Computer Science and Application 计算机科学与应用, 2026, 16(5), 392-405
Published Online May 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.165192

Hans汉斯

基于高效时空建模的多帧融合视频去雾模型

邵玉娇，魏伟波*，潘振宽

青岛大学计算机科学技术学院，山东 青岛

收稿日期：2026年4月23日；录用日期：2026年5月22日；发布日期：2026年5月29日

摘　要

针对视频去雾任务中时空信息利用不足导致去雾后视频连贯性差的问题，文章提出了一种基于编码器–解码器的类U-Net网络的新型视频去雾模型(UnDehazeNet)。该模型借鉴先进时空建模思想，构建局部与全局协同的特征学习机制，无需依赖显式光流计算即可捕捉帧间运动规律与雾气分布特性，有效控制算力开销。同时，在编码器–解码器部分集成可变形三维卷积，强化多尺度特征的挖掘与融合能力，充分发挥类U-Net架构在图像恢复领域的优势，该模型能够在有效实现高质量视频去雾的同时，保证视频的连贯性。在REVIDE与HazeWorld数据集上的实验结果表明，UnDehazeNet在峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)、结构相似性指数(Structural Similarity Index Measure, SSIM)两项核心定量指标及定性可视化效果中均表现更优，综合性能显著提升。

关键词

视频去雾，时空建模，深度学习，3D卷积，帧间融合，神经网络

Multi-Frame Fusion Video Dehazing Model Based on Efficient Spatiotemporal Modeling

Yujiao Shao, Weibo Wei*, Zhenkuan Pan

College of Computer Science and Technology, Qingdao University, Qingdao Shandong

Received: April 23, 2026; accepted: May 22, 2026; published: May 29, 2026

Abstract

To address the problem of insufficient spatiotemporal information utilization in video dehazing tasks, which often results in poor temporal coherence of dehazed videos, this paper proposes a novel video dehazing model based on an encoder-decoder U-Net-like architecture, termed UnDehazeNet. Drawing on advanced spatiotemporal modeling concepts, the model constructs a feature learning mechanism that coordinates local and global interactions, enabling it to capture inter-frame motion patterns and

*通讯作者。

文章引用: 邵玉娇, 魏伟波, 潘振宽. 基于高效时空建模的多帧融合视频去雾模型[J]. 计算机科学与应用, 2026, 16(5): 392-405. DOI: 10.12677/csa.2026.165192

fog distribution characteristics without relying on explicit optical flow computation, thereby effectively controlling computational overhead. Meanwhile, deformable 3D convolutions are integrated into the encoder-decoder to enhance the extraction and fusion of multi-scale features, fully leveraging the inherent advantages of the U-Net-like architecture in image restoration. Consequently, the proposed model achieves high-quality video dehazing while ensuring temporal coherence. Experimental results on the REVIDE and HazeWorld datasets demonstrate that UnDehazeNet outperforms comparative methods in both core quantitative metrics (PSNR and SSIM) as well as qualitative visualizations, with significantly improved overall performance.

Keywords

Video Dehazing, Spatiotemporal Modeling, Deep Learning, 3D Convolution, Inter-Frame Fusion, Neural Network

1. 引言

智能交通、无人驾驶等领域对视频图像清晰度的需求日益严苛[1]，图像细节完整性直接决定目标识别、场景理解等下游任务的执行精度。然而，真实场景中的雾天环境会严重削弱图像的对比度和清晰度，不仅导致视觉系统识别率显著下降，还会对后续计算机视觉任务(如目标检测、图像分割、人员再识别)的研究造成负面影响[2]，甚至可能产生安全事故与经济损失。因此，开展雾天视频的清晰化恢复研究，实现有雾视频中清晰帧的有效还原，具有重要的实际意义和应用价值。在图像与视频去雾领域，大气散射模型(Atmospheric Scattering Model, ASM) [3]是阐释雾效成像机制的经典理论基础，被学术界广泛认可。该模型指出，雾天相机接收的图像信号是物体表面反射光与大气颗粒散射背景光的叠加产物，这种叠加效应是雾天图像模糊、细节丢失的核心原因，也为去雾算法设计提供了物理依据。

早期去雾研究主要围绕单幅图像展开，经历了从传统图像增强到物理模型驱动，再到深度学习[4]主导的演进过程。初期方法多依赖直方图均衡化等直接增强手段[5]，虽能提升对比度，但未契合雾天成像物理规律，易出现颜色失真、细节过度增强等问题。物理模型驱动阶段以暗通道先验理论[6]的提出为标志，通过挖掘无雾图像特征估计去雾参数，实现了更贴合物理逻辑的去雾效果，却在强光照以及大面积天空复杂场景中存在明显适用局限。深度学习技术的兴起，推动数据驱动型去雾方法成为主流，形成两条核心技术路径：一是物理模型引导型[7]通过神经网络预测投射图、大气光等中间参数，结合 ASM 模型重建无雾图像；二是端到端映射型[8] [9]，直接学习有雾与无雾图像的非线性映射关系，跳过中间参数估计，场景适配性更强。

单幅图像去雾技术的成熟为视频去雾提供了基础支撑，早期视频去雾多直接采用单幅图像方法进行逐帧处理，该方法忽略了帧间时序互补信息，导致处理后的视频出现帧间闪烁、伪影等时间不一致问题。针对这一问题，研究人员提出逐帧去雾 + 后处理矫正框架，在单帧去雾后通过时序优化修正一致性偏差[10]-[12]，该方法将视频去雾向整体化推进。随后在深度学习的驱动下，研究重心从后处理矫正转向原生时序建模，核心突破在于强化模型对时空信息的融合能力，以此从根源上提升去雾效果与帧间连贯性[13]。相关研究一方面通过深度挖掘视频时序特征优化透射图估计，甚至将去雾网络与目标检测任务联合优化，实现雾天视觉恢复与下游任务精度的同步提升；另一方面融合语义分割技术，基于相邻帧物理特性假设

完成跨帧特征融合，引入语义先验强化复杂场景下的去雾鲁棒性[14]。数据集构建与算法创新的协同推进，是视频去雾技术落地的重要支撑[15]。真实场景数据集的出现，为模型实际性能验证提供了可靠依据，配套算法通过置信度引导机制进一步提升了去雾稳定性；无监督学习思路也逐步应用于该领域，通过引入深度信息模拟跨帧雾气变化与运动状态，利用合成帧优化时空一致性[16]；同时，部分研究通过记忆模块融入物理先验特征，强化长期时序信息的挖掘与利用，丰富了视频去雾的技术体系[17] [18]。尽管现有技术已取得显著进展，但仍存在核心瓶颈：时序建模层面，部分方法仅通过简单帧拼接实现时序关联捕捉，未能深度挖掘特征提取过程中的帧间依赖，一致性优化效果有限；物理驱动层面，方法性能强依赖透射率 t(x)和大气光 A 的精准估计，在复杂光照、遮挡、动态场景中参数估计易产生误差；泛化能力层面，多数方法在浓雾、雾气分布不均场景中恢复效果欠佳，但面对雾强度、摄像环境变化的适配性不足，且对参数误差敏感，易出现雾残留、色彩偏差等问题。

针对上述问题，本文提出一种基于编码器－解码器的类 U-Net 视频去雾网络模型(UnDehazeNet)，通过模块化架构设计突破现有技术瓶颈。该模型首先创新性设计了基于 Transformer 的高效时空建模模块(Efficient Spatiotemporal Modeling Module, ESMM)，其中 ESMM 模块的设计思路受 Uniformer [19] [20]启发。Uniformer 作为融合卷积与自注意力机制的经典通用视觉框架，其在视频领域拓展的核心优势在于解决传统 3D 卷积感受野有限、纯 Transformer 全局注意力冗余的问题，为视频时序建模提供了高效的技术范式，在视频分类、目标检测等多种视觉任务中取得优异性能。本文借鉴 Uniformer 卷积与自注意力融合的时序建模核心思路，通过针对性设计其中的全局与局部注意力机制，使其适配于视频去雾。其次本文在编码器与解码器结构中集成可变形三维卷积(3D Deformable Convolution, D3D)，并结合跨阶段特征融合机制，三者协同作用，显著提升雾天视频的清晰化恢复效果。

2. 方法

2.1. 整体网络架构

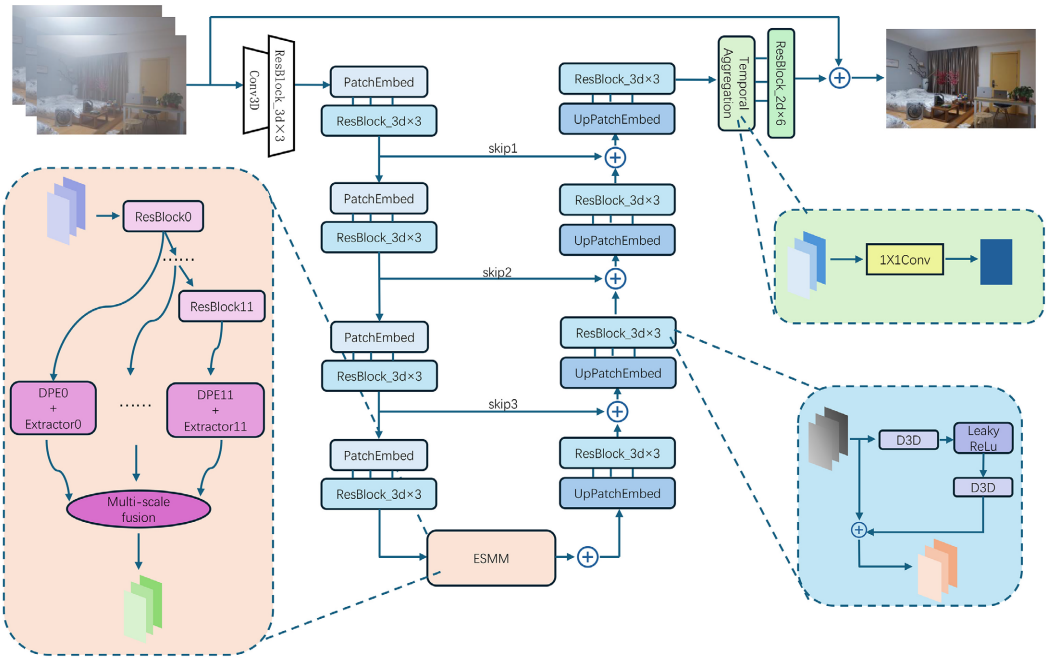

Figure 1. UnDehazeNet network model structure diagram
图 1. UnDehazeNet 网络模型结构图

基于高效时空建模模块的多帧融合视频去雾方法 UnDehazeNet 的网络模型结构图如图 1 所示。具体而言，UnDehazeNet 模型采用基于编码器–解码器结构的类 U-Net 网络架构，主要由编码器、ESMM 模块以及解码器三个部分组成。该模型设计的核心在于通过编码器逐步提取输入视频帧的多尺度时空特征，经由 ESMM 模块对编码特征进行全局–局部协同的时空关系建模与增强，最后由解码器将处理后的特征逐步恢复为清晰的视频帧，从而实现端到端的视频去雾处理。

编码器部分由 D3D 与下采样操作构成。其中，D3D 能够在时空维度上对视频序列进行自适应采样，从而更加有效地利用帧间时空信息，具备较强的时空特征建模能力以及更高的运动感知灵活性。在此基础上，通过多级编码结构逐步完成特征提取与分辨率压缩，并借助跳跃连接将不同层级的特征信息传递至解码阶段，以缓解深层网络中的信息丢失问题，同时保留丰富的低层细节特征。

在完成浅层特征提取之后，编码特征被输入至 ESMM 模块。该模块通过多头自注意力机制对全局上下文信息进行建模，同时结合时序注意力机制显式刻画视频帧之间的依赖关系，从而进一步增强网络对时空相关特征的表达能力，提高对动态场景中雾气分布变化的建模能力。解码器部分采用基于 PixelShuffle [21]的非对称上采样结构，并结合特征融合模块对来自不同层级的多尺度特征进行有效整合，以逐步恢复高分辨率特征表示并增强细节重建能力。与传统转置卷积上采样方式相比，PixelShuffle 通过像素重排操作实现特征图分辨率的提升，能够有效避免棋盘伪影等问题，生成更加自然流畅的图像细节。最后，对多帧去雾结果进行自适应融合，并通过残差连接将融合特征与输入信息相结合，输出最终的去雾图像，在保证去雾效果的同时保留原始图像的底层结构信息。

为验证本文所提 UnDehazeNet 模型的计算效率与资源占用情况，本节从参数量、单帧推理速度、峰值 GPU 显存占用三个维度，将其与当前主流视频去雾模型 MAP-Net、DCL 进行定量对比。所有对比模型均在相同硬件环境与测试配置下运行，保证结果公平可靠。对比结果如表 1 所示。

Table 1. Comparison of the complexity of different models
表 1. 不同模型的复杂度对比

模型	参数量(M) ↓	单帧 FPS ↑	峰值 GPU 显存(MB) ↓
MAP-Net	28.75	18.30	276.57
DCL	26.22	14.76	451.86
UNDehazeNet	**11.29**	**13.11**	**194.08**

从表 1 可以看出，UnDehazeNet 的参数量仅为 MAP-Net 的 39.3%、DCL 的 43.1%；峰值显存比 MAP-Net 降低 29.8%，比 DCL 降低 57.1%。推理速度方面，UnDehazeNet 达到 13.11 FPS，与 DCL (14.76 FPS) 相近，略低于 MAP-Net (18.30 FPS)。上述结果表明，本文模型在保持可接受推理速度的同时，显著降低了存储与内存开销。

2.2. ESMM 模块

ESMM 模块是解决视频边缘闪烁、保证帧间时序连贯性的核心组件。其以 Transformer 形式融合 3D 卷积与时空自注意力，构建局部–全局协同的时空建模机制，实现视频特征的高效表达。

网络结构中，ESMM 嵌入编码器与解码器之间，接收多尺度特征序列。通过残差注意力块(ResBlock)搭建协同框架：空间域利用多头自注意力(Multi-Head Self-Attention, MHSA)建立跨区域像素关联，提取多尺度雾浓度特征；时间域通过局部时空关系注意力(Local Spatiotemporal Relation Attention, LSRA)结合分组 3D 卷积，在时空邻域内高效建模帧间运动模式，隐式学习动态变化规律。

模块进一步通过深度位置编码(Depth-Wise Positional Encoding, DPE)的 3D 卷积实现特征校准，显式编码空间与时序位置信息。帧间特征提取模块(Extractor)实现双向注意力交互：对中间帧分别计算其与前向帧、后向帧的注意力响应，自适应聚合相邻帧特征，增强时序关联。最后，引入通道注意力的多级特征融合模块，对不同层级特征自适应加权，输出高时空一致性的深层特征序列，为解码器重建无雾帧提供支撑。

1) 全局时空建模

全局建模通过标准多头注意力实现，主要用于空间全局建模，每个位置计算与同一帧内所有位置的关系，对于帧 t 中的位置 i，其输出表示为：

$$Output_i = \sum_{j\in \text{同一帧 } t} \alpha_{ij} \cdot V_j \tag{1}$$

其中：

$$\alpha_{ij} = \text{softmax}\left(\frac{Q_x K_y^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{2}$$

其中，α_{ij} 表示位置 i 与位置 j 之间的注意力权重，用于衡量两个位置特征的关联强度；V_j 表示位置 j 的值向量张量，d_k 为查询向量与键向量的维度，用于归一化注意力权重，避免数值过大导致的 softmax 饱和问题。

2) 局部时空关系注意力模块

时空局部窗口内捕获雾气浓度变化，有效适配雾气非均匀分布特性，隐式学习帧间运动模式，同时 LSRA 采用分组卷积与通道降维(dw_reduction = 1.5)减少 3D 卷积计算量，其计算过程分为以下 4 个阶段：

① 输入归一化：对输入特征张量进行三维批量归一化处理，消除通道间的分布差异，加速模型训练收敛，计算公式如下：

$$X_{norm} = BN_3(X) \tag{3}$$

其中，$X \in R^{B\times C\times T\times H\times W}$ 是输入特征张量，B 为批量大小，C 为特征通道数，T 为时间步数(连续帧数量)，H 和 W 分别为特征图的高度与宽度；BN_3 表示针对三维特征张量的批量归一化运算。

② 降维映射：采用 $1\times1\times1$ 三维卷积对归一化后的特征进行通道降维，减少后续计算量，计算公式如下：

$$X_{red} = W_{1*3D} X_{norm} \tag{4}$$

其中，$1\times1\times1$ 卷积核将通道数从 C 减少至 $C' = \lfloor C/r \rfloor$，$r = 1.5$ 为通道降维比率，$W_1 \in R^{C'\times C\times1\times1\times1}$ 为三维卷积核函数。

③ 时间维度建模：采用深度可分离卷积在时间维度上建模帧间关联，捕捉雾气的动态变化，计算公式如下：

$$X_{temp} = W_{2*3D} X_{red} \tag{5}$$

其中，$W_{2*3D} \in R^{C'\times1\times k_t\times1\times1}$ 为深度可分离卷积核，$k_t = 3$ 为时间卷积核大小。为保持时间维度 T 不变，设置填充 $p = \lfloor kt/2 \rfloor$，确保输出与输入的时间长度一致，便于后续特征融合。

④ 恢复通道维度：再次采用 $1\times1\times1$ 三维卷积将降维后的特征通道数恢复至原始维度 C，得到最终的位置嵌入特征，计算公式如下：

$$E_{pos} = W_{3*3D} X_{temp} \tag{6}$$

其中，$W_{3*3D} \in R^{C'\times C\times1\times1\times1}$ 为逐点卷积核，最终输出位置嵌入特征 $E_{pos} \in R^{B\times C\times T\times H\times W}$。该模块与全局时空建模

模块通过残差连接实现局部与全局特征的深度融合，既保证了全局雾浓度分布的捕捉，又兼顾了局部雾景细节与帧间运动模式的建模。

3) 帧间特征提取模块

Extractor 模块用于实现相邻帧之间的双向注意力交互，强化帧间时序关联，其核心是通过跨模态注意力机制隐式建模帧间依赖关系，最终输出融合时序信息的当前帧特征。该模块接收两个输入张量：当前帧特征 x 与相邻帧特征 y，核心注意力计算过程如下：

$$\text{Attention}(x)=\text{softmax}\left(\frac{Q_x K_y^{\mathrm{T}}}{\sqrt{d_k}}\right)V_y \tag{7}$$

其中，查询向量 Q_x、键向量 K_y 与值向量 V_y 分别来自不同的输入分支：查询向量 Q_x 由当前帧特征生成，即 $Q_x=\text{LayerNorm}_1(x)W_Q$，$W_Q$ 表示查询向量的投影矩阵张量；键向量 K_y 与值向量 V_y 由相邻帧特征生成，即 $K_y=\text{LayerNorm}_3(y)W_K$、$V_y=\text{LayerNorm}_3(y)W_V$，$W_K$、$W_V$ 分别为键向量与值向量的投影矩阵张量。LayerNorm 层用于对输入特征进行归一化，提升注意力机制的稳定性与建模效果。

2.3. 可变形三维卷积

D3D 通过将可变形卷积与传统三维卷积的创新性结合，实现了对视频序列时空特征的高效建模。D3D 操作示意图如图 2 所示。与传统固定采样网格的三维卷积不同，D3D 可通过学习偏移量自适应调整采样位置，能够更好地匹配雾气非均匀分布与场景运动带来的形变。

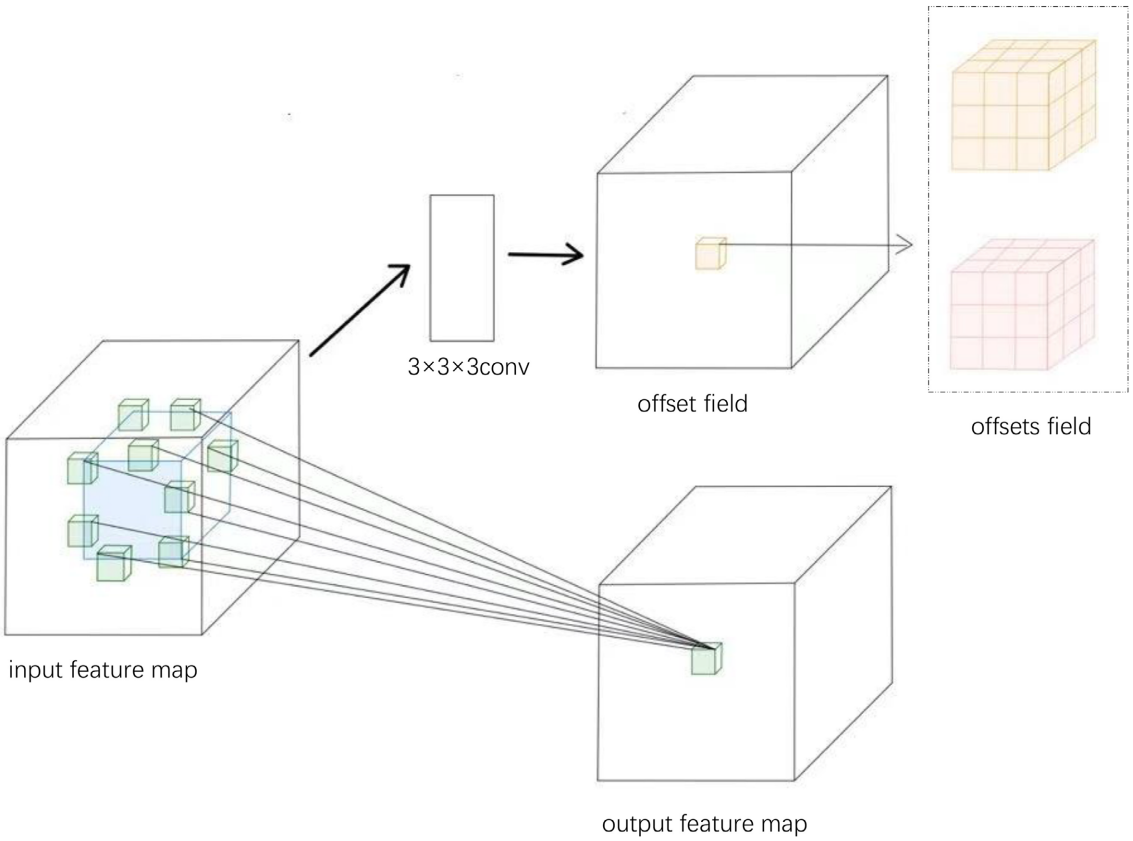

Figure 2. Schematic diagram of deformable 3D convolution
图 2. D3D 操作示意图

传统三维卷积只能在规则网格上提取特征，而 D3D 能够聚焦于雾浓度变化剧烈区域与运动边缘，实现更精准的时空特征建模，从而提升去雾的完整性与细节保持能力。D3D 被公式化可表示为：

$$y(p_0)=\sum_{n=1}^{N} w(p_n)\cdot x(p_0+p_n+\Delta p_n) \tag{8}$$

其中，p_0 表示输出特征图的位置坐标，p_n 枚举了标准 3 × 3 × 3 卷积核的 27 个采样位置(即 N = 27)，而 Δp_n 则是通过网络学习得到的空间偏移量。

D3D 通过辅助的 3 × 3 × 3 卷积层预测 2D 空间偏移场，在保持时间维度连续性的同时，实现采样网格的空间形变。这一动态学习偏移机制可表示为：

$$\Delta p = f_{\text{offset}}(x;\theta_{\text{offset}}) \tag{9}$$

其中，f_{offset} 是偏移生成网络，θ_{offset} 是其参数。偏移量通常为分数坐标，因此需采用双线性插值实现可微的特征采样，确保梯度可传播。此外，D3D 仅对空间维度进行形变，既保留时间维度的自然连续性先验，又将计算复杂度控制在一定范围内。通过以上设计实现动态时空建模。在视频去雾任务中，D3D 能同时建模雾气的动态扩散(时间维度)和非均匀分布(空间维度)，通过融合时空维度信息，在运动感知和特征表达能力方面均取得显著提升。

为了直观验证 D3D 是否能够有效捕捉非均匀雾气分布，本文提取了 D3D 模块的输出特征图，并计算其通道方差，得到每帧的特征响应热力图。图 3 展示了其中一帧的可视化结果，左侧为原始输入帧，右侧为 D3D 模块特征响应热力图。从图中可以看出，D3D 的特征响应与雾气分布高度一致：浓雾区域呈现高响应，清晰区域呈现低响应。这表明 D3D 通过学习到的空间偏移场，能够自适应地关注非均匀雾气的分布模式。

Figure 3. D3D module feature response heatmap
图 3. D3D 模块特征响应热力图

2.4. 损失函数

本文采用多损失函数联合优化策略以提升视频去雾效果的视觉质量和色彩准确性。损失函数由三部分组成：像素级的 L1 损失(L1 Loss) [22]、衡量图像结构相似性的 SSIM 损失(Structural Similarity Loss, SSIM Loss) [23]以及基于 VGG 特征空间的感知损失(Perceptual Loss) [24]。

1) L1 损失。通过计算预测帧与真实帧的像素级绝对误差，约束模型实现精确的颜色还原，避免因像素偏差导致的整体色彩失真，其表达式为：

$$L_{\text{L1}}=\frac{1}{N}\sum_{i=1}^{N}\left|J_i-\hat{J}_i\right| \tag{10}$$

其中，J_i 表示真实无雾帧的第 i 个像素值，$\hat{J}_i$ 表示模型预测帧的第 i 个像素值，N 为单帧图像的总像素

数。

2) 为了进一步约束图像结构信息的保持能力，本文引入结构相似性损失。SSIM 能够从亮度、对比度以及结构信息三个方面衡量两幅图像之间的相似程度，相较于简单的像素误差指标，其更加符合人类视觉感知特性。SSIM 的计算公式如下：

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \tag{11}$$

其中，μ_x 与 μ_y 分别表示图像 x 与 y 的平均亮度值，σ_x^2 与 σ_y^2 表示方差，σ_{xy} 表示协方差，C_1 与 C_2 为防止分母为零的常数。基于此，结构相似性损失可定义为：

$$L_{\mathrm{SSIM}}=1-\mathrm{SSIM}(J,\hat{J}) \tag{12}$$

其中，$\mathrm{SSIM}(J,\hat{J})$ 表示真实无雾帧 J 与预测帧 $\hat{J}$ 的结构相似性指数，取值范围为[0, 1]。该损失值越小，说明预测帧与真实无雾帧的结构一致性越好，去雾结果的细节保留越完整、视觉效果越自然。

3) 感知损失。基于预训练 VGG16 网络的高层特征空间计算，通过约束预测帧与真实帧的语义特征相似度，避免像素级损失可能导致的视觉逼真但语义失真问题，使去雾结果更符合人眼对自然场景的感知预期。本文选用 VGG16 网络的 conv3_3 层输出特征(该层对图像纹理、轮廓等中层语义信息的表征能力最优)，其表达式为：

$$L_{\mathrm{Perceptual}}=\frac{1}{C\times H\times W}\left\|\phi(J)-\phi(\hat{J})\right\|_2 \tag{13}$$

其中，$\phi(\cdot)$ 表示 VGG16 网络 conv3_3 层的特征提取函数，$\phi(J)\in R^{C\times H\times W}$ 为真实无雾帧的特征图(C、H、W 分别为特征图的通道数、高度和宽度)，$\|\cdot\|$ 表示 L2 范数。

最后，最终的损失函数可以定义为：

$$L_{\mathrm{total}}=\lambda_{\mathrm{L1}}\cdot L_{\mathrm{L1}}+\lambda_{\mathrm{SSIM}}\cdot L_{\mathrm{SSIM}}+\lambda_{\mathrm{Perceptual}}\cdot L_{\mathrm{Perceptual}} \tag{14}$$

其中，权重系数 λ_{L1}、λ_{SSIM}、$\lambda_{\mathrm{Perceptual}}$ 分别为 L1、SSIM 和感知损失权重。

3. 实验

3.1. 实验设置

本文算法的训练与性能验证均基于两大经典视频去雾基准数据集完成，分别为真实室内视频去雾数据集 REVIDE 与大规模户外合成视频去雾数据集 HazeWorld，两类数据集覆盖不同场景、不同雾浓度的雾天视频数据，能够全面验证算法在合成与真实雾景下的去雾性能与场景泛化能力。

本文将本章算法与经典的图像去雾算法和近几年的视频去雾算法进行定量和定性分析，选取的具体模型包括经典图像去雾算法 DCP，以及近年来的视频去雾算法 PM-Net [25]、MAP-Net 和 DCL。

Table 2. Hyperparameter settings in the ESMM module
表 2. ESMM 模块中超参数设置

层数	不同阶段特征维度	多头注意力头数
12	[24, 48, 72, 96]	12

实验过程中，输入视频帧被随机裁剪为 224 × 224 大小，并通过随机旋转和翻转操作进行数据增强。实验采用 AdamW 优化器，初始学习率设置为 1×10^{-4}，总迭代次数为 20 K，批量大小设为 4。该模型在

NVIDIA GeForce GTX 1080 Ti 显卡上基于 Pytorch 框架实现，通过 SSIM 和 PSNR 指标对模型性能进行客观评价与分析。实验所用 ESMM 模块的超参数如表 2 所示，UnDehazeNet 模型的整体超参数设置如表 3 所示。

表 2 中，ESMM 模块的帧内帧间特征提取模块均采用 12 层结构，搭配[24, 48, 72, 96]的多尺度特征维度，适配模型多尺度特征提取需求，12 头多头注意力则用于强化模块的全局－局部时空建模能力。

Table 3. Hyperparameter settings in the UnDehazeNet model
表 3. UnDehazeNet 模型中超参数设置

学习率	批量大小	迭代次数	L1 损失权重	SSIM 损失权重	感知损失权重
1×10^{-4}	4	20 K	1	0.2	0.3

表 3 中，模型损失函数采用 L1 损失、SSIM 损失与感知损失加权融合的方式，其中 L1 损失权重设为 1、SSIM 损失权重设为 0.2、感知损失权重设为 0.3，可有效平衡去雾效果、图像细节保留与视觉一致性，其余超参数(学习率、批量大小、迭代次数)与前文实验设置保持一致，确保实验的连贯性与可复现性。

3.2. 对比实验结果分析

1) 定量比较。UnDehazeNet 模型和比较方法在 REVIDE 和 HazeWorld 上的定量结果分别见表 4、表 5。

Table 4. Quantitative comparison on the REVIDE dataset
表 4. REVIDE 数据集的定量比较

模型	PSNR ↑	SSIM ↑
DCP	11.03	0.7285
PM-Net	22.01	0.8759
MAP-Net	24.16	0.9043
DCL	24.52	0.9067
UNDehazeNet	**26.67**	**0.9153**

Table 5. Quantitative comparison on the HazeWorld dataset
表 5. HazeWorld 数据集的定量比较

模型	PSNR ↑	SSIM ↑
DCP	20.49	0.8126
PM-Net	24.70	0.9259
MAP-Net	27.12	0.9349
DCL	27.56	0.9415
UNDehazeNet	**29.35**	**0.9676**

从表 4 的实验结果可以看出，在 REVIDE 数据集上，传统先验方法 DCP 的 PSNR 仅为 11.03 dB，SSIM 为 0.7285，明显低于基于深度学习的方法，这表明基于简单先验的传统方法难以应对真实雾天视频的复杂性。基于深度学习的 PM-Net 取得了 22.01 dB 的 PSNR 和 0.8759 的 SSIM，相比传统方法有显著提升。MAP-Net 和 DCL 进一步提升了性能，分别达到 24.16 dB/0.9043 和 24.52 dB/0.9067。本文提出的

UnDehazeNet 实现了较好的效果，PSNR 达到 26.67 dB，SSIM 达到 0.9153。与次优的 DCL 相比，UnDehazeNet 的 PSNR 提高了 2.15 dB，SSIM 从 0.9067 提高到 0.9153。这一显著提升验证了本文提出的 ESMM 模块和 D3D 在真实雾天视频去雾任务中的有效性。

从表 5 的实验结果可以看出，在 HazeWorld 数据集上，各方法的性能普遍高于 REVIDE 数据集，这主要是因为 HazeWorld 为合成数据集，其雾的分布更符合大气散射模型的理想假设，且场景内容与训练数据分布更为一致。DCP 方法在合成数据集上的表现较真实数据集有明显提升，PSNR 达到 20.49 dB，SSIM 为 0.8126，但仍明显低于深度学习方法。PM-Net、MAP-Net 和 DCL 分别取得了 24.70 dB/0.9259、27.12 dB/0.9349 和 27.56 dB/0.9415 的效果。本文方法 UnDehazeNet 在所有对比方法中表现较好，PSNR 达到 29.35 dB，SSIM 达到 0.9676。与次优的 DCL 相比，UnDehazeNet 的 PSNR 提高了 1.79 dB，SSIM 从 0.9415 提高到 0.9676。在两个不同性质的数据集上取得一致的性能提升，充分证明了本文方法具有良好的鲁棒性，能够适应不同场景、不同雾浓度条件下的视频去雾任务。

2) 定性比较。为了直观评估不同方法的去雾效果，在 REVIDE 数据集和 HazeWorld 数据集的视频帧上，对本文提出的方法与多种经典去雾算法进行了定性对比分析，分别如图 4、图 5 所示。

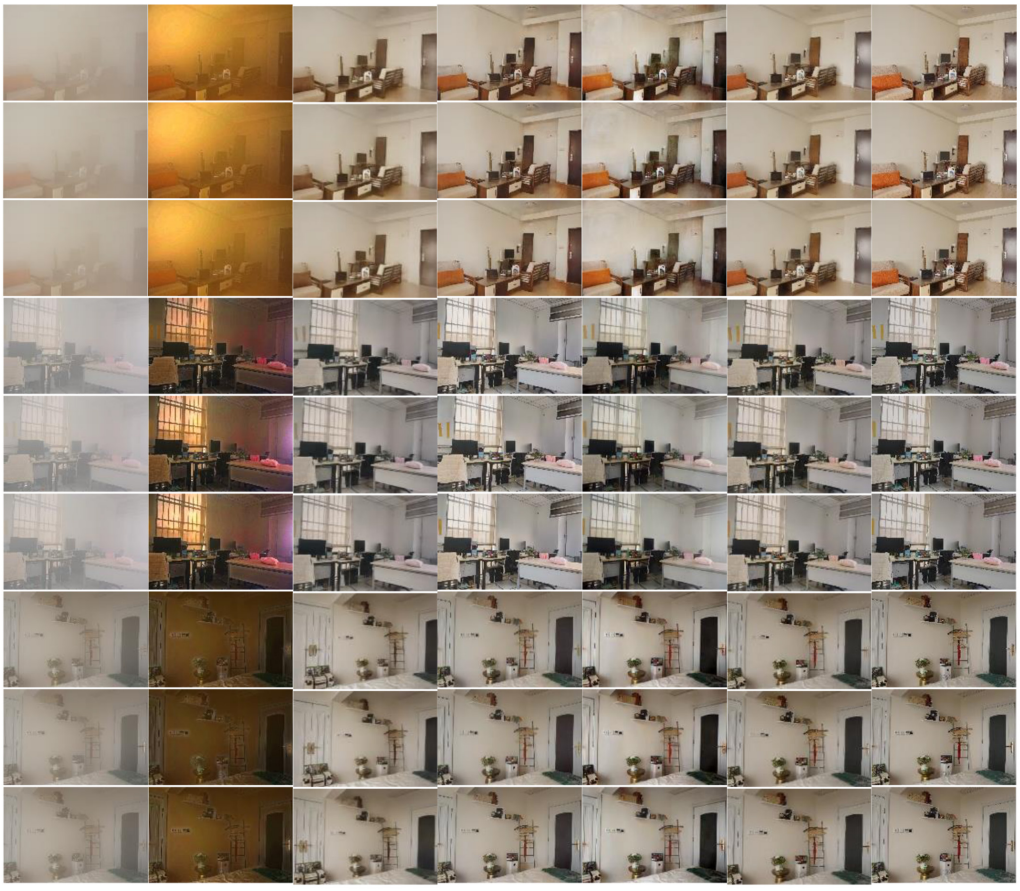

Figure 4. Qualitative comparison on the REVIDE dataset
图 4. REVIDE 数据集上的定性比较

从图 4 可以看出，在室内场景 REVIDE 数据集中，原始有雾输入帧受到明显雾气遮挡影响，整体对比度较低，物体边缘模糊，部分纹理细节难以辨识。传统先验方法 DCP 虽然能够在一定程度上去除部分

雾气，但仍存在较为明显的雾残留现象，例如在墙壁边缘及家具轮廓处仍可观察到较重的雾气痕迹，并且整体表现出去雾不彻底、亮度偏暗、色彩失真等问题。同时，该方法恢复后的图像整体偏暗，对比度和亮度均有所不足，导致视觉效果仍然较为模糊。在视频去雾方法中，MAP-Net 恢复结果具有较高的颜色饱和度，但在雾气较为集中的区域(如图像角落部分)仍然存在去雾不彻底的问题；而 DCL 虽然能够较好地恢复图像结构信息，但其输出结果存在轻微的颜色偏冷现象，同时整体对比度与饱和度偏高，导致部分区域出现视觉上的不自然现象。

从图中所示的连续三帧视频图像可以看出，本文提出的 UnDehazeNet 在去雾效果和视觉质量方面均表现出更优的性能。该方法不仅能够有效去除图像各区域中的雾气成分，还能够较好地恢复场景中的细节信息。例如，在门框及其周围区域的光影纹理恢复方面，所提出方法的结果与对应的无雾参考帧最为接近。同时，恢复后的连续帧之间过渡自然，未出现明显的闪烁现象或结构伪影，说明该方法在提升去雾质量的同时，也能够较好地保持视频序列的时序一致性。

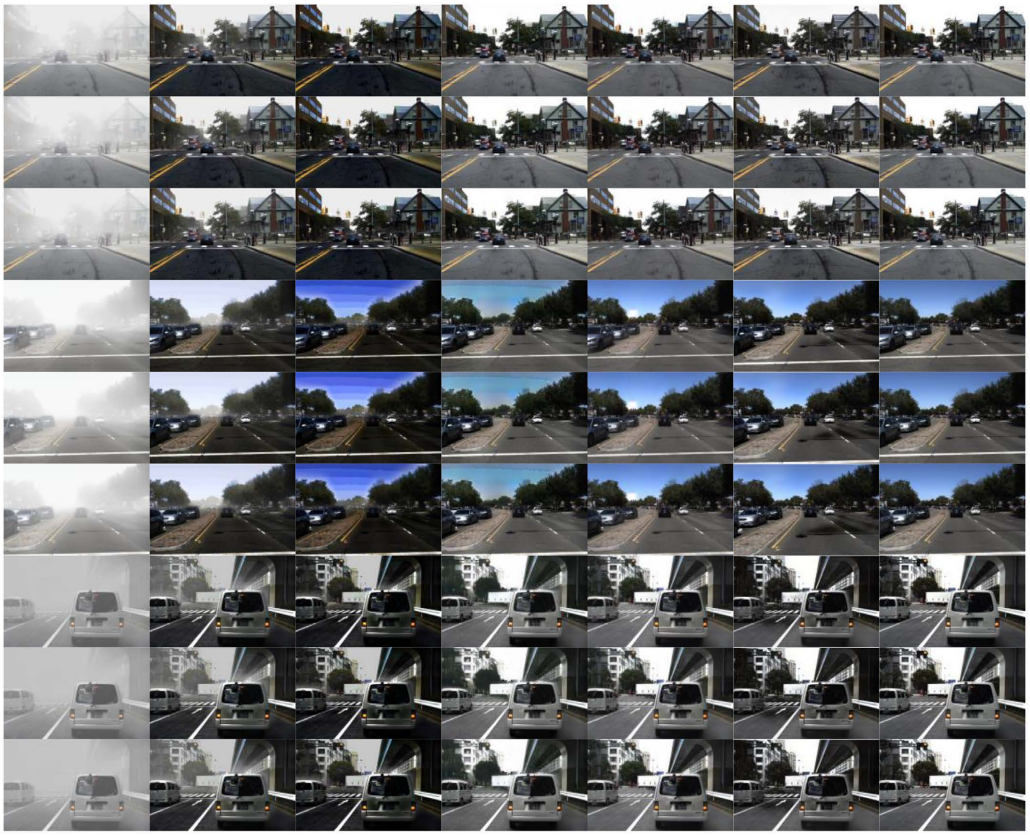

Figure 5. Qualitative comparison on the HazeWorld dataset
图 5. HazeWorld 数据集上的定性比较

从图 5 可以看出，在 HazeWorld 数据集的室外场景中，有雾输入帧受浓雾与动态环境影响，远处物体完全模糊，近景物体边缘虚化。对比方法中，DCP 算法在近处可以有效去除雾气，但是对于远处雾气去除效果较差；PM-Net 虽能提升图像清晰度，但存在明显的颜色失真(如天空颜色饱和度过高)；MAP-Net 的去雾效果较好，但蓝色天空区域颜色以及纹理恢复效果较差；DCL 在该数据集中表现较好，但是天空处出现白色光斑。本文方法在该场景下表现出更强的适应能力：一方面，能有效穿透浓雾，恢复远处物

体的细节(如远处建筑的窗户、道路标识)；另一方面，精准还原了户外场景的自然色彩(天空的蓝色)，无明显的亮度或颜色跳变。

3.3. 消融实验结果分析

为系统验证 UnDehazeNet 中核心模块的有效性，本文在 REVIDE 数据集上开展了消融实验，重点分析 ESMM 模块与 D3D 模块对网络性能的增益。以仅包含编码器－解码器结构的类 U-Net 网络作为基准模型(Basic)，逐步添加 ESMM 与 D3D 模块，构建多组消融模型。所有模型均采用相同的训练配置(迭代轮次、学习率、优化器、损失函数等)，并以 PSNR 和 SSIM 作为定量评价指标，实验结果如表 6 所示。从定量结果可见，相较于基准模型 Basic，单独引入 ESMM 模块使 PSNR 提升 3.78 dB、SSIM 提升 0.0392，验证了 ESMM 对视频去雾性能的正向增益；单独引入 D3D 模块使 PSNR 提升 4.70 dB、SSIM 提升 0.1115，体现了三维可变形卷积对空间去雾精度的显著提升作用。当两个模块协同集成时，完整 UnDehazeNet 模型取得最优性能，PSNR 较基准模型提升 6.11 dB，SSIM 提升 0.1532，充分证明了二者的互补性与协同增益。

Table 6. Ablation experiments on the REVIDE dataset
表 6. REVIDE 数据集的消融实验

Basic	ESMM	D3D	PSNR ↑	SSIM ↑
√			20.56	0.7621
√	√		24.34	0.8013
√		√	25.26	0.8736
√	√	√	26.67	0.9153

为进一步直观展示各模块对去雾效果与时序连贯性的影响，图 6 给出了消融模型在典型视频序列上的定性对比结果(图中展示连续三帧中的中间帧)。其中，(a) 为输入有雾帧，(b) 为 Basic 模型输出，(c) 为 Basic + ESMM 输出，(d) 为 Basic + D3D 输出，(e) 为完整模型输出，(f) 为真实无雾参考帧(GT)。

Figure 6. Qualitative comparison of ablation experiments on the REVIDE dataset
图 6. REVIDE 数据集上消融实验的定性对比

从视觉效果可见，Basic 模型输出的图像在物体边缘与细节区域存在明显的模糊和伪影。引入 ESMM 模块后，边缘细节的一致性提升，有效抑制了闪烁伪影；引入 D3D 后，模型对浓雾区域的去雾精度明显

提高，去雾效果更彻底，且避免了过度增强导致的色彩失真。完整模型融合了两者的优势，在保证时序连贯性的同时实现了较好的细节还原与色彩保真，其去雾结果与 GT 的视觉一致性较好。从效果层面分析，ESMM 模块通过局部－全局协同的时空建模与跨尺度特征融合，有效抑制了视频帧间的边缘闪烁问题，显著提升了去雾结果的时序连贯性，同时多尺度特征的整合也有助于保留图像的边缘与纹理细节；D3D 模块则凭借对雾气密度分布的三维空间感知能力，提升了网络对不同浓度雾气区域的自适应去雾精度，减少了局部去雾不彻底或过度去雾的现象。二者协同作用，不仅强化了网络的特征表达能力，还在时序一致性与去雾精度之间实现了更好的平衡，充分验证了所提模块设计的合理性与必要性。

4. 结论

本文提出了一种基于解码器－编码器的类 U-Net 网络的新型视频去雾网络模型(UnDehazeNet)，通过引入局部－全局协同的高效时空建模模块和可变形三维卷积，显著提升了视频去雾的性能和时序一致性。此外，在室内数据集 REVIDE 和室外数据集 HazeWorld 上进行了大量的实验，均证明了 UnDehazeNet 及其组件的有效性。与其他方法相比，UnDehazeNet 在 PSNR 和 SSIM 以及颜色还原度上均有所提升。在接下来的研究中，针对极端雾天场景下偶发的色彩失真问题，深入探究雾浓度与光照变化对色彩还原的影响机制，通过引入自适应色彩校准模块与雾特性感知机制，进一步提升去雾结果的色彩自然度与准确性。

基金项目

本文受国家自然科学基金(批准号：11472144)、国家自然科学基金(批准号：12472040)资助。

参考文献

[1] 赵世吉, 张金钊, 林立飞, 等. 基于 FFA-Net 与 YOLOv5 的雾天行车障碍检测技术研究[J]. 中阿科技论坛(中英文), 2022(9): 141-144.

[2] 贾童瑶, 卓力, 李嘉锋, 等. 基于深度学习的单幅图像去雾研究进展[J]. 电子学报, 2023, 51(1): 231-245.

[3] Narasimhan, S.G. and Nayar, S.K. (2002) Vision and the Atmosphere. *International Journal of Computer Vision*, **48**, 233-254. https://doi.org/10.1023/a:1016328200723

[4] Zhang, H. and Patel, V.M. (2018) Densely Connected Pyramid Dehazing Network. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 3194-3203. https://doi.org/10.1109/cvpr.2018.00337

[5] 刘姝廷, 孙诚志, 娄浩云, 等. 基于直方图均衡化和 Retinex 的图像去雾研究[J]. 信息与电脑(理论版), 2023, 35(15): 172-175.

[6] He, K.M., Sun, J. and Tang, X.O. (2011) Single Image Haze Removal Using Dark Channel Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**, 2341-2353. https://doi.org/10.1109/tpami.2010.168

[7] Cai, B., Xu, X., Jia, K., Qing, C. and Tao, D. (2016) DehazeNet: An End-to-End System for Single Image Haze Removal. *IEEE Transactions on Image Processing*, **25**, 5187-5198. https://doi.org/10.1109/tip.2016.2598681

[8] Li, B., Peng, X., Wang, Z., Xu, J. and Feng, D. (2017) AOD-Net: All-in-One Dehazing Network. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 4780-4788. https://doi.org/10.1109/iccv.2017.511

[9] 边宇霄. 基于深度学习的端到端图像去雾算法研究[D]: [硕士学位论文]. 长春: 吉林大学, 2024.

[10] Borkar, K. and Mukherjee, S. (2018) Video Dehazing Using LMNN with Respect to Augmented MRF. *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, Hyderabad, 18-22 December 2018, 1-9. https://doi.org/10.1145/3293353.3293395

[11] Zhang, J., Li, L., Zhang, Y., Yang, G., Cao, X. and Sun, J. (2011) Video Dehazing with Spatial and Temporal Coherence. *The Visual Computer*, **27**, 749-757. https://doi.org/10.1007/s00371-011-0569-8

[12] 宁贝, 杨明. 基于多尺度引导滤波的实时视频去雾算法[J]. 中北大学学报(自然科学版), 2024, 45(4): 439-447.

[13] Kim, J., Jang, W., Park, Y., Lee, D., Sim, J. and Kim, C. (2012) Temporally X Real-Time Video Dehazing. 2012 19*th IEEE International Conference on Image Processing*, Orlando, 30 September-3 October 2012, 969-972. https://doi.org/10.1109/icip.2012.6467023

[14] Ren, W., Zhang, J., Xu, X., Ma, L., Cao, X., Meng, G., *et al*. (2019) Deep Video Dehazing with Semantic Segmentation. *IEEE Transactions on Image Processing*, **28**, 1895-1908. https://doi.org/10.1109/tip.2018.2876178

[15] Zhang, X., Dong, H., Pan, J., Zhu, C., Tai, Y., Wang, C., *et al*. (2021) Learning to Restore Hazy Video: A New Real-World Dataset and a New Method. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 9235-9244. https://doi.org/10.1109/cvpr46437.2021.00912

[16] Yang, Y., Guo, C. and Guo, X. (2024) Depth-Aware Unpaired Video Dehazing. *IEEE Transactions on Image Processing*, **33**, 2388-2403. https://doi.org/10.1109/tip.2024.3378472

[17] 林志鹏, 秦佳, 秦品乐, 等. 基于物理先验引导的记忆增强视频去雾算法[J]. 中北大学学报(自然科学版), 2025, 46(6): 726-733.

[18] Xu, J., Hu, X., Zhu, L., Dou, Q., Dai, J., Qiao, Y., *et al*. (2023) Video Dehazing via a Multi-Range Temporal Alignment Network with Physical Prior. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 18053-18062. https://doi.org/10.1109/cvpr52729.2023.01731

[19] Li, K., Wang, Y., Gao, P., *et al*. (2022) UniFormer: Unified Transformer for Efficient Spatiotemporal Representation Learning. arXiv: 2201.04676.

[20] Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Wang, L., *et al*. (2023) UniFormerV2: Unlocking the Potential of Image Vits for Video Understanding. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 1632-1643. https://doi.org/10.1109/iccv51070.2023.00157

[21] 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

[22] Zhao, H., Gallo, O., Frosio, I. and Kautz, J. (2017) Loss Functions for Image Restoration with Neural Networks. *IEEE Transactions on Computational Imaging*, **3**, 47-57. https://doi.org/10.1109/tci.2016.2644865

[23] Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, **13**, 600-612. https://doi.org/10.1109/tip.2003.819861

[24] Johnson, J., Alahi, A. and Li, F.F. (2016) Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *Computer Vision—ECCV* 2, Springer, 694-711. https://doi.org/10.1007/978-3-319-46475-6_43

[25] Liu, Y., Wan, L., Fu, H., Qin, J. and Zhu, L. (2022) Phase-Based Memory Network for Video Dehazing. *Proceedings of the* 30*th ACM International Conference on Multimedia*, Lisboa, 10-14 October 2022, 5247-5435. https://doi.org/10.1145/3503161.3547998