IMAGE RETARGETING: A DATASET AND METRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image retargeting adaptively resizes images for different display formats while preserving important content and perceptual quality. However, its development and evaluation remain limited by the lack of a comprehensive dataset with annotations and task-specific metrics. To address this gap, we construct a benchmark containing 5,000 input images, 20,000 instance-level annotations, and 20,000 pixel-level reference outputs. The dataset covers diverse input aspect ratios, instance counts, foreground layouts, and retargeting-sensitive content. To enable comprehensive and diagnostic evaluation, we further introduce a protocol comprising four scores. Semantic Faithfulness (SMF) and Layout Faithfulness (LOF) use the instance-level annotations to evaluate foreground preservation and reference-layout fidelity, while Scene and Texture Consistency (STC) and Structural and Visual Fidelity (SVF) assess aspects of visual quality not captured by these annotations. To the best of our knowledge, this is the first comprehensive image-retargeting benchmark to jointly provide instance-level annotations, pixel-level reference outputs, and annotation-driven evaluation metrics. Experiments on seven representative methods demonstrate that the proposed metrics agree more strongly with human judgments than existing evaluation measures. Our dataset and code will be made publicly available to the community.

1 INTRODUCTION

Image retargeting adapts an image to a target resolution or aspect ratio while preserving its important content and perceptual quality. It is a core component of cross-device content adaptation, with applications ranging from digital-media production to camera–projector systems (Mademlis et al., 2016; Holliman et al., 2011). Over the past decades, a broad spectrum of techniques has been developed, including cropping, scaling, seam carving, warping, multi-operator optimization, and more recent generative approaches (Avidan & Shamir, 2023; Zhang et al., 2005; Rubinstein et al., 2008; Hashemzadeh et al., 2019; Shocher et al., 2019; Shen et al., 2024). These methods make different trade-offs when adapting content to a target canvas. Cropping may remove important boundary content, global scaling preserves all pixels but distorts object geometry, and content-aware or generative methods may alter spatial organization or introduce visual artifacts. **A diverse dataset and an evaluation system capable of comparing methods from multiple perspectives are therefore essential for method development and practical deployment.**

However, both aspects remain underexplored. Regarding data resources, as shown in Table 1, existing datasets, including RetargetMe (Rubinstein et al., 2010), CUHK (Ma et al., 2012), and NRID (Hsu et al., 2014), contain fewer than 100 input images and offer limited diversity in source aspect ratios and scene content. More importantly, none of them provides annotations that describe how salient content should be preserved under different target aspect ratios. This lack of annotations makes it difficult to quantitatively evaluate newly generated retargeting results.

To fill this data gap, we construct a benchmark containing 5,000 input images and four target aspect ratios (1:1, 3:4, 16:9, and 9:16). We increase the diversity of source aspect ratios, instance quantities, foreground arrangements, and retargeting-sensitive visual attributes. Although existing methods follow different technical routes, apply to different scenes, and may produce multiple acceptable results, high-quality retargeted images generally share two common properties. First, semantically important content, particularly salient foreground instances, should remain recognizable and sufficiently complete, as also reflected by the widespread use of saliency or importance maps in existing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

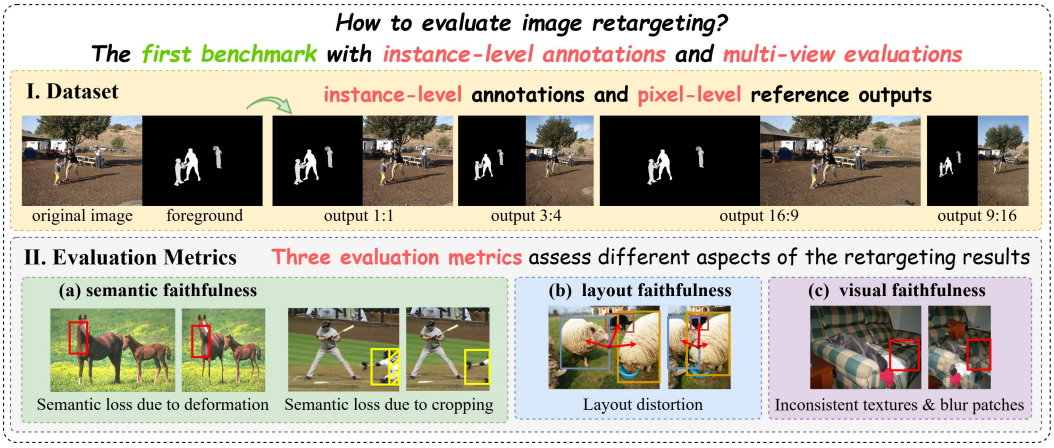


Figure 1: We introduce an annotated benchmark for image retargeting evaluation. It contains 5,000 input images, 20,000 instance-level annotations, and 20,000 pixel-level reference outputs under four target aspect ratios. The benchmark evaluates retargeted results along three complementary dimensions: semantic integrity, reference-layout fidelity, and visual plausibility.

methods (Zhang et al., 2005; Avidan & Shamir, 2023; Hashemzadeh et al., 2019). Second, the spatial organization of these instances should remain coherent after adaptation to the target canvas. In line with the core objective of image retargeting described above, for each input–target-ratio pair, we provide an instance-level annotation that specifies the salient foreground instances to be preserved and their spatial organization on the target canvas. This process yields 20,000 instance-level annotations for 4 aspect ratios. We additionally provide 20,000 pixel-level reference outputs, generated from the input images and their instance-level annotations and selected through manual review. Each provides one possible realization for the corresponding input and target ratio. These paired outputs offer a resource for future supervised retargeting studies. **Unlike existing small-scale datasets without reusable output annotations, our dataset covers diverse source aspect ratios, instance quantities, foreground arrangements, and retargeting-sensitive visual attributes, more importantly, provides explicit annotations for method comparison and quantitative evaluation.**

A comprehensive benchmark requires not only a diverse dataset with comprehensive annotations, but also an evaluation protocol for quantitative assessments. However, an effective and widely adopted evaluation protocol for this task is lacking. Existing studies frequently rely on subjective user evaluations (Apostolidis & Mezaris, 2021) or measures adapted from image quality assessment and related visual tasks (Fang et al., 2016; Givkashi et al., 2024; Cho et al., 2017). Subjective studies are costly to reproduce for every newly developed method, while existing objective measures typically assess only selected properties or aggregate heterogeneous degradations into a single quality prediction. As a result, they provide limited evidence about why a retargeted result fails—for example, whether it loses an important foreground instance, disrupts the organization of multiple instances, or introduces implausible textures and structural artifacts.

We therefore introduce an evaluation protocol that characterizes image-retargeting quality along three complementary dimensions: **semantic integrity**, **reference-layout fidelity**, and **visual plausibility**. Semantic integrity is assessed by Semantic Faithfulness (**SMF**), which considers whether salient foreground content remains recognizable and sufficiently complete after retargeting. Layout Faithfulness (**LOF**) measures agreement with the annotated target layout through saliency-weighted positional deviations. Visual plausibility is assessed by Scene and Texture Consistency (**STC**) and Structural and Visual Fidelity (**SVF**), which evaluate scene and texture consistency, structural integrity, and foreground–background harmony.

Together, these components form the benchmark illustrated in Fig. 1, covering diverse scenes and aspect ratios. Instance-level annotations support both quantitative evaluation and supervised retargeting: AIR-DR (Dong et al., 2026) uses them to train its instance relocation module, whose effectiveness is supported by ablation studies. Pixel-level reference outputs offer potential supervision for future studies. We then use our dataset and evaluation metrics to evaluate seven representative

Table 1: Comparison of image-retargeting datasets.

Resource	Input images	Source aspect ratios	Instance-level annotation	Pixel-level reference
RetargetMe	80	10	✗	✗
CUHK	57	7	✗	✗
NRID	35	7	✗	✗
Ours	**5,000**	**21**	**20,000**	**20,000**

methods—Cropping, Scaling, Seam Carving, PruneRepaint, InGAN, GPDM, and GPNN—at all four target aspect ratios. Quantitative and qualitative experiments show strong agreement between the proposed scores and human judgments on the evaluated outputs.

Our contributions can be summarized as follows:

(1) We introduce an annotated image-retargeting benchmark with 5,000 input images, 20,000 instance-level annotations, and 20,000 pixel-level reference outputs across four target aspect ratios. Our benchmark provides instance-level annotations for quantitative evaluation and supervised retargeting, together with pixel-level reference outputs for future supervised learning.

(2) We develop an interpretable evaluation protocol for foreground preservation, reference-layout fidelity, and visual plausibility. Output-level experiments show strong agreement with corresponding human judgments across seven methods and four target aspect ratios.

(3) Building on the proposed dataset and evaluation protocol, we quantitatively benchmark representative image-retargeting methods. The results reveal distinct trade-offs among content preservation, foreground organization, and visual quality that are consistent with the underlying mechanisms of the evaluated methods.

2 Related Work

2.1 Image Retargeting Methods

Classical image retargeting methods resize images by selecting, removing, or deforming existing content. Content-aware cropping retains important subregions (Zhang et al., 2005; Santella et al., 2006), seam carving removes low-energy paths (Avidan & Shamir, 2023), and content-aware scaling and warping distribute deformation according to local structure (Liang et al., 2012; Hu et al., 2014). Multi-operator approaches further combine these transformations. Because these methods remove and deform content differently, they may cause salient-content loss, object deformation, or disruption of spatial relationships. Recent approaches extend retargeting to learned transformation and content generation. MOIR (Zhou et al., 2020) and Cycle-IR (Tan et al., 2019) learn to select or compose resizing operations, while InGAN (Shocher et al., 2019), GPDM (Elnekave & Weiss, 2022), and GPNN (Granot et al., 2022) exploit internal image statistics to generate retargeted results. PruneRepaint (Shen et al., 2024) combines content removal with generative completion. However, existing datasets are too small and insufficiently diverse to support comprehensive evaluation of these methods; their lack of large-scale annotations also hinders supervised learning and further improvements in model performance.

2.2 Retargeting Datasets

Existing image-retargeting datasets are limited in scale and provide different forms of auxiliary information. RetargetMe (Rubinstein et al., 2010) contains precomputed retargeting results and pairwise human preferences, CUHK (Ma et al., 2012) provides mean opinion scores for a fixed set of method outputs, and NRID (Hsu et al., 2014) includes subjective rankings and binary saliency maps. However, their subjective labels are tied to fixed outputs, while the saliency maps describe only important regions in the input image. None provides reusable annotations specifying which foreground instances should be preserved or how they should be organized under different target aspect ratios. In contrast, our benchmark contains 5,000 input images, 20,000 instance-level anno-

tations, and 20,000 pixel-level reference outputs across four target aspect ratios. The instance-level annotations support evaluation of new results, while the reference outputs offer potential supervision for future studies.

2.3 Retargeting Quality Assessment

Subjective studies remain a common evaluation protocol because retargeting quality is inherently perceptual and multiple outputs may be acceptable for the same target ratio (Apostolidis & Mezaris, 2021; Shen et al., 2024; Dong et al., 2026). However, such studies are costly, evaluator-dependent, and difficult to reproduce at benchmark scale. Objective measures have therefore been developed to quantify content loss, geometric deformation, structural inconsistency, and saliency preservation (Liu et al., 2011; Hsu et al., 2014; Fang et al., 2014). Later approaches combine multiple factors, including salient-region preservation, artifacts, global structure, location-based aesthetics, and symmetry (Liang et al., 2016); compare retargeted images using representations at different semantic levels (Zhang et al., 2017a;b); or directly measure the amount of salient content discarded during retargeting (Shen et al., 2024).

Retargeting studies have also adopted general-purpose quality and perceptual-similarity models (Fang et al., 2016), as well as measures based on bidirectional patch similarity (Simakov et al., 2008) and dense feature correspondence (Liu et al., 2008). These measures provide useful quality signals, but even methods that consider multiple factors typically combine different types of degradation into a single prediction. An overall score therefore does not reveal whether a result loses salient content, disrupts the foreground layout, or introduces visual artifacts. Our metrics separately assess content preservation, reference-layout fidelity, and visual quality, showing strong agreement with corresponding human judgments on the evaluated outputs.

3 Dataset

3.1 Data Collection and Diversity

Most input images are drawn from the Saliency-Ranking dataset (Liu et al., 2021), which provides the instance masks and saliency rankings required for annotation. Because it is dominated by images with an aspect ratio of approximately 4:3, we supplement it with images of varied aspect ratios from additional datasets (Fan et al., 2019; Li et al., 2014; Movahedi & Elder, 2010).

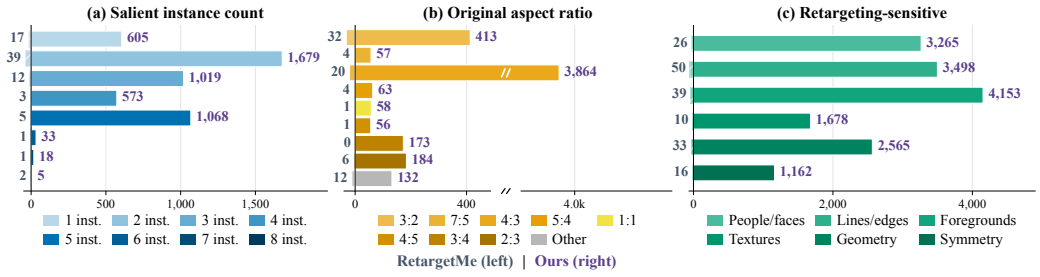


Figure 2: Dataset diversity in instance counts, aspect ratios, and retargeting-sensitive content.

We curate the input images according to three factors that affect retargeting difficulty: **instance count, input aspect ratio, and content sensitivity**. Multi-instance scenes require preserving both individual objects and their spatial relationships, while diverse input aspect ratios introduce different degrees of geometric adaptation. To account for content sensitivity during image selection, we ensure coverage of six retargeting-sensitive attributes identified in prior work (Rubinstein et al., 2010): people or faces, lines or clear edges, salient foreground objects, textures or repeating patterns, geometric structures, and symmetry.

The resulting benchmark contains **5,000 input images** with diverse foreground configurations and retargeting-sensitive content, as summarized in Fig. 2.

3.2 INSTANCE-LEVEL ANNOTATION

Although retargeting methods may produce different acceptable outputs, high-quality results generally preserve salient foreground content and maintain its spatial organization on the target canvas. The widespread use of saliency and importance maps in existing methods further supports foreground preservation as a common requirement (Zhang et al., 2005; Avidan & Shamir, 2023; Hashemzadeh et al., 2019). We therefore construct an instance-level annotation for each input–target-ratio pair, subject to two criteria: i) salient instances remain recognizable and sufficiently complete; and ii) their spatial relationships and overall organization remain coherent under the target aspect ratio.

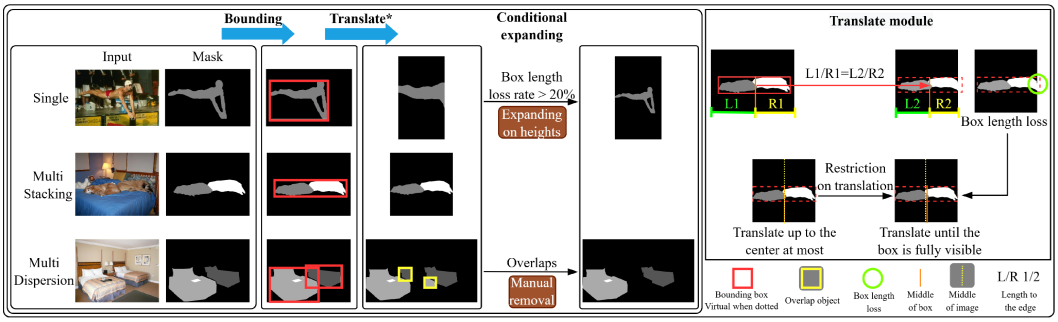


Figure 3: Annotation pipeline for single, stacked, and dispersed objects.

We adopt a human-in-the-loop pipeline based on the instance masks and saliency rankings provided by the source datasets. Automatic preprocessing initializes each annotation, after which annotators inspect and refine instance completeness, boundary handling, and foreground organization. The procedure consists of three steps, as illustrated in Fig. 3.

Step I: Bounding. We first extract the salient foreground instances using the masks provided by the source datasets, including Saliency-Ranking (Liu et al., 2021). These datasets also provide saliency rankings that reflect the relative perceptual importance of different foreground instances. A minimum enclosing square is constructed around each salient mask to determine its spatial extent and center. When multiple instances are spatially stacked or too closely connected to be transformed independently, they are treated as a single group and assigned a shared bounding region.

Step II: Translating and Trimming. Once the bounding boxes are established, we translate and trim the masks to fit the target canvas. The translation strategy depends on the spatial relationships among instances.

For single-object and stacked-object scenes, all masks and bounding boxes undergo the same horizontal translation to preserve their relative spatial positions. When the available space is insufficient, the frame is horizontally expanded according to the original image height. Objects extending beyond the frame are shifted inward by the overflow distance, with translation limited to the image center. If length-wise cropping of a bounding box exceeds 20%, we consider the resulting semantic loss excessive; therefore, the frame is expanded to fully contain the object.

For dispersed-object scenes, each bounding box is repositioned individually, with overlap and clipping ratios recorded. The frame is expanded when box overlap exceeds 20%.

Step III: Human Review and Final Adjustment. The review workload was divided among three annotators, who inspected every initialized annotation for instance completeness, boundary clipping, unintended overlap, and layout coherence. When necessary, they refined instance positions, mask regions, grouping, or frame expansion, prioritizing the most salient instance in unavoidable conflicts. Ambiguous overlapping cases were automatically flagged for focused inspection. Overall, **35.3%** of the initialized annotations required manual adjustment. Only annotations that satisfied both the instance-completeness and layout-coherence criteria were retained as final instance-level annotations. Representative results are shown in Fig. 4.

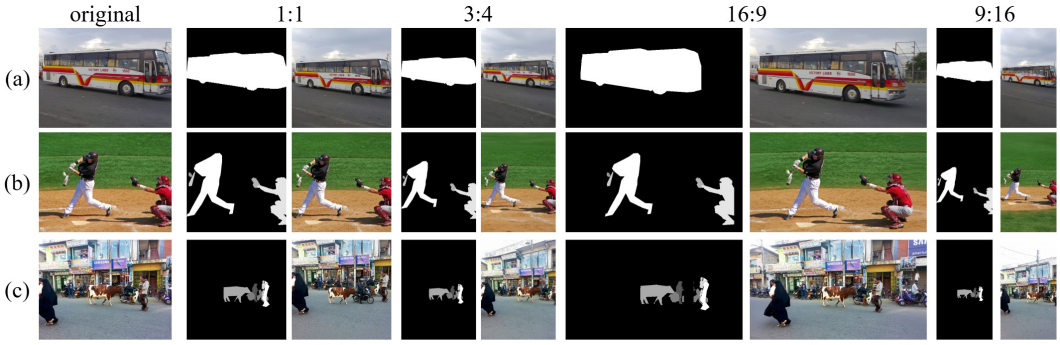


Figure 4: From left to right: the input image, instance-level annotations, and pixel-level reference outputs at different target aspect ratios. From top to bottom: single-object, multiple-dispersed-object, and multiple-stacked-object scenes.

3.3 Pixel-Level Reference Output Generation

We generate pixel-level reference outputs using *Gemini 2.5 Flash* (Comanici et al., 2025), conditioned on each input image and its corresponding instance-level annotation. For every input–target-ratio pair, five candidates are generated and independently rated by three researchers. The candidate with the highest mean rating is retained as the reference output; if none satisfies the quality criteria, additional candidates are generated using revised prompts. Each selected output represents one possible visual realization of the instance-level annotation. These pairs offer potential supervision for retargeting models; their training utility is not evaluated in this work.

Overall, our benchmark contains **5,000** input images, **20,000** instance-level annotations, and **20,000** pixel-level reference outputs. Representative examples are shown in Fig. 4.

4 Evaluation Metrics

4.1 Overview

Our instance-level annotations specify the salient foreground instances and their organization on the target canvas. Based on these annotations, we introduce two scores for the corresponding aspects of retargeting quality. **Semantic Faithfulness (SMF)** measures the preservation of salient foreground content, while **Layout Faithfulness (LOF)** measures agreement with the annotated foreground organization. We first extract salient instances from the retargeted output and match them with the annotated instances. These correspondences are used to compute both scores, with saliency rankings reflecting the relative importance of instances.

Instance-level annotations describe only foreground masks and layouts but cannot fully capture texture consistency, background coherence, or perceptual artifacts. We therefore complement SMF and LOF with two vision-language-model scores: **Scene and Texture Consistency (STC)** and **Structural and Visual Fidelity (SVF)**. Together, these scores evaluate retargeted outputs along three complementary dimensions: **semantic integrity**, **reference-layout fidelity**, and **visual plausibility**.

4.2 Instance-wise Matching

To select a reliable salient-instance extractor, we evaluated multiple SOR methods and adopted SeqRank (Guan & Lau, 2024), which exhibited the most stable performance. Given a retargeted image, SeqRank extracts its salient foreground instances. We then establish correspondences between the extracted instances and those recorded in the instance-level annotation by jointly considering shape and positional consistency.

Specifically, the matching coefficient is computed as a weighted combination of Intersection over Union (IoU) and the cosine similarity between instance-center vectors. An iterative strategy based on this coefficient progressively matches the instances. Detailed formulations and algorithmic pro-

cedures are provided in Appendix A. The matching procedure achieves **95.3%**, **93.1%**, and **93.8%** accuracy on samples containing 2, 3, and 4 instances, respectively, demonstrating its reliability. Moreover, most remaining unmatched cases arise from foreground loss in the retargeted outputs rather than incorrect correspondence, thereby providing an indirect indication of semantic integrity.

4.3 Semantic Faithfulness

Each matched mask pair is aligned by translation and isotropic scaling before IoU computation, removing positional and size differences while preserving aspect-ratio and shape differences. Following SOR, we assign greater weight to more salient instances to reflect their perceptual importance.

SMF is defined as $SMF = \sum_{i=1}^{n} IoU_i \cdot w_i$, where IoU_i is the IoU of the i-th matched pair after translation and isotropic scale normalization and w_i is its saliency-derived weight. The weights follow an approximately arithmetic progression. Higher SMF indicates better preservation of salient foreground content.

4.4 Layout Faithfulness

LOF measures agreement with the annotated target layout. Before computing instance positions, we resize the retargeted output and its masks to the reference canvas resolution, placing both in a common coordinate system. As shown in Fig. 5, $\mathbf{r}_{1i}$ and $\mathbf{r}_{2i}$ are vectors from the canvas center to the corresponding instance centroids in the annotation and resized output, respectively. Using the same weights as SMF, we aggregate their Euclidean distances as $LOF = \sum_{i=1}^{n} w_i \cdot \|\mathbf{r}_{2i} - \mathbf{r}_{1i}\|$. Lower LOF indicates closer agreement with the annotated target layout. On the evaluated outputs, this agreement is positively associated with human layout ratings.

4.5 Visual Plausibility

Because SMF and LOF rely on instance masks, they cannot assess texture consistency, background coherence, or perceptual artifacts. We therefore use the vision-language model (VLM) *Gemini 2.5 Pro* (Comanici et al., 2025) to evaluate visual plausibility from two perspectives:

Scene and Texture Consistency (STC). Given the original image and its retargeted result, the VLM is prompted to evaluate the consistency of the overall scene appearance and texture details.
Structural and Visual Fidelity (SVF). Given only the retargeted result, the VLM is prompted to assess structural integrity, artifacts and distortions, foreground–background harmony, and overall visual quality.

5 Experiment

5.1 Overview

For consistent benchmarking, we evaluate seven representative methods on the same set of 200 input images: Cropping, Scaling, Seam Carving (Avidan & Shamir, 2023), PruneRepaint (Shen et al., 2024), InGAN (Shocher et al., 2019), GPDM (Elnekave & Weiss, 2022), and GPNN (Granot et al., 2022). This set size balances evaluation coverage and computational cost, since InGAN requires separate self-supervised training for each image. Each method produces outputs at four target aspect ratios: 1:1, 3:4, 16:9, and 9:16.

To assess agreement with human judgments, we first collect subjective ratings for retargeted outputs (Sec. *Subjective Evaluation*). We then examine the agreement between the proposed scores and human judgments through quantitative correlation analysis and qualitative comparisons (Sec. *Agreement with Human Judgments*). After confirming their agreement with human judgments, we apply the proposed scores to the complete 200-image evaluation set to benchmark and analyze the seven methods (Sec. *Benchmark Results*).

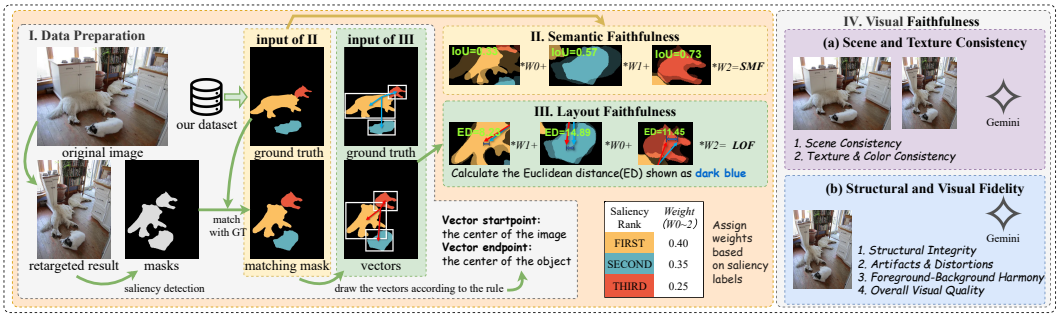


Figure 5: Evaluation pipeline. Mask colors indicate saliency ranks and instance matches; annotated masks serve as references for SMF and LOF.

5.2 Subjective Evaluation

We select 50 images from the 200-image evaluation set and generate outputs using all seven methods at four target aspect ratios. Eight professional evaluators rate each output on semantic integrity, layout coherence, scene and texture consistency, and structural and visual fidelity. All criteria use a five-point scale, where higher is better. For each output and criterion, we average the eight evaluators' ratings to obtain a mean opinion score (MOS), yielding 350 scored outputs per target aspect ratio (50 input images $\times$ seven methods). An example questionnaire and the mean subjective scores are provided in Appendix B.

5.3 Agreement with Human Judgments

We assess agreement between the proposed scores and subjective ratings through quantitative correlations and qualitative comparisons. For semantic preservation, we compare **SMF** with Bidirectional Similarity (BDS) (Simakov et al., 2008) and Saliency Discard Ratio (SDR) (Shen et al., 2024); for foreground layout, we compare **LOF** with SIFT Flow (Liu et al., 2008) and the location-based aesthetics component of Liang et al. (Liang et al., 2016), termed **LiangL**. These baselines represent objective measures commonly used in image-retargeting evaluation.

Table 2: Output-level correlations with MOS for 350 evaluated outputs per target ratio. Each cell reports SRCC / PLCC (higher is better). BDS, SDR, SIFT, LiangL, and LOF are negated before correlation analysis. Dagger† denotes a proposed score.

	Semantic evaluation			Layout evaluation			Visual quality	
Ratio	BDS	SDR	SMF†	SIFT	LiangL	LOF†	STC†	SVF†
1:1	.193/.134	.617/.616	**.702/.749**	.118/.267	.292/.240	**.698/.735**	.708/.774	.795/.729
3:4	.206/.123	.721/.741	**.827/.842**	.431/.404	.411/.326	**.708/.725**	.786/.806	.833/.744
16:9	.280/.188	.784/.815	**.805/.856**	.280/.240	.242/.180	**.800/.811**	.785/.837	.778/.730
9:16	.516/.460	.710/.777	**.792/.848**	.422/.393	.382/.355	**.794/.758**	.781/.795	.858/.773

5.3.1 Quantitative Evaluation

Following (Zhang et al., 2021), we use SRCC (Spearman, 1961) and PLCC (Pearson, 1896) to assess agreement with human judgments. For each target aspect ratio, we evaluate 350 individual outputs (50 input images $\times$ seven methods) and compute correlations over valid objective-score–MOS pairs for each criterion. Scores are not averaged within methods before computing correlations. For BDS, SDR, SIFT Flow, LiangL, and LOF, we report correlations with the negated objective scores to align their direction with MOS. Table 2 reports these output-level correlations.

SMF and LOF yield higher reported SRCC and PLCC than their respective comparison metrics across all four target ratios. LOF achieves reported SRCC of 0.698–0.800 and PLCC of 0.725–

0.811, supporting agreement with human layout judgments on the evaluated outputs. STC and SVF achieve SRCC of 0.708–0.786 and 0.778–0.858, respectively.

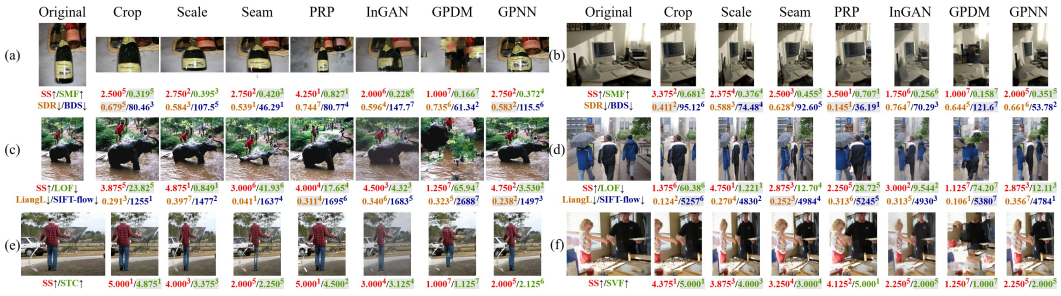


Figure 6: Qualitative comparison of our metrics and subjective scores (**SS**). Examples (a)–(b), (c)–(d), (e), and (f) correspond to SMF, LOF, STC, and SVF, respectively. Metric and subjective scores are shown in green and red. Superscripts denote ranks among seven methods; gray backgrounds indicate agreement between the two rankings.

5.3.2 Qualitative Evaluation

Fig. 6 compares metric and subjective rankings. SDR measures changes in salient-region width and may miss deformation along the orthogonal dimension. In Fig. 6(a), it favors Seam Carving despite PRP better preserving object structure, whereas SMF reflects this difference. BDS measures local patch similarity and favors GPDM despite its global distortion in the same example; SMF ranks the results more consistently with human judgment.

LiangL evaluates salient-region centers using composition rules rather than annotated instance locations, potentially favoring well-positioned but damaged content (Fig. 6(c)). SIFT Flow relies on dense local-descriptor correspondences and yields similar scores for perceptually distinct results under large displacements (Fig. 6(c)–(d)). LOF compares matched positions with the reference layout and better distinguishes these cases in agreement with subjective ratings. STC and SVF likewise closely follow subjective rankings in Fig. 6(e)–(f).

Table 3: Average results across four target aspect ratios. Best scores are shown in bold.

Metric	Crop	Scale	Seam	PRP	InGAN	GPDM	GPNN
SMF ↑	0.616	0.568	0.560	**0.659**	0.490	0.359	0.537
LOF ↓	67.35	**50.97**	65.39	67.48	62.69	94.97	58.47
STC ↑	**4.860**	4.800	3.578	4.060	2.455	1.485	3.310
SVF ↑	**2.925**	2.915	2.165	2.220	1.555	1.130	1.955

5.4 Benchmark Results

Table 3 shows complementary strengths: PruneRepaint achieves the highest SMF, Scaling the lowest LOF, and Cropping the highest STC and SVF, closely followed by Scaling. GPDM performs worst across all four metrics. No method leads in every dimension, supporting multidimensional evaluation. Fig. 6 illustrates these trade-offs.

6 Conclusion

We introduce a foreground-centric retargeting benchmark with 5,000 inputs, 20,000 instance-level annotations, and 20,000 pixel-level reference outputs across four target ratios. On outputs from seven methods, SMF, LOF, STC, and SVF show strong agreement with human judgments, with SMF and LOF outperforming their respective comparison metrics in correlation. These findings support annotation-driven evaluation of foreground preservation and reference-layout fidelity on the tested outputs, complemented by visual-quality assessment.

AI USE STATEMENT

Generative AI tools assisted manuscript writing and revision. As described in the paper, Gemini models were used to generate candidate reference outputs and compute the STC and SVF scores. Generated reference candidates were manually reviewed and selected. The authors take responsibility for the final text, methodology, and reported results.

REFERENCES

Konstantinos Apostolidis and Vasileios Mezaris. A fast smart-cropping method and dataset for video retargeting. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2618–2622. IEEE, 2021.

Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 609–617. 2023.

Donghyeon Cho, Jinsun Park, Tae-Hyun Oh, Yu-Wing Tai, and In So Kweon. Weakly-and self-supervised learning for content-aware deep image retargeting. In *Proceedings of the IEEE international conference on computer vision*, pp. 4558–4567, 2017.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Zhitong Dong, Chao Li, Yongjian Deng, and Hao Chen. Air-dr: Adaptive image retargeting with instance relocation and dual-guidance repainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3668–3676, 2026.

Ariel Elnekave and Yair Weiss. Generating natural images with direct patch distributions matching. In *European Conference on Computer Vision*, pp. 544–560. Springer, 2022.

Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8554–8564, 2019.

Yuming Fang, Kai Zeng, Zhou Wang, Weisi Lin, Zhijun Fang, and Chia-Wen Lin. Objective quality assessment for image retargeting based on structural similarity. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 4(1):95–105, 2014.

Yuming Fang, Zhijun Fang, Feiniu Yuan, Yong Yang, Shouyuan Yang, and Neal N Xiong. Optimized multioperator image retargeting based on perceptual similarity measure. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(11):2956–2966, 2016.

MohammadHossein Givkashi, MohammadReza Naderi, Nader Karimi, Shahram Shirani, and Shadrokh Samavi. Supervised deep learning for content-aware image retargeting with fourier convolutions. *Multimedia Tools and Applications*, pp. 1–17, 2024.

Niv Granot, Ben Feinstein, Assaf Shocher, Shai Bagon, and Michal Irani. Drop the gan: In defense of patches nearest neighbors as single image generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13460–13469, 2022.

Huankang Guan and Rynson WH Lau. Seqrank: Sequential ranking of salient objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1941–1949, 2024.

Mahdi Hashemzadeh, Bahareh Asheghi, and Nacer Farajzadeh. Content-aware image resizing: an improved and shadow-preserving seam carving method. *Signal Processing*, 155:233–246, 2019.

Nicolas S Holliman, Neil A Dodgson, Gregg E Favalora, and Lachlan Pockett. Three-dimensional displays: a review and applications analysis. *IEEE transactions on Broadcasting*, 57(2):362–371, 2011.

Chih-Chung Hsu, Chia-Wen Lin, Yuming Fang, and Weisi Lin. Objective quality assessment for image retargeting based on perceptual geometric distortion and information loss. *IEEE Journal of Selected Topics in Signal Processing*, 8(3):377–389, 2014.

Wenyu Hu, Zhongxuan Luo, and Xin Fan. Image retargeting via adaptive scaling with geometry preservation. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 4(1):70–81, 2014.

Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 280–287, 2014.

Yun Liang, Zhuo Su, and Xiaonan Luo. Patchwise scaling method for content-aware image resizing. *Signal Processing*, 92(5):1243–1257, 2012.

Yun Liang, Yong-Jin Liu, and Diego Gutierrez. Objective quality prediction of image retargeting algorithms. *IEEE Transactions on Visualization and Computer Graphics*, 23(2):1099–1110, 2016.

Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. Sift flow: Dense correspondence across different scenes. In *European conference on computer vision*, pp. 28–42. Springer, 2008.

Nian Liu, Long Li, Wangbo Zhao, Junwei Han, and Ling Shao. Instance-level relative saliency ranking with graph reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8321–8337, 2021.

Yong-Jin Liu, Xi Luo, Yu-Ming Xuan, Wen-Feng Chen, and Xiao-Lan Fu. Image retargeting quality assessment. In *Computer Graphics Forum*, volume 30, pp. 583–592. Wiley Online Library, 2011.

Lin Ma, Weisi Lin, Chenwei Deng, and King Ngi Ngan. Image retargeting quality assessment: A study of subjective scores and objective metrics. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):626–639, 2012.

Ioannis Mademlis, Anastasios Tefas, Nikos Nikolaidis, and Ioannis Pitas. Multimodal stereoscopic movie summarization conforming to narrative characteristics. *IEEE Transactions on Image Processing*, 25(12):5828–5840, 2016.

Vida Movahedi and James H Elder. Design and perceptual validation of performance measures for salient object segmentation. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pp. 49–56. IEEE, 2010.

Karl Pearson. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187):253–318, 1896.

Michael Rubinstein, Ariel Shamir, and Shai Avidan. Improved seam carving for video retargeting. *ACM transactions on graphics (TOG)*, 27(3):1–9, 2008.

Michael Rubinstein, Diego Gutierrez, Olga Sorkine, and Ariel Shamir. A comparative study of image retargeting. In *ACM SIGGRAPH Asia 2010 papers*, pp. 1–10. 2010.

Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. Gaze-based interaction for semi-automatic photo cropping. In *SIGCHI*, pp. 771–780, 2006.

Feihong Shen, Chao Li, Yifeng Geng, Yongjian Deng, and Hao Chen. Prune and repaint: Content-aware image retargeting for any ratio. In *NeurIPS*, 2024.

Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. Ingan: Capturing and retargeting the "dna" of a natural image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4491–4500, 2019. doi: 10.1109/ICCV.2019.00459.

Denis Simakov, Yaron Caspi, Eli Shechtman, and Michal Irani. Summarizing visual data using bidirectional similarity. In *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8. IEEE, 2008.

Charles Spearman. The proof and measurement of association between two things. 1961.

Weimin Tan, Bo Yan, Chuming Lin, and Xuejing Niu. Cycle-ir: Deep cyclic image retargeting. *IEEE Transactions on Multimedia*, 22(7):1730–1743, 2019.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Mingju Zhang, Lei Zhang, Yanfeng Sun, Lin Feng, and Weiying Ma. Auto cropping for digital photographs. In *ICME*, pp. 4–pp. IEEE, 2005.

Yabin Zhang, Weisi Lin, Qiaohong Li, Wentao Cheng, and Xinfeng Zhang. Multiple-level feature-based measure for retargeted image quality. *IEEE Transactions on image processing*, 27(1): 451–463, 2017a.

Yichi Zhang, King Ngi Ngan, Lin Ma, and Hongliang Li. Objective quality assessment of image retargeting by incorporating fidelity measures and inconsistency detection. *IEEE Transactions on Image Processing*, 26(12):5980–5993, 2017b.

Zicheng Zhang, Wei Sun, Xiongkuo Min, Tao Wang, Wei Lu, and Guangtao Zhai. A full-reference quality assessment metric for fine-grained compressed images. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–4. IEEE, 2021.

Ya Zhou, Zhibo Chen, and Weiping Li. Weakly supervised reinforced multi-operator image retargeting. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1):126–139, 2020.

APPENDIX

This appendix provides additional details to complement the main paper.

- **Instance Matching Algorithm.** Describes the instance-wise matching algorithm used to compute our evaluation metrics, including the calculation of matching coefficients and the iterative procedure for pairing instances between original and retargeted images.
- **Detailed Experimental Data.** Provides extended quantitative results for different retargeting methods across various aspect ratios, including both objective metric scores (SMF, LOF, STC, SVF) and subjective ratings.
- **Benchmarking Retargeting Models.** Presents qualitative and quantitative analyses of representative retargeting methods using concrete visual examples, covering single-object, two-object, and multi-object scenarios.

Additional implementation details and visual examples are included where necessary.

A INSTANCE MATCHING ALGORITHM

Given a retargeted image, we first apply SeqRank Guan & Lau (2024) to extract foreground instances. For each candidate instance pair, the matching process evaluates shape and positional consistency to obtain a matching coefficient. We then use these coefficients to iteratively match the instances.

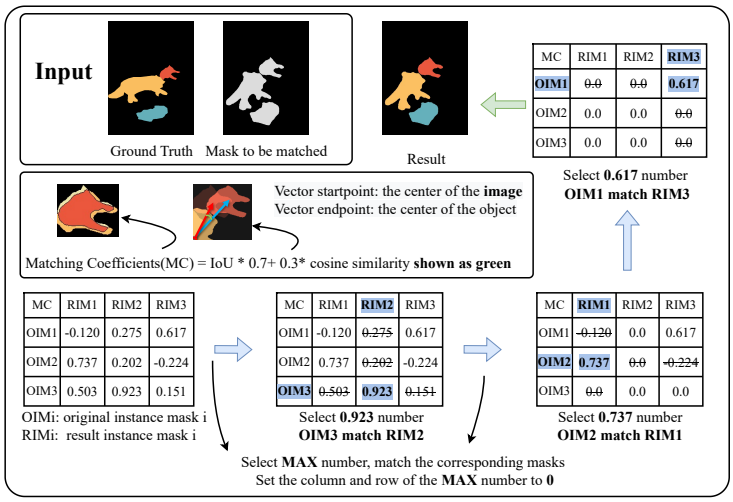


Figure 7: Detailed matching flowchart. Matched instances per iteration are highlighted in dark blue.

Matching coefficient. The matching coefficient consists of two components: shape consistency between instances, measured by intersection over union (IoU), and positional consistency ($PosCons$), determined by the cosine similarity between the "instance vectors." Each vector points from the image center to the center of an instance's minimum bounding box. The components are combined as

$$MatchCoef = 0.7 \times IoU + 0.3 \times PosCons. \tag{1}$$

We assign a higher weight to IoU because shape consistency is more important for matching.

Iterative matching. The detailed workflow is illustrated in Figure 7. We progressively match the ground-truth masks from our dataset to the retargeted masks using the matching coefficients. For multiple instances, we construct a two-dimensional score matrix whose columns correspond to ground-truth masks and whose rows correspond to retargeted masks. Each matrix entry is the matching coefficient for the corresponding mask pair. At each iteration, we select the pair with the highest coefficient as a match and remove its row and column from the matrix. This process

continues until no valid candidate pairs remain. Our matching method achieves matching accuracies of **95.3%**, **93.1%**, and **93.8%** on samples containing two, three, and four instances, respectively, demonstrating its reliability.

B Detailed Experimental Data

Table 4: Detailed results of our four metrics across seven methods. Higher is better for SMF, STC, and SVF, while lower is better for LOF. "PRP": PruneRepaint; "Seam": Seam Carving. Best scores are in bold.

Metric	Ratio	Crop	Scale	Seam	PRP	InGAN	GPDM	GPNN
SMF ↑	1:1	0.681	0.632	0.618	**0.709**	0.559	0.403	0.615
	3:4	0.604	0.569	0.554	**0.657**	0.480	0.289	0.533
	16:9	0.684	0.620	0.618	**0.734**	0.544	0.520	0.587
	9:16	0.493	0.449	0.448	**0.534**	0.376	0.225	0.413
LOF ↓	1:1	48.87	**38.78**	52.53	50.20	46.83	69.12	40.68
	3:4	66.06	**48.05**	60.53	62.91	55.75	92.13	54.46
	16:9	68.52	**56.53**	70.99	70.47	66.91	93.04	65.57
	9:16	85.95	**60.53**	77.50	86.35	81.25	125.6	73.16
STC ↑	1:1	**4.98**	4.94	4.06	4.32	2.90	1.60	3.62
	3:4	**5.00**	4.90	3.05	3.96	2.32	1.26	3.24
	16:9	**4.64**	4.60	3.90	4.28	2.36	1.98	3.80
	9:16	**4.82**	4.76	3.30	3.68	2.24	1.10	2.58
SVF ↑	1:1	**3.58**	3.52	2.84	2.58	1.90	1.20	2.38
	3:4	**3.02**	2.84	2.04	2.26	1.54	1.08	1.94
	16:9	2.66	**2.86**	2.14	2.18	1.44	1.18	1.94
	9:16	**2.44**	**2.44**	1.64	1.86	1.34	1.06	1.56

Table 5: Detailed subjective scores for the four evaluation perspectives across seven methods. Higher is better. "PRP": PruneRepaint; "Seam": Seam Carving. Best scores are shown in bold.

Criterion	Ratio	Crop	Scale	Seam	PRP	InGAN	GPDM	GPNN
Semantic faithfulness	1:1	4.48	3.96	3.91	**4.61**	3.70	2.41	3.75
	3:4	3.89	3.74	3.26	**4.24**	3.19	1.46	3.04
	16:9	4.17	3.80	3.69	**4.28**	3.51	2.78	3.60
	9:16	3.27	3.33	2.74	**3.61**	2.45	1.15	2.30
Layout faithfulness	1:1	3.77	**4.93**	4.35	3.90	4.60	3.43	4.64
	3:4	3.38	**4.87**	4.17	3.56	4.33	2.45	4.20
	16:9	3.95	**4.82**	4.33	4.10	4.36	3.51	4.37
	9:16	2.87	**4.75**	3.85	3.14	3.68	1.57	3.41
Scene and texture consistency	1:1	**4.78**	4.17	3.83	4.39	3.62	2.28	3.61
	3:4	**4.75**	4.00	3.37	4.16	3.19	1.55	3.05
	16:9	**4.73**	3.92	3.66	4.17	3.38	2.63	3.43
	9:16	**4.63**	3.64	2.92	3.68	2.57	1.30	2.39
Structural and visual fidelity	1:1	**4.64**	4.14	3.75	4.33	3.58	2.25	3.57
	3:4	**4.51**	3.90	3.26	4.11	3.09	1.45	3.04
	16:9	**4.51**	3.87	3.63	4.17	3.34	2.55	3.42
	9:16	**4.28**	3.49	2.83	3.60	2.40	1.27	2.29

B.1 Quantitative Metrics

The detailed objective metric results are presented in Table 4. Comparing them with the subjective scores in Table 5 reveals a high degree of consistency, supporting the alignment between our evaluation metrics and human perceptual judgments.

B.2 Subjective Ratings

Please provide the scores for the 1:1 retargeting ratio



	Crop	Scale	Seam Carving	PruneRepaint	InGAN	GPDM	GPNN
Semantics Preservation	Please select	Please select	Please select	Please select	Please select	Please select	Please select
Layout Preservation	Please select	Please select	Please select	Please select	Please select	Please select	Please select
Scene and Texture Consistency	Please select	Please select	Please select	Please select	Please select	Please select	Please select
Structural and Visual Fidelity	Please select	Please select	Please select	Please select	Please select	Please select	Please select

Figure 8: Illustrative questionnaire showing the structure used in the study. Participants rated each "Please Select" item on a 1–5 scale.

The questionnaire used in the subjective study is shown in Figure 8. For each retargeted image, evaluators selected a score from 1 to 5 using a drop-down menu. The detailed subjective results are reported in Table 5.

C Benchmarking Retargeting Models

C.1 Quantitative Analysis

We evaluate seven representative image-retargeting methods on our proposed dataset using three groups of metrics: Semantic Faithfulness (SMF), Layout Faithfulness (LOF), and visual-plausibility metrics (STC and SVF). This evaluation measures the methods' effectiveness in preserving semantic content, following the annotated target layout, and avoiding visual distortions and artifacts.

As shown in Table 4, traditional resizing-based methods (**Cropping**, **Scaling**, and **Seam Carving**) generally achieve higher visual consistency but differ in semantic and layout preservation. **Cropping** directly removes pixels without deformation, yielding the highest STC and SVF scores. Its higher LOF indicates greater positional deviations from the annotated layout. **Scaling** uniformly rescales the entire image and achieves the lowest LOF, indicating the closest agreement with annotated instance positions, though semantic details may be slightly diluted. In contrast, **Seam Carving** Avidan & Shamir (2023), which relies solely on low-level energy maps, tends to remove salient semantic regions with repetitive textures, causing structural and texture distortions.

Among learning-based approaches, **PruneRepaint** Shen et al. (2024) achieves the best overall SMF by leveraging semantics-aware pruning and repainting, effectively preserving important content while maintaining moderate reference-layout fidelity. The three patch-based generative methods—**InGAN** Shocher et al. (2019), **GPDM** Elnekave & Weiss (2022), and **GPNN** Granot et al. (2022)—synthesize new pixels based on local patch statistics rather than explicit spatial constraints. While

GPNN achieves relatively lower LOF, InGAN and especially GPDM exhibit severe global structural disruption, leading to low SMF, STC, and SVF scores. Overall, semantics-aware optimization and global geometric consistency remain the most reliable strategies for faithful retargeting, whereas purely generative patch-based approaches struggle to balance semantic, spatial, and visual plausibility simultaneously.

C.2 Qualitative Analysis

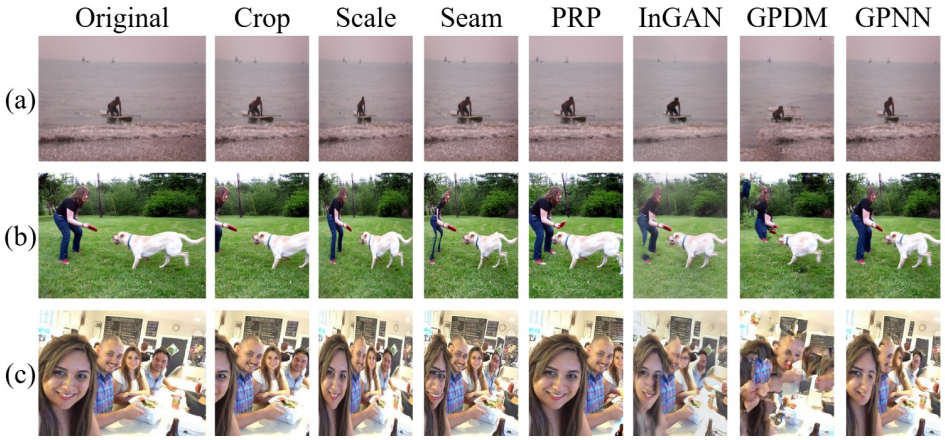


Figure 9: From (a) to (c), retargeting difficulty increases. The examples highlight characteristic behaviors of different methods. Detailed scores are reported in Table 6.

Table 6: Metric scores for cases (a)–(c) in Figure 9. Higher is better for SMF, STC, and SVF; lower is better for LOF. Best scores are bold. "PRP: PruneRepaint; "Seam: Seam Carving.

	case (a)				case (b)				case (c)			
Method	SMF	LOF	STC	SVF	SMF	LOF	STC	SVF	SMF	LOF	STC	SVF
Crop	**0.759**	8.559	**5.000**	**5.000**	0.287	29.32	**5.000**	**5.000**	0.420	103.9	**5.000**	**5.000**
Scale	0.460	**1.118**	4.000	4.000	0.467	10.67	4.000	3.000	0.484	57.97	4.000	3.000
Seam	0.677	5.852	**5.000**	4.000	0.395	**9.689**	3.000	3.000	0.495	69.83	3.000	3.000
PRP	0.730	2.500	**5.000**	**5.000**	**0.609**	15.09	**5.000**	4.000	**0.616**	**32.97**	**5.000**	**5.000**
InGAN	0.599	7.071	**5.000**	4.000	0.560	20.99	2.000	2.000	0.446	71.96	2.000	1.000
GPDM	0.374	20.83	3.000	3.000	0.303	24.58	2.000	1.000	0.157	79.55	2.000	1.000
GPNN	0.545	**1.118**	**5.000**	4.000	0.525	14.15	2.000	2.000	0.429	74.48	3.000	2.000

We analyze the strengths and weaknesses of the seven methods using concrete visual examples, grouping the test images into three scene types of increasing difficulty.

Scene I: Single-Object Scene. When the image contains only one foreground object and the background occupies most of the area (Figure 9(a)), the outputs are generally comparable and visually satisfactory. Retargeting is relatively easy when the background is large, homogeneous, and can be safely replicated or removed.

Scene II: Two-Object Scene. PruneRepaint prioritizes important foreground regions and retains most semantic information, achieving the highest SMF. **Crop** directly removes boundary regions, causing partial loss of foreground content and moderate SMF performance. **Scale** achieves relatively low LOF, indicating that matched instance positions remain close to the annotated positions, although Seam achieves the lowest LOF in this example. Cropping produces larger positional deviations from the reference layout and therefore higher LOF. The three patch-based generative models synthesize pixels through local patch-distribution matching instead of explicit structural constraints. Consequently, they are prone to artifacts and shape inconsistencies, leading to low-to-moderate STC and SVF scores, as illustrated in Figure 9(b).

Scene III: Multi-Object Scene. As shown in Figure 9(c), multiple people occupy most of the frame and create a complex scene structure. **PruneRepaint** performs the best: it avoids severe distortions and preserves the most important semantic regions. Although **Crop** causes no geometric deformation, it lacks PruneRepaint's semantic preservation capability. **Scale** maintains the layout but introduces an unnatural appearance due to heavy compression. Seam Carving and the patch-based generative methods degrade significantly. Seam Carving removes low-energy seams based on gradient magnitude rather than semantic content, often tearing or warping important objects in dense scenes. The generative methods also struggle, frequently producing artifacts and strong structural distortions.

Summary. Across the three difficulty levels, PruneRepaint consistently delivers robust performance by preserving essential semantics and avoiding artifacts. Cropping does not deform the retained pixels, but it inevitably discards semantic content and alters the image composition. Scale preserves spatial relationships but suffers from unnatural compression. Seam Carving cannot reliably identify true semantic importance and therefore often causes unnatural deformation. The patch-based generative approaches perform well on simple scenes but tend to generate artifacts and shape distortions in scenes with multiple objects or complex structures.