Supplementary Material

7.1. Lenviz Capture

As mentioned in our the paper, the LENVIZ dataset has been captured using 3 camera module. While our S5K4H7YX03-FGX9 camera is better suited for close range captures, S5KJN1SQ03 and S5KJNS camera features (showcased in Table 1) allow for richer capture of details at medium and long range.

7.1.1. Long-Exposure Time Calculation

Figure 1. Long-exposure Illuminance time experimental setup. During the captures the room lighting was controlled and all the background lights were turned off to maintain the darkroom (low-light conditions).

To empirically determine the values for the long-exposure parameters, a controlled experiment was conducted in a darkroom environment. This setup was designed to systematically measure and calibrate the relationship between scene illuminance and camera-captured image quality.

Experimental Setup All experiments were performed in a controlled darkroom to eliminate external light sources and ensure precise illuminance levels. A high precision lux meter was used to accurately measure the illuminance of the scene at the subject's position, providing ground-truth illuminance values for calibration. For test subject, a standardized test chart (e.g., DXO standard chart along with deadleaves chart) and a a mannequin with a color palette were used, as indicated in the Fig 1. These provided a consistent reference for evaluating sharpness, texture and color fidelity the image quality that from our study was highly effected in low-light conditions. Regarding camera configurations, the phones with camera modules listed in Table 1 was mounted on a tripod to ensure stability and was configured to capture long-exposure images. Simultaneously, a professional grade DSLR camera (Canon EOS R6 Mark II Mirrorles & Canon RF 15-35mm f/2.8 L IS USM Lens) was placed on a separate tripod with its field of view overlapping the phone's. The DSLR was set to auto exposure mode to provide a consistent, high-quality reference for image metrics.

Experimental Procedure First, for illuminance calibration, a lux-meter was used to measure and verify the illuminance levels for each scene. The lights in the darkroom were adjusted to create a series of controlled illuminance settings raging from 0.1 lux to 50 lux, with 1 lux increments as measure by external lux-meter. During image capture, images were captured simultaneously with both the phone and the DSLR at each illuminance level. This process was repeated for long-exposure with the phone's shutter speed ranging from 2 seconds to 30 seconds in 2 second increments resulting in 15 sets (each containing scenes between lux 0.1 to 50). The DSLR images was captured on the first set. So the total of 16 sets (15 from phone and 1 from DSLR) were then submitted to DXOMark (professional imaging lab) for analysis. The lab conducted a detailed evaluation of key camera tests, including Modulation Transfer Function (MTF), deadleaves chart analysis, and noise characteristics. The final parameter tuning for long-exposure was based on the reports of camera testing from DXOMark, along with expert visual inspection against DSLR reference images, were used to tune the gamma parameter of Eq 2. These parameters were adjusted to the closest possible match in terms of sharpness, texture, and color fidelity between the phone's camera long-exposure output and the professional DSLR's reference images based on the test results for each images from DXOMARK . The example of the DSLR reference image as well as before and after tuned GT output is shown in Fig 2.

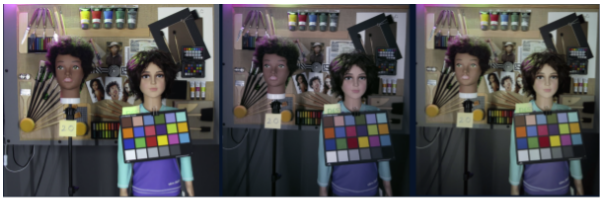

Figure 2. Examples of images captured during the empirical tuning of long-exposure time. Left image is DSLR captured as the control at lux meter reading=2 (post in note in scene indicates the lux reading of scene), middle image represents the long-exposure capture at exposure setting of 4 seconds, right image represents the final tuned long-exposure captured at time based on Eq 2.

7.1.2. Luminance VS Exposure time

The estimated exposure time for our shots is plotted against the estimated illuminance, as the estimated scene illuminance increases, the exposure time reduces accordingly to minimize the occurrence of over-exposed regions.

Table 1. Camera Module Specifications

Camera module Name	Resolution (MP)	Aperture	Pixel-size (um)	Camera-Size	Focus	FOV (Diag)
S5K4H7YX03-FGX9	8	f/2.0	1.12	1/3	Fixed (27cm∼39.3cm)	78°
S5KJN1SQ03	50	f/1.8	0.64	1/2.76”	Auto (10cm∼INF)	74.26°
S5KJNS	50	f/1.8	0.64	1/2.76”	Auto (10cm∼INF)	74.26°

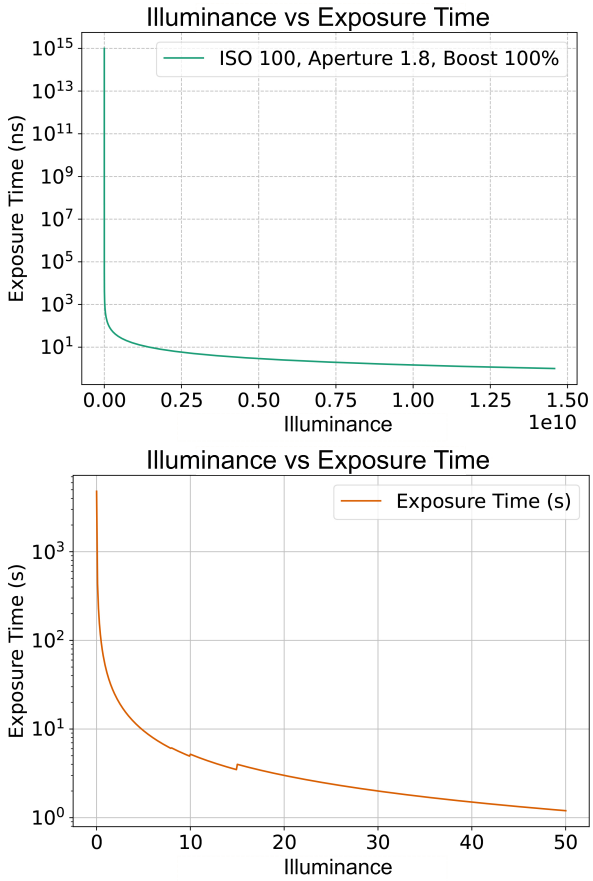

Figure 3. Illuminance vs exposure time (log scale) [**Top**] For standard values of ISO 100, aperture 1.8 and no additional boost. [**Bottom**] For long-exposure shot.

7.2. LENVIZ Additional Properties

7.2.1. Content class

Fig 4 provides an overlook at the distribution of the different object categories identified by the AWS object detection algorithm. In general, the detected objects within our images can be classified into 27 broad categories, ranging from ”Plants” and ”Animals” to ”Buildings” and ”Vehicles”. Each of these categories are further divided into 230 object-specific labels like ”Chair” or ”Couch” for the general ”Furniture” Category. Section 7.2 Table 2 provides a breakdown of each of the 230 object labels and their respective categories. We include the list of object labels identified

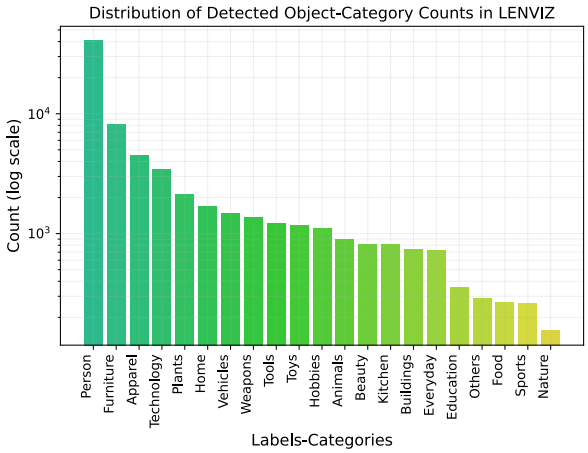

Figure 4. Demonstration of **27 unique object categories** comprising the **230 object labels** throughout the LENVIZ dataset.

for each given scene as well as their bounding boxes in our dataset release.

7.2.2. Illuminance distribution

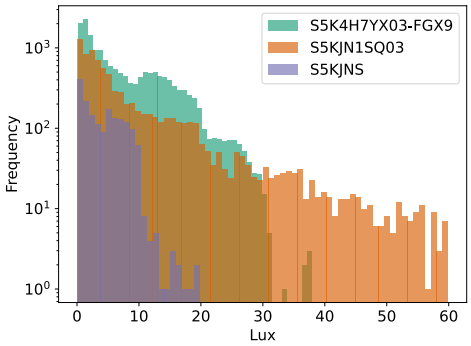

Figure 5. LENVIZ Illuminance Histogram

7.2.3. Feature Embedding Analysis

To understand the underlying structure and relationships within our LENVIZ dataset, we analyzed the feature distribution of our scenes in a latent space representation and compared it against other benchmakring datasets in the field, namely, LOL [30] (Single-Exposure), and SICE [5] (Multi-Exposure). We extracted a deep feature representa-

Table 2. Object Categories and Labels Breakdown

Category	**Objects**	**Object labels**
Person Description	12	Girl, Man, Boy, Person, Baby, Female, Adult, Male, Child, Woman, Teen, Bride
Plants and Flowers	3	Plant, Rose, Fungus
Technology and Computing	14	Laptop, Disk, Speaker, Microphone, VR Headset, Mobile Phone, Monitor, Earbuds, Remote Control, QR Code, Computer Keyboard, Camera, Headphones, Tablet Computer
Toys and Gaming	1	Doll
Furniture and Furnishings	12	Chandelier, Ceiling Fan, Chair, Rug, Bench, Dining Table, Door, Photo Frame, Lamp, Desk, Painting, Couch
Beauty and Personal Care	3	Toothbrush, Lipstick, Tattoo
Kitchen and Dining	5	Fork, Plate, Shaker, Spoon, Cup
Buildings and Architecture	5	Windmill, Tower, Clock Tower, Building, Gate
Tools and Machinery	11	Power Drill, Hammer, Baton, Switch, Brush, Blow Dryer, Screwdriver, Scissors, Tape, Shovel, Screw
Apparel and Accessories	32	Shirt, Shorts, Bridal Veil, Box, High Heel, Glasses, Hat, Wristwatch, Sunglasses, Sweater, Overcoat, Coat, Suit, Wallet, Tie, Glove, Diamond, Handbag, Belt, Bracelet, Shoe, Ring, Necklace, Razor, Sock, Helmet, Locket, Perfume, Backpack, Jacket, Jeans, Scarf
Home and Indoors	19	Swimming Pool, Hot Tub, Sink Faucet, Crib, Staircase, Fireplace, Package, Mailbox, Bathtub, Toilet, Bed, Sink, Shower Faucet, Mixer, Lawn Mower, Washer, Cooktop, Refrigerator, Microwave
Weapons and Military	13	Dagger, Mace Club, Spear, Axe, Bow, Mortar Shell, Gun, Crossbow, Dynamite, Sword, Grenade, Arrow, Knife
Vehicles and Automotive	11	Wheel, Boat, Bus, Airplane, Gas Pump, Pickup Truck, Train, Truck, E-scooter, Motorcycle, Car
Food and Beverage	16	Pear, Egg, Lobster, Milk, Burger, Beer, Orange, Ice Cream, Bread, Apple, Hot Dog, Banana, Pineapple, Can, Ketchup, Pizza
Hobbies and Interests	9	Toy, Bicycle, Violin, Clapperboard, Piano, Book, Teddy Bear, Guitar, Smoke Pipe
Nature and Outdoors	1	Moon
Symbols and Flags	2	Flag, Cross
Sports	18	Field Hockey Stick, Ice Hockey Puck, Rugby Ball, Ping Pong Paddle, Baseball (Ball), Soccer Ball, Baseball Glove, Volleyball (Ball), Cricket Bat, Scoreboard, Ice Hockey Stick, Skateboard, Baseball Bat, Tennis Ball, Basketball (Ball), American Football (Ball), Chess, Cricket Ball
Animals and Pets	25	Dinosaur, Honey Bee, Dog, Spider, Insect, Turtle, Fish, Giraffe, Kangaroo, Mouse, Lion, Chicken, Antelope, Elephant, Cat, Bird, Tiger, Bear, Sheep, Lizard, Horse, Shark, Snake, Pig, Zebra
Education	1	Blackboard
Text and Documents	4	Business Card, Credit Card, Passport, Driving License
Everyday Objects	3	Disposable Cup, Candle, Pen
Offices and Workspaces	1	White Board
Transport and Logistics	3	Traffic Light, Road Sign, Utility Pole
Events and Attractions	3	Hanukkah Menorah, Balloon, Snowman
Medical	2	First Aid, Pill
Public Safety	1	Fire Hydrant

tion for each of the scenes of all three datasets using the output of the last convolutional block in the VGG16 model given its well known suitability for feature extraction and its ability to recognize a vast range of visual patterns. To

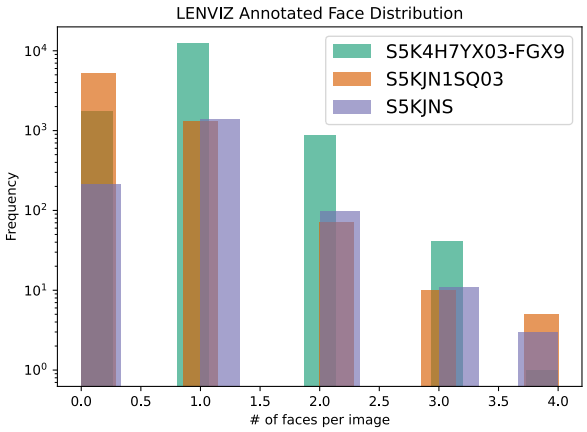

Figure 6. Number of Faces Per Scene Histogram

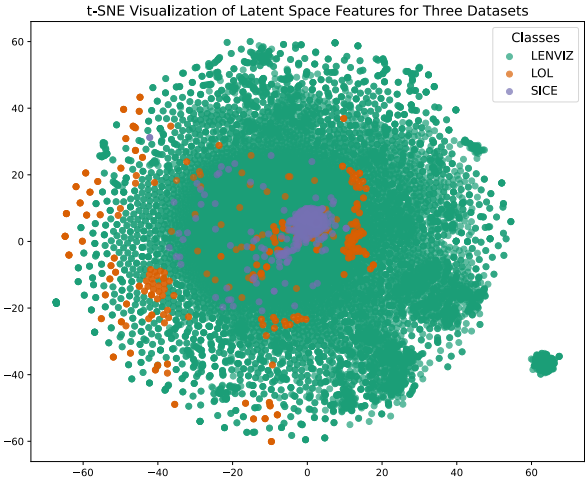

Figure 7. Feature Embedding comparison of LENVIZ Long-exposure frames vs LOL vs SICE

reduce the dimensionality of these features and visualize their distribution, we applied t-SNE to map these high-dimensional data points to a lower-dimensional space while preserving their local structure. As seen in Fig 7, the t-SNE visualization reveals that both LOL and SICE have some degree of overlap. The distribution of features from the LOL datasets seems to be relatively widespread in comparison to SICE. Despite this, our dataset is able to encompass a much wider feature distribution than these two benchmark datasets, as such allowing for models to learn from a wider set of features when using LENVIZ.

7.2.4. Test Dataset

To introduce illumination variations, we systematically adjusted illuminance levels from 0 to 20 for rear imagers (S5KJN1SQ03 & S5KJNS) and from 0 to 30 for front imagers (S5K4H7YX03-FGX9). Furthermore, we included both flash-on and flash-off conditions to replicate authentic low-light scenarios encountered in everyday photography.

For a comprehensive evaluation of background influence and model robustness, we analyzed images captured by renowned testing platforms like DXOmark[5] and our internal units. Drawing inspiration from these sources, we incorporated diverse background elements into our test dataset, including DXOmark charts, solid-colored wallpaper, polka dots (for ringing artifact analysis), text, and more. With all these variations we are additionally providing a reference (human-generated low-light enhancement ground truth) as well as no-reference types of test dataset. This is helpful to initially evaluate the model IQ using no-reference dataset and later use the reference as stage two evaluation quantitatively and qualitatively. Fig 8 illustrates representative test scenes for both indoor and outdoor environments.

Our test dataset distinguishes itself from existing benchmarks not only by its broader range of scene types but also by its emphasis on model robustness. These meticulously designed scenes are intended to challenge the trained enhancement models, assessing their ability to maintain consistent performance under varying low-light conditions and ensuring stable reproducibility.

7.3. Lenviz Additional Application

7.3.1. SOTA Models

Single Exposure Approaches: Zero-DCE++ [12], leverages a zero-reference deep curve estimation technique to enhance images without the need for paired ground-truth data, focusing on real-time illumination adjustment to improve brightness and contrast. LLFormer [29], incorporates axis-based multi-head self-attention and cross-layer attention fusion. Finally, ExpoMamba [2] introduces a novel architecture that integrates frequency state space components in a Mamba (a state space model family) to tackle real-time processing challenges.

Multi-Exposure Approaches: MEFNet [19] employs a multi-exposure fusion network that predicts the fusion weight maps at a low resolution for fast processing, HoLoCo [16] introduces contrastive learning with a holistic and local contrastive constraint to multi-exposure image fusion to recover details and allow for uniform illumination in over and under exposed regions. MobileMEF [11] is a lightweight multi-exposure fusion network for real-time processing on mobile platforms.

For training, we used our entire set of Human-edited Ground Truth Training data scenes (13,067 scenes total). All low-light methods were trained for 100 epochs using the original implementations and hyper-parameters provided by their respective authors. When training Single exposure

[5] https://corp.dxomark.com/

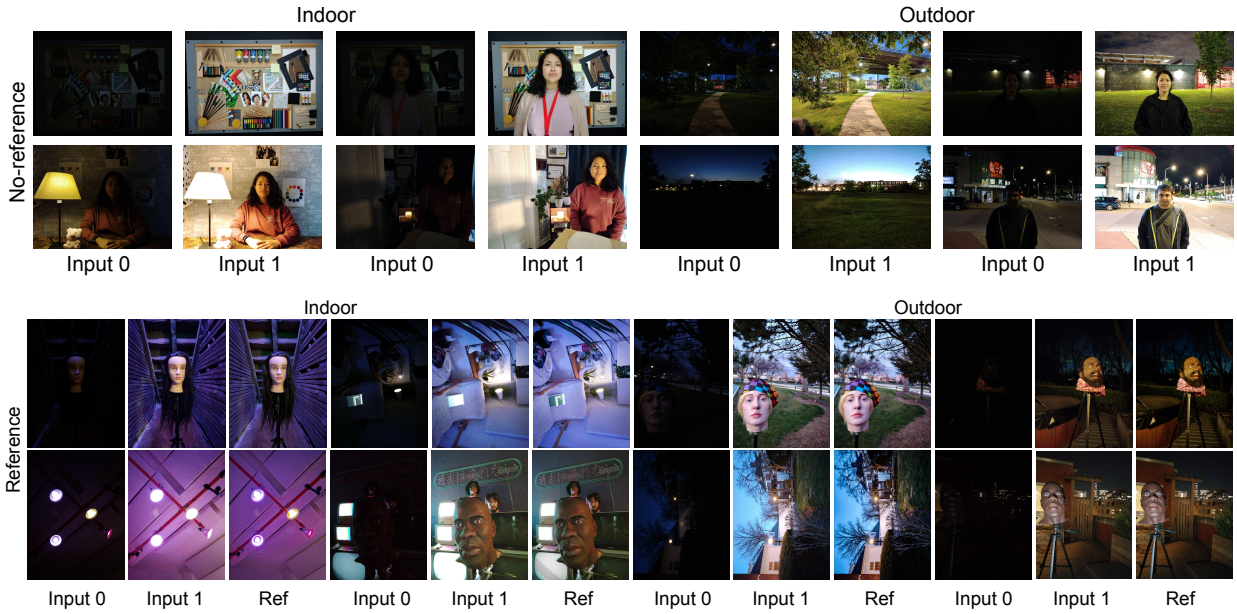

Figure 8. Example of LENVIZ No-Reference [Top] and Reference [Bottom] Test data

Table 3. Quantitative Evaluation results for SOTA single exposure (SE) methods and Multi-exposure fusion (ME) methods on LENVIZ dataset. Here, higher the PSNR and SSIM value and lower the LPIPS score indicated that better the output quality

Type	**SOTA**	**Test Dataset**	**↑ PSNR**		**↑ SSIM**		**↓ LPIPS**	
			LENVIZ trained	LOL/SICE trained	LENVIZ trained	LOL/SICE trained	LENVIZ trained	LOL/SICE trained
SE	ZeroDCE++ [12]	LENVIZ	16.35	**16.38**	0.359	**0.437**	0.663	**0.611**
		LOL	14.75	14.75	0.518	0.518	0.328	0.328
	LLFormer [28]	LENVIZ	**21.13**	17.05	**0.665**	0.551	**0.358**	0.498
		LOL	19.33	**19.79**	**0.757**	0.772	**0.240**	0.277
	ExpoMamba [2]	LENVIZ	**19.8**	17.04	**0.585**	0.534	**0.524**	0.599
		LOL	**23.65**	18.55	**0.816**	0.759	**0.169**	0.291
ME	MEFNet [19]	LENVIZ	20.71	**20.77**	**0.609**	0.606	**0.457**	0.458
		SICE	**21.13**	20.94	0.612	0.612	**0.358**	0.361
	HoLoCo [16]	LENVIZ	21.31	**20.77**	**0.613**	0.606	0.689	0.689
		SICE	13.78	**13.90**	0.614	**0.615**	**0.526**	0.529
	MobileMEF [11]	LENVIZ	**20.93**	19.47	**0.626**	0.613	**0.492**	0.561
		SICE	13.65	**14.36**	**0.637**	0.632	0.484	**0.452**

methods, we used frames captured at a low exposure value (EV -20) as the input for the models. For Multi exposure approaches, we provided two frames as input, one captured at low exposure (EV -20), and one at medium exposure (EV 0). The selection of these exposure values was done to closely follow the observed illumination of the input frames provided in the LOL and SICE training data.

To measure the performance of these methods, we conducted a quantitative and qualitative evaluation to assess their image quality when trained with our dataset in contrast with the results obtained when training with existing benchmark datasets. We further evaluate the generalization capabilities of each approach by performing cross-dataset evaluation. For our quantitative evaluation we included evaluation metrics such as PSNR, SSIM, and LPIPS. Our qualitative evaluation consisted of a human study to evaluate the perceived image quality of each method's outputs. We incorporated the feedback of 238 users who rated the outputs and provided insights on their perceived quality based on eight features: naturalness, brightness, blur, details, colorfulness, noise, contrast, and skin tone accuracy.

7.4. Quantitative evaluation

To complement our extensive user study and provide a comprehensive quantitative assessment, we evaluated the performance of six state-of-the-art low-light enhancement models when trained on our dataset and on two leading benchmark datasets (LOL and SICE). The results summarized in Table 3, highlight the superior performance of models trained on our data across key metrics.

The evaluation included three primary metrics: LPIPS (Learned Perceptual Image Patch Similarity), SSIM (Structural Similarity Index Measure) and PSNR (Peak Signal-to-Noise Ratio). LPIPS is a perceptual metric that uses a deep learning model to measure the similarity of two images as a human would perceive it. SSIM evaluates image quality based on structural information, brightness, and contrast, providing a more human-centric score than tradition pixel based metrics. PSNR, in contrast, is a simple pixel-by-pixel comparison that is highly sensitive to small pixel shifts or variations.

Our findings reveal an overwhelming trend of superior performance in LPIPS and SSIM for models trained on our dataset, both on our test set and in cross-dataset evaluations. This indicates that our dataset is exceptionally effective at training models that produce perceptually pleasing images with high structural and aesthetic fidelity. With only minor exceptions, such as ZeroDCE++ and MobileMEF obtaining slightly better results when trained on LOL and SICE (0.05 and 0.03 respectively). This performance is almost entirely mirrored in the SSIM metric, with LENVIZ-trained models consistently outperforming their counterparts, with slightly more variation observed among multi-exposure methods trained on SICE.

In contrast to the perceptual metrics, the results for PSNR were more mixed. While models trained on our dataset still secured superior scores in a significant portion of the test cases, the overall distribution of scores was less consistent across methods and datasets. This is because PSNR is a strict, pixel-by-pixel metric that is sensitive to subtle differences in color, brightness, and alignment that a human eye would not notice.

7.4.1. Failure cases analysis

We also include a dedicated analysis of failure cases to provide a more comprehensive understanding of LENVIZ data trained model's limitations and to guide future research. Fig 9 showcases these instances with a side-by-side comparison of our data trained model output against the benchmark data trained model output. In these specific cases, the user study revealed a preference for the benchmark trained model output due to its superior color, contrast, detail, and lower noise. Upon analysis, we observed that while the outputs of our data trained model output and the benchmark are visually quiet similar, the quality difference is largely attributable to our model's early stopping. Due to time constraints, the model was trained only 100 epoch using our dataset and with sheer the size of the dataset the training time was much longer not allowing us to wait for full convergence. Based on our prior experience with similar models, we are confident that continued training would enable our data trained model to produce results that not only match but also surpass the benchmark trained model output quality.

7.4.2. Model Generalizability

The examples in Fig 10 showcase the improved performance of the SOTA models when trained on LENVIZ data vs benchmark (LOL/SICE) dataset. For fairness of comparison, the test data used in this study was benchmark (LOL/SICE). We can observe that the model output when trained on LENVIZ data demonstrates improved brightness, contrast, texture, and sharpness. This provides string empirical evidence that our dataset, being captured with 3 different camera module under fixed camera settings, enables models to learn highly robust and generalizable features. The unprecedented scale and diversity of our dataset are key factors in its effectiveness as a training tool for low-light enhancement models across a variety of camera hardware and scene types.

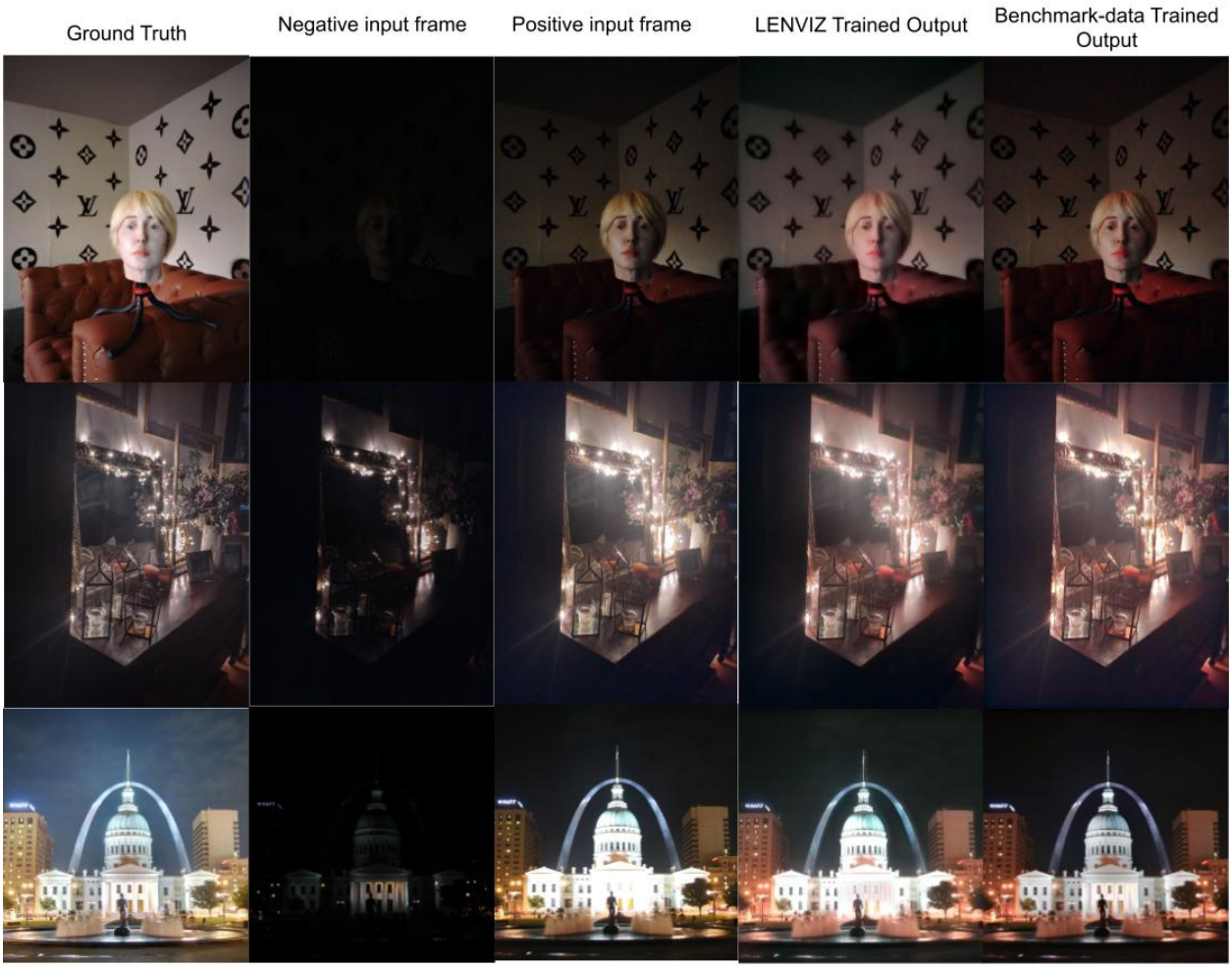

Figure 9. Example of failed cases where users preferred Benchmark-trained model output (MobileMEF) compared to LENVIZ trained model output supported with GT and input frames (dark and bright).

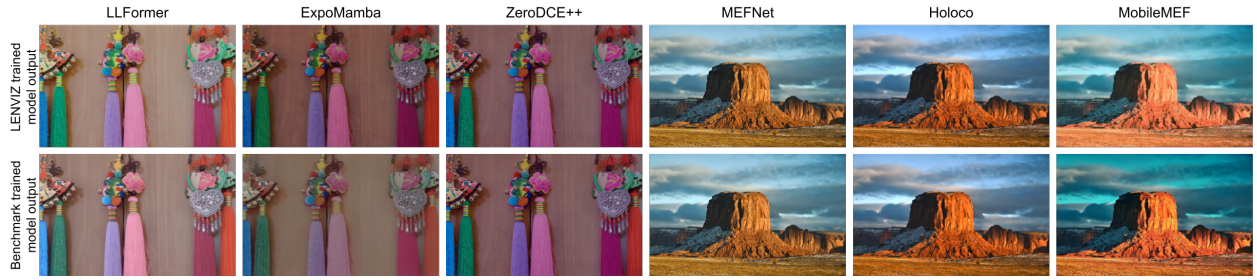

Figure 10. Output samples from models trained with LENVIZ data vs Benchmark-data (LOL/SICE) using the benchmark test data. Model's trained on LENVIZ demonstrate comparable or superior image quality even when tested on data captured by different (unseen) cameras, showcasing the model's generalizability after the model is trained using LENVIZ dataset