ELVIS: Enhance Low-Light For Video Instance Segmentation in the Dark

Joanne Lin [1] Ruirui Lin [1] Yini Li [1] David Bull [1] Nantheera Anantrasirichai [1]

[1]Visual Information Laboratory, University of Bristol

CVPR JUNE 3-7, 2026 DENVER COLORADO

Abstract

Video instance segmentation (VIS) in low-light videos is highly challenging due to degradations such as noise and blur. Existing VIS methods struggle in such conditions due to the lack of large-scale annotated data and limited synthetic pipelines with realistic temporal degradations. We propose **ELVIS** (**E**nhance **L**ow-**L**ight for **V**ideo **I**nstance **S**egmentation), the **first** framework that enables domain-adaption of SOTA VIS models to low-light scenarios.

Contributions:

- A new synthetic low-light video pipeline which incorporates illumination adjustment, noise and blur.
- Adding an enhancement decoder head into existing architectures to disentangle degradations from content features.
- An unsupervised network, VDP-Net, which estimates degradation profiles for a calibration-free synthesis of realistic low-light videos.

Synthetic Low-Light Video Pipeline

Illumination

Illumination adjustment using images converted to XYZ color space.

$$X' = 2^{\epsilon} X, \tag{1}$$

Blur

Synthesize camera-shake motion blur using a multivariate Gaussian kernel.

$$H[i,j] = \frac{\exp\left(-\frac{1}{2}\mathbf{r}_{ij}^{\top}\Sigma^{-1}\mathbf{r}_{ij}\right)}{\sum_{i,j}\exp\left(-\frac{1}{2}\mathbf{r}_{ij}^{\top}\Sigma^{-1}\mathbf{r}_{ij}\right)}, \tag{2}$$

Noise

Physics-based statistical distributions to synthesize common noise types.

$$N_r \sim \mathcal{N}(0,\sigma_r^2), \quad N_r \in \mathbb{R}^{T\times C\times H\times W}, \tag{3}$$

$$(X+N_s) \sim \mathcal{P}(\frac{X}{K})K, \quad N_r \in \mathbb{R}^{T\times C\times H\times W}, \tag{4}$$

$$N_q \sim \mathcal{U}(0,\lambda_q), \quad N_q \in \mathbb{R}^{T\times C\times H\times W}, \tag{5}$$

$$N_b \sim \begin{cases} \mathcal{N}(0,\sigma_b^2), & N_b \in \mathbb{R}^{T\times C\times 1\times W}, \quad \text{if } \theta_b = 0 \\ \mathcal{N}(0,\sigma_b^2), & N_b \in \mathbb{R}^{T\times C\times H\times 1}, \quad \text{if } \theta_b = 1 \end{cases} \tag{6}$$

Overall

$$X^{low} = Deg(X^{high},\phi) = H * (2^{\epsilon}X^{high}) + N, \tag{7}$$

$$\phi = \{\epsilon, \sigma_r, K, \lambda_q, \sigma_b, \theta_b, \sigma_{H_x}, \sigma_{H_y}, \theta_H\}. \tag{8}$$

Overview of ELVIS Framework

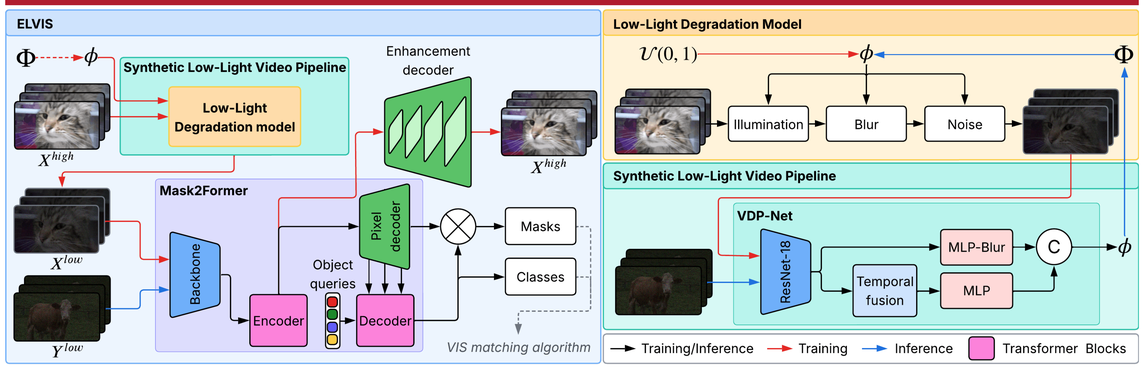

Figure 1. Overview of the proposed ELVIS framework, which consists of three main components: (i) the unsupervised synthetic low-light model, (ii) the architecture of VDP-Net for estimating degradation profiles, and (iii) the addition of the enhancement decoder to VIS methods for a regularization loss.

Quantitative Results

Table 1. Evaluation on the synthetic YouTube-VIS 2019 [10] validation set using several SOTA online VIS methods and backbone networks, each trained on synthetic low-light videos generated with our synthetic pipeline.

Method	Backbone	ELVIS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_1	AR_{10}
MinVIS [1]	ResNet-50	✗	36.4	**57.3**	36.4	13.3	30.6	**49.4**	36.5	44.4
GenVIS [2]	ResNet-50	✗	39.1	58.4	42.7	16.2	34.8	**55.2**	40.3	48.4
DVIS++ [3]	ResNet-50	✗	38.8	59.9	42.8	23.8	38.5	51.4	39.5	49.6
MinVIS [1]	ResNet-50	✓	**37.2**	57.0	**39.6**	**15.1**	**34.5**	30.9	**37.8**	**45.7**
GenVIS [2]	ResNet-50	✓	**41.0**	**59.8**	**46.2**	**18.5**	**38.0**	54.2	**42.0**	**51.2**
DVIS++ [3]	ResNet-50	✓	**42.5**	**63.8**	**46.6**	**27.6**	**42.5**	**54.3**	**41.7**	**51.9**
MinVIS [1]	SWIN-L	✗	51.8	73.8	57.9	28.0	43.0	69.8	46.7	57.5
GenVIS [2]	SWIN-L	✗	53.7	74.8	58.7	**27.6**	44.9	71.0	49.1	59.6
DVIS++ [3]	ViT-L	✗	55.2	77.2	62.1	**29.5**	51.6	70.5	49.7	61.5
MinVIS [1]	SWIN-L	✓	**54.2**	**78.3**	**61.6**	**28.4**	**47.6**	**71.7**	**49.8**	**60.1**
GenVIS [2]	SWIN-L	✓	**55.3**	**79.3**	**61.1**	26.6	**49.6**	**72.3**	**49.5**	**61.2**
DVIS++ [3]	ViT-L	✓	**56.9**	**78.7**	**65.3**	25.5	**54.3**	**74.8**	**50.3**	**62.7**

Table 2. Evaluation on ELVIS-S (real low-light videos)

Method	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Unsupervised Synthetic Pipelines					
Lv et al. [4]	53.5	92.7	**43.5**	45.3	54.1
Cui et al. [5]	51.1	**96.8**	43.0	45.3	51.9
Lin et al. [6]	35.1	81.4	25.0	32.8	35.0
Ours (random)	39.9	70.6	26.6	35.6	45.0
Ours	**54.5**	**96.8**	43.0	**46.6**	**55.3**
Two-stage baselines					
SDSD-Net [7]	46.7	96.0	26.6	41.9	47.8
StableLLVE [8]	57.3	**96.8**	**47.7**	47.2	57.8
DarkIR [9]	55.9	**96.8**	45.5	47.2	56.6
ELVIS	**58.0**	**96.8**	47.6	**48.1**	**58.8**

References:

[1] *Wang et al.* MinVIS: A Minimal Video Instance Segmentation Framework without Video-based Training, NeurIPS 2022. [2] *Heo et al.* A Generalized Framework for Video Instance Segmentation, CVPR 2023. [3] *Zhang et al.* DVIS++: Improved Decoupled Framework for Universal Video Segmentation, TPAMI 2025. [4] *Lv et al.* Attention guided low-light image enhancement with a large scale low-light simulation dataset, IJCV 2021. [5] *Cui et al.* Multitask AET With Orthogonal Tangent Regularity for Dark Object Detection, ICCV 2021. [6] *Lin et al.* Towards a General-Purpose Zero-Shot Synthetic Low-Light Image and Video Pipeline, ACMMM McGE Workshop 2025. [7] *Wang et al.* Seeing Dynamic Scene in the Dark: High-Quality Video Dataset with Mechatronic Alignment, ICCV 2021. [8] *Zhang et al.* Learning Temporal Consistency for Low Light Video Enhancement From Single Images, CVPR 2021. [9] *Feijoo et al.* DarkIR: Robust Low-Light Image Restoration, CVPR 2025. [10] *Yang et al.* Video Instance Segmentation, ICCV 2019. [11] *Wang et al.* Multi-Object Tracking in the Dark, CVPR 2024.

Qualitative Results

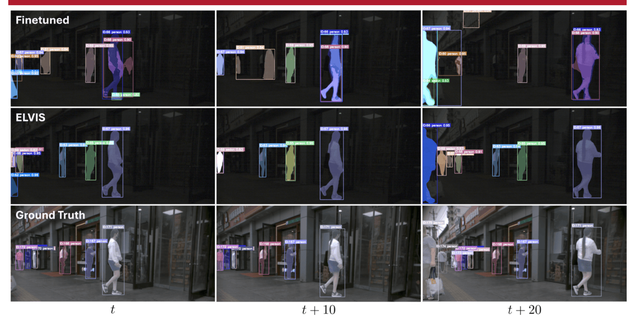

Figure 2. Visual comparison of VIS results on the LMOT-S dataset using GenVIS R50. The columns show every 10 frames from frame t in the example video to show the tracking performances.

(a) SDSD-Net [7] (b) StableLLVE [8] (c) DarkIR [9] (d) Ours

Figure 3. Visual comparison of VIS predictions using two-stage baselines versus ELVIS on ELVIS-S.

Links

(a) ArXiV (b) Project Page (c) LinkedIn

joanne.lin@bristol.ac.uk