

VAE kompression med AVIF

DVA305

**Information – kunskap – vetenskap – etik
[VT] 2026 Grupp 4**



Författare:

Ayman, Martin, Patrik och Salah

Ort och datum:
Västerås, 29 januari 2026

Sammanfattning

Det här avsnittet ska helt enkelt vara en sammanfattning av hela rapporten. En lämplig omfattning är c:a 200 – 250 ord. En bra tumregel är att sammanfattningen ska hållas så kort det går, den ska vara kompakt men tydlig, informativ och väcka intresse.

Summa allt det som är väsentligt i rapporten och presentera de viktigaste resultaten och slutsatserna. Följande bör ingå:

- Presentation och introduktion av området för arbetet.
- Översiktlig presentation av frågeställningen.
- Generell beskrivning av arbetet, hur ni angripit frågeställningen och vad ni har gjort.
- Sammanfattning av resultat och slutsatser.

Det ska inte finnas någon beskrivning av hur rapporten är uppställd och inte några detaljer i sammanfattningen. Sammanfattningen ska kunna läsas helt fristående från resten av rapporten och vända sig till en ganska bred målgrupp läsare. Den ska ge en bra grund för att en läsare ska kunna bedöma om hen är intresserad av att läsa hela rapporten.

När hela rapporten är klar bör ni granska och vid behov revidera sammanfattningen så att den överensstämmer med rapporten.

Innehåll

Inledning	3
Bakgrund	4
Frågeställning	4
Metod	4
Forskningsetiskt ställningstagande	4
Utförande	4
Resultat	5
Diskussion	5
Relaterat arbete	5
Slutsatser	5
Framtida arbete	5
Referenser	6
Referenser	6
A Bilagor	7

Inledning

Behovet att optimera lagring och läsning av bilder har sedan länge varit en stor och viktig del inom datavetenskapen och således föremål för intensiv forskning [?]. I och med informationssamhällets framväxt har det uppstått ett allt större behov av att kunna lagra filerna så kompakt som möjligt för att vi ska kunna hantera denna växande mängd av data på våra enheter [?]. Utvecklingen har framgångsrikt lett till nya algoritmer som på effektiva sätt kunnat inhämta och presentera bildinformationen från olika lagrings- eller strömningsmedia [?].

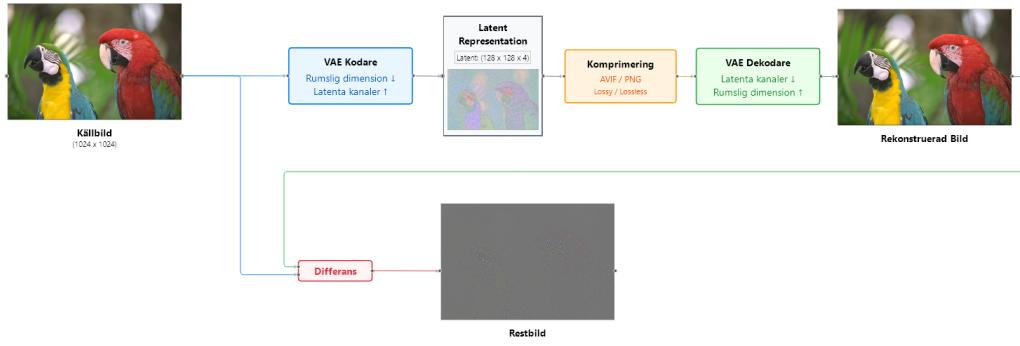
Traditionellt sett har bilder mestadels redigerats i bildbehandlingsprogram såsom Photoshop eller Gimp. Bilden har behandlats utifrån sin rumsliga struktur med pixeln som viktigaste beståndsdel. Bildredigering har i dessa program framförallt handlat om att skala, rotera eller ändra färgen på enskilda pixlar eller grupper av pixlar. Huvudsyftet har varit att möjliggöra ett intuitivt sätt att arbeta med bilder för oss mänskor [?]. I detta sammanhang har populära filformat såsom JPG, PNG och AVIF sakta utformats för att tjäna detta syfte på bästa sätt [?].

Med framväxten av maskininlärning (ML) och djupinlärning (DL) har nya sätt att bearbeta bilder introducerats [?]. Istället för att manipulera pixlar direkt i det visuella planet finns nu möjligheter att arbeta med bildens semantiska innehåll i en s.k. latent representation [?]. En ML-modell kan med denna latenta representation som grund generera en, eller flera visuellt lika (eller olika) bilder på skärmen [?]. Eftersom ML-modeller inte behöver tolka bilder utifrån samma strukturella ordning som mänskor, kan bildinformationen omstruktureras på nya sätt som bättre passar modellens syfte [?]. Målet med dessa ML-baserade metoder är inte i första hand att minimera filstorleken, utan genom att komprimera bort överflödig information underlättar man beräkningsarbetet och ökar på så sätt även prestandan i modellerna [?]. Ser man enbart till filstorlek kontra bildkvalitet är dagens etablerade kompressionsalgoritmer redan mycket effektiva. Till exempel kan dessa med lätthet krympa bildens filstorlek upp till 60 gånger jämfört med originalet [?]. Detta med en kvalitet som för blotta ögat är nästintill omöjlig att särskilja från originalbilden. Dessutom kan sådana komprimerade bildfiler öppnas på bara några millisekunder. Detta till trots finns ingen egentlig fördel med att använda dessa bildformat i ML-modeller. I komprimerat tillstånd är bildinformationen kodad på ett sätt som gör det omöjligt att bearbeta bilden vidare utan att först återföra den till sin fulla storlek, varpå alla fördelarna har gått förlorade. [?].



Figur 1: Exempel på latent

Tack vare att ML-modeller som tidigare nämnts inte behöver ha bilder i ett direkt visuellt format under bearbetningen forskas det på att skapa mer kompakte representationer av bilddata som samtidigt behåller semantisk information. Mycket av den information som krävs enbart för att återge bilden kan utelämnas tills modellen är klar och den slutliga bilden ska visualiseras [?]. Resultatet blir ett kompakt dataformat som ändå behåller det viktigaste för att beskriva vad bilden föreställer. Med ett sådant format kan bilder både ta mindre plats på hårddisken och matas direkt in i ML-modeller utan tunga mellanliggande steg för bearbetning och rekonstruktion [?]. Detta kan i så fall visa sig särskilt värdefullt när man arbetar med väldigt stora bilder eller stora mängder bilddata, såsom till exempel högupplösta fotografier eller video där traditionell hantering skulle vara ineffektiv [?].



Figur 2: Översikt av flödet

Bakgrund

Frågeställning

Uppvisar SDXL:s latenta representation tillräckligt med 'bildlika' egenskaper för att vanliga bildkomprimeringssmetoder (AVIF/PNG) ska fungera för att minska lagringsstorleken utan att bildkvaliteten faller samman vid rekonstruktion?

Metod

Vi väljer 20 bilder från Flickr 8k Datfaset (eller alternativt The Kodak Lossless True Color Image Suite).

Bilderna konverteras med SDXL:s VAE (madebyollin/sdxl-vae-fp16-fix) för att omvandla varje bild till dess latenta rymd, exempelvis 128x128x4@f32 om bilden är 1024x1024xRGB (HxWxC).

Istället för flyttal kvantifierar vi värderna som heltal med 8/10/12/16 bitar och sparar som vanligt bildformat (RGBA) inför nästa steg.

Bilddatan komprimeras med PNG respektive AVIF i tre olika kvalitetslägen, låg/medel/hög (vilket ger olika filstorlekar). AVIF väljs då detta format stödjer fyra kanaler (RGBA), till skillnad från JPG (RGB).

Slutligen återskapas varje latent representation från de komprimerade filerna och omvandlas tillbaka med samma VAE till en RGB**□**bild.

Vi mäter skillnaden uttryckt i PSNR och SSIM mellan den ursprungliga bilden och varje rekonstruktion och jämför dessa med respektive filstorlek.

Forskningsetiskt ställningstagande

Utförande

(Obs: Ni ska inte använda rubriken ovan, utan ersätta den med lämplig rubrik beroende på ert arbete.)

Resultat

Diskussion

Relaterat arbete

Slutsatser

Framtida arbete

Referenser

- [1] T. Kouzelis, I. Kakogeorgiou, S. Gidaris och N. Komodakis, “EQ-VAE: Equivariance Regularized Latent Space for Improved Generative Image Modeling”, i *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, arXiv:2502.09509, vol. 267, Vancouver, Canada: PMLR, 2025.
- [2] A. Horé och D. Ziou, “Image Quality Metrics: PSNR vs. SSIM”, i *Proceedings of the 20th International Conference on Pattern Recognition*, Istanbul, Turkey: IEEE, 2010. DOI: [10.1109/ICPR.2010.579](https://doi.org/10.1109/ICPR.2010.579)
- [3] D. Podell m. fl., “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis”, i *International Conference on Learning Representations*, 2024.

A Bilagor

Övrig information som inte direkt passar in i texten, men som kan vara intressant för en läsare som vill fördjupa sig, placeras ni som bilagor sist i rapporten. Det kan t.ex. vara källkod till en implementation, testdata, bilder av en prototyp etc. Var samtidigt noga med att rapporten ska gå att läsa utan bilagorna.