Faculty of Science and Technology, Keio University, Kanagawa, Japan
sshiba@keio.jp
2Department of EECS, Technische Universit¨ at Berlin, Berlin, Germany
3Einstein Center Digital Future and SCIoI Excellence Cluster, Berlin, Germany
Abstract. Event cameras respond to scene dynamics and o■er advantages to estimate motion. Following recent image-based deep-learning achievements, optical ■ow estimation methods for event cameras haverushed to combine those image-based methods with event data. However, it requires several adaptations (data conversion, loss function, etc.) as they have very di■erent properties. We develop a principled methodto extend the Contrast Maximization framework to estimate optical ■ow from events alone. We investigate key elements: how to design the objective function to prevent over■tting, how to warp events to deal betterwith occlusions, and how to improve convergence with multi-scale raw events. With these key elements, our method ranks ■rst among unsupervised methods on the MVSEC benchmark, and is competitive on the DSEC benchmark. Moreover, our method allows us to expose the issues of the ground truth ■ow in those benchmarks, and produces remarkableresults when it is transferred to unsupervised learning settings. Our code is available at https://github.com/tub-rip/event based optical ■ow.

****

Towards E■cient and Scale-Robust Ultra-High-De■nition Image Demoir´ eing
Xin Yu1, Peng Dai1,W e n b oL i2,L a nM a3, Jiajun Shen3,J i aL i4, and Xiaojuan Qi1(B)

1The University of Hong Kong, Hong Kong, China
xjqi@eee.hku.hk
2The Chinese University of Hong Kong, Hong Kong, China
3TCL AI Lab, Hong Kong, China
4Sun Yat-sen University, Guangzhou, China
Abstract. With the rapid development of mobile devices, modern
widely-used mobile phones typically allow users to capture 4K resolution
(i.e., ultra-high-de■nition) images. However, for image demoir´ eing, a chal-
lenging task in low-level vision, existing works are generally carried out on
low-resolution or synthetic images. Hence, the e■ectiveness of these meth-o d so
n4 Kr e s o l u t i o ni m a g e si ss t i l lu n k n o w n .I nt h i sp a p e
r ,w ee x p l o r e
moir´e pattern removal for ultra-high-de■nition images. To this end, we
propose the ■rst ultra-high-de■nition demoir´ eing dataset (UHDM), which
contains 5,000 real-world 4K resolution image pairs, and conduct a bench-
mark study on current state-of-the-art methods. Further, we present an
e■cient baseline model ESDNet for tackling 4K moir´ e images, wherein
we build a semantic-aligned scale-aware module to address the scale vari-
ation of moir´ e patterns. Extensive experiments manifest the e■ectiveness
of our approach, which outperforms state-of-the-art methods by a largemargin whi
le being much more lightweight. Code and dataset are avail-
able at https://xinyu-andy.github.io/uhdm-page .
Keywords: Image demoir´ eing
·Image restoration ·
Ultra-high-de■nition

****

ERDN: Equivalent Receptive Field
Deformable Network for Video
Deblurring

Bangrui Jiang1,2, Zhihuai Xie2(B), Zhen Xia2, Songnan Li2, and Shan Liu2
1Tsinghua Shenzhen International Graduate School,
Tsinghua University, Beijing, China
2Tencent Media Lab, Shenzhen, China
{zhihuaixie,zhenxia,sunnysnli,shanl }@tencent.com
Abstract. Video deblurring aims to restore sharp frames from blurry
video sequences. Existing methods usually adopt optical ■ow to compen-
sate misalignment between reference frame and each neighboring frame.However, in
accurate ■ow estimation caused by large displacements will
lead to artifacts in the warped frames. In this work, we propose an equiv-
alent receptive ■eld deformable network (ERDN) to perform alignmentat the featur
e level without estimating optical ■ow. The ERDN intro-
duces a dual pyramid alignment module, in which a feature pyramid is
constructed to align frames using deformable convolution in a cascadedmanner. Sp
eci■cally, we adopt dilated spatial pyramid blocks to predict
o■sets for deformable convolutions, so that the theoretical receptive ■eld
is equivalent for each feature pyramid layer. To restore the sharp frame,we prop
ose a gradient guided fusion module, which incorporates struc-
ture priors into the restoration process. Experimental results demon-
strate that the proposed method outperforms previous state-of-the-artmethods on
multiple benchmark datasets. The code is made available at:
https://github.com/TencentCloud/ERDN .
Keywords: Video deblurring
·Deformable convolution ·Receptive
■eld

****

Rethinking Generic Camera Models
for Deep Single Image Camera Calibration
to Recover Rotation and Fisheye

Distortion
Nobuhiko Wakai1(B), Satoshi Sato1, Yasunori Ishii1,
and Takayoshi Yamashita2
1Panasonic Holdings, Osaka, Japan
{wakai.nobuhiko,sato.satoshi,ishii.yasunori} @jp.panasonic.com
2Chubu University, Aichi, Japan
{takayoshi@isc.chubu.ac.jp}
Abstract. Although recent learning-based calibration methods can pre-
dict extrinsic and intrinsic camera parameters from a single image, theaccuracy
of these methods is degraded in ■sheye images. This degra-
dation is caused by mismatching between the actual projection and
expected projection. To address this problem, we propose a generic cam-
era model that has the potential to address various types of distor-
tion. Our generic camera model is utilized for learning-based methodsthrough a c
losed-form numerical calculation of the camera projection.
Simultaneously to recover rotation and ■sheye distortion, we propose a
learning-based calibration method that uses the camera model. Further-more, we p
ropose a loss function that alleviates the bias of the magni-
tude of errors for four extrinsic and intrinsic camera parameters. Exten-
sive experiments demonstrated that our proposed method outperformedconventional
methods on two large-scale datasets and images captured
by o■-the-shelf ■sheye cameras. Moreover, we are the ■rst researchers to
analyze the performance of learning-based methods using various typesof projecti
on for o■-the-shelf cameras.
Keywords: Camera calibration
·Fisheye camera ·Recti■cation

****
ART-SS: An Adaptive Rejection
Technique for Semi-supervised
Restoration for Adverse
Weather-A■ected Images
Rajeev Yasarla(B), Carey E. Priebe, and Vishal M. Patel
Johns Hopkins University, Baltimore, MD 21218, USA
{ryasarl1,cep,vpatel36 }@jhu.edu
Abstract. In recent years, convolutional neural network-based single
image adverse weather removal methods have achieved signi■cant perfor-
mance improvements on many benchmark datasets. However, these meth-
ods require large amounts of clean-weather degraded image pairs for train-ing, w
hich is often di■cult to obtain in practice. Although various weather
degradation synthesis methods exist in the literature, the use of synthet-
ically generated weather degraded images often results in sub-optimalperformance
on the real weatherdegraded images due to the domain gap
between synthetic and real world images. To deal with this problem, var-
ious semi-supervised restoration (SSR) methods have been proposed forderaining o
r dehazing which learn to restore clean image using synthet-
ically generated datasets while generalizing better using unlabeled real-
world images. The performance of a semi-supervised method is essentially
based on the quality of the unlabeled data. In particular, if the unlabeled
data characteristics are very di■erent from that of the labeled data, thenthe pe
rformance of a semi-supervised method degrades signi■cantly. We
theoretically study the e■ect of unlabeled data on the performance of an
SSR method and develop a technique that rejects the unlabeled imagesthat degrade
the performance. Extensive experiments and ablation study
show that the proposed sample rejection method increases the perfor-
mance of existing SSR deraining and dehazing methods signi■cantly. Codeis availa
ble at: https://github.com/rajeevyasarla/ART-SS .
Keywords: Semi-supervision
·Deraining ·Dehazing ·Rejection
techniique

1
****
Fusion from Decomposition:
A Self-Supervised Decomposition
Approach for Image Fusion
Pengwei Liang1, Junjun Jiang1(B), Xianming Liu1, and Jiayi Ma2
1Harbin Institute of Technology, Harbin 150001, China
{jiangjunjun,csxm }@hit.edu.cn
2Wuhan University, Wuhan 430072, China
Abstract. Image fusion is famous as an alternative solution to generate
one high-quality image from multiple images in addition to image restora-
tion from a single degraded image. The essence of image fusion is tointegrate co
mplementary information or best parts from source images.
The current fusion methods usually need a large number of paired sam-
ples or sophisticated loss functions and fusion rules to train the super-vised o
r unsupervised model. In this paper, we propose a powerful image
decomposition model for fusion task via the self-supervised representa-
tion learning, dubbed Decomposition for Fusion (DeFusion). With-
out any paired data or sophisticated loss, DeFusion can decompose the
source images into a feature embedding space, where the common and
unique features can be separated. Therefore, the image fusion can beachieved wit
hin the embedding space through the jointly trained recon-
struction (projection) head in the decomposition stage even without any
■ne-tuning. Thanks to the development of self-supervised learning, wecan train t
he model to learn image decomposition ability with a brute
but simple pretext task. The pretrained model allows for learning very
e■ective features that generalize well: the DeFusion is a uni■ed versatile
framework that is trained with an image fusion irrelevant dataset and
can be directly applied to various image fusion tasks. Extensive exper-iments de
monstrate that the proposed DeFusion can achieve compara-
ble or even better performance compared to state-of-the-art methods
(whether supervised or unsupervised) for di■erent image fusion tasks.
Keywords: Image fusion
·Self-supervised learning ·Image
decomposion
1
****
Learning Degradation Representations
for Image Deblurring
Dasong Li1, Yi Zhang1, Ka Chun Cheung2, Xiaogang Wang1,4,
Hongwei Qin3(B), and Hongsheng Li1,4,5(B)
1MMLab, CUHK, Hong Kong, China
dasongli@link.cuhk.edu.hk, hsli@ee.cuhk.edu.hk
2NVIDIA AI Technology Center, Santa Clara, USA
3SenseTime Research, Hong Kong, China
qinhongwei@sensetime.com
4Centre for Perceptual and Interactive Intelligence Limited, Hong Kong, China
5Xidian University, Xi’an, China
Abstract. In various learning-based image restoration tasks, such as
image denoising and image super-resolution, the degradation represen-
tations were widely used to model the degradation process and handlecomplicated
degradation patterns. However, they are less explored in
learning-based image deblurring as blur kernel estimation cannot per-
form well in real-world challenging cases. We argue that it is particu-larly nec
essary for image deblurring to model degradation representa-
tions since blurry patterns typically show much larger variations than
noisy patterns or high-frequency textures. In this paper, we propose
a framework to learn spatially adaptive degradation representations of
blurry images. A novel joint image reblurring and deblurring learningprocess is
presented to improve the expressiveness of degradation rep-

resentations. To make learned degradation representations e■ective in reblurring and deblurring, we propose a Multi-Scale Degradation Injec-tion Network (MSDI-Net) to integrate them into the neural networks. With the integration, MSDI-Net can handle various and complicated blurry patterns adaptively. Experiments on the GoPro and RealBlurdatasets demonstrate that our proposed deblurring framework with the learned degradation representations outperforms state-of-the-art methods with appealing improvements. The code is released at https://github.com/dasongli1/Learning degradation .
Keywords: Image deblurring ·Degradation representations
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19797-0 42.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13678, pp. 736–753, 2022.https://doi.org/10.1007/978-3-031-19797-0 _4
****