WarpI2I: Image Warping for Image-to-Image Translation

Shen Zheng, Anurag Ghosh, Gaurav Parmar, and Srinivasa Narasimhan

Carnegie Mellon University

Abstract. Image-to-image (I2I) translation has achieved strong results in tasks like human relighting and driving scene translation using latent diffusion models (LDMs). However, compact LDMs often struggle to preserve fine-grained structures because the encoder compresses high-resolution inputs into a spatially downsampled latent space. To address this issue, we propose a simple saliency-guided warp–unwarp framework that reallocates spatial representation toward salient regions before encoding, enabling better preservation of structural details without increasing latent resolution. The warped image is processed by the original diffusion model and then mapped back via an inverse warp. In addition, we propose a simple and efficient outpainting-based synthetic data generation pipeline to produce high-quality paired data for image relighting. Our method is model-agnostic, requires no architectural modification, and introduces negligible computational overhead. Experiments on human relighting, driving scene relighting, and translation demonstrate improved structural preservation, lighting faithfulness, and image quality, with our framework extending naturally to video via frame-by-frame application with good temporal stability. Project Webpage: https://shenzheng2000.github.io/WarpI2I.github.io/

Keywords: Image Warping · Image-to-Image Translation · Relighting

1 Introduction

Image-to-image (I2I) translation has achieved remarkable progress in tasks such as human relighting and driving scene weather transformation. Latent diffusion models (LDMs) [15,34,38] have emerged as the dominant paradigm due to their strong photorealism, stable optimization, and scalability.

While high-capacity commercial systems (e.g., Midjourney [30], DALL·E 3 [3]) demonstrate impressive performance, they are closed-source and computationally intensive. Thus, lightweight open-source LDMs like Stable Diffusion [38] and its scaled variants [15, 34] are widely adopted for practical deployment.

However, compact LDMs often struggle to preserve fine-grained foreground structures, particularly facial details and small objects. This limitation stems from the encoder's spatial compression, which maps high-resolution inputs into a downsampled latent space and discards high-frequency structural information.

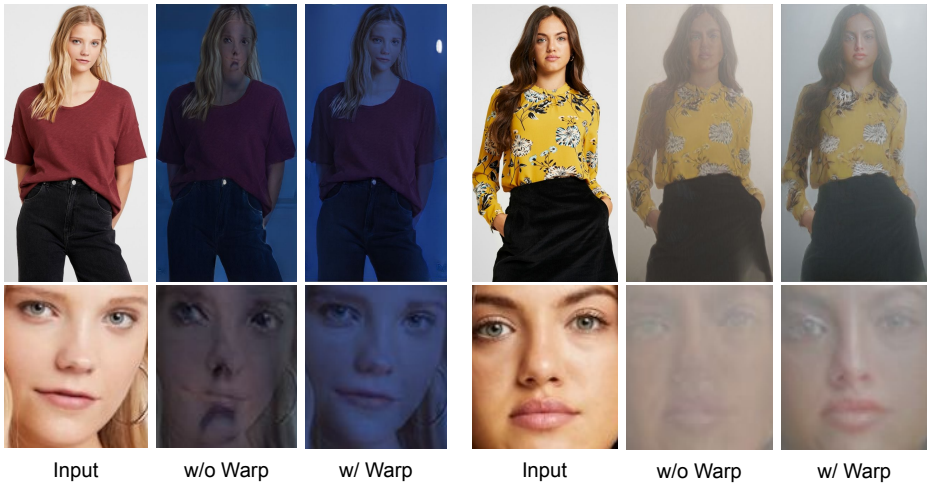


Fig. 1: Our warping significantly improves fine details of face regions, compared to the baseline img2img-turbo [32]. Real human image is from VITON-HD [9] dataset using moonlight (*left*) and foggy (*right*) as prompts for human relighting.

A straightforward remedy is to decrease the encoder's compression ratio or increase the input image resolution to obtain a larger latent spatial size. However, this strategy significantly increases computational memory and inference time, as the computational cost scales quadratically with latent spatial size, making it impractical for real-world deployment. Alternative strategies such as multi-scale diffusion pipelines [1, 22, 39] or test-time refinement [12, 26] have been explored, but these approaches typically introduce additional computational overhead during training or inference, and often require architecture modification.

In this work, instead of enlarging the latent space, we preserve the original encoder compression ratio and input resolution, and counter the spatial shrinking caused by the encoder using a content-aware spatial image warping before encoding. This warping locally enlarges salient (important) regions, giving more representation in the latent space. This is followed by inverse warping (unwarp) after image translation to restore the original geometry (Figure 2). This warp-unwarp strategy is model-agnostic, requires no architectural changes and little additional computation, and can be integrated into any latent image-to-image translation framework. This paper makes the following contributions.

- We propose a simple, yet effective, saliency-guided warp–unwarp framework that reallocates spatial representation to salient regions before encoding, counteracting latent spatial shrinkage without enlarging the latent space.
- Our approach is model-agnostic, requires no architectural modification, and is compatible with existing latent image-to-image translation frameworks.
- Our warping introduces negligible computational overhead (3.0 ms per image) without any additional learnable parameters.

– To produce high-quality paired data for human and driving scene relighting, we propose a simple and efficient synthetic data generation pipeline.

Extensive experiments demonstrate strong performance across multiple image-to-image translation tasks, including human relighting, driving scene relighting, and weather or time-of-day translation. Quantitative evaluation using user studies and standard metrics (FID [20], KID [5], Clean-FID [33], DINO-Struct [46], ArcFace [13], masked LPIPS [53], CLIP score [19, 35]) shows that our warping-based method consistently outperforms recent state-of-the-arts, including pix2pix-Turbo, CycleGAN-Turbo [32], IC-Light [52] and DreamLight [29]. Although primarily designed for static images, our method can also be applied frame-by-frame to video, demonstrating good temporal consistency and stability (See project webpage for video results). Even without warping, our data generation pipeline for creating high-quality paired images could benefit downstream tasks like scene understanding and image rendering. *Additional results are in Supp.*

2 Related Works

2.1 Image Warping

Non-uniform spatial transformations were originally developed to correct geometric distortions, as studied in Feature-Based Image Metamorphosis [2] and Digital Image Warping [47]. With deep learning, Spatial Transformer Networks [25] enabled end-to-end learned transformations that intentionally distort images to improve performance. Learning to Zoom [36] showed the benefit of over-sampling salient regions for multi-resolution reasoning, motivating extensions to detection and segmentation. Following this line, Fovea [44] introduces a dataset-level static & temporal prior, while Two-Plane Prior [16] employs an image-level geometric prior to guide warping. More recently, InstanceWarp [55] leverages an instance-level object prior for domain-adaptive detection and segmentation.

In comparison, we are the first to apply saliency-guided image warping to generative tasks with three key insights. *(a) Model-Agnostic.* Prior discriminative warping methods like [45, 55] unwarp in 2D feature space, making them incompatible with 1D token sequences from ViT/DiT backbones, whereas our method unwarps in image space, making it model-agnostic; *(b) Compute-Efficient.* Even if 2D feature maps are generated by the encoder, prior methods must retrain the encoder for warp-compatible features, which is expensive, whereas our method requires only LoRA [23] fine-tuning; *(c) Improved Fine-Grained Details.* Prior approaches construct saliency maps from coarse-grained signals (e.g., scene-level [16, 44, 45] or instance-level priors [55]), whereas we derive saliency from fine-grained, part-level cues which show better details (See Figure 1), while remaining compatible with coarser signals when needed.

2.2 Text-Guided Image Relighting

Text-guided relighting leverages diffusion models to control illumination via natural language, bypassing the need for physical lighting representations which are

costly and often unavailable in the wild [14]. Text2Relight [6] frames portrait illumination as a text-conditional generation problem, while IC-Light [52] enforces consistent light transport and DreamLight [29] improves global coherence. However, they often struggle to preserve fine-grained details and spatial consistency.

Our approach differs from prior relighting methods in three key aspects. *(a) Lightweight Data Pipeline.* Prior approaches [6, 29, 52] rely on complex, closed-source, and time-consuming synthetic data pipelines, whereas we leverage off-the-shelf text-to-image models to build a simple, open-source, and efficient pipeline that automatically generates high-quality paired data for diverse relighting conditions; *(b) Data-Efficient.* IC-Light [52] and DreamLight [29] require over 10M and 1M training images respectively, whereas our pipeline achieves competitive performance using only 10K images; *(c) Driving Scene Relighting.* Prior methods focus mainly on human or general object relighting, whereas we demonstrate strong performance on the long-tail driving scenario of roadwork.

2.3 Weather and Time-of-Day Driving Scene Translation

Transforming weather and time-of-day conditions in driving scenes predominantly relies on unpaired image-to-image (I2I) translation, as acquiring perfectly paired, pixel-aligned images of the same dynamic scene under drastically different conditions (e.g., clear day vs. heavy fog) is impractical. Existing approaches broadly fall into GAN-based [18] and diffusion-based paradigms [21]. GAN-based methods address domain discrepancies through cycle consistency (DualGAN [49], CycleGAN [57]), shared latent spaces (UNIT [28], MUNIT [24]), and contrastive objectives (CUT [31], MoNCE [51], AS-IntroVAE [7], TPSeNCE [56]), whereas diffusion-based frameworks (Unit-DDPM [42], DDIBs [43], img2img-turbo [32]) have improved photorealism and stable optimization. Unlike many existing I2I methods that struggle to preserve foreground structures under weather and time-of-day transformations, we leverage image warping to maintain structural consistency while enabling realistic scene-level changes.

3 Method

3.1 Overview

Figure 2 illustrates the motivation and architecture of our warp-unwarp framework, which addresses the loss of fine details caused by spatial downsampling in latent diffusion-based image translation by enlarging salient regions before encoding and restoring the original geometry after translation.

In Section 3.2, we first construct high-quality synthetic paired data for human relighting and driving scene relighting using a unified FLUX-based pipeline. In Section 3.3, we define a part-level saliency formulation that identifies semantically important regions. In Section 3.4, we describe the image warping and unwarping mechanism driven by this saliency map. Finally, in Section 3.5, we show how the proposed warp–relight–unwarp pipeline integrates seamlessly into standard image-to-image translation frameworks.

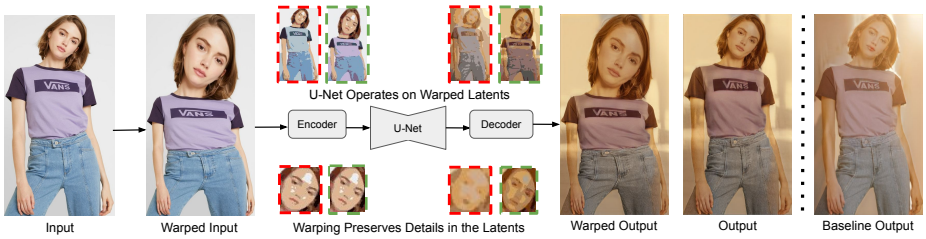


Fig. 2: Overview of our warp-unwarp approach for latent diffusion-based image-to-image translation. Latent diffusion models encode the input into a spatially downsampled latent space (typically 8×), leaving small but detail-rich salient regions like faces with too few latent pixels to reconstruct details like eyes and expressions. We apply a content-aware warp before encoding to enlarge salient regions, and an inverse warp after translation to restore the original geometry. This warp-unwarp strategy is model-agnostic, requires no architectural changes, very little additional computation, and can be integrated into any latent image-to-image translation framework. Original latents in **red**, and warped latents in **green**. Best viewed when zoomed in.

3.2 Synthetic Paired Data Generation

Real paired relighting data is difficult to obtain, as capturing the same scene under multiple controlled lighting conditions while preserving geometry and foreground consistency is labor-intensive and often infeasible. Consequently, recent relighting approaches such as IC-Light [52], Text2Relight [6], and DreamLight [29] rely on synthetic paired datasets generated through complex and computationally intensive pipelines that are time-consuming and resource-heavy.

To address these limitations, we develop a simple and efficient pipeline to generate high-quality synthetic data for training paired I2I models such as pix2pix-Turbo [32] under two relighting scenarios: human relighting and driving scene relighting. Both scenarios are built upon a unified FLUX-based [27] generation framework. The pipeline illustrated in Figure 3 is designed for human relighting, where we aim to preserve the same foreground while generating a different background under new lighting conditions. In contrast, for driving scene relighting, we remove the outpainting and depth estimation steps and directly apply FLUX text-to-image generation, producing images with the same foreground and background but different lighting.

Empirically, the paired I2I model achieves robust relighting performance using only 10K synthetic training pairs. These pairs can be generated in approximately one day using eight NVIDIA RTX A6000 GPUs. Example synthetic images generated from our pipeline are shown in Figure 4.

Image Outpainting via FLUX. We use outpainting to generate relit variants for controlled lighting changes while keeping the original foreground unchanged. Specifically, we apply `FLUX.1-Fill-dev` on an expanded canvas with a foreground mask. A base prompt is derived from a ChatGPT description of each image using a structured template that specifies the foreground person's attributes (pose, clothing, accessories, hair, and expression) and the background

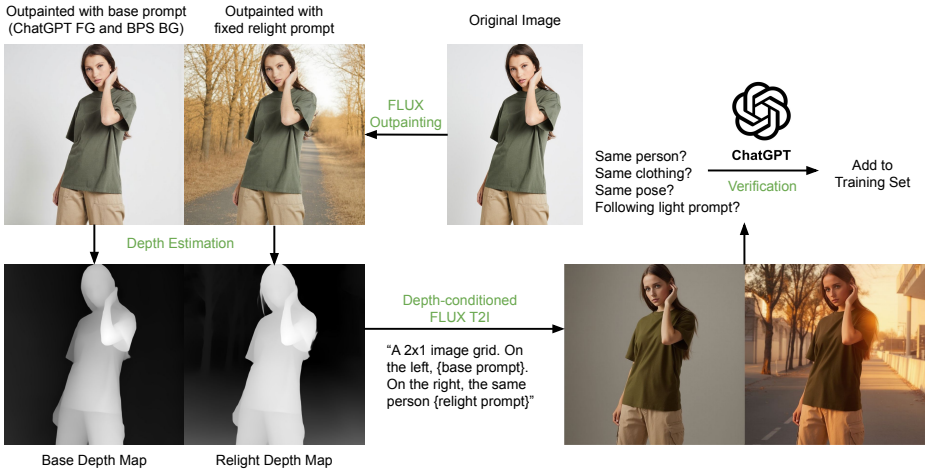


Fig. 3: Our pipeline for generating synthetic paired training images. We first perform FLUX outpainting to expand the scene while GPT provides text annotations that explicitly separate foreground and background descriptions. We then apply Background Prompt Substitution (BPS) to replace the annotated background with a neutral, descriptive prompt. Next, we run Depth Anything on the outpainted images to obtain depth maps, and feed them into FLUX depth-conditioned generation to synthesize a 2×1 image pair depicting the same person under different backgrounds and lighting. Finally, we use ChatGPT to verify whether the generated pair preserves identity, pose, and prompt fidelity, and keep only approved pairs for training.

scene description, and a fixed relighting prompt is used to produce the relit background for each lighting condition.

Background Prompt Substitution (BPS). The VITON-HD [9] dataset contains images with a plain white background. When captions are generated automatically from ChatGPT, this background is typically described as *"empty background with white studio light."* Consequently, the generated base images also contain white background. Training an I2I model to translate such images into relit outputs with diverse backgrounds creates an ill-posed one-to-many mapping, which can destabilize training. To mitigate this issue, we propose Background Prompt Substitution (BPS), which replaces this simple background description with a fixed but neutral and more descriptive background prompt: *"background of a neutral indoor studio with visible walls and a flat floor plane, softly lit with even ambient light."*

Depth Estimation via Depth Anything. We compute depth maps for both the outpainted base and relit image using the Depth Anything-Large model [48]. These depth maps provide geometric conditioning for the subsequent generation step while maintaining consistent foreground structure across lighting variants.

Depth-Conditioned Image Generation via FLUX. Although the outpainted images share the same foreground and different backgrounds, they are not ideal training pairs because the foreground in the relit image is not explicitly relit by

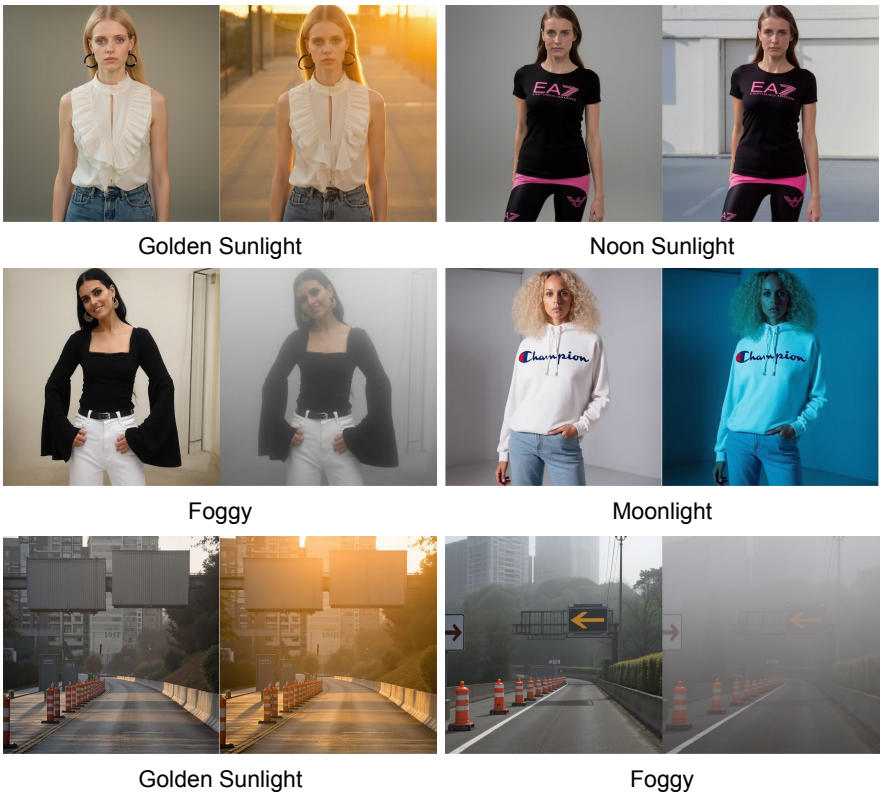


Fig. 4: Synthetic images generated from our pipeline. Our pipeline produces high-quality training pairs for both human (top two rows) and driving scenes (bottom row). For each pair, the left image is a synthetic base image and the right image is the synthetically relit result under the target lighting condition. Our image pairs preserve foreground identity, maintain strong visual quality, and follow the lighting prompts. For driving scenes, we keep the background consistent without outpainting, while human images allow background outpainting for more flexible appearance.

the relighting prompt. To address this, we concatenate the base and relit depth maps into a 2×1 grid (left–right) and use this combined depth map as the conditioning input. We pair it with a structured prompt (*"A 2×1 image grid; on the left, base prompt; on the right, the same person relight prompt"*) and generate the final image using `FLUX.1-Depth-dev`, preserving identity, clothing, and pose while producing globally different lighting across the two halves.

ChatGPT Image Filtering. Not all generated images are high-quality enough to serve as pseudo ground truth for training, and manually filtering low-quality samples at scale would be impractical. Therefore, we use ChatGPT to automatically filter the generated base images and relit images by verifying lighting consistency with the target description and checking identity, clothing, and pose consistency. Samples are retained only when all criteria are satisfied.

3.3 Part-Level Saliency Guidance

Higher saliency corresponds to stronger spatial warping, resulting in greater magnification of the selected region. Therefore, the saliency map should be derived from semantically important cues and tailored to the target task. For human-centric relighting, we assign higher saliency to fine-grained, part-level regions such as faces, especially the eyes. For driving scenes, saliency focuses on object-level regions rather than small part-level details.

The static prior [44] relies on a single dataset-level saliency map for all images, causing warping to incorrectly emphasize average object locations when a specific image deviates from this distribution, while the geometric prior [16] depends on vanishing point cues, not applicable in portrait or non-perspective scenes and tends to oversample distant regions that are not necessarily important.

Our approach is most similar to the instance prior in Instance-Warp [55], but extends saliency guidance to fine-grained parts (e.g., eyes and faces) rather than being limited to whole objects. Unlike Instance-Warp, we do not rely on dataset statistics (e.g., object size distributions), as we found that a simple saliency bandwidth of 128 is sufficient for effective warping and unwarping.

Given an image, we aim to oversample important parts like faces, but inside the faces, we want to oversample *more* for *very* important parts like eyes. Following prior work [16, 44, 55], we use kernel density estimation to model the bounding box saliency as,

$$S = \sum_{(c_i, w_i, h_i) \in y_a} N\left(c_i, b \begin{bmatrix} w_i & 0 \\ 0 & h_i \end{bmatrix}\right) \tag{1}$$

Where c_i, w_i, and h_i represent the center, width, and height of the bounding box, $N(\mu, \Sigma)$ is the normal distribution, and b is the warping bandwidth. Intuitively, larger bandwidth leads to less intense warping.

3.4 Image Warping and Unwarping

Image Warping. Let I denote the input image and S denote the saliency map defined over the output resolution. We parameterize the warp through an inverse mapping (T_S^{-1}) and obtain the warped image by backward resampling,

$$I'(u) = W_T(I) = I\big(T_S^{-1}(u)\big) \tag{2}$$

where u indexes pixel locations. Image warping is efficient, since it introduces an additional 3.0 ms latency per image and has no additional learned parameters, which is negligible for most image translation applications.

Image Unwarping. I2I Translation demands pixel-level fidelity. Unlike prior warping approaches that operate in feature space (e.g., LZU [45], Instance-Warp [55]) or prediction space (e.g., Fovea [44], TPP [16]), we perform unwarping directly in image space to better preserve fine-grained details.

Fig. 5: Warping followed by unwarping introduces minimal information loss. From left to right, we show the original cropped image, the warped image, the reconstructed image after unwarping, and the corresponding pixel-wise difference heatmap, which indicates minimal information loss.

Table 1: Task-specific configurations of our warping framework. For different tasks, we adopt different base I2I models, saliency regions, and saliency signals. The saliency signal is used to derive a saliency map over the specified region, which guides image warping and unwarping. When available, ground-truth annotations are used as the saliency signal; otherwise, detector-generated bounding boxes are employed.

Task	Type	Base I2I Model	Saliency Region	Saliency Signal
Human Relight	Paired	pix2pix-Turbo	Face & Eyes	InsightFace (buffalo_l)
Driving Relight	Paired	pix2pix-Turbo	Object	YOLO-World (yolov8x)
Weather/ToD Driving	Unpaired	CycleGAN-Turbo	Object	GT bboxes

Since the inverse warp T^{-1} has no closed-form solution, we follow LZU [45] and approximate it with $\tilde{T}^{-1}$, a piecewise tiling of locally invertible bilinear mappings, ensuring that both forward and inverse transformations exist.

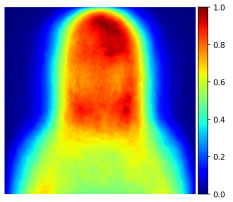
$$T(u) \approx \tilde{T}(u) = \begin{cases} \tilde{T}_{ij}^{-1}(u), & \text{if } u \in \text{Range}(\tilde{T}_{ij}) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The final image is obtained via backward resampling using $\tilde{T}^{-1}$, effectively reversing the warping operation. Our insight is that this unwarping operation, while being an approximation, is reasonable for inversion while losing little fidelity (See Figure 5). Moreover, as the inversion operation is differentiable, the diffusion model learns to compensate for any artifacts caused by warping and unwarping. Finally, unwarping remains efficient, introducing a negligible 3.0 ms latency per image and has no learned parameters.

3.5 Adding Warping into Image-to-Image Translation

Our warping framework is model-agnostic, and integrates seamlessly into both paired and unpaired image-to-image translation settings. As summarized in Table 1, the only *task-specific* components are the base I2I model and the choice of saliency region and signal. All other steps, including saliency map construction, and the warp–relight–unwarp pipeline, remain identical across tasks. Letting the task dictate which regions to warp is often desirable, since it focuses detail where it matters most for that specific problem (e.g., face for human, car for driving).

That said, task-specific saliency isn't required. For a *task-independent* approach, choices are: *(a)* task-independent detector, where bounding-box size relative to image size (e.g., 10%) is used to flag small salient regions; *(b)* pixel-wise error (input - neutral-prompt output) indicates regions for improvement. For human photos, the face region has high error (right) and can guide saliency.



4 Experiments

4.1 Human Relighting

Datasets. We evaluate human relighting on the VITON-HD [9] and StreetTryOn [11] datasets. For dataset preparation, we first annotate textual relighting prompts using ChatGPT (annotation format in Supp.) and generate foreground masks with Grounded-SAM2 [37]. Using these annotations, we construct synthetic paired data by running our text-to-image pipeline on VITON-HD.

Training and Testing. We train pix2pix-Turbo [32] on synthetic pairs with our warp–relight–unwarp framework, where saliency is computed using detected bounding boxes of the face and eyes. For testing, we evaluate on real images from the VITON-HD test split and the StreetTryOn [11] test set with no fine-tuning.

Evaluation Metrics. We conduct a user study with 60 participants with undergraduate computer vision background, as evaluating technical aspects such as semantic consistency and lighting faithfulness requires domain knowledge. Participants rate results on a 1–5 scale (worst to best) across four criteria: 1) person identity preservation, 2) clothing identity preservation, 3) overall image quality, and 4) lighting faithfulness (i.e., alignment between lighting and text prompt). Additional quantitative metrics such as ArcFace identity similarity [13], masked LPIPS [53], CLIP score [19, 35], and GPT-4o evaluation are in Supp.

Results. Qualitative comparisons are shown in Figure 6, and quantitative results are summarized in Table 2. Our method achieves consistently strong performance across all four criteria, with notably better person and cloth identity preservation, higher image quality, and stronger lighting faithfulness to the relit prompt, outperforming IC-Light [52] and DreamLight [29] by a clear margin.

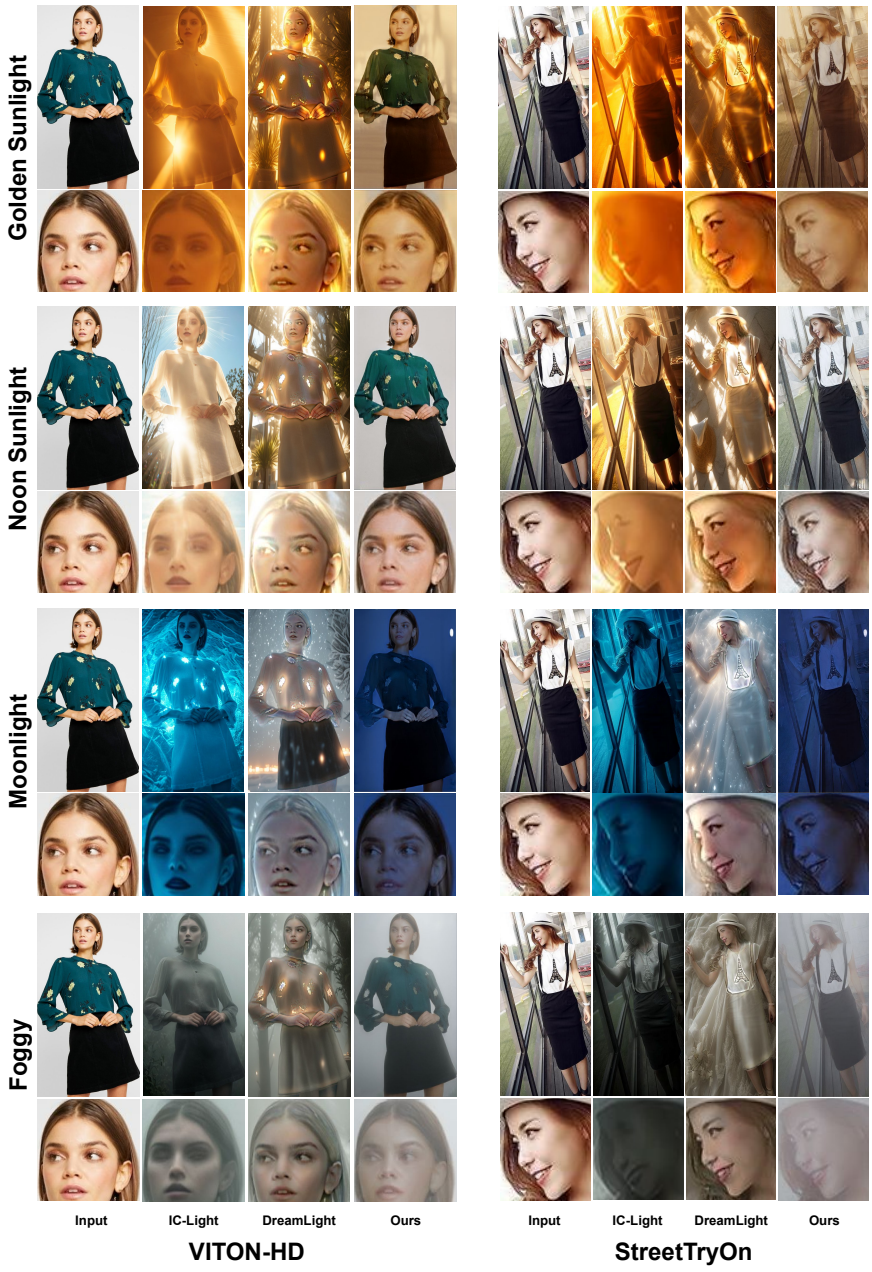


Fig. 6: Qualitative Comparison on Human Relighting. We show one example from VITON-HD [9] and another from StreetTryOn [11] under different lighting prompts. IC-Light [52] and DreamLight [29] fail to preserve clothing identity and exhibit poor lighting faithfulness to the prompts. IC-Light also alters person identity due to distorted facial features. Our method better preserves identity and lighting fidelity, resulting in higher image quality. Please see detailed relighting prompts in Supp.

Table 2: User study results on human relighting. Our method utilize a pix2pix-Turbo baseline trained with synthetic pairs generated by our pipeline and evaluated on VITON-HD [9] and StreetTryOn [11] datasets. Scores are averaged over four lighting prompts: golden sunlight, noon sunlight, moonlight, and foggy conditions. The proposed warping and Background Prompt Substitution (BPS) improve performance.

Methods	VITON-HD					StreetTryOn				
	Person Identity↑	Cloth Identity↑	Image Quality↑	Lighting Faithfulness↑	Avg.	Person Identity↑	Cloth Identity↑	Image Quality↑	Lighting Faithfulness↑	Avg.
IC-Light [52]	3.43	3.46	3.47	3.51	3.47	3.33	3.20	3.25	3.33	3.28
DreamLight [29]	3.52	3.50	3.50	3.41	3.48	3.16	3.13	3.07	3.12	3.12
Ours (No BPS)	3.70	3.72	3.62	3.64	3.67	3.60	3.59	3.49	3.53	3.55
Ours (No Warp)	3.60	3.62	3.48	3.54	3.56	3.60	3.63	3.49	3.55	3.57
Ours	**4.36**	**4.43**	**4.31**	**4.21**	**4.33**	**4.31**	**4.29**	**4.18**	**4.14**	**4.23**

Table 3: Inference latency for 768×1024 resolution images (VITON dataset).

Model	ChatGPT img2img	IC-Light	DreamLight	pix2pix-Turbo	**pix2pix-Turbo + Warp-Unwarp**
Latency (s)	∼30.0	1.237	1.954	0.892	**0.898 (+0.006)**

4.2 Ablation Studies

We conduct ablations to demonstrate the effectiveness of our warping and Background Prompt Substitution (BPS). User study scores are reported in Table 2, showing that both BPS and warping improve image relighting. Additional ablations are in Supp.: *(a)* Our method's robustness to detection error and bandwidth values; *(b)* pix2pix-Turbo with our warping vs. ChatGPT image-to-image.

4.3 Inference Latency

As shown in Table 3, our proposed warp-unwarp module introduces a negligible additional latency of only 0.006 s (3 ms warping, 3 ms unwarping), making it a promising plug-and-play module suitable for real-time applications.

4.4 Weather and Time-of-Day Driving Scene Translation

Datasets. We evaluate on BDD100K [50] (clear day, clear night, rainy day), Cityscapes [10], DarkZurich [40], ACDC [41] (foggy), and Dense Fog [4].
Training and Testing. We train CycleGAN-Turbo [32] with our warp–unwarp framework, where saliency is computed using ground-truth bounding boxes of foreground objects. Models are evaluated on each dataset's official test split.
Evaluation Metrics. We report FID [20], KID [5], Clean-FID [33], and DINO-Struct [46] to measure both visual fidelity and structural consistency.
Results. Quantitative results are summarized in Table 4 for intra-dataset experiments and Table 5 for cross-dataset experiments. Our warping consistently

Table 4: Quantitative results on weather and time-of-day driving scene translation (intra-dataset). KID is reported ×1000 and DINO-Struct ×100.

Methods	BDD100K (Day → Night)				BDD100K (Night → Day)				BDD100K (Clear → Rainy)			
	FID↓	KID↓	Clean-FID↓	DINO Struct.↓	FID↓	KID↓	Clean-FID↓	DINO Struct.↓	FID↓	KID↓	Clean-FID↓	DINO Struct.↓
CycleGAN-Turbo [32]	19.2	8.08	31.3	3.00	41.5	27.7	45.2	3.80	57.7	9.47	57.2	1.50
+Warp Det bbox	17.7	6.70	17.3	2.95	36.8	24.1	35.6	3.55	56.0	6.93	56.3	0.86
+Warp GT bbox	**17.5**	**6.61**	**17.2**	**2.92**	**36.3**	**23.8**	**35.4**	**3.51**	**55.9**	**6.91**	**55.9**	**0.84**

Table 5: Quantitative results on weather and time-of-day driving scene translation (cross-dataset). KID is reported ×1000 and DINO-Struct ×100.

Methods	Cityscapes → DarkZurich				DarkZurich → Cityscapes				Cityscapes → ACDC Foggy			
	FID↓	KID↓	Clean-FID↓	DINO Struct.↓	FID↓	KID↓	Clean-FID↓	DINO Struct.↓	FID↓	KID↓	Clean-FID↓	DINO Struct.↓
CycleGAN-Turbo [32]	159.5	98.1	160.3	4.35	221.4	157.8	209.6	**2.22**	308.8	340.9	286.5	9.18
+Warp Det bbox	154.1	93.6	155.0	3.62	213.7	146.6	206.5	2.99	182.7	108.7	169.5	3.72
+Warp GT bbox	**153.3**	**93.2**	**153.4**	**3.20**	**212.6**	**144.9**	**204.6**	2.75	**169.2**	**96.6**	**154.4**	**2.53**

improves FID, KID, and Clean-FID, indicating better alignment with real images in the target domain. DINO-Struct also improves in most cases, indicating better preservation of structural information from the source image. The only exception is DarkZurich → Cityscapes, where the metric is less reliable due to the reduced robustness of DINO features under low-light conditions [54]. Finally, replacing GT bboxes with YOLO-World [8] detected bboxes yields comparable gains, showing our method does not rely on ground-truth annotations. Qualitative results are in Supp.

4.5 Driving Scene Relighting

Datasets. We use the ROADWork [17] dataset due to its rich ground-truth text annotations for generating relighting training pairs. We construct synthetic paired data by running our text-to-image pipeline on ROADWork (Pittsburgh).

Training and Testing. We train pix2pix-Turbo [32] on synthetic pairs with our warp–relight–unwarp framework, where saliency is computed using YOLO-World [8] detected object bounding boxes. Evaluation is at ROADWork (Boston).

Evaluation Metrics. We follow the same user-study protocol as in human relighting, but replacing person and clothing identity with semantic consistency.

Results. Qualitative comparisons are shown in Figure 7, and user study results are shown in Table 6. Our method, especially with warping, shows better semantic consistency, image quality, and lighting faithfulness, compared with the baseline IC-Light [52] and DreamLight [29].

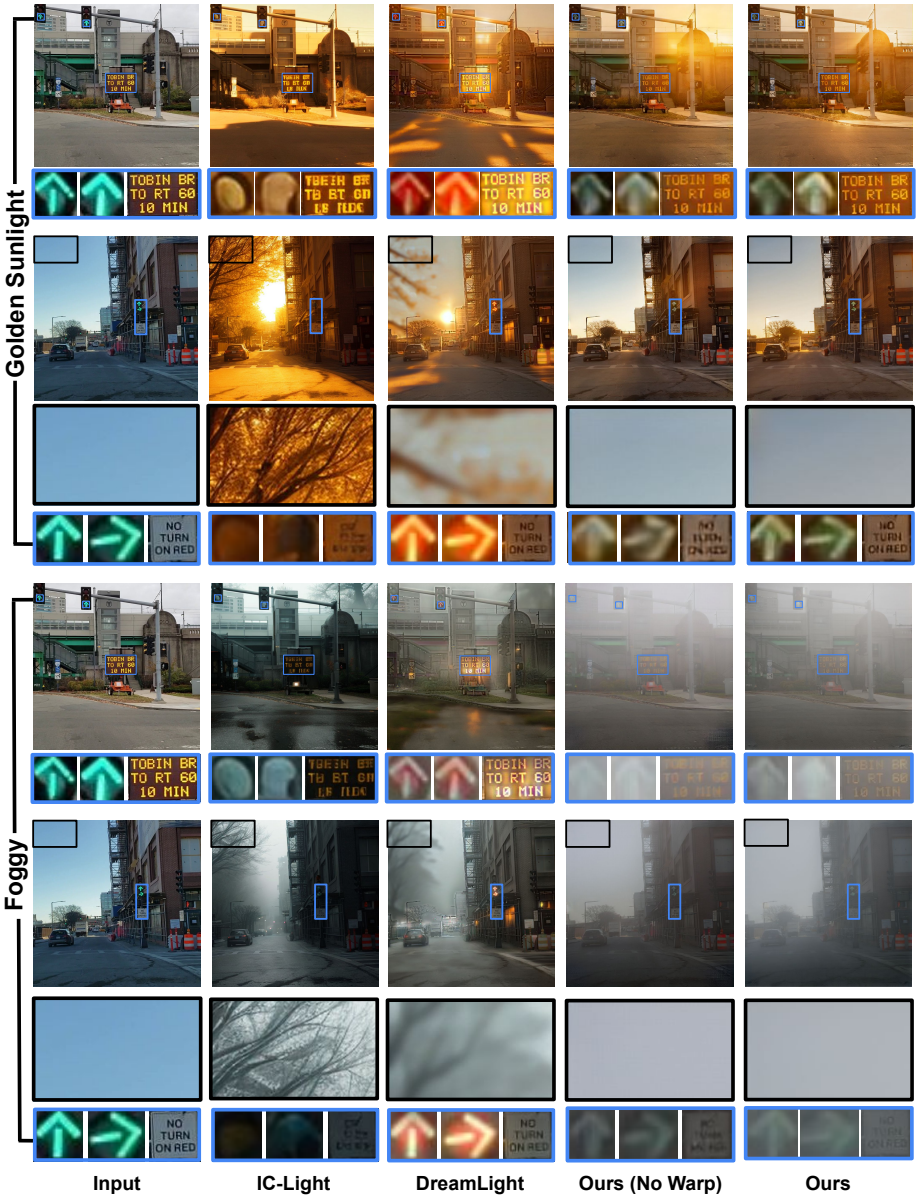


Fig. 7: Qualitative comparison on driving scene relighting. IC-Light [52] and DreamLight [29] produce lower image quality and less faithful lighting, and often fail to maintain semantic consistency. For example, IC-Light turns green traffic arrows into circular lights or removes them, while DreamLight changes green arrows into red or yellow ones. Our method has better semantic consistency, image quality, and lighting faithfulness. With our warping, small details such as traffic arrows and the 'No Turn on Red' sign become more clearly visible.

Table 6: User study results on driving scene relighting. Our method utilizes a pix2pix-Turbo baseline trained with synthetic pairs generated by our pipeline and evaluated on the ROADWork (Boston) dataset under two lighting prompts: golden sunlight and foggy. For driving scene relighting, Background Prompt Substitution (BPS) is not applied because we keep the background consistent and therefore do not perform image outpainting when generating the synthetic training pairs. The proposed warping improves performance, particularly on semantic consistency.

Methods	ROADWork (Golden Sunlight)				ROADWork (Foggy)			
	Semantic Consistency↑	Image Quality↑	Lighting Faithfulness↑	Avg.	Semantic Consistency↑	Image Quality↑	Lighting Faithfulness↑	Avg.
IC-Light [52]	3.63	3.65	3.82	3.70	3.82	3.87	3.92	3.77
DreamLight [29]	3.97	3.90	3.95	3.94	3.85	3.87	3.90	3.91
Ours (No Warp)	4.17	4.30	4.28	4.25	4.18	4.28	4.37	4.26
Ours	**4.53**	**4.55**	**4.55**	**4.54**	**4.48**	**4.45**	**4.48**	**4.51**

5 Conclusion and Limitations

In this paper, we propose a simple saliency-guided warp–unwarp framework that reallocates spatial representation toward salient regions before encoding, enabling better preservation of structural details without increasing latent resolution. The warped image is processed by the original diffusion model and then mapped back via an inverse warp. In addition, we propose a simple and efficient outpainting-based synthetic data generation pipeline to produce high-quality paired data for image relighting. Our method is model-agnostic, requires no architectural modification, and introduces negligible computational overhead. Experiments on human relighting, driving scene relighting and translation demonstrate improved structural preservation, lighting faithfulness, and overall image quality. Beyond static images, our framework extends naturally to video via frame-by-frame application with good temporal stability.

Limitations. Although our warping is model-agnostic, we currently evaluate it only on one-step diffusion baselines (e.g., pix2pix-Turbo and CycleGAN-Turbo [32]) due to computational constraints. Future work could explore integrating our approach with larger, multi-step diffusion models. Additionally, while our framework extends to video on a frame-by-frame basis, it may flicker under fast motion, where stronger temporal priors such as optical flow could improve consistency.

Acknowledgments. This research was supported in parts by SpreeAI (human relighting), General Motors - Research (driving scenes) and by U.S. Department of Transportation Grant 69A3552344811 and 69A3552348316 through the Carnegie Mellon University Safety21 University Transportation Center.

References

1. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
2. Beier, T., Neely, S.: Feature-based image metamorphosis. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2 (2023)
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf (2023)
4. Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., Heide, F.: Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In: CVPR (2020)
5. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
6. Cha, J., Ren, M., Singh, K.K., Zhang, H., Hold-Geoffroy, Y., Yoon, S., Jung, H., Yoon, J.S., Baek, S.: Text2relight: Creative portrait relighting with text guidance. In: AAAI (2025)
7. Changjie, L., Shen, Z., Zirui, W., Omar, D., Gaurav, G.: As-introvae: Adversarial similarity distance makes robust introvae. In: ACML (2023)
8. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: Real-time open-vocabulary object detection. In: CVPR (2024)
9. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: CVPR (2021)
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
11. Cui, A., Mahajan, J., Shah, V., Gomathinayagam, P., Liu, C., Lazebnik, S.: Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In: WACV (2025)
12. Dalal, D., Vashishtha, G., Mishra, U., Kim, J., Kanda, M., Ha, H., Lazebnik, S., Ji, H., Jain, U.: Constructive distortion: Improving mllms with attention-guided image warping. arXiv preprint arXiv:2510.09741 (2025)
13. Deng, J., Guo, J., Niannan, X., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)
14. Einabadi, F., Guillemaut, J.Y., Hilton, A.: Deep neural models for illumination estimation and relighting: A survey. In: CGF (2021)
15. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
16. Ghosh, A., Reddy, N.D., Mertz, C., Narasimhan, S.G.: Learned two-plane perspective prior based image resampling for efficient object detection. In: CVPR (2023)
17. Ghosh, A., Zheng, S., Tamburo, R., Vuong, K., Alvarez-Padilla, J., Zhu, H., Cardei, M., Dunn, N., Mertz, C., Narasimhan, S.G.: Roadwork: A dataset and benchmark for learning to recognize, observe, analyze and drive through work zones. In: ICCV (2025)
18. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NeurIPS (2014)
19. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: EMNLP (2021)

20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS (2017)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
22. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. JMLR (2022)
23. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR (2022)
24. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: ECCV (2018)
25. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. NeurIPS (2015)
26. Khan, M.A.H., Jain, Y., Bhattacharyya, S., Vineet, V.: Test-time prompt refinement for text-to-image models. In: ICCV (2025)
27. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
28. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. NeurIPS (2017)
29. Liu, Y., Xiao, W., Wang, Q., Chen, J., Wang, S., Wang, Y., Wu, X., Tang, Y.: Dreamlight: Towards harmonious and consistent image relighting. NeurIPS (2026)
30. Midjourney, I.: Midjourney. `https://www.midjourney.com` (2024), accessed: 2026-03-03
31. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: ECCV (2020)
32. Parmar, G., Park, T., Narasimhan, S., Zhu, J.Y.: One-step image translation with text-to-image models. arXiv preprint arXiv:2403.12036 (2024)
33. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: CVPR (2022)
34. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
36. Recasens, A., Kellnhofer, P., Stent, S., Matusik, W., Torralba, A.: Learning to zoom: a saliency-based sampling layer for neural networks. In: ECCV (2018)
37. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
39. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS (2022)
40. Sakaridis, C., Dai, D., Gool, L.V.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: ICCV (2019)

41. Sakaridis, C., Dai, D., Van Gool, L.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: ICCV (2021)
42. Sasaki, H., Willcocks, C.G., Breckon, T.P.: Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. arXiv preprint arXiv:2104.05358 (2021)
43. Su, X., Song, J., Meng, C., Ermon, S.: Dual diffusion implicit bridges for image-to-image translation. arXiv preprint arXiv:2203.08382 (2022)
44. Thavamani, C., Li, M., Cebron, N., Ramanan, D.: Fovea: Foveated image magnification for autonomous navigation. In: ICCV (2021)
45. Thavamani, C., Li, M., Ferroni, F., Ramanan, D.: Learning to zoom and unzoom. In: CVPR (2023)
46. Tumanyan, N., Bar-Tal, O., Bagon, S., Dekel, T.: Splicing vit features for semantic appearance transfer. In: CVPR (2022)
47. Wolberg, G.: Digital image warping, vol. 10662. IEEE computer society press Los Alamitos, CA (1990)
48. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
49. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: ICCV (2017)
50. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: CVPR (2020)
51. Zhan, F., Zhang, J., Yu, Y., Wu, R., Lu, S.: Modulated contrast for versatile image synthesis. In: CVPR (2022)
52. Zhang, L., Rao, A., Agrawala, M.: Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In: ICLR (2025)
53. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
54. Zhang, X., Xu, Z., Tang, H., Gu, C., Chen, W., El Saddik, A.: Wakeup-darkness: When multimodal meets unsupervised low-light image enhancement. ACM TOMM (2025)
55. Zheng, S., Ghosh, A., Narasimhan, S.: Instance-warp: Saliency guided image warping for unsupervised domain adaptation. In: WACV (2025)
56. Zheng, S., Lu, C., Narasimhan, S.G.: Tpsence: Towards artifact-free realistic rain generation for deraining and object detection in rain. In: WACV (2024)
57. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)