확산 모델 기반 비전-언어 검색 정렬을 통한 단일 이미지 초해상화

이다인 [1], 고재균 [2], 임병완 [3], 김태현 [4]
[1] 한양대학교 지능융합학과 석사과정
[2] 한양대학교 컴퓨터소프트웨어학과 박사 후 연구원
[3] 한양대학교 컴퓨터소프트웨어학과 박사과정
[4] 한양대학교 컴퓨터소프트웨어학과 교수

dainlee@hanyang.ac.kr, rhworbs1124@hanyang.ac.kr, pook0612@hanyang.ac.kr, taehyunkim@hanyang.ac.kr

Diffusion-based Vision–Language Retrieval Alignment for Single-Image XSR

Dain Lee[1], Jae Kyun Ko[2], Byung Wan Lim[2], Tae Hyun Kim[2]
[1]Dept. of Intelligence and Convergence, Hanyang University
[2]Dept. of Computer Science, Hanyang University

요 약

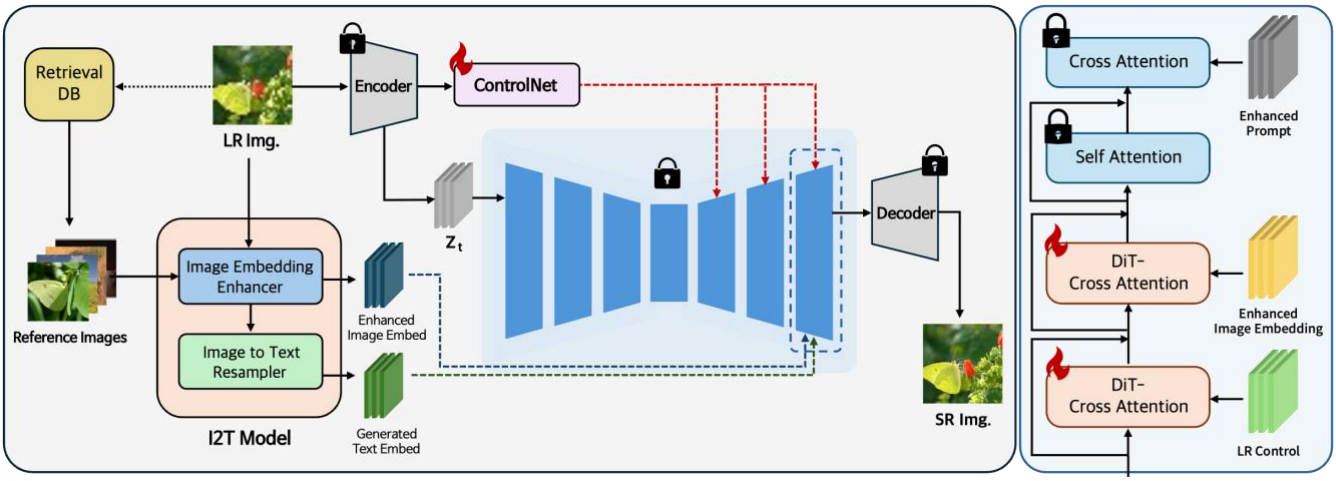
단일 저해상도(LR) 한 장으로 기존 초해상화 모델을 사용하여 고배율의 초해상화(XSR)를 수행하면, 입력 내부 신호만으로는 고주파 복원이 근본적으로 부족해 과도한 환각, 평활화가 발생한다. 이러한 문제를 해결하기 위해 본 논문은 2-Stage 파이프라인을 제안한다. Stage-1: (i) RFA(Reference-Finding Aligner)가 LR 표현을 HR 임베딩 공간(CLIP)으로 정렬해 대규모 HR DB 에서 의미적으로 근접한 참조를 안정적으로 검색하고, (ii) 검색 참조와 LR 을 융합하는 I2T(Embedding Enhancer)가 강화된 이미지/텍스트 임베딩을 산출한다. Stage-2: Stable Diffusion 기반 U-Net 과 ControlNet 에 LR 힌트와 강화 임베딩(이미지/텍스트)을 AdaIN/교차어텐션 조절자로 주입해 HR 을 합성한다. ImageNet 데이터셋 기반 ×16 설정(32→512)에서, 기존 모델들 대비 다양한 매트릭에서 개선된 성능을 보였다.

1. 서론

단일 이미지 초해상화(SR)은 저해상도 입력에서 고해상도 출력을 복원하는 문제로, 상실된 고주파 성분의 복원이 핵심이다. 특히 x16 과 같은 Extreme 배율에서는 저해상도(LR) 내부 신호만으로는 원본 복원이 근본적으로 불충분하며, 과도한 환각 또는 과도한 평활화가 발생한다. 최근 확산 모델과 대규모 비전-언어 모델(VLM)은 강력한 사전 지식(prior)을 제공하지만[1, 2, 10, 17, 20], 단일 LR 만으로는 해당 prior 를 안정적으로 정렬하여 활용하는 방법은 덜 탐색되었다.

본 연구의 목표는 LR 이미지 한장만으로 수행하는 고배율 초해상화이다. 생성 모델의 표현력을 빌리되, LR 에서 추출한 정보가 손실 없이 보존되도록 2 stage 구조를 설계하였다(그림 1). 기여는 다음과 같다.

- LR→HR 임베딩 정렬로 검색 이전 단계의 표현 불일치를 학습적으로 해소하여, 단일 LR 에서도 참조 Top-K 적합도/정확도를 정량 향상.
- 대규모 언어 모델의 프롬프트 없이 조회된 HR 참조 4 장 + LR 토큰을 다중 어텐션으로 직접 융합, 이미지 · 텍스트 이중 조건을 생성하여 강한 열화에서도 안정적 조건화 제공.
- 2-stage 확산 파이프라인에서 ControlNet[23]의 LR 구조 힌트와 I2T 임베딩을 AdaIN 기반으로 교차어텐션 Q/K · LayerNorm 통계에 주입하는 경량 어댑터를 설계, 사전학습 Stable Diffusion 과 호환적으로 결합.



(그림 1) 모델 전체 프레임워크 (좌) Stage-1 과 Stage-2 의 전반적 그림. (우) 3 중 조건이 확산 모델에 주입되는 방식.

2. 관련 연구

확산 기반 복원/초해상화는 LDM으로 잠재공간에서의 고해상 합성 · SR이 가능해지고[17], SR3가 초기 확산-SR 틀을 제시했으며[18], 이후 DiffBIR는 열화 제거→세부 재생의 2단 구성으로 분리[12], SeeSR는 열화 인지 soft/hard 프롬프트로 의미 보존을 강화, ControlNet[23]은 엣지, 포즈, 세그 등 공간 조건 주입을 표준화해 다양한 SR 어댑터에 쓰인다[22,23].

한편 VLM 계열에서는 CLIP의 이미지-텍스트 대조 정렬[15], ViT의 패치 기반 트랜스포머[6], BERT의 텍스트 백본[5]이 기반을 이루고, Flamingo, BLIP-2는 대규모 사전학습 백본을 동결한 채 경량 어댑터(Q-Former/Resampler 등)로 효율을 높였다[1,2,10].

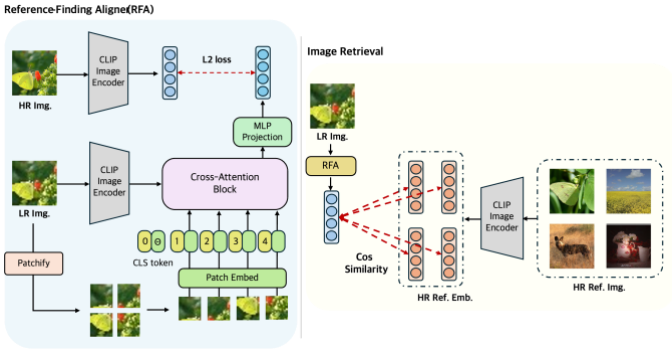
NLP의 Retrieval-augmented Generation(RAG)은 외부 검색을 결합해 정확도, 근거성을 높였고 [9], Self-RAG, In-Context RALM은 언제/무엇을 검색할지를 학습적으로 제어한다 [3,16]. 복원으로의 확장에서는 CoSeR가 인지 임베딩 기반 참조 생성, 주입을, ReFIR가 최근접 참조의 무학습 주입을 제안한다[19,8].

3. 제안 방법

3.1. 전체 개요

제안 파이프라인은 그림 1 에 요약된다. Stage-1 에서 LR 1 장을 입력으로, (a) 레퍼런스 검색과 (b) 이미지 임베딩 강화(I2T)와 텍스트 임베딩 생성을 수행한다. Stage-2 는 Stable Diffusion[17]과 ControlNet[23]에 두 임베딩과 LR 힌트를 AdaIN 기반 모듈로 주입하여 SR 이미지를 생성한다.

3.2. Stage-1 (a): Reference-Finding Aligner (RFA)



(그림 2) RFA 전체 개요: (좌) 학습 단계, (우) 추론 및 참조 검색 단계.

LR 이미지를 CLIP 이미지 인코더에 통과시켜 패치 토큰 256개와 학습 가능한 CLS 토큰 1개로 구성된 시퀀스 F_{LR} 를 얻는다. RFA는 교차 어텐션 기반의 소형 트랜스포머로, 입력 F_{LR} 의 CLS 벡터를 $\hat{C}_{LR} \in R^{1024}$ 로 산출한다. 학습 시에는 동일 이미지의 HR 임베딩 CLS C_{HR} 과의 정렬 손실을 다음과 같이 최소화한다(그림 2-좌).

$$L_{align} = \left|\hat{C}_{LR} - C_{HR}\right|_2^2$$

추론 단계에서는 $\hat{C}_{LR}$ 을 HR DB의 CLS들과 cosine 유사도로 비교하여 Top-K 참조를 검색한다. 이는 ReFIR [8]과 달리, LR→HR 표현 정렬을 통해 32×32와 같은 극 저해상도에서도 의미 기반 검색을 가능하게 한다.

3.3. Stage-1 (b): Image-To-Text model(I2T)

검색된 HR 참조 4 장의 CLIP 토큰 집합과 LR 토큰을 입력으로 받아, Image Embedding Enhancer 는 LR 토큰을 쿼리로, LR+참조 토큰을 키/밸류로 다중 어텐션으로 융합하여 LR 신호를 참조 문맥과 함께 해석한다. 이 블록의 출력 강화 이미지 임베딩은 CLIP 이미지 인코더 표현과 코사인 유사도 정렬로 학습되어, GT HR 의 임베딩과 최대한 가깝게 맞춰진다.

동시에, 강화 토큰을 토큰-단위 풀링으로 얻은 요약 벡터를 사전 증류된 경량 텍스트 디코더(GPT-2 distill)의 조건으로 넣어 텍스트 토큰을 생성하고, CLIP 텍스트 인코더를 통해 얻은 임베딩을 BLIP-2[10]로 사전 생성한 프롬프트의 CLIP 임베딩과 코사인 정렬로 감독한다. (모든 CLIP 인코더는 동결됨)

3.4. Stage-2: 삼중 조건 주입 확산 SR

기반 모델은 사전 학습된 Stable Diffusion 의 U-Net 이며, 입력 LR 로부터 추출한 구조 힌트는 ControlNet 을 통해 주입한다(그림 3). 우리는 Stage-1 에서 얻은 강화 이미지/텍스트 임베딩을 확산 백본의 정규화와 교차 어텐션 경로에 통계적으로 정렬되도록 넣는 경량 어댑터를 설계하였다(그림 1-우). 구체적으로, Adaptive Instance Normalization(AdaIN)을 사용해 교차 어텐션의 Q, K 및 LayerNorm 출력 통계를 제어 임베딩 방향으로 치환하고, 블록별 학습 가능한 게이트 α, β, γ를 교차어텐션 전 · 후 정규화와 FFN 경로에 배치해 주입 강도를 조절한다. 학습에서는 확산 U-Net 과 텍스트 인코더를 대부분 동결하고, 제안 어댑터와 ControlNet 의 일부만 미세 조정하여 안정성과 데이터 효율을 확보하였다. 이 구성은 사전학습 생성 프라이어를 보존하면서 LR 구조 힌트, 검색 증강 임베딩을 일관된 분포 정렬 방식으로 결합해 환각을 억제하고 세부 복원을 강화한다.

4. 실험

데이터셋은 ImageNet 약 900k 를 사용하였다. (학습 100k, 참조 DB 800k, 테스트 2k). 저해상도 입력은 bicubic 으로 32×32, 고해상도 타깃은 512×512 로 구성하였다. 비교를 위해 SeeSR, StableSR, CoSeR 3 가지 모델을 본 모델과 동일한 세팅에서 재학습을 진행하였다.

4.1. 정량 평가

정량 평가는 FID / DISTS / LPIPS / CLIP-Score / MANIQA / MUSIQ 을 사용하였다. FID 는 생성 분포와 실제 HR 분포 간 거리를 Inception 특성의 정규분포 근사로 계산하며, 생성 품질의 분포적 타당성을 평가한다. DISTS 와 LPIPS 는 각각 구조, 텍스처 통합 유사도와 딥특징 기반 지각 거리로서, 인간 지각과의 상관이 높다. CLIP-Score 는 CLIP 임베딩 유사도로 텍스트 조건과의 정합도를 참조 없이 평가한다. MANIQA 와 MUSIQ 는 무참조(NR-IQA) 계열로, 초해상 결과의 전반적 지각 품질을 측정한다.

Ours_base 세팅은 ControlNet 에 LR 구조 힌트 + 우리가 만든 텍스트 임베딩만 주입한 텍스트-온리 베이스라인이다. 반면 AdaIN 은 본 논문에서 제안한 삼중 주입을 적용한 경량 어댑터를 포함한다.

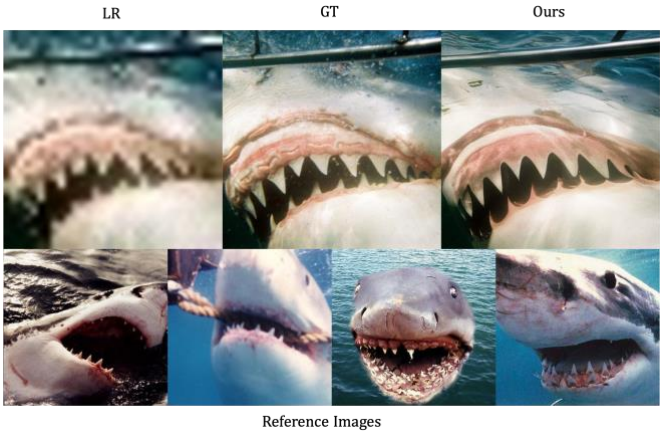
	FID↓	DISTS↓	LPIPS↓	CLIP-Score↑	MANIQA↑	MUSIQ↑
SeeSR(x16)	33.15	0.1940	0.3914	0.7162	0.3781	64.7572
StableSR(x16)	46.39	0.2163	0.4010	0.7116	0.6374	65.7558
CoSeR(x16)	37.97	0.2088	0.4356	0.6910	0.3890	63.3140
Ours_Base	34.61	0.2066	0.4587	0.7551	0.6730	69.8380
AdaIN	29.45	0.1883	0.3936	0.7683	0.6470	68.0969

(표 1) 해당 매트릭들에 대한 정량적 수치.

제안법은 Ours_base 대비 FID↓, DISTS↓, LPIPS↓로 분포, 지각 거리 모두 개선되었고, CLIP-Score↑로 텍스트 조건과의 정합이 높아졌다. 재학습한 SeeSR/CoSeR/CoSeR 와 비교해도, 분포 유사도(FID)와 지각 거리(DISTS/LPIPS)에서 일관되게 우수하거나 경쟁적이며, 텍스트 정합과 NR-IQA 지표에서도 견조한 성능을 확인했다

4.2. 정성 평가

ImageNet 이미지 2,000 장을 테스트로 사용하여 얻은 초해상화 결과와 해당 이미지를 SR 하는 과정에서 사용된 참조 이미지들을 함께 제시한다.



(그림 3) (위) LR/GT/Inference 결과 (아래) 참조 이미지

아래는 재학습한 SeeSR/CoSeR/CoSeR(×16)과의 정성적 비교를 진행한 결과이다.

(그림 4) 왼쪽부터 순서대로 LR, GT 그리고 SeeSR, CoSeR, StableSR, Ours 의 x16 배 SR 결과이다.

참고문헌

[1] J.-B. Alayrac *et al.*, “Flamingo: a visual language model for few-shot learning,” arXiv preprint arXiv:2204.14198, 2022.

[2] X. Chen *et al.*, “PaLI: A jointly-scaled multilingual language-image model,” ICLR, 2023.

[3] A. Asai *et al.*, “Self-RAG: Learning to retrieve, generate, and critique through self-reflection,” ICLR, 2024.

[4] A. Asai, S. Min, Z. Zhong, D. Chen, “Retrieval-based language models and applications,” ACL, 2023.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” NAACL, 2019.

[6] A. Dosovitskiy *et al.*, “An image is worth 16×16 words: Transformers for image recognition at scale,” ICLR, 2021.

[7] W. Fedus, B. Zoph, N. Shazeer, “Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity,” JMLR, 2022.

[8] H. Guo *et al.*, “ReFIR: Grounding large restoration models with retrieval augmentation,” NeurIPS, 2024.

[9] P. Lewis *et al.*, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” NeurIPS, 2020.

[10] J. Li, D. Li, S. Savarese, S. Hoi, “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” ICML, 2023.

[11] S. Li *et al.*, “Enhancing retrieval-augmented generation: A study of best practices,” arXiv preprint arXiv:2501.07391, 2025.

[12] X. Lin *et al.*, “DiffBIR: Toward blind image restoration with generative diffusion prior,” ECCV, 2024.

[13] A. Mallen *et al.*, “When not to trust language models: Investigating effectiveness of parametric and non-parametric memories,” ACL, 2023.

[14] S. Min *et al.*, “FActScore: Fine-grained atomic evaluation of factual precision in long form text generation,” EMNLP, 2023.

[15] A. Radford *et al.*, “Learning transferable visual models from natural language supervision,” ICML, 2021.

[16] O. Ram *et al.*, “In-context retrieval-augmented language models,” TACL, 2023.

[17] R. Rombach *et al.*, “High-resolution image synthesis with latent diffusion models,” CVPR, 2022.

[18] C. Saharia *et al.*, “Image super-resolution via iterative refinement,” PAMI, 2022.

[19] H. Sun *et al.*, “CoSeR: Bridging image and language for cognitive super-resolution,” CVPR, 2024.

[20] A. Vaswani *et al.*, “Attention is all you need,” NeurIPS, 2017.

[21] J. Wang *et al.*, “Exploiting diffusion prior for real-world image super-resolution,” IJCV, 2024.

[22] R. Wu *et al.*, “SeeSR: Towards semantics-aware real-world image super-resolution,” CVPR, 2024.

[23] L. Zhang, A. Rao, M. Agrawala, “Adding conditional control to text-to-image diffusion models,” ICCV, 2023.