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

Supplementary Material

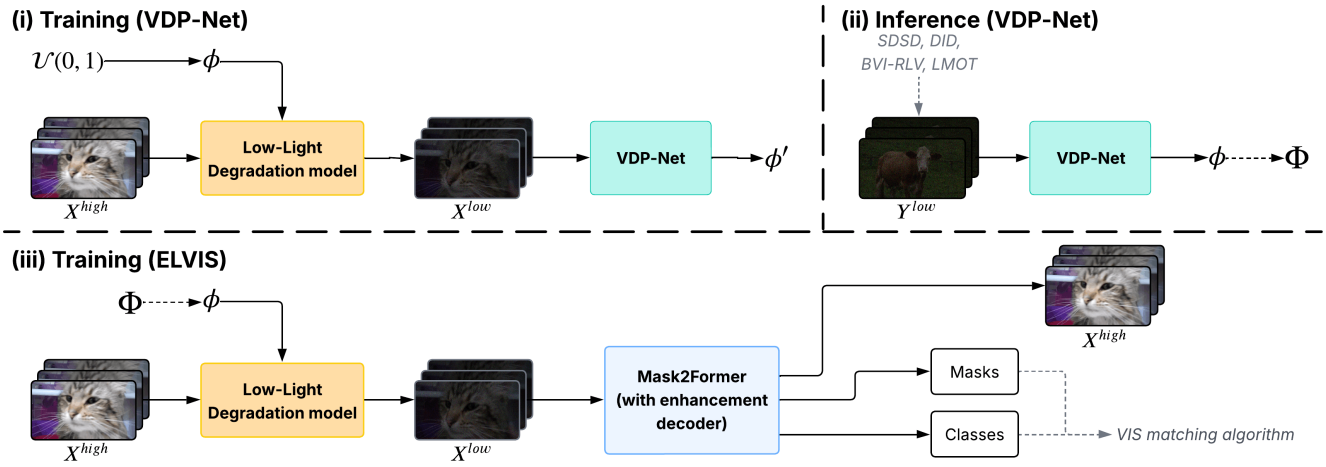


Figure S1. Step-by-step process for using the ELVIS framework: (i) train VDP-Net to predict degradation profile ϕ, (ii) build Φ from real low-light datasets [14, 27, 39, 40], and (iii) use degradation profiles from Φ to create synthetic data to train ELVIS.

S1. ELVIS Details

Fig. S1 shows a step-by-step visualization of each process required for the ELVIS framework for clarity. Refer to the main paper for the implementation details of each step.

S2. LMOT-S Benchmark

S2.1. Construction details

In order to construct a large low-light VIS benchmark for evaluation, we had to ensure that the Segment Anything Model (SAM) [20] was capable of producing high-quality segmentation masks for the dataset. LMOT [40] is well-suited to our requirements as the dataset provides both video object tracking annotations *and* paired low-light and normal-light frames. The paired normal-light frames allows us to collect segmentation masks without the performance impact arising from low-light degradations, while the video object tracking annotations allows us to easily prompt Segment Anything to produce segmentation masks for the object of interest.

We construct LMOT-S by downsampling the low-light sRGB frames from the original LMOT [40] validation set to 720p resolution and dividing the videos in clips of 51 frames each (22 even segments per video). This enables us to evaluate on 88 videos; allowing us to determine the robustness of our method across a large scale. We then leverage the provided bounding box annotations, to feed into SAM [20] along with the paired normal-light frames to acquire pseudo ground truth segmentation annotations. Finally we process the segmentation results with the original annotations to match with the YouTube-VIS 2019 [42] annotation format, allowing for easy evaluation. Due to the mismatch in categories, we rename all 'car' instances in LMOT-S to 'sedan', rename all 'motorcycle' instances to 'motorbike' and remove all 'bicycle' and 'bus' instances. The IDs of each category are re-mapped to align with the IDs from YouTube-VIS 2019.

Table S1. Statistics of count of instances from YouTube-VIS [42] versus LMOT-S validation sets.

Dataset	Min	Max	Mean
YouTube-VIS 2019 [42]	1	6	1.7
LMOT-S	9	75	31.5

Table S2. LMOT-S results on GenVIS [16] (ResNet-50) trained on our synthetic low-light YouTube-VIS 2019 videos. Original pretrained normal-light results are shown for comparison.

Method	Light	AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AR_{100}
GenVIS [16]	Low	6.6	14.5	5.2	5.4	9.8	11.7
GenVIS [16] + ELVIS	Low	6.7	15.5	4.4	5.3	12.1	14.0
GenVIS [16] (pre-trained)	Normal	11.7	21.4	10.5	8.8	16.8	21.0

S2.2. Extra results

As we had increased the number of detections to 100, the AR_{100} becomes a meaningful metric for understanding a method's performance at detecting and segmenting a large number of instances. We report these results in Tab. S3 which shows our method also has the best AR_{100} scores.

Table S3. Further evaluation on LMOT-S for two-stage baselines and synthetic pipelines. **Bold** indicates the best performance.

Two-Stage Baselines	
Method	AR_{100}
SDSD-Net [39]	5.4
StableLLVE [47]	8.1
DarkIR [12]	8.0
ELVIS	**14.0**

Synthetic Pipelines	
Method	AR_{100}
Lv *et al*. [31]	9.5
Cui *et al*. [9]	10.2
Lin *et al*. [26]	5.7
Ours (random)	7.4
Ours	**11.7**

Table S4. Evaluation of SDSD-Net [39] trained with physics-based low-light synthetic pipelines across SDSD [39] and DID [14] using PSNR (↑), SSIM (↑), and LPIPS (↓). Outputs are histogram-matched. **Bold** indicates the best performances.

Method	SDSD			DID		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Lv *et al*. [31]	21.421	0.464	0.563	16.606	0.846	0.476
Cui *et al*. [9]	19.199	0.645	0.380	22.042	0.824	0.220
Lin *et al*. [26]	**23.088**	0.645	0.393	17.354	0.624	0.450
Ours	20.703	**0.688**	**0.282**	**22.794**	**0.846**	**0.197**

S2.3. Limitations of LMOT-S benchmark

Although LMOT-S [40] offers a useful preliminary evaluation for real-world settings, it is a far more challenging benchmark than YouTube-VIS 2019 [42]. One main reason for the notably lower performances when evaluating on the LMOT-S dataset can be attributed to the high instance counts (see Tab. S1). LMOT-S contains at least 9 instances per video, whereas the YouTube-VIS 2019 validation set contains at most 6; existing VIS methods are not designed to handle such large number of instances.

Another challenge is the cross-over issue, which leads to a high number of identity switches. Combined with the fact that these methods are designed to track fewer instances, this results in frequent identity switches and a failure to form new tracklets. An example can seen in Fig. S2, where in the finetuned GenVIS [16] model, it transferred the sedan identity (ID: 196) onto the motorcycle as it drove past the car, and created a new identity (ID: 193) for the car. ELVIS reduces the occurrences of such issues, but they still persist (see the truck in frame t). In Tab. S2, we also compare against GenVIS with off-the-shelf weights provided by the authors, on the paired normal-light frames of LMOT-S, to emphasize the limitations of the benchmark.

As such, we encourage considering results on both ELVIS-S and LMOT-S to obtain a more comprehensive understanding of real low-light performance.

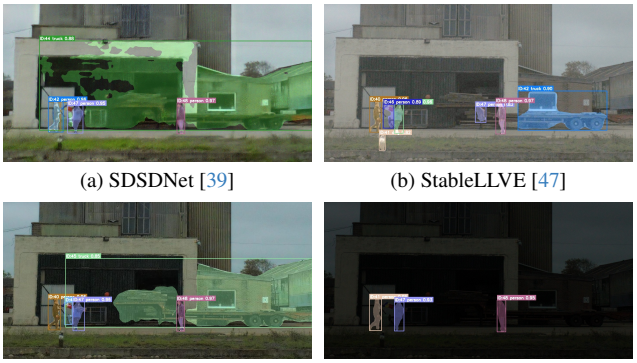
S3. Two-Stage Baseline Outputs

Fig. S3 shows a visual example of how the enhancement outputs of [12, 39, 47] resulted in lower VIS performances. We find that the enhanced outputs produced many false positive detections. Despite our method being unsuccessful in detecting the truck, all of the two-stage baselines incorrectly segmented the buildings as part of the truck. SDSD-Net [39] performed the worst consistently, likely attributing to the artifacts it frequently produces, and heavily oversegmented for the truck. StableLLVE [47] outputted a perceptually realistic output but when passed into GenVIS R50, resulted in incorrect detections such as a dog and extra people. DarkIR [12] produced the highest quality enhanced output but still incorrectly classified the building as part of the truck. We would like to highlight that while DarkIR generally produced the perceptually most accurate outputs, it underperforms StableLLVE on both ELVIS-S and LMOT-S (Tab. 3), suggesting that its enhancement process degrades or suppresses semantic features necessary for downstream video instance segmentation. Despite being the prevailing paradigm for low-light downstream tasks, two-stage baselines exhibit clear limitations, indicating the need for further research into LLVE methods that also consider the semantic features of the scene.

S4. Low-Light Video Enhancement Results

We find that the outputs from SDSD-Net [39] show severe artifacts (see Fig. S4 and Fig. S5); this is an observation also mentioned by Fu *et al*. [14]. This explains the modest performances across all synthetic pipelines in Tab. 6. For further analysis, we conduct additional LLVE experiments focusing on the denoising component of SDSD-Net by using different synthetic training data. Before computing the evaluation metrics, we apply histogram matching between the enhanced outputs and the ground truths. The results reflect the noise synthesis accuracy of each method. The results are shown in Tab. S4. While PSNR and SSIM show noticeable improvements, we find that LPIPS scores worsen with histogram matching, likely because it can amplify the artifacts, reducing perceptual similarity. We would like to mention that histogram matching is not the perfect solution for alleviating the artifacts in evaluation (see Fig. S4).

Figure S2. Visual comparison of segmentation and tracking success and failure cases on the LMOT-S dataset using GenVIS [16] method with a ResNet-50 backbone finetuned on our synthetic data (top row) versus implementing our ELVIS framework (middle row), with ground truth (bottom row) for reference. The columns represent frames in the example video, sampled every 10 frames from time t, to show the tracking performances.



(a) SDSDNet [39] (b) StableLLVE [47]

(c) DarkIR [12] (d) Ours

Figure S3. Visual comparison of VIS predictions using two-stage baselines versus ELVIS on ELVIS-S, with the enhanced outputs.

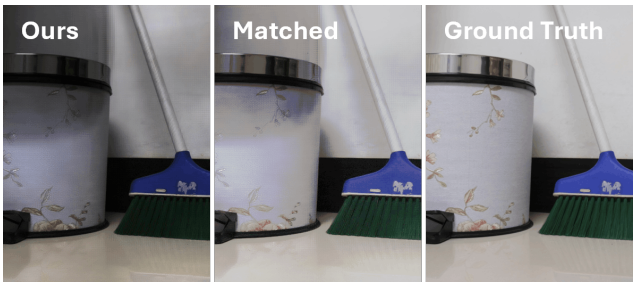


Figure S4. Example of artifacts in the enhanced output which is not removed with histogram matching. (Left) the enhanced output from SDSD-Net [39] trained on our synthetic pipeline. (Center) the histogram-matched output to the ground truth. (Right) is the ground truth normal-exposed frame.

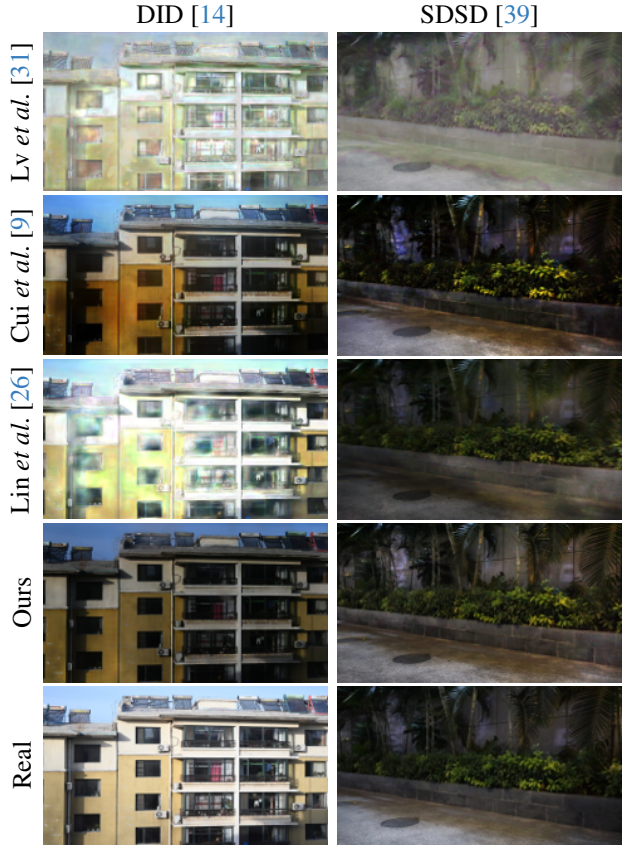


Figure S5. Qualitative comparison of the enhanced outputs of SDSD-Net [39] trained on different synthetic pipelines from the SDSD [39] and DID [14] datasets.