KPLM-STA: Physically-Accurate Shadow Synthesis for Human Relighting via Keypoint-Based Light Modeling

Xinhui Yin[1,3], **Qifei Li**[2,3], **Yilin Guo**[4], **Hongxia Xie**[1,2,3*], **Xiaoli Zhang**[1,2,3*]

[1]College of Software Engineering, Jilin University, China
[2]College of Computer Science and Technology, Jilin University, China
[3]Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Jilin University, China
[4]School of Computer Science, Peking University, China

Abstract

Image composition aims to seamlessly integrate a foreground object into a background, where generating realistic and geometrically accurate shadows remains a persistent challenge. While recent diffusion-based methods have outperformed GAN-based approaches, existing techniques, such as the diffusion-based relighting framework IC-Light, still fall short in producing shadows with both high appearance realism and geometric precision, especially in composite images. To address these limitations, we propose a novel shadow generation framework based on a Keypoints Linear Model (KPLM) and a Shadow Triangle Algorithm (STA). KPLM models articulated human bodies using nine keypoints and one bounding block, enabling physically plausible shadow projection and dynamic shading across joints, thereby enhancing visual realism. STA further improves geometric accuracy by computing shadow angles, lengths, and spatial positions through explicit geometric formulations. Extensive experiments demonstrate that our method achieves state-of-the-art performance on shadow realism benchmarks, particularly under complex human poses, and generalizes effectively to multi-directional relighting scenarios such as those supported by IC-Light.

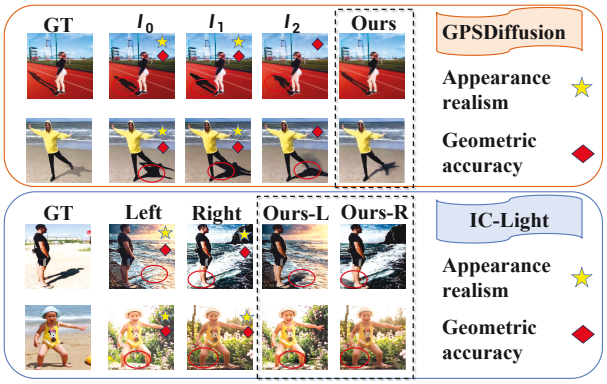

Figure 1: **Comparison with GPSDiffusion (Zhao et al. 2025) and IC-Light (Zhang, Rao, and Agrawala 2025).** The first row and columns $2-4$ are the random different results of GPSDiffusion, the second row and columns $2-3$ are the different light results of IC-Light. It can be seen that each image presents problems of appearance realism (shadow color is slightly darker) and geometric accuracy (limb shape is incorrect). However, our method (ours) has effectively solved these two problems.

Introduction

Image composition (Niu et al. 2021) aims to combine foreground images and background images to generate composite images and has a wide range of applications. The problem addressed in this paper has broad practical application prospects, especially in the generation of composite image shadows in multiple scenarios and with multiple characters, such as virtual game scene transformation, obstacle shadow recognition in autonomous driving of automobiles, and estimation of building height based on shadows in remote sensing and geographic information systems, which are of great significance (Liasis and Stavrou 2016; Kadhim and Mourshed 2017; Song et al. 2023; Liu et al. 2024). Therefore, this issue not only has academic research value but also has a promoting effect on practical application.

Diffusion-based methods (Liu et al. 2024; Zhao et al. 2025) can leverage rich prior knowledge from pre-trained foundational models, consequently outperforming GAN approaches significantly and producing more realistic shadow effects. Recent studies show that diffusion-based methods are effective in producing shadows with appearance realism, yet they struggle to capture geometric accuracy. As shown in Figure 1, the problems of appearance realism and geometric accuracy are clearly visible. **Appearance realism** aims to make the generated shadows visually consistent with the physical properties of real shadows (such as color, transparency, softness and hardness, edge blurriness, etc.) and conform to the natural performance under lighting conditions, and **geometric accuracy** aims at the shape, position, and proportion of the shadow generated in three-dimensional space to be strictly consistent with the geometric relationships of objects and light sources in the scene. Imposing Consistent Light (IC-Light) (Zhang, Rao, and Agrawala 2025) has been established as a rapid diffusion-based composite image relighting technology. However, due to generating backgrounds via prompts before compositing

*Corresponding author

with the foreground, IC-Light can preserve scene texture details after relighting, but still suffers from issues in appearance realism and geometric accuracy of ground shadows.

The task of generating shadows based on composite images has been a research focus in recent years. **Appearance realism** and **geometric accuracy** have always been the difficulties in the shadow generation task. To improve this problem, we propose KPLM-STA. The Keypoints Linear Model (KPLM) module precisely localizes key limb coordinates of the human body, ensuring the realism of shadow shape generation. Shadow Triangle Algorithm (STA) then estimates the corresponding shadow length and proportion for each limb based on varying viewpoints, light directions, poses, and ground angles. Through KPLM and STA, shadow angle, shape, length, and scale can be accurately controlled, resulting in realistic and geometrically consistent shadows.

In summary, our contributions are as follows:

• We propose to equip diffusion-based shadow generation model with KPLM, which is a key point to enhance the apperance realism.

• We introduce a shadow triangle algorithm to predict the shadow geometry for the foreground shadow.

• We conducted experiments on the images processed by DESOBA (Hong, Niu, and Zhang 2022), DESOBAv2 (Liu et al. 2024), and IC-Light to verify the effectiveness of our method. Experimental results show that our method achieves the *state-of-the-art* on three datasets, and it has a greater level of authenticity than previous methods.

Related Work

Shadow Generation

Shadow generation methods can be broadly categorized into **GAN-based, geometry-guided, and diffusion-based paradigms**. **GAN-based approaches**, such as Mask-ShadowGAN (Hu et al. 2019) and AR-ShadowGAN (Liu et al. 2020), aim to synthesize visually realistic shadows via adversarial training. While effective in texture generation, they often suffer from structural inconsistencies—such as distorted contours or unnatural deformation under novel lighting—which limit geometric plausibility.

To address this, **geometry-guided methods** introduce structural priors to enforce alignment between objects and their shadows. DMASNet (Tao et al. 2024) and SGRNet (Hong, Niu, and Zhang 2022) leverage bounding-box regression and physical illumination constraints to improve fidelity. However, these methods are primarily designed for rigid or simplified objects, making them unsuitable for articulated human figures.

Recently, **diffusion-based approaches** have gained attention for their flexibility and controllability. SGDiffusion (Liu et al. 2024) and GPSDiffusion (Zhao et al. 2025) incorporate sketch priors or geometric embeddings to enhance spatial consistency. Yet, they largely focus on static objects or primitive shapes, without addressing the complexity of articulated human shadows. Moreover, most existing methods tend to optimize for either appearance realism or geometric accuracy—rarely both.

Meanwhile, **text-conditioned relighting frameworks** offer fine-grained lighting control. IC-Light (Zhang, Rao, and Agrawala 2025) leverages a latent diffusion model to manipulate light direction and tone via prompts, achieving high-quality relighting with contextual consistency. However, it lacks explicit modeling of object-ground interaction and thus cannot generate cast shadows, undermining both photorealism and physical correctness.

To address this gap, we extend IC-Light by explicitly modeling cast shadows for articulated human figures. Our method enables end-to-end generation of shadows that are both light-consistent and structurally grounded, advancing physically-aware, human-centric shadow synthesis.

Human Model

Human shadow generation requires structured body modeling. OpenPose (Cao et al. 2019; Wei et al. 2016; Simon et al. 2017; Cao et al. 2017) and MediaPipe (Google 2020) offer efficient 2D pose estimation from monocular images, providing keypoint landmarks that can represent body geometry. SMPL (Loper et al. 2023), a parametric 3D mesh model, offers high-fidelity shape and pose reconstruction, but is often computationally intensive and unsuitable for lightweight relighting tasks. In our framework, we adopt a simplified 2D body abstraction using 9 keypoints and one bounding block, which captures sufficient spatial structure while enabling efficient shadow geometry inference.

Method

Our method consists of three stages: **geometry-aware preprocessing**, **latent diffusion-based (Rombach et al. 2022) shadow generation**, and **GAN-based (Goodfellow et al. 2014) post-processing refinement**. Given a composite image I_c without foreground shadow, we firstly extract geometric priors using Keypoints Linear Model (KPLM) and Shadow Triangle Algorithm (STA), which provide keypoint and directional shadow information. These priors are encoded and injected into a latent diffusion model (LDM) (Rombach et al. 2022) via ControlNet (Zhang, Rao, and Agrawala 2023) following SGDiffusion (Liu et al. 2024). Finally, we employ a trainable GAN-based post-processing network to refine the generated image and alleviate color shift and background variation. An overview of our pipeline is shown in Figure 2.

KPLM-STA

In the first stage, KPLM and STA perform synchronized preprocessing to ensure geometric accuracy. As shown in Figure 2, both modules require two sets of computations. Given the input image I_c, KPLM determines the keypoint data of the foreground object through individual mapping of 9 key points and 1 key block, while STA models shadow triangles to estimate the overall position and length of the shadow, including missing limb parts. These geometric features are organized and encoded to control key factors such as shadow color and shape.

In the second stage, given the ground-truth shadow image I_g, noise ϵ, and optional text input T_0, the latent representation is obtained via $Z_0 = E_r(I_g)$, and forward diffusion is

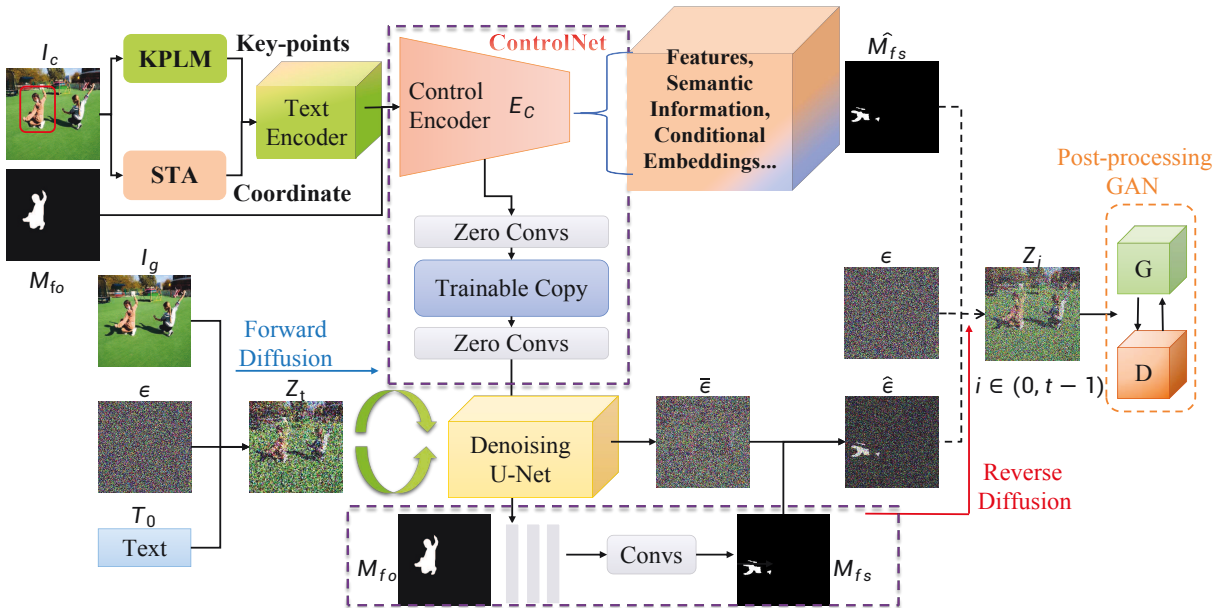

Figure 2: **The framework of our KPLM-STA.** In the first stage, we use KPLM to get key point coordinates, use STA to obtain the shadow angle. In the second stage, we use Control Encoder and Diffusion to generate image. Finally, we use a post-processing GAN for realistic processing.

computed by $Z_t = \epsilon + Z_0$. Following SGDiffusion, we apply ControlNet to inject spatial priors into the denoising U-Net. Specifically, the processed I_c and M_{fo} are concatenated and fed into the control encoder to produce conditional features, which are combined with Z_t during denoising. Additionally, ControlNet uses M_{fo} to predict the foreground shadow mask M_{fs}, which regulates the reverse diffusion process and guides the generation of intermediate latents Z_i. After decoding Z_i, the predicted shadow image $\hat{I}_g$ is obtained.

In the third stage, $\hat{I}_g$ is further refined by a multi-task GAN-based post-processing network that incorporates the foreground mask M_{fo} and composite image I_c to improve realism and remove color shift or background disparity.

Keypoints Linear Model (KPLM)

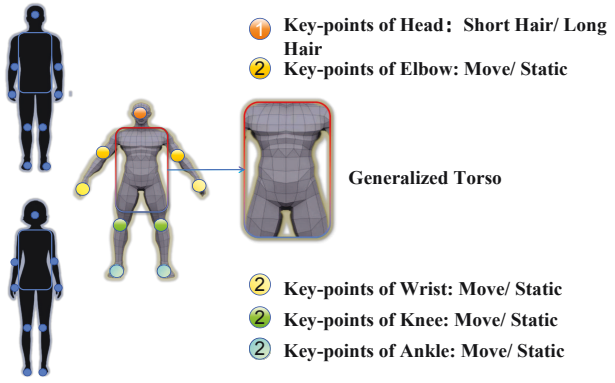

Figure 3: KPLM modeling schematic diagram.

For the task of generating human body shadows in two-dimensional images, there are often problems such as missing limbs, incorrect positions and postures. Previous work such as Pose2Light (Liu, Qiu, and Shen 2022) estimated the lighting direction of human pose in coarse ambient scenes. We leverage 2D keypoints to model the light direction via a linear mapping. KPLM regards the trunk as a universal skeleton frame, which changes with human posture and angle. Unlike pose capture or other human-based tasks, the human body corresponding to the shadow does not require more detailed keypoint capture. KPLM can determine the general posture and position of the people in the picture, and it can achieve better results at a lower cost.

For human body postures, we adopt 9 key points and 1 key block for modeling. Moreover, we can abstract the changes in human shadow postures into the changes of 9 key points as shown on the right side of Figure 3, that is, the generation of human shadow is driven by 9 key points. As shown in the figure, our 9 key points are Head-1, Elbow-2, Wrist-2, Knee-2, Ankle-2 respectively, and the trunk can be regarded as a key block, which can be obtained through STA rotation scaling. As shown in Figure 4, our method was compared with the most universal methods, OpenPose and MediaPipe. The object mask of KPLM is closer to the real object mask than OpenPose, and KPLM uses fewer points than MediaPipe.

In addition, the four vertices of the block change as the human body rotates, which ensures that the universal skeleton frame transforms as the human body rotates. The skeleton frame and key points of each character can be determined through KPLM to determine the human body posture and position. Although in our current implementation KPLM is treated as a pre-processing step, it can extracts

Figure 4: KPLM compared with the mask of object in the most universal method OpenPose, MediaPipe.

structural pose representations from the input image. In this form, KPLM can be integrated into the full diffusion framework, enabling end-to-end training. We leave this extension for future work.[1]

Shadow Triangle Algorithm (STA)

Prior work has explored physically-based shadow rendering by simulating light projection over estimated height maps, such as Pixel Height Maps (Sheng et al. 2022), which enable hard shadow synthesis under geometric constraints. However, such methods assume the human body stands vertically on a flat ground plane, limiting applicability to upright poses and restricting scene diversity. In contrast, STA supports arbitrary body positions and complex poses by enabling triangle-based shadow modeling for each limb individually.

Recent state-of-the-art methods like GPSDiffusion (Zhao et al. 2025) treat the human and its shadow as two separate bounding boxes, deriving the shadow via rotation and translation. While effective at a coarse level, this lacks the granularity needed for fine structural alignment. STA overcomes this by leveraging keypoints extracted from KPLM to perform triangle-based limb-level shadow projection, allowing precise control over the orientation and length of each limb's shadow.

STA models the person and their shadow as geometric

[1]More details see (Yin et al. 2025).

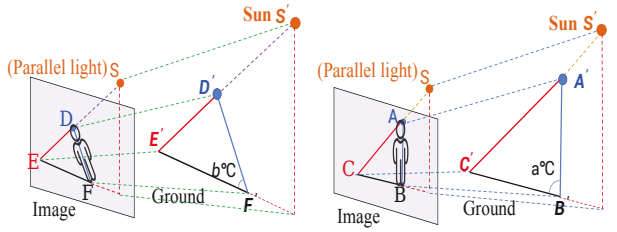

Figure 5: **Shadow Triangle Construction.** Given a human limb segment AB and the corresponding shadow endpoint C, the triangle $\triangle ABC$ is formed under parallel light from the sun. A' and B' denote projected shadow points on the ground. The angle θ is formed between the light direction and the ground plane.

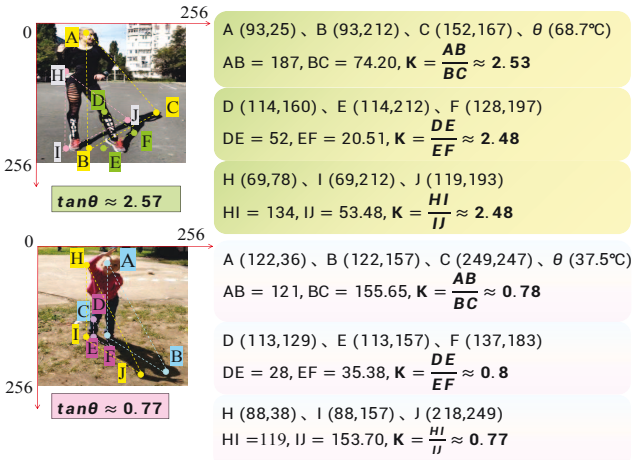

Figure 6: **Proof of STA in Physical Formulas.** K is the proportionality coefficient between body segments and shadows; θ is the angle between the light and the ground.

segments. According to the principle of light projection, the light source direction is uniquely determined by the alignment of the person and the shadow. For points A and D, let C_1 and C_2 be the corresponding circles of their heads, with centers r_1 and r_2, respectively by $\mathrm{Ar}_1 \perp \mathrm{CS}, \quad \mathrm{Dr}_2 \perp \mathrm{ES}$. The triangle formed by light direction, body, and shadow is called a shadow triangle. As shown in Figure 5, triangles $\triangle ABC$ and $\triangle DEF$ correspond to different light angles a° and b°. The limbs and their shadows obey a linear scaling relationship that

$$\mathrm{BC} = \mathrm{K} \cdot \mathrm{AB}, \quad \mathrm{FE} = \mathrm{K} \cdot \mathrm{DF}. \tag{1}$$

Since sunlight can be treated as parallel light, the shadow length of each limb is computed by scaling the limb vector with a coefficient K. To model this process differentiably, we formulate STA as an affine geometric projection. For a limb defined by two keypoints $P_1, P_2 \in \mathrm{R}^2$ on the image plane (e.g., elbow to wrist), we aim to compute its projected shadow endpoint P_3, forming a proportional relationship:

$$\mathrm{K} = \frac{\|\mathrm{P}_1 - \mathrm{P}_2\|_2}{\|\mathrm{P}_2 - \mathrm{P}_3\|_2}, \tag{2}$$

where $\|\cdot\|_2$ denotes Euclidean distance. This coefficient K encodes the ratio between the limb and its shadow, varying with light direction, body orientation, and camera viewpoint. Under orthographic projection and directional lighting, we approximate $K \approx \tan\theta$, where θ is the angle between the light and the ground. Given K, we compute the projected shadow position using:

$$\text{ShadowPos} = P_2 + \frac{1}{K} \cdot \|P_1 - P_2\|_2 \cdot \begin{bmatrix}\cos(\theta)\\ \sin(\theta)\end{bmatrix}. \quad (3)$$

Eq. 3 simulates the projection of a limb along the light direction. After computing projected shadow endpoints for all limbs, we construct shadow limbs by connecting each original limb to its corresponding shadow. These projected limbs form the basis of the shadow mask used in the reverse diffusion process.

To validate STA, we examine both image-space consistency and physical correctness. First, we compare the scaling factor K across the full human body, left elbow, and right knee in 2D image space. Results show all values of K are approximately equal, with a maximum deviation of 0.05 due to keypoint approximation and sampling noise. Similarly, for another sample, the deviation among K values is below 0.02.

Second, in the physical world, shadow formation obeys $\tan\theta = \frac{h}{L}$, where h is the object height and L is shadow length. We empirically verify that $K \approx \tan\theta$ by extracting keypoint coordinates, computing limb and shadow lengths, and measuring θ. As shown in Figure 6, the real-world angle θ aligns with that in the image, confirming that STA faithfully maps 3D shadow behavior to 2D projection, with a maximum observed error of 0.09, which is acceptable given the measurement noise.

Loss Function

Our training objective consists of three main components that correspond to distinct stages in our generation pipeline: geometry-guided denoising diffusion, and control mask prediction.

Geometry-guided diffusion loss. Following SGDiffusion (Liu et al. 2024), we supervise the denoising network using a geometry-aware conditional control feature extracted by KPLM and STA. These modules provide spatial priors that are injected via ControlNet. The diffusion noise prediction loss is formulated as:

$$\mathcal{L}_{\text{mwsg}} = \mathrm{E}_{t,\epsilon\sim\mathcal{N}(0,1)}\left[\|W_{fs} \circ (\epsilon - \hat{\epsilon}_\psi(z_t, t, c_{\text{geo}}))\|_2^2\right], \quad (4)$$

where c_{geo} denotes the concatenated geometric condition derived from KPLM and STA, $\hat{\epsilon}_\psi$ is the noise predicted by the denoising network at time step t, W_{fs} is the soft foreground shadow mask, highlighting training focus on shadow regions.

Foreground shadow mask loss. To ensure that the generated shadows are correctly located with respect to the foreground object, we supervise the predicted mask $\hat{M}_{fs}$ with an L_1 loss against the ground-truth M_{fs}:

$$\mathcal{L}_{\text{mask}} = \left\|\hat{M}_{fs} - M_{fs}\right\|_1. \quad (5)$$

Optionally, dice loss or binary cross-entropy can be added to stabilize training, but we use L_1 loss by default. Finally, we summarize the mask prediction loss in Eq. 4 and weighted noise loss in Eq. 5 as $\mathcal{L}_{all} = \mathcal{L}_{mswg} + \lambda\mathcal{L}_{mask}$, where λ is a trade-off parameter.

Experiments

Experimental Setting

For KPLM, we need to obtain the key point coordinates of each image and store them in json form. Each json includes information such as image number, pose: front, side, key point coordinates, etc. For STA, we need Ground Truth as the guide, and we need to obtain the coordinates of points A, B, and C for each Ground Truth. Following (Zhao et al. 2025), we divide DESOBAv2 into 21,088 training images with 27,718 tuples (composite image, foreground object mask, ground-truth image) and 487 test images with 855 tuples, including both BOS (with background object-shadow pairs) and BOS-free images in the test set. Our Method is implemented with PyTorch (Paszke et al. 2019). We use AdamW optimizer (Loshchilov and Hutter 2017) with a fixed learning rate of 0.0001. All experiments are conducted on 4 NVIDIA RTX 4090 GPUs.

Experiments with IC-Light

The goal of this paper is to generate physically accurate shadows for images relit by IC-Light. We deployed the open-source IC-Light model on an NVIDIA RTX 4090, keeping the output resolution consistent with that of the input image. To ensure generalizability, we evaluated the scene types, seasonal conditions, and lighting directions in various ways. Although IC-Light excels in the relighting appearance, it does not produce realistic ground shadows. Our method, when applied to the relit outputs, can synthesize plausible and geometry consistent shadows. For quantitative evaluation, we use ground-truth images as reference and compare PSNR, SSIM, and LPIPS between IC-Light and our method, as shown in Table 3. IC-Light's scores on these metrics are naturally lower due to the lack of shadow data post-relighting, but this is to be expected and does not reflect its original relighting performance. Figure 8 visualizes the shadows added by our method. Since our pipeline is conditioned on both the relit image and the original structure, the generated shadows align well with the actual scene semantics. As demonstrated in Figure 9, our KPLM module remains effective in various lighting directions.

Experiments on DESOBA

The results of the quantitative comparison are summarized in Table 2. In contrast to DESOBAv2, our comparisons include Pix2Pix (Isola et al. 2017) and its residual variant, Pix2Pix-Res. Our method consistently outperforms all baseline approaches across all metrics. Although the DESOBA dataset is smaller in scale compared to DESOBAv2, our approach still achieves state-of-the-art performance. Specifically, our method yields the lowest GRMSE and LRMSE,

Figure 7: Visual comparison with state-of-the-art methods on DESOBAv2 dataset. From left to right, we show composite image, foreground object mask, results of ShadowGAN(Zhang, Liang, and Wang 2019), AR-SG(Liu et al. 2020), SGRNet(Hong, Niu, and Zhang 2022), DMASNet(Tao et al. 2024), SGDiffusion(Liu et al. 2024), GPSDiffusion(Zhao et al. 2025), Ours, Ground-truth.

Method	BOS Test Image						BOS-free Test Image					
	GR ↓	LR ↓	GS ↑	LS ↑	GB ↓	LB ↓	GR ↓	LR ↓	GS ↑	LS ↑	GB ↓	LB ↓
ShadowGAN	7.511	67.464	0.961	0.197	0.446	0.890	17.325	76.508	0.901	0.060	0.425	0.842
Mask-SG	8.997	79.418	0.951	0.180	0.500	1.000	19.338	94.327	0.906	0.044	0.500	1.000
AR-SG	7.335	58.037	0.961	0.241	0.383	0.761	16.067	63.713	0.908	0.104	0.349	0.682
SGRNet	7.184	68.255	0.964	0.206	0.301	0.596	15.596	60.350	0.909	0.100	0.271	0.534
DMASNet	8.256	59.380	0.961	0.228	0.276	0.547	18.725	86.694	0.913	0.055	0.297	0.574
SGDiffusion	6.098	53.611	0.971	0.370	0.245	0.487	15.110	55.874	0.913	0.117	0.233	0.452
GPSDiffusion	5.896	46.713	0.966	0.374	0.213	0.423	13.809	55.616	0.917	0.166	**0.197**	0.384
Ours	**4.486**	**46.190**	**0.972**	**0.376**	**0.205**	**0.418**	**12.620**	**54.243**	**0.941**	**0.247**	0.198	**0.382**

Table 1: The results of different methods on DESOBAv2 dataset. The best results are highlighted in boldface.**GSSIM (GS), LSSIM (LS), and LBer** are indicators that can measure appearance realism, while **GRMSE (GR), LRMSE (LR)** and **GBer** are indicators that can measure geometric accuracy.

as well as the highest GSSIM and LSSIM* scores, clearly demonstrating its effectiveness. In fairness, all methods were evaluated under the same pre-processing protocol, and the visualization results for DESOBA are included in the supplementary materials.

Experiments on DESOBAv2

As shown in Table 1, our method achieves the best performance across most key metrics, including GRMSE, GSSIM, LSSIM, and LB, demonstrating a clear overall advantage. Although our GBer score is slightly higher than the GPSdiffusion (Zhao et al. 2025), this marginal difference of 0.001 falls within the range of experimental variation and has negligible impact on overall performance. This minor increase in boundary error may result from our model's design, which emphasizes structural consistency and semantic alignment with human posture. Such a design choice may lead to slightly softer transitions at certain boundary regions. Nevertheless, our superior LB score indicates that our method still maintains accurate delineation in critical boundary ar-

Method	BOS Test Image				BOS-free Test Image			
Evaluation matric	GRMSE ↓	LRMSE ↓	GSSIM ↑	LSSIM* ↑	GRMSE ↓	LRMSE ↓	GSSIM ↑	LSSIM* ↑
Pix2Pix	7.659	75.346	0.927	0.249	18.875	81.444	0.856	0.110
Pix2Pix-Res	18.305	81.966	0.901	0.107	5.961	76.046	0.971	0.253
ShadowGAN	5.985	78.413	0.986	0.240	19.306	87.017	0.918	0.078
Mask-ShadowGAN	8.287	79.212	0.953	0.245	19.475	83.457	0.891	0.109
ARShadowGAN	6.481	75.099	0.983	0.251	18.723	81.272	0.917	0.109
SGRNet	4.754	61.762	0.988	0.380	15.128	61.439	0.927	0.183
GPSDiffusion	3.613	39.843	0.991	0.415	12.603	60.312	0.931	0.197
Ours	**3.012**	**38.720**	**0.997**	**0.423**	**11.974**	**59.108**	**0.933**	**0.198**

Table 2: The results of different methods on DESOBA dataset. The best results are highlighted in boldface.**GSSIM, LSSIM** are indicators that can measure appearance realism, while **GRMSE, LRMSE** are indicators that can measure geometric accuracy.

Figure 8: Visualization results of shadow generation for IC-Light relighting images by our method.

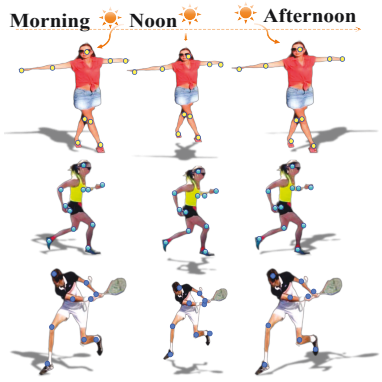

Figure 9: KPLM with Different Light.

eas. Therefore, we believe our method strikes a better balance between structural precision and visual realism.

Ablation

We conduct ablation studies on BOS test images to investigate the impact of each module using four metrics: GR, LR, GB, and LB. Firstly, we trained the base model of ControlNet, which had no geometry before, as shown in the first row of Table 4. Secondly, we injected STA into ControlNet

Prompt	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Different Scene	IC-Light	27.961	0.166	0.612
	Ours	**28.140**	**0.235**	**0.590**
Different Light	IC-Light	27.869	0.182	0.582
	Ours	**27.972**	**0.207**	**0.563**
Different Season	IC-Light	27.759	0.200	0.584
	Ours	**27.767**	**0.249**	**0.558**

Table 3: Different prompts' comparison on DESOBAv2.

to explore the effect. The result in the second line confirmed that the shadow triangle algorithm is useful. Compare the third line and the first line to verify the validity of KPLM. As mentioned in Line 5, our mature model achieves the best performance. To further demonstrate our effectiveness, we evaluated the metrics on the DESOBAv2 test set.

Row	Base	KPLM	STA	GR ↓	LR ↓	GB ↓	LB ↓
1	+	−	−	14.317	61.735	0.214	0.441
2	+	−	+	13.172	55.384	0.201	0.397
3	+	+	−	13.213	55.744	0.207	0.391
4	+	+	+	12.620	54.243	0.198	0.382

Table 4: **Ablation studies of our method on BOS test images from DESOBAv2 dataset.** "Base" means ControlNet base model.

Conclusion

In this paper, our framework for generating shadows after IC-Light relighting is presented. The Keypoints Linear Model (KPLM) and Shadow Triangle Algorithm (STA) are combined to identify the spatial coordinates of keypoints that correspond to shadows induced by the human body. Both modules obtain all necessary geometric information during preprocessing and inject it directly into the control encoder. Then performed using a diffusion-based model to generate shadow. In future work, we will explore more lightweight diffusion models, aiming to maintain generation quality while significantly reducing model complexity. Additionally, we will investigate further structural refinements to enhance both the efficiency and fidelity of the shadow.

Acknowledgements

The work was supported by Jilin Provincial Key Research and Development Program (NO. 20220203035SF).

References

Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; and Sheikh, Y. A. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, 2672–2680.

Google. 2020. MediaPipe: A Framework for Building Perception Pipelines. https://mediapipe.dev. Accessed: 2025-07-11.

Hong, Y.; Niu, L.; and Zhang, J. 2022. Shadow Generation for Composite Image in Real-world Scenes. *AAAI*.

Hu, X.; Jiang, Y.; Fu, C.-W.; and Heng, P.-A. 2019. Mask-ShadowGAN: Learning to Remove Shadows from Unpaired Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Kadhim, N.; and Mourshed, M. 2017. A shadow-overlapping algorithm for estimating building heights from VHR satellite images. *IEEE Geoscience and remote sensing letters*, 15(1): 8–12.

Liasis, G.; and Stavrou, S. 2016. Satellite images analysis for shadow detection and building height estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 119: 437–450.

Liu, D.; Long, C.; Zhang, H.; Yu, H.; Dong, X.; and Xiao, C. 2020. ARShadowGAN: Shadow Generative Adversarial Network for Augmented Reality in Single Light Scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Liu, D.; Qiu, Q.; and Shen, Y. 2022. Pose2Light: Learning to Estimate Illumination from Human Pose. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4): 2135–2148.

Liu, Q.; You, J.; Wang, J.; Tao, X.; Zhang, B.; and Niu, L. 2024. Shadow generation for composite image using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8121–8130.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Niu, L.; Cong, W.; Liu, L.; Hong, Y.; Zhang, B.; Liang, J.; and Zhang, L. 2021. Making images real again: A comprehensive survey on deep image composition. *arXiv preprint arXiv:2106.14490*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695. ArXiv preprint arXiv:2112.10752 [cs.CV].

Sheng, Y.; Liu, Y.; Zhang, J.; Yin, W.; Oztireli, A. C.; Zhang, H.; Lin, Z.; Shechtman, E.; and Benes, B. 2022. Controllable shadow generation using pixel height maps. In *European Conference on Computer Vision*, 240–256. Springer.

Simon, T.; Joo, H.; Matthews, I.; and Sheikh, Y. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Song, Z.; He, Z.; Li, X.; Ma, Q.; Ming, R.; Mao, Z.; Pei, H.; Peng, L.; Hu, J.; Yao, D.; et al. 2023. Synthetic datasets for autonomous driving: A survey. *IEEE Transactions on Intelligent Vehicles*, 9(1): 1847–1864.

Tao, X.; Cao, J.; Hong, Y.; and Niu, L. 2024. Shadow Generation with Decomposed Mask Prediction and Attentive Shadow Filling. arXiv:2306.17358.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional pose machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yin, X.; Li, Q.; Guo, Y.; Xie, H.; and Zhang, X. 2025. KPLM-STA: Physically-Accurate Shadow Synthesis for Human Relighting via Keypoint-Based Light Modeling. arXiv:2511.08169.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, L.; Rao, A.; and Agrawala, M. 2025. Scaling In-the-Wild Training for Diffusion-based Illumination Harmonization and Editing by Imposing Consistent Light Transport. In *The Thirteenth International Conference on Learning Representations*.

Zhang, S.; Liang, R.; and Wang, M. 2019. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5: 105–115.

Zhao, H.; Liu, Q.; Tao, X.; Niu, L.; and Zhai, G. 2025. Shadow Generation Using Diffusion Model with Geometry Prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7603–7612.