Computer Science and Application 计算机科学与应用, 2026, 16(1), 198-204
Published Online January 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.161016

Hans汉斯

基于结构引导Transformer的单视图三维重建去模糊方法

张媛梦，林立霞，曹　鹏

北京印刷学院信息工程学院，北京

收稿日期：2025年12月7日；录用日期：2026年1月9日；发布日期：2026年1月20日

摘　要

随着XR与AR等交互式应用的迅速发展，利用图像进行三维重建在计算机视觉领域展现出重要价值。然而，实际拍摄图像过程中普遍存在的运动模糊会削弱纹理与结构信息，显著降低三维重建的几何一致性与细节完整度。为此，本文提出了一种面向单视图三维重建任务的结构引导Transformer去模糊网络。该方法引入了显式结构先验，通过结构引导前馈网络增强Transformer在模糊区域的边缘辨识能力；同时使用多头卷积自注意力模块降低传统自注意力的计算复杂度并加强局部空间建模能力。为了验证结构恢复对三维几何推断的有效性，本文将去模糊结果输入3D Gaussian Splatting的单视图重建框架中进行评估。实验结果显示，所提方法在多项指标上均取得更优表现。

关键词

Transformer，三维重建，3D Gaussian Splatting

A Structure-Guided Transformer-Based Single-View 3D Reconstruction Deblurring Method

Yuanmeng Zhang, Lixia Lin, Peng Cao

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: December 7, 2025; accepted: January 9, 2026; published: January 20, 2026

Abstract

With the rapid development of interactive applications such as XR and AR, 3D reconstruction has demonstrated significant value in the field of computer vision. However, motion blur, which is

文章引用: 张媛梦, 林立霞, 曹鹏. 基于结构引导 Transformer 的单视图三维重建去模糊方法[J]. 计算机科学与应用, 2026, 16(1): 198-204. DOI: 10.12677/csa.2026.161016

prevalent in practice, weakens texture and structural information, significantly reducing the geometric consistency and detail integrity of 3D reconstruction. To address this, this paper proposes a structure-guided Transformer deblurring network for 3D reconstruction tasks. This method introduces an explicit structural prior and enhances the Transformer's edge recognition ability in blurred regions through a structure-guided feedforward network; simultaneously, it uses a multi-head convolutional self-attention module to reduce the computational complexity of traditional self-attention and strengthen local spatial modeling capabilities. To verify the effectiveness of structural recovery for 3D geometric inference, the deblurring results are evaluated using a single-view reconstruction framework based on 3D Gaussian Splatting. Experimental results show that the proposed method achieves superior performance on multiple metrics.

Keywords

Transformer, 3D Reconstruction, 3D Gaussian Splatting

1. 引言

近年来，随着 XR、AR 等交互式应用的发展，三维内容的制作和理解正在从专业领域逐渐走向大众场景。为了降低模型构建门槛，提高数据处理效率，三维重建技术例如神经辐射场[1] (Neural Radiance Fields, NeRF)、3D 高斯溅射[2] (3D Gaussian Splatting,3D GS)成为当前学术界研究的重点方向。其中，得益于深度学习模型在先验建模能力上的提升，近年的三维重建方法已能够从单视图中推测出具有一定几何一致性的三维结构，并在多类物体上取得稳定效果。同时，单视图三维重建[3]因其仅依赖单张图像即可推断物体的三维形状，而在空间建模、机器人感知以及移动端应用中展现出较高的应用价值。

然而，单视图三维重建高度依赖输入图像的质量。实际场景中，由于拍摄者移动、光线不足或设备限制，图像常出现不同程度的运动模糊。模糊会掩盖边缘、削弱纹理，进而破坏深度与法线估计的可靠性，使网络难以恢复正确的物体几何形状。在单视图条件下，这一问题尤其突出，运动模糊会导致单张视图提供的信息受损，使得模型难以有效进行三维重建。当前，已有单视图三维重建方法[4]-[6]通常假设输入图像清晰，几乎不考虑退化情况下的重建问题。

因此，本文考虑了使用单张模糊视图进行三维重建的场景，提出一个适用于三维重建任务的单图像去模糊模型。考虑到三维重建对边缘和结构一致性的依赖更高，而现有去模糊模型更关注视觉效果本身，本文提出了一种结构引导的 Transformer 去模糊框架。首先，对 Transformer 的前馈网络(Feedforward Neural Network, FFN)进行了结构感知改进，通过提取单幅图像的梯度作为明确的结构先验，并在 Transformer 的前馈层注入结构信息，以增强模糊区域的边缘恢复能力。此外，本文引入了多头卷积自注意力模块[7]，该模块相较于 Transformer 的自注意力模块，具有更低的计算复杂度。最后，为了验证模型恢复的结构是否真正有利于三维重建，本文进一步将去模糊结果输入单视图 3D GS 渲染模型[8]中进行三维重建。最后，在真实物体数据集上进行实验的结果表明，该方法在多项指标上的效果均具有竞争力。

本文的主要工作包括：

1. 提出一个面向单视图三维重建任务的结构引导的 Transformer 去模糊框架；

2. 为了降低计算复杂度，引入了多头卷积自注意力模块。同时，为了使去模糊图像更适用于三维重建任务，设计了结构感知前馈网络，使模型在恢复模糊细节时更关注边缘与几何相关特征；

3. 在 3D Gaussian Splatting 渲染框架下验证去模糊结果的实际三维重建性能，实验证明模型能够显著提升重建质量。

2. 模型架构

Transformer 模型最初是为自然语言任务中的序列处理而开发的。当前，它已经被适应于许多视觉任务[9]。本文所提出的基于 Transformer 的去模糊网络模型架构如图 1 所示，其中图 1 上半部分为完整系统流程，包括去模糊网络模块和单视图三维重建模型。去模糊网络引入了输入图像的结构先验信息，有助于网络在图像恢复过程中细化边缘信息，经去模糊网络恢复后的图像输入至单视图 3D GS 重建模型中进行三维重建。单视图三维重建模型采用文献[8]提出的基于 U-Net 的 3D GS 模型，并使用其公开的预训练权重作为三维渲染后端。本文重点讨论去模糊网络模型，三维重建模型具体实现细节超出了本文所讨论的范围。图 1 下半部分为去模糊网络中 Transformer 层具体实现细节，包括多头卷积自注意力模块、结构引导模块和结构引导前馈网络，其中多头卷积自注意力模块相较于传统的自注意力机制具有更小的计算复杂度，结构引导模块通过 Sobel 卷积提取结构先验信息，结构引导前馈网络通过融合图像深层特征和边缘信息，在特征变换过程中引入结构位置约束。

2.1. 结构引导的 Transformer 去模糊网络

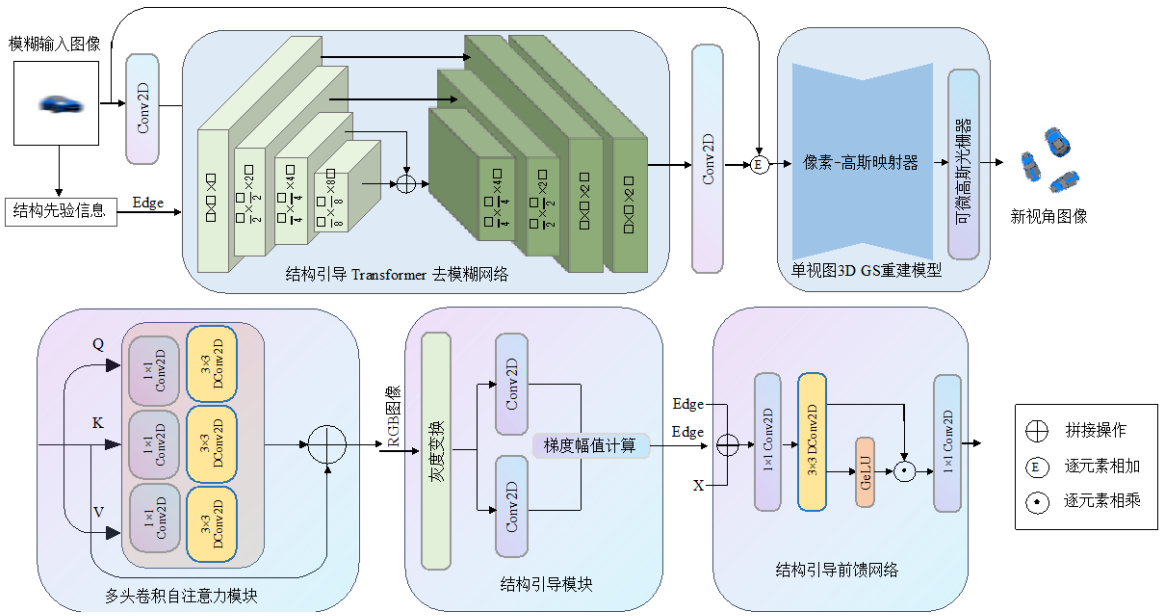

Figure 1. Structure-guided transformer deblurring network model architecture
图 1. 结构引导 Transformer 去模糊网络模型架构

去模糊网络采用经典的编码器－解码器架构。输入模糊图像首先通过一个二维卷积提取浅层特征表示，随后输入到去模糊网络的四级编码器。每一级编码器由若干个 Transformer 层组成，层数分别为 4，6，6 和 8。编码过程中，编码器逐步减少空间尺寸，提高特征嵌入维度。解码器同样由 4 层网络构成，每层包含与编码器对应的 Transformer 层，层数分别为 8，6，6 和 4。第四层编码器的输出作为解码器输入，解码器逐步恢复空间尺寸和特征嵌入维度。每层网络的上采样和下采样采用文献[10]所提出的像素混洗方法。此外，本文使用残差连接将对应级别编解码器的输出特征进行聚合，以保留输入图像中的精细结构和纹理细节。最后，对细化后的特征进行卷积，生成残差图像与退化图像进行逐元素相加得到恢复图像。

2.2. 多头卷积自注意力模块

为了减少 Transformer 自注意力机制的计算复杂度，本文引入了多头卷积自注意力模块，如图 1 所示，原始自注意力机制的计算量随空间分辨率呈二次增长，在图像恢复任务中成本极高。而该模块通过在通道维度而非空间维度上计算 Attention，从根本上避免了大规模空间交互，从而将计算复杂度由空间分辨率主导转化为由通道数主导，显著降低了时间与显存开销。同时，在生成查询(Q)、键(K)与值(V)向量之前，模块采用 1 × 1 点卷积增强跨通道表达能力，并通过 3 × 3 深度可分离卷积引入局部空间上下文，使得注意力既具备全局感受野，又保留了局部感知能力。

2.3. 结构引导前馈网络

本文提出的结构引导前馈网络，目标是在保持前馈网络层高效的同时，提升其对图像结构的感知能力，使得去模糊网络输出结果适用于三维重建任务。为了让网络在特征变换过程中显式关注图像结构，首先利用输入图像生成其 Sobel 边缘图。Sobel 算子能够通过局部梯度响应有效刻画图像中的亮度变化方向，对物体轮廓和结构边界具有较强的敏感性，同时，边缘图作为一种稳定的局部结构先验，与深层特征相比更不易受模糊退化影响，因此能够为网络提供可靠结构约束。具体实现如图 1 所示，结构引导模块通过灰度映射与两个 3 × 3 的 Sobel 卷积在水平方向和垂直方向上计算一阶梯度响应。随后，对两个方向梯度进行融合，生成单通道的结构引导特征图，用于为后续特征建模提供稳定的结构先验。

结构引导前馈网络具体实现细节如图 1 所示，首先，编解码器输出的深层特征与由结构引导模块提取的边缘特征 Edge 通过一个二维卷积层进行融合，随后通过一个门控机制，门控机制由两个平行路径的逐元素乘积实现，其中一条路径由 GeLU 函数激活。其中，深度卷积可以对相邻像素的信息进行编码，这有利于模型学习局部图像结构。从整体来看，结构引导前馈网络通过控制编解码器不同层级的信息交互，它使得去模糊模型的每一层网络聚焦于自身特定的任务的同时，与其他层网络信息互补。

2.4. 损失函数

在以三维重建为最终目标的去模糊任务中，损失函数的设计不仅需要关注图像空间上的复原精度，更必须兼顾结构与几何细节的稳定恢复。由于三维重建过程对图像局部梯度和边缘一致性高度敏感，单纯依赖像素域的 $L1$ 损失往往导致模型在整体亮度和纹理上收敛良好，但在结构细节上依旧存在模糊或偏移。因此，本文将重建质量与结构约束结合，构建了由像素一致性与梯度一致性共同组成的损失框架，如式(1)所示。

$$L = L_{L1} + \lambda L_{grad} \tag{1}$$

其中 $L_{L1} = \left\| \hat{I} - I_{gt} \right\|_1$，$\hat{I}$ 表示模糊后的结构图像，I_{gt} 代表原始图像。梯度一致性损失基于 Sobel 算子计算图像的水平与垂直梯度，其中 $L_{grad} = \left\| \nabla_x \hat{I} - \nabla_x I_{gt} \right\|_1 + \left\| \nabla_y \hat{I} - \nabla_y I_{gt} \right\|_1$，$\nabla_x$ 和 ∇_y 分别代表图像水平方向与垂直方向的离散梯度算子。超参数 λ 用于控制梯度约束的强度。

3. 实验结果与分析

3.1. 数据集和运动模糊建模

去模糊网络训练使用 GoPro 数据集[11]，该数据集包含上千张运动模糊图像。同时，每张模糊图像都与其对应的清晰图像成对存在，便于监督学习模型的训练。针对测试阶段的三维重建任务测试数据集，本文采用谷歌扫描数据集，谷歌扫描数据集包含上千种真实家用物品模型。

由于缺乏大规模，具备真实运动模糊的多视角图像及其对应相机参数的公开数据集，为模拟真实场

景中的运动模糊现象，与文献[12]类似，本文使用在所选择的测试数据上合成了运动模糊。首先，本文基于图像退化数学模型对数据集中图像进行模糊处理，生成具有线性运动模糊特征的图像，其计算方式为：

$$G(u,v)=H(u,v)F(u,v) \tag{2}$$

其中，$F(u,v)$ 和 $G(u,v)$ 分别对应清晰图像与模糊图像的频域表示，传递函数 $H(u,v)$ 定义为:

$$H(u,v)=\frac{T}{\pi(ua+vb)}\sin\left[\pi(ua+vb)\right]e^{-j\pi(ua+vb)} \tag{3}$$

该模型将运动模糊建模为图像与运动核的卷积过程，能够有效模拟相机在曝光期间的线性位移引起的模糊效应。其中，T 表示曝光时间，参数 a 和 b 分别表示在相机曝光时间 T 内，图像在水平方向和竖直方向上的位移距离。

3.2. 评估指标

本文将新视角图像合成的表现作为衡量三维重建模型质量的核心依据，通过多项指标综合评估不同方法的重建精度与可感知一致性。具体而言，峰值信噪比(Peak Signal to Noise Ratio, PSNR)、结构相似性指数(Structural Similarity Index Measure, SSIM)以及学习感知图像块相似度(Learned Perceptual Image Patch Similarity, LPIPS)共同构成了图像质量评价体系。PSNR 依托均方误差(Mean Square Error, MSE)度量预测图像与参考图像在数值层面的偏差；SSIM 从亮度、对比度与结构三方面模拟人类视觉感知，用于检验图像的结构性保真度；LPIPS 则基于深度特征空间计算两图像之间的感知距离，更贴近真实视觉主观体验，

3.3. 结果分析

为了验证所提方法的有效性，表 1 中展示了本文的方法与主流去模糊方法的性能对比，其中 Restormer 模型[7]作为基于 Transformer 的代表性去模糊网络，在图像去模糊任务中展现出了优异的性能。SI 为图 1 中展示的基于 3D GS 的单视图三维重建模型，OpenLRM 为一种基于 Transformer 架构的三维重建模型。从表 1 中可以看出，在谷歌扫描数据集上，在输入为清晰图像时，SI 模型在三项指标上均具有最优的表现，而在模糊输入下，其性能表现下降严重，原因在于 SI 模型完全在纯净的输入样本上进行训练，对模糊图像缺少鲁棒性。在使用 Restormer 模型对模糊图像进行去模糊后，SI 模型的重建性能略微提升，但整体性能仍弱于本文方法，其原因在于本文的模型在设计上针对三维重建任务做了特定优化，同时考虑像素恢复和几何特征约束，能够更有效的恢复关键结构信息。此外，OpenLRM 本身重建性能不及 SI 模型，因此即使在 Restormer 对图像进行去模糊后，结果依旧不理想。

Table 1. Comparison of experimental metrics on the Google Scan dataset
表 1. 在谷歌扫描数据集下实验指标对比

Method	PSNR	SSIM	LPIPS
模糊输入 + SI	15.51	0.81	0.205
Restormer + SI	17.46	0.84	0.156
Restormer + OpenLRM	16.42	0.79	0.278
Ours + SI	**18.04**	**0.85**	**0.147**
GT 输入 + SI	20.62	0.93	0.122

图 2 展示了三维重建模型在新视角渲染结果上的可视化对比。由图可见，在输入图像存在明显模糊

的情况下，Restormer 去模糊后的 OpenLRM 与 SI 模型仍难以完整恢复物体的细节结构，其生成结果在鞋面纹理、局部颜色区域上均存在不同程度的缺失或偏移。相比之下，本文方法能够更为准确地重建鞋面白色条纹、黄色鞋舌等细节，同时在包含小男孩鞋子的区域也表现出更优的结构一致性。

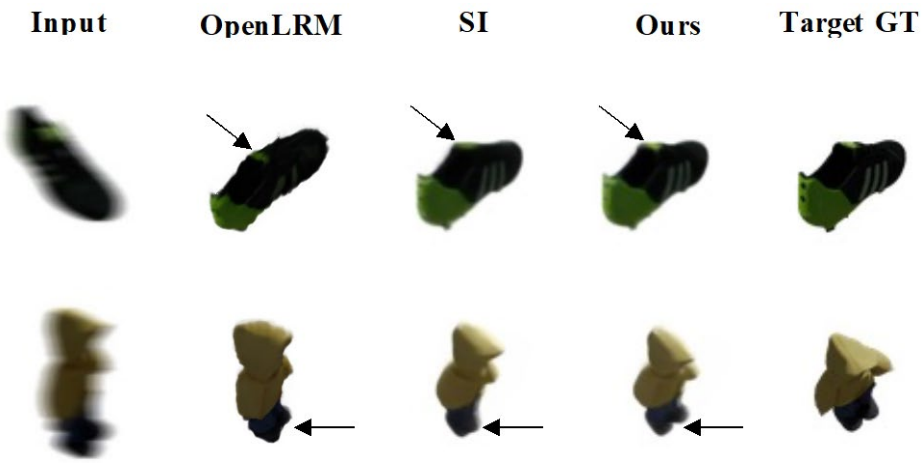

Figure 2. Google Scans dataset: A new perspective on generating image comparisons
图 2. 谷歌扫描数据集新视角生成图像对比

消融实验结果如表 2 所示，其中 Ours w/o SGFN 代表无结构引导的去模糊 Transformer 模型，Ours w/o L_grad 代表不包含梯度一致性的损失函数框架。从表 2 中可以看出，本文所提出的模型具有最优的重建性能，Ours w/o SGFN 由于缺乏对图像结构信息的感知，在特征变换过程中难以显式关注图像结构，导致其最终重建效果并不理想。Ours w/o L_grad 在损失函数层面未对梯度一致性进行约束，使得模型在图像恢复中难以兼顾几何细节。Ours w/o SGFN，L_grad 模型既无法有效感知图像结构信息，又缺乏对局部几何细节的约束，其三维重建效果下降最为显著。

Table 2. Comparison of ablation experiment metrics on the Google Scan dataset
表 2. 在谷歌扫描数据集下消融实验指标对比

Method	PSNR	SSIM	LPIPS
Ours w/o SGFN, L_grad + SI	17.46	0.84	0.156
Ours w/o SGFN + SI	17.71	0.84	0.153
Ours w/o L_grad + SI	17.89	0.85	0.149
Our s+ SI	**18.04**	**0.85**	**0.147**

4. 结语

本文围绕单视图三维重建任务在运动模糊场景下的性能下降问题，提出了一种结构引导的 Transformer 去模糊框架，以结构先验为核心、结合多头卷积自注意力机制与结构感知前馈网络，增强了模型在模糊区域的边缘恢复能力与几何一致性。在真实物体数据集上的实验结果表明，该方法在单视图三维重建任务中显著提升了新视角合成的表现。与主流去模糊模型相比，本文方法在 PSNR、SSIM 与 LPIPS 等指标上均表现出更强的性能。未来的工作将致力于研究能够直接处理模糊输入的三维重建模型，使去模糊与重建过程在同一网络中协同学习。

参考文献

[1] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R. and Ng, R. (2021) NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, **65**, 99-106. https://doi.org/10.1145/3503250

[2] Kerbl, B., Kopanas, G., Leimkuehler, T. and Drettakis, G. (2023) 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, **42**, 1-14. https://doi.org/10.1145/3592433

[3] Fu, K., Peng, J., He, Q. and Zhang, H. (2020) Single Image 3D Object Reconstruction Based on Deep Learning: A Review. *Multimedia Tools and Applications*, **80**, 463-498. https://doi.org/10.1007/s11042-020-09722-8

[4] Yang, S., Zhang, H., Ren, J., Tang, Z., Zhao, M. and Liu, Y. (2025) Zero-1-to-3DGS: A Single Image to 3D Gaussian by Consistent Multi-View Generation. 2025 *IEEE International Conference on Multimedia and Expo* (*ICME*), Nantes, 30 June-4 July 2025, 1-6. https://doi.org/10.1109/icme59968.2025.11209455

[5] Xu, D., Jiang, Y., Wang, P., Fan, Z., Shi, H. and Wang, Z. (2022) SinNeRF: Training Neural Radiance Fields on Complex Scenes from a Single Image. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M. and Hassner, T., Eds., *Computer Vision—ECCV* 2022, Springer, 736-753. https://doi.org/10.1007/978-3-031-20047-2_42

[6] Yu, A., Ye, V., Tancik, M. and Kanazawa, A. (2021) PixelNeRF: Neural Radiance Fields from One or Few Images. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 4576-4585. https://doi.org/10.1109/cvpr46437.2021.00455

[7] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S. and Yang, M. (2022) Restormer: Efficient Transformer for High-Resolution Image Restoration. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 5718-5729. https://doi.org/10.1109/cvpr52688.2022.00564

[8] Szymanowicz, S., Rupprecht, C. and Vedaldi, A. (2024) Splatter Image: Ultra-Fast Single-View 3D Reconstruction. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 10208-10217. https://doi.org/10.1109/cvpr52733.2024.00972

[9] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., *et al.* (2023) A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 87-110. https://doi.org/10.1109/tpami.2022.3152247

[10] Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A.P., Bishop, R., *et al.* (2016) Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 1874-1883. https://doi.org/10.1109/cvpr.2016.207

[11] Nah, S., Kim, T.H. and Lee, K.M. (2017) Deep Multi-Scale Convolutional Neural Network for Dynamic Scene Deblurring. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 257-265. https://doi.org/10.1109/cvpr.2017.35

[12] Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D. and Matas, J. (2018) DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8183-8192. https://doi.org/10.1109/cvpr.2018.00854