AIR-DR: Adaptive Image Retargeting with Instance Relocation and Dual-guidance Repainting

Zhitong Dong[1, 2], **Chao Li**[3], **Yongjian Deng**[4], **Hao Chen**[1, 2*]
[1]Southeast University
[2]Key Laboratory of New Generation Artificial Intelligence Technology
[3]Alibaba Group
[4]Beijing University of Technology
dongzt@seu.edu.cn, lllcho.lc@alibaba-inc.com, yjdeng@bjut.edu.cn, haochen303@seu.edu.cn

Abstract

Image retargeting aims to adjust the aspect ratio of images to accommodate various display devices. While existing methods consider both foreground semantics and background inpainting, their Seam-carving-based framework is inherently destructive, often compromising the structural integrity of foreground instances. Furthermore, conventional inpainting models struggle to achieve pixel-level accuracy with global-only guidance, leading to local inconsistencies and background distortions. To address these challenges, we reformulate image retargeting as a instance-level re-layout task. By **A**daptive **I**nstance **R**elocation and **D**ual-guidance **R**epainting (AIR-DR), our method preserves the structural integrity of the foreground and recovers the background with consistent details. Additionally, we introduce an adaptive retargeting decision that maintains robustness across challenging retargeting scenarios and any ratios. Extensive experiments on multiple public datasets across various aspect ratios demonstrate that our approach consistently outperforms existing methods in both objective metrics and subjective evaluations. Comprehensive ablation studies further validate the effectiveness of each component.

Code — https://github.com/Dongzt/AIR-DR

1 Introduction

As multi-device ecosystems continue to expand across PCs, smartphones, and tablets, the demand for visual content that can seamlessly adapt to a wide range of aspect ratios has grown significantly. Image retargeting (Vaquero et al. 2010; Fan et al. 2024; Rubinstein et al. 2010) addresses this challenge by adjusting image aspect ratios while preserving key semantics and perceptual quality. Despite its practical significance, this task remains relatively underexplored.

The primary challenges in image retargeting lie in: (1) preserving the original pixel-level content, (2) preventing deformation and distortion of salient objects, and (3) maintaining overall visual aesthetics. The traditional approaches include Scaling and Cropping (Figure 1 (b)), offer simple solutions but suffer from inherent limitations. Scaling (Andreadis and Amanatiadis 2005) retains all image content but often introduces noticeable geometric distortions. Cropping (Zhang et al. 2005) preserves semantic integrity within certain regions, but offen discards important visual information, leading to semantic loss. To address these limitations, Seam-carving-based techniques (Avidan and Shamir 2023; Rubinstein, Shamir, and Avidan 2008) compute an energy map over the image and iteratively remove seams with the lowest energy, thereby preserving visually important content. However, due to the absence of explicit semantic guidance, these approaches often fail to retain critical semantic structures, especially in complex scenes.

With the advancement of deep learning, several approaches (Wu et al. 2019; Liu et al. 2018) have integrated deep semantic features to guide pixel deletion or retention within traditional pixel-shifting frameworks. While these methods improve semantic awareness, they primarily focus on foreground regions, often resulting in fragmented or inconsistent backgrounds and compromised visual aesthetics. Alternatively, some works (Shocher et al. 2018; Mei et al. 2021; Cho et al. 2017) tackle the retargeting problem from a generative perspective. However, global image generation frequently struggles to preserve struc-

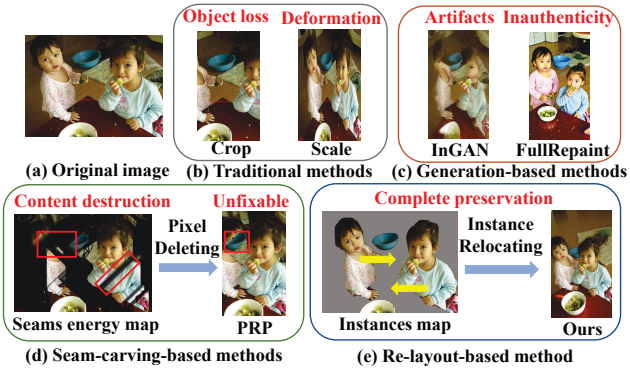

Figure 1: Traditional methods (b) lack generalization, generative approaches (c) suffer from low fidelity, and pixel-deletion methods (d) cause irreversible damage. Re-layout-based method (e) offers both robustness and realism while preserving instance integrity.

*Corresponding author.

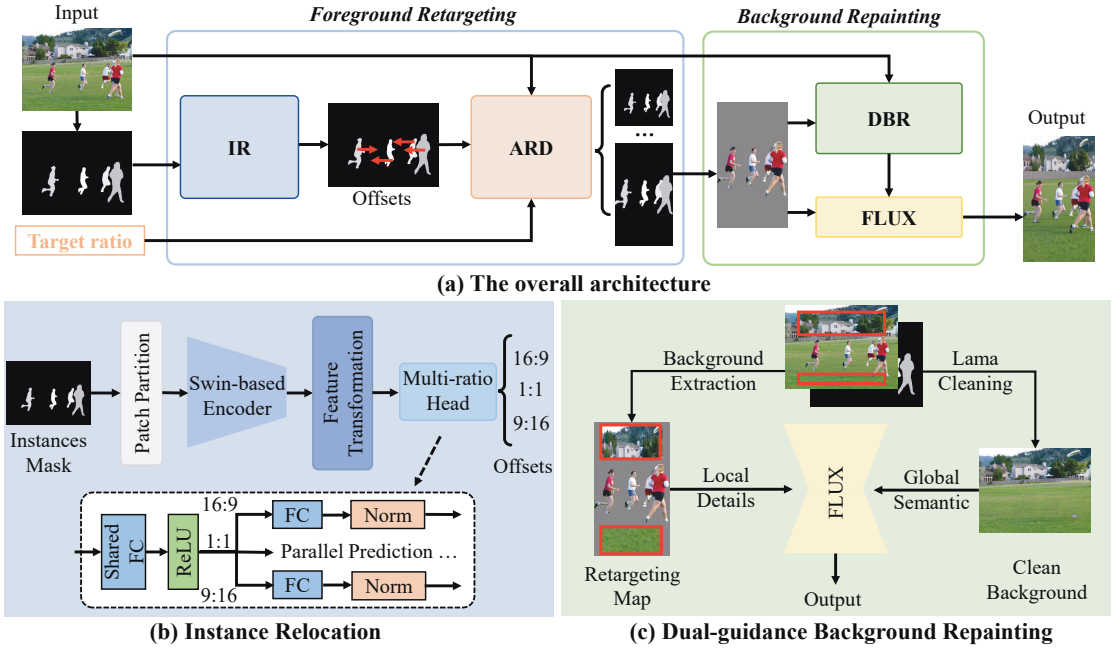

Figure 2: The overall architecture of our proposed AIR-DR. The input consists of an RGB image and a target ratio. Instance segmentation is first applied to obtain instance masks, based on which the instance relocation module predicts the retargeting offsets. Then adaptive retargeting decision generalizes the retargeting to any ratios. Finally, Global and local background guidance derived from the original image constrain the FLUX model to achieve pixel-wise consistent background.

tural realism and content fidelity, as illustrate in Figure 1 (c).

Although the current state-of-the-art method PRP (Shen et al. 2024), which operates within the Seam-carving framework and performs retargeting through content-aware selective pruning and repainting, considers both foreground semantic constraints and background inpainting, it still suffers from two critical limitations: (1) **Destructive nature of seam-carving.** Despite the incorporation of saliency constraints, the underlying energy function still relies heavily on low-level gradient information. As shown in Figure 1 (d), the energy flow is misdirected from the human face toward the visually prominent blue basin, leading to structural distortion of the instance. (2) **Limited guidance from diffusion models for local editing.** Diffusion-based repainting model struggles to leverage global semantics for precise local inpainting along seams, often causing color inconsistencies and structural artifacts in background regions.

In summary, most existing approaches formulate image retargeting as a **pixel-level prune task**, where pixels deemed less important are removed based on saliency or semantic relevance. However, in complex or ambiguous scenes, such strategies often remove pixels within important foreground instances, making it difficult to accurately reconstruct their structure. Consequently, essential semantic information may be irreversibly lost.

In this work, we fundamentally address this issue by reformulating image retargeting as a **instance-level re-layout task**, as illustrated in Figure 1 (e). By adaptively relocating instance positions for retargeting, our method preserves the integrity of salient objects without the need to remove any critical pixels. We propose **A**daptive **I**nstance **R**elocation and **D**ual-guidance **R**epainting (AIR-DR), a two-stage framework that first retargets instances to preserve their relative spatial relationships, followed by constrained background repainting to generate the pleasing and semantically consistent retargeted image.

To this end, we predict the displacement of each instance during the retargeting process, as shown in Figure 2 (b). Instance Relocation (IR) module first extracts complete instances from the original image and then models their spatial relationships to predict horizontal displacement for each instance. Combined with the Adaptive Retargeting Decision (ARD), our approach supports flexible retargeting to arbitrary aspect ratios while maintaining robustness across diverse scene configurations. To synthesize natural and visually coherent backgrounds, we further propose a hierarchical Dual-guidance Background Repainting (DBR) module, as shown in Figure 2 (c). On the one hand, DBR removes the foreground and performs repainting to obtain a complete background for global semantic guidance. On the other hand, DBR applies local background constraints to generate pixel-level realistic background details.

Our contributions can be summarized as follows:

- We introduce a novel retargeting paradigm that reformulates the task as instance-level layout refactoring, fundamentally differing from conventional approaches.
- We propose a dual-guidance background repainting strategy that jointly enhances global semantic consistency and local visual realism.

- Extensive experiments across diverse aspect ratios demonstrate that our method outperforms state-of-the-art approaches in both objective metrics and human perceptual evaluations. Comprehensive ablation studies further validate the effectiveness of each proposed component.

2 Related Work

2.1 Image Retargeting

Existing image retargeting methods primarily pursue two objectives: preserving essential content and avoiding visual artifacts. Early approaches often struggle to achieve a satisfactory balance between these goals. For instance, Scaling (Andreadis and Amanatiadis 2005) seeks to retain global image structure by uniformly removing pixels, but fails under significant aspect ratio changes, often causing severe deformation of salient objects. Cropping (Zhang et al. 2005; Santella et al. 2006) preserve structural integrity by selecting an optimal window of the target size, but inevitably discard important content outside the selected region. Seam-carving (Avidan and Shamir 2023) removes low-energy seams to better balance content preservation and visual quality. Several improvements (Dong et al. 2009; Zhou, Chen, and Li 2020; Kajiura et al. 2020) based on Seam-carving enhance generalization across different scenarios. However, due to the absence of explicit semantic understanding, these methods often produce distortions in foreground regions, especially in complex scenes with cluttered backgrounds.

The rise of deep learning (LeCun et al. 1998) has introduced semantic information to image retargeting. DeepIR (Lin, Zhou, and Chen 2019) adopts pretrained VGG (Simonyan and Zisserman 2014) to explicitly extract semantic information and retargets the image from a coarse semantic space to fine pixel space. SmartScale (Dickman et al. 2023) utilizes existing object detection model to assist Seam-carving. However, neglecting the background can lead to discontinuities in background pixels, thereby affecting the aesthetic appeal of the image. For aesthetics and local smoothness, some methods (Mei et al. 2021; Cho et al. 2017) adopt Generative Adversarial Networks (GANs) (Goodfellow et al. 2020) to generate the retargeting results. InGAN (Shocher et al. 2018) and SinGAN (Shaham, Dekel, and Michaeli 2019) divide the image into patches and learn the internal distribution of patches, destroying the overall semantics. MCGAN (Dy et al. 2023) introduces mask to highlight importance areas. However, due to the limitations of implicit semantic expression, these methods preserve the global semantics but destroy the details, resulting in inconsistent appearance with the original image.

PRP (Shen et al. 2024) attempt to balance semantic preservation and artifact avoidance by combining selective pruning with globally guided inpainting. However, their destructive Seam-carving-based pruning often breaks the structural integrity of instances, making it difficult to fix during subsequent guided reconstruction. In contrast, our method does not require pixel removal. It preserves the structural integrity of instances while maintaining realistic and coherent backgrounds through locally guided background preservation.

2.2 Conditional Image Inpainting

Conditional Image Inpainting (Elharrouss et al. 2020) and Image Retargeting (Fan et al. 2024) are both image generation processes that follow certain constraints. Diffusion models (Rombach et al. 2022; Podell et al. 2023) generate realistic and controllable images by progressively denoising in the latent space. Some methods (Ye et al. 2023; Zhang, Rao, and Agrawala 2023) based on diffusion models have been widely applied to controllable image inpainting task.

However, it is difficult to directly transfer these methods to image retargeting task, because they can usually only achieve semantic consistency and it is difficult to achieve pixel-level accurate repainting. Therefore, additional guiding conditions are necessary for constraints.

To enhance the controllability of inpainting, PRP (Shen et al. 2024) utilizes IP-Adapter to guide the repainting process. The IP-Adapter (Ye et al. 2023) introduces image prompts via a decoupled cross-attention adapter branch for conditional control. However, applying global guidance within the local inpainting region often results in inconsistency and distortion. To address this issue, we design a local background guidance module, achieving pixel-level accuracy in repainting.

3 Methodology

3.1 Overview

The overall architecture of AIR-DR is illustrated in Figure 2 (a). It consists of three main components: Instance Relocation (IR), Adaptive Retargeting Decision (ARD), and Dual-guidance Background Repainting (DBR).

3.2 Instance Relocation

Previous retargeting methods are predominantly driven by energy-based pixel removal. Even with the incorporation of saliency constraints, the energy computation remains heavily reliant on low-level gradient cues, limiting the model's ability to capture high-level semantic structures.

In contrast to traditional approaches, our method directly predicts instance-wise displacements without removing pixels. These displacements serve as explicit and explainable alternatives to conventional energy map computations. As illustrated in the Figure 2 (b), the input to the Instance Relocation module is a grayscale mask $M \in \mathbb{R}^{H \times W}$ generated by a pre-trained instance segmentation model (Liu et al. 2021). This mask encodes the size, shape, and spatial location of each instance, while effectively filtering out background regions. We adopt a lightweight Swin Transformer (Liu et al. 2021), denoted as $\mathcal{F}(\cdot)$, as our vision backbone. Its hierarchical window-based attention mechanism enables the model to simultaneously capture global instance distributions and fine-grained local features, while maintaining robustness in complex visual scenes. Formally, this process can be expressed as:

$$F = \mathcal{F}(M), \quad F \in \mathbb{R}^{C \times H' \times W'}, \tag{1}$$

where F denotes the feature map output by the backbone network, with C representing the number of channels and

$H' \times W'$ the feature resolution).

To enable retargeting across multiple aspect ratios, we design a parallel prediction architecture. The features extracted by the backbone are shared across all branches, while three independent displacement heads decode instance-wise offsets corresponding to three base aspect ratios: 16:9, 9:16, 1:1. Based on these ratios, our Adaptive Retargeting Decision module in Section 3.3 can extend the results to any ratio. During implementation, we first reduce the dimensionality of the backbone features and transform them into a latent representation Z via a projection layer $\phi(\cdot)$. This process can be written as:

$$Z = \text{Flatten}(\phi(F)), \quad Z \in \mathbb{R}^{C'}. \tag{2}$$

Each prediction branch $\psi_r(\cdot)$ then independently estimates the displacement δ_r and normalizes the result to the range $[-1, 1]$:

$$\delta_r = \psi_r(Z), \quad r \in \{16{:}9,\ 9{:}16,\ 1{:}1\}. \tag{3}$$

$$D = \tanh\left(\begin{bmatrix}\delta_{16:9}\\ \delta_{9:16}\\ \delta_{1:1}\end{bmatrix}\right), \quad D \in [-1, 1]^{3\times N}. \tag{4}$$

The final output is a $3 \times N$ instance-level horizontal displacement vector, where N denotes the number of detected instances, and δ_r represents the relative displacement required for each instance under the target aspect ratio r. During training, we compute the Mean Squared Error (MSE) loss only over valid instances to ensure accurate supervision.

3.3 Adaptive Retargeting Decision

To achieve retargeting at any ratio, we select a base ratio that is closest to the target ratio. This base ratio captures the primary directional trend of the retargeting transformation and provides guidance on how instances should be spatially shifted. Subsequently, instances are relocationed on a canvas corresponding to the target aspect ratio, according to the predicted displacements.

In relatively simple retargeting scenarios, basic methods such as cropping may be sufficient. However, in more complex cases, even state-of-the-art approaches often struggle to preserve structural integrity. As shown in the Figure 3, the ARD module adaptively selects appropriate retargeting strategies, enhancing robustness across a wide range of image layouts and aspect ratio variations.

Specifically, we calculate the difference between the original and target aspect ratios as: $\Delta r = |r_{\text{ori}} - r_{\text{target}}|$. If Δr is below a threshold ε, set to 0.3 in our experiment, instance relocation is deemed unnecessary. When sufficient background context is available, the image is directly cropped to the target ratio for optimal results. Otherwise, we apply the background repainting strategy described in Section 3.4 to fill in the cropped margins, thereby achieving the desired aspect ratio while preserving content coherence. This can be formalized as:

$$ARD = \begin{cases}\text{Cropping} & \text{if } S_{\text{space}} \le \tau \\ \text{Background Extension} & \text{if } S_{\text{space}} > \tau\end{cases}, \tag{5}$$

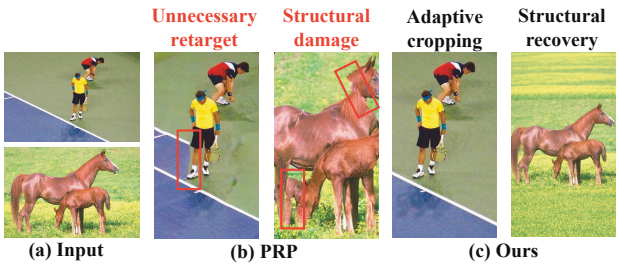

Figure 3: The adaptive retargeting strategy enhances robustness from multiple perspectives.

where τ denotes the smallest region that satisfies the target aspect ratio while encompassing all instances. When the aspect ratio difference Δr exceeds the threshold ε, we further evaluate whether the semantic structure of the image has been damaged. This is done by computing the instance retention rate, defined as:

$$\overline{R} = \sum_{i=1}^{n}\left(w_i \times \frac{A_i^{\text{cur}}}{A_i^{\text{ori}}}\right). \tag{6}$$

For each instance i, A_i^{ori} is its original area, A_i^{cur} is the area preserved in the retargeted image, and w_i the instance weight. The weight is calculated as:

$$w_i = \frac{A_i^{\text{bbox}}}{\sum_{j=1}^{n} A_j^{\text{bbox}}}, \tag{7}$$

where A_i^{bbox} denotes the area of the bounding box of instance i, reflecting its relative importance among all instances.

If the instance retention rate $\overline{R}$ is less than the threshold γ, set to 0.95 in our experiment, we expand the image size to restore the semantic integrity of the scene. This enlarged image then serves as the canvas for the subsequent background repainting stage.

3.4 Dual-guidance Background Repainting

As shown in the Figure 2 (c), after obtaining the retargeting map from Section 3.3, we perform background repainting guided by both global and local contexts. For global guidance, in order to remove foreground objects and reconstruct the complete background, we utilize the object removal tool LaMa (Suvorov et al. 2022) to erase the masked regions in the original image. The clean background helps eliminate interference from foreground instances during background repainting process.

To further enhance consistency with the original background, we introduce the local background guidance inpainting. Specifically, we analyze the surrounding regions (top, bottom, left, and right) of each foreground instances in both the original image and the retargeting map to determine whether sufficient background context is available. The spatial threshold η is defined as:

$$\eta = \alpha \cdot \max(\text{height}, \text{weight}), \tag{8}$$

where α is a proportional coefficient which is set to 0.2. If sufficient surrounding context is available, we extract the corresponding local background regions from the original image and paste them into the appropriate locations in the retargeting map.

Finally, by combining global and local background guidance, we employ the FLUX inpainting model (Labs 2024) to synthesize pixel-level accurate and visually pleasing background content. Pretrained image generation models possess strong aesthetic priors and a certain level of world knowledge, which help achieve harmonious integration between foreground objects and background. This is particularly beneficial when there are interactive relationships between them, as the diffusion process naturally refines the composition during generation. The generation formula is as follows:

$$y = M_f \odot y^k + M_b \odot y^k + (1 - M_f - M_b) \odot y^{unk}, \quad (9)$$

where $\odot$ denotes element-wise multiplication, M_f and M_b denote the foreground and background masks, respectively. The known area y^k and the unknown area y^{unk} are integrated to generate the final image y.

4 Experiments

4.1 Implementation Details

Model Setting For the Instance Relocation model detailed in Section 3.2, we adopt a lightweight Swin Transformer for feature encoding, utilizing a 24-dimensional embedding space and a simplified layer depth. This design maintains robust feature extraction capabilities while significantly reducing the model's parameter count to just 0.98M, which is 27 times smaller than Swin-T. The network takes a 640×640 single-channel grayscale image as input, producing feature map F with a shape of $192 \times 20 \times 20$, as shown in Equation (1). These features are processed by a convolutional layer for dimensionality reduction and flattened into a one-dimensional vector Z with a length of 25600, as shown in Equation (2). Finally, each prediction head outputs instance displacement vector δ_r, with a shape of 1×8 through parallel fully connected layers.We limit the number of valid instances to be no more than the maximum number of instances in the dataset, i.e., $N \leq 8$.

In Section 3.4, we integrate FLUX.1-Fill and FLUX.1-Redux (Labs 2024) to perform repainting process. FLUX.1-Fill is a text-guided inpainting model, while FLUX.1-Redux provides guidance constraints based on background semantics. By combining both, our framework is capable of handling high-resolution retargeting tasks with aesthetic demands, such as HD wallpaper generation.

Training and Inference The weakly supervised image retargeting dataset RetarSet (SEU-WSY 2025) provides annotations for the instance distribution after retargeting. We calculate the positional difference between these annotations and the original instance locations to derive the ground truth displacement labels for training.

Our method is implemented in PyTorch and trained on a NVIDIA L20 GPU. We train the network using the AdamW (Loshchilov and Hutter 2017) optimizer with a learning rate of 1×10^{-4}. The model is trained for 300 epochs with a batch size of 16. During both training and inference, we maintain a consistent instance ordering strategy to ensure a one-to-one correspondence between each instance and its displacement.

4.2 Experiment Design

Objective Evaluation Metric Traditional metrics (Manjunath et al. 2001; Liu et al. 2011; Hsu et al. 2014) for image retargeting often diverge from human perception, as they typically fail to distinguish between important content and background regions. To intuitively assess the effectiveness of retargeting methods, we use the *Saliency Discard Ratio* (SDR) (Shen et al. 2024) as metric to evaluate semantic preservation. The SDR is computed as follows:

$$SDR = \frac{\left| W_{ori}^{union} - W_{out}^{union} \right|}{W_{ori}^{union}}, \quad (10)$$

where W_{ori}^{union} and W_{out}^{union} denote the union of salient instance widths in the original and retargeting images, respectively. We use the mask width obtained from the pre-trained instance segmentation model (Liu et al. 2021) to calculate the value of SDR. To ensure reproducibility, we fix the random seed for all generation-based methods. Hyperparameters are set according to values recommended in prior work or manually tuned. The reported results are the average of three repeated runs.

User Study Metric Considering the subjective nature of retargeting results, we adopt manual scoring as an additional evaluation method. We invite 10 volunteers to score the results from 0 to 3 in four aspects: Content completeness, Deformation, Local smoothness, and Aesthetics. **Content completeness** assesses whether the visual content suffers from incompleteness or inconsistency; **Deformation** examines the degree of distortion within crucial regions; **Local smoothness** evaluates the continuity of local areas in the image; and **Aesthetics** measures the overall harmony and visual appeal of the composition. Regarding the score, '0' indicates that the retargeting image severely compromises the attributes of the original image, while '3' denotes that the retargeting image retains the original attributes fine. As illustrated in the Figure 4, the original image is displayed on the left, while the anonymized results of the eight retargeting methods under the challenging 16:9 aspect ratio are shown on the right. See the Supplementary Materials for the investigation process.

Generalization Validation We evaluate our method on the widely used RetargetMe (Rubinstein et al. 2010) dataset (80 images, 41% multi-object scenes) and the more challenging RetarSet (SEU-WSY 2025) dataset (1000 images, 87% multi-object scenes). In addition to the common aspect ratios, we also tested untrained ratios of 4:3 and 3:4 to demonstrate the generalization capability of our approach. Additional results under more extreme aspect ratios (e.g., 21:9 and 9:21) are provided in the Supplementary Materials.

4.3 Comparative Evaluation

We compare our proposed model with six popular image retargeting methods, namely, Scaling, Cropping, Seam-

Datasets	RetarSet					RetargetMe				
ratios	16:9	1:1	9:16	4:3	3:4	16:9	1:1	9:16	4:3	3:4
Scale (Andreadis and Amanatiadis 2005)	0.526	0.237	0.315	0.403	0.017	0.611	0.378	0.292	0.500	0.154
Crop (Zhang et al. 2005)	0.429	0.150	0.065	0.289	0.017	0.529	0.339	0.129	0.396	0.099
SC (Avidan and Shamir 2023)	0.449	0.180	0.269	0.329	0.027	0.533	0.303	0.346	0.427	0.098
InGAN (Shocher et al. 2018)	0.524	0.239	0.294	0.394	0.050	0.582	0.415	0.273	0.500	0.172
PRP (Shen et al. 2024)	0.354	0.124	0.103	0.227	0.040	0.448	0.264	0.216	0.320	0.097
FR (Labs 2024)	0.308	0.330	0.405	0.372	0.205	0.299	0.326	0.450	0.370	0.242
PRP + FLUX (Shen et al. 2024; Labs 2024)	0.346	0.118	0.174	0.223	0.049	0.418	0.239	0.254	0.284	0.120
Ours	**0.113**	**0.073**	**0.031**	**0.100**	**0.013**	**0.114**	**0.117**	**0.056**	**0.138**	**0.071**

Table 1: Comparison of SDR values with other methods on two datasets with different aspect ratios. Lower values indicate better semantic completeness. The best results are highlighted in **bold**.

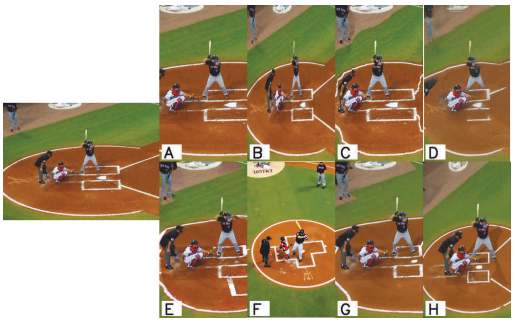

Figure 4: Example of User study.

Settings	CCS ↑	DS ↑	LS ↑	AS ↑	**AVS ↑**
Scale	2.740	1.644	2.158	1.474	2.004
Crop	1.556	2.786	2.726	2.280	2.337
SC	2.282	1.260	1.306	1.078	1.482
InGAN	2.022	0.998	1.000	0.818	1.210
PRP	2.184	2.464	2.416	2.138	2.301
FR	2.192	2.424	2.604	2.448	2.417
PRP + FLUX	2.212	2.470	2.414	2.206	2.326
Ours	**2.864**	**2.884**	**2.920**	**2.792**	**2.865**

Table 2: Subjective comparison with other methods in aspect ratio 16:9. 'CCS', 'DS', 'LS', 'AS', and 'AVS' denote *Content Completeness Score*, *Deformation Score*, *Local Smoothness Score*, *Aesthetic Score*, and *Average Score*, respectively. ↑ indicates that larger are better. The best results are highlighted in **bold**.

carving (SC) (Avidan and Shamir 2023), InGAN (Shocher et al. 2018), PRP (Shen et al. 2024) and FLUX-based full repainting (FR) (Labs 2024). To ensure a fair comparison, we additionally replace PRP's original generation model with the same FLUX model used in our method. We quantitatively evaluate the performance of our method using the objective metric and subjective assessments.

Table 1 presents the performance of various methods based on the objective metric. The difficulty of image retargeting varies across aspect ratios, in the 3:4 setting, simple strategies like scaling and cropping are often sufficient to achieve strong results. Our method also achieves the best performance in this case by adaptively adjusting the retargeting strategy. When facing the most challenging 16:9 setting, both traditional methods and previous approaches based on seam-carving or generative models struggle. In contrast, our method significantly reduces the distortion of salient regions, primarily due to the instance-level retargeting strategy. By predicting the spatial distribution of instances and relocating them accordingly, our model better preserves the structural integrity within salient objects. Furthermore, thanks to its adaptive scalability, our method also achieves the best performance on unseen aspect ratios such as 3:4 and 4:3, demonstrating strong generalization ability.

Table 2 presents the performance of various methods based on user subjective evaluations. The Scale method achieves a high Content Completeness score, as it avoids removing any regions. The Crop method obtains higher scores in Deformation and Local Smoothness due to its preservation of geometric integrity without introducing distortions. Methods based on generative models generally achieve higher Aesthetic scores, benefiting from their strong aesthetic priors. Our method maintains content integrity while reducing distortion and local artifacts, achieving the highest scores in Aesthetics and overall average. Although FR ranks second on average, its lack of fine-grained guidance makes it difficult to maintain identity consistency during generation. As a result, facial features and other fine details often deviate from the original image, diminishing its practical value.

To qualitatively evaluate our method, we visually compare it with the above approaches across various target ratios. Figures 5 and 6 illustrate the retargeting results under two extreme target ratios. We can clearly observe that traditional and pixel-level methods often lead to the destruction of key content, such as the deformation and loss of the bus in Figure 6. While generation-based methods tend to introduce visual distortion, such as the inconsistency of facial features produced by FR in Figure 5. In contrast, our method retargets the key content at the instance level to minimize destruction. The dual-guidance mechanism enhances the fidelity of the repainting regions, resulting in pixel-level accurate and visually pleasing outcomes. More experiments see Supplementary Materials.

Figure 5: Visual comparison to other methods on ratio 9:16.

Figure 6: Visual comparison to other methods on ratio 16:9.

4.4 Ablation Study

In this section, we conduct comprehensive ablation studies on five commonly used aspect ratios to validate the effectiveness of each component in our proposed model.

RetarSet	16:9	1:1	9:16	4:3	3:4
Pixel Deletion	0.354	0.124	0.103	0.227	0.040
IR	0.198	0.116	0.099	0.156	0.038
IR + ARD	0.131	0.092	0.032	0.118	0.013
IR + ARD + DBR	**0.113**	**0.073**	**0.031**	**0.100**	**0.013**
RetargetMe	16:9	1:1	9:16	4:3	3:4
Pixel Deletion	0.448	0.264	0.216	0.320	0.097
IR	0.244	0.143	0.175	0.149	0.096
IR + ARD	0.143	0.124	0.072	0.147	0.073
IR + ARD + DBR	**0.114**	**0.117**	**0.056**	**0.138**	**0.071**

Table 3: Ablation results of SDR across five aspect ratios on two datasets. The best results are highlighted in **bold**.

Effectiveness of Instance Relocation As shown in Table 3, compared to pixel deletion method (represented by PRP), Instance Relocation (IR) reduces the SDR by relocating salient objects at the instance level. This advantage is particularly evident under the 16:9 aspect ratio, where pixel deletion methods tend to prune the main subject, resulting in the loss of critical content. As illustrated in Figure 7, IR preserves the car content more effectively than pixel deletion methods. This is attributed to IR's better preservation of the structural integrity of salient regions.

Effectiveness of Adaptive Retargeting Decision As shown in Table 3, the comparison between IR and 'IR + ARD' demonstrates that the Adaptive Relocation Decision (ARD) module brings significant improvements under extreme aspect ratios (e.g., 16:9 and 9:16). As illustrated in Figure 7, ARD identifies the overlap among the three individuals and adaptively expands the image to restore a reasonable layout. This highlights the module's effectiveness in enhancing the model's robustness across diverse scenarios and aspect ratios.

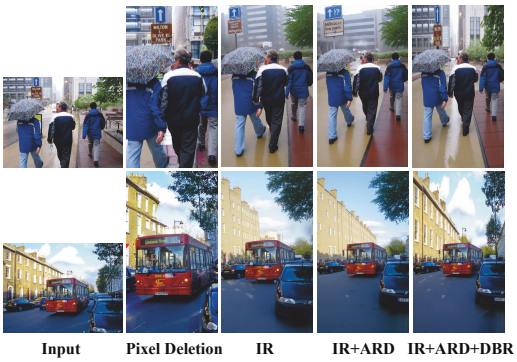

Figure 7: Visualization to demonstrate the effectiveness of each component in our method.

Effectiveness of Dual-guidance Background Repainting As shown in Table 3, introducing Dual-guidance constraints into the background repainting further reduces the SDR. This is because global semantic guidance alone may lack fine-grained details and can even introduce unwanted artifacts. In contrast, the local background constraint better preserves background details, resulting in more visually appealing and high-fidelity outputs. As illustrated in Figure 7, although 'IR + ARD' can generate natural background, difference still exists compared to the original image. Through an additional local background guidance module, DBR generates street and sky that are almost identical to those in the original image.

4.5 Discussions

Recently, several advanced open-source image editing models, such as FLUX.1-Kontext (Labs et al. 2025), BAGEL (Deng et al. 2025), and commercial systems like GPT-4o (OpenAI 2025), have emerged in the field of image generation. These models aim to unify understanding and generation, enabling image manipulation through textual instructions. While they often support specific output aspect ratios, the weak constraint of natural language makes it difficult to consistently satisfy the strict requirements of image retargeting, such as enforcing precise aspect ratios and preserving high visual fidelity. As a result, despite their impressive capabilities, these models are not yet suitable for image retargeting tasks.

5 Conclusion

Our paper proposes a novel image retargeting approach AIR-DR. It preserves the structural integrity of salient objects by retargeting instances rather than removing pixels. By integrating adaptive decision and dual-guidance background inpainting, our method achieves both robustness and visual fidelity. Extensive experiments demonstrate the effectiveness of our design and the superiority of the method.

Acknowledgments

The work is jointly supported by the National Natural Science Foundation of China (NSFC) under Grant 62261160576 and Grant 62203024, the Beijing Natural Science Foundation (4252026), the Research and Development Program of Beijing Municipal Education Commission (KM202310005027), and the Zhishan Young Scholar Program of SEU, Fundamental Research Funds for the Central Universities of China. This research work is also supported by the Big Data Computing Center of Southeast University.

References

Andreadis, I.; and Amanatiadis, A. 2005. Digital Image Scaling. In *2005 IEEE Instrumentationand Measurement Technology Conference Proceedings*.

Avidan, S.; and Shamir, A. 2023. Seam carving for content-aware image resizing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*.

Cho, D.; Park, J.; Oh, T.-H.; Tai, Y.-W.; and So Kweon, I. 2017. Weakly-and self-supervised learning for content-aware deep image retargeting. In *ICCV*.

Deng, C.; Zhu, D.; Li, K.; Gou, C.; Li, F.; Wang, Z.; Zhong, S.; Yu, W.; Nie, X.; Song, Z.; Shi, G.; and Fan, H. 2025. Emerging Properties in Unified Multimodal Pretraining. arXiv:2505.14683.

Dickman, E.; Diefenbach, P.; Burlick, M.; and Stockton, M. 2023. Smart Scaling: A Hybrid Deep-Learning Approach to Content-Aware Image Retargeting. In *ACMSIGGRAPH*.

Dong, W.; Zhou, N.; Paul, J.-C.; and Zhang, X. 2009. Optimized image resizing using seam carving and scaling. *ACM Transactions on Graphics (TOG)*.

Dy, J. B.; Virtusio, J. J.; Tan, D. S.; Lin, Y.-X.; Ilao, J.; Chen, Y.-Y.; and Hua, K.-L. 2023. MCGAN: mask controlled generative adversarial network for image retargeting. *Neural. Comput. Appl*.

Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S.; and Akbari, Y. 2020. Image inpainting: A review. *Neural Processing Letters*.

Fan, X.; Zhang, Z.; Sun, L.; Xiao, B.; and Durrani, T. S. 2024. A comprehensive review of image retargeting. *Neurocomputing*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*.

Hsu, C.-C.; Lin, C.-W.; Fang, Y.; and Lin, W. 2014. Objective quality assessment for image retargeting based on perceptual geometric distortion and information loss. *IEEE J. Sel. Top. Signal Process*.

Kajiura, N.; Kosugi, S.; Wang, X.; and Yamasaki, T. 2020. Self-play reinforcement learning for fast image retargeting. In *MM*.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE*.

Lin, J.; Zhou, T.; and Chen, Z. 2019. DeepIR: A deep semantics driven framework for image retargeting. In *ICMEW*.

Liu, S.; Wei, Z.; Sun, Y.; Ou, X.; Lin, J.; Liu, B.; and Yang, M.-H. 2018. Composing semantic collage for image retargeting. *IEEE Transactions on Image Processing*.

Liu, Y.-J.; Luo, X.; Xuan, Y.-M.; Chen, W.-F.; and Fu, X.-L. 2011. Image retargeting quality assessment. In *Computer Graphics Forum*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.

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

Manjunath, B. S.; Ohm, J.-R.; Vasudevan, V. V.; and Yamada, A. 2001. Color and texture descriptors. *TCSVT*.

Mei, Y.; Guo, X.; Sun, D.; Pan, G.; and Zhang, J. 2021. Deep supervised image retargeting. In *ICME*.

OpenAI. 2025. Introducing 4o Image Generation.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Rubinstein, M.; Gutierrez, D.; Sorkine, O.; and Shamir, A. 2010. A comparative study of image retargeting. In *SIGGRAPH*.

Rubinstein, M.; Shamir, A.; and Avidan, S. 2008. Improved seam carving for video retargeting. *ACM transactions on graphics (TOG)*.

Santella, A.; Agrawala, M.; DeCarlo, D.; Salesin, D.; and Cohen, M. 2006. Gaze-based interaction for semi-automatic photo cropping. In *SIGCHI*.

SEU-WSY. 2025. Image Retargeting: A Dataset and Metrics. https://github.com/SEU-WSY/Image-Retargeting-ADataset.

Shaham, T. R.; Dekel, T.; and Michaeli, T. 2019. Singan: Learning a generative model from a single natural image. In *ICCV*.

Shen, F.; Li, C.; Geng, Y.; Deng, Y.; and Chen, H. 2024. Prune and Repaint: Content-Aware Image Retargeting for any Ratio. In *NeurIPS*.

Shocher, A.; Bagon, S.; Isola, P.; and Irani, M. 2018. Ingan: Capturing and remapping the" dna" of a natural image. *arXiv preprint arXiv:1812.00231*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *WCACV*.

Vaquero, D.; Turk, M.; Pulli, K.; Tico, M.; and Gelfand, N. 2010. A survey of image retargeting techniques. In *Applications of digital image processing*.

Wu, J.; Xie, R.; Song, L.; and Liu, B. 2019. Deep Feature Guided Image Retargeting. In *VCIP*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Zhang, M.; Zhang, L.; Sun, Y.; Feng, L.; and Ma, W. 2005. Auto cropping for digital photographs. In *ICME*.

Zhou, Y.; Chen, Z.; and Li, W. 2020. Weakly supervised reinforced multi-operator image retargeting. *IEEE Transactions on Circuits and Systems for Video Technology*.