Multi-Modal Masked Autoencoders for Learning Image-Spectrum Associations for Galaxy Evolution and Cosmology

Morgan Himes[1]
morganhimes@ucla.edu

Samiksha Krishnamurthy[2]
samikris24@ucla.edu

Andrew Lizarraga[3]
andrewlizarraga@ucla.edu

Srinath Saikrishnan[4]
srinathsai22@ucla.edu

Vikram Seenivasan[1]
vikrams25@ucla.edu

Jonathan Soriano[1]
jonathansoriano@astro.ucla.edu

Ying Nian Wu[3]
ywu@stat.ucla.edu

Tuan Do[1]
tdo@astro.ucla.edu

[1] Department of Physics and Astronomy, UCLA, Los Angeles, CA 90095
[2] Department of Electrical and Computer Engineering, UCLA, Los Angeles, CA 90095
[3] Department of Statistics and Data Science, UCLA, Los Angeles, CA 90095
[4] Department of Computer Science, UCLA, Los Angeles, CA 90095

Abstract

Upcoming surveys will produce billions of galaxy images but comparatively few spectra, motivating models that learn cross-modal representations. We build a dataset of 134,533 galaxy images (HSC-PDR2) and spectra (DESI-DR1) and adapt a Multi-Modal Masked Autoencoder (MMAE) to embed both images and spectra in a shared representation. The MMAE is a transformer-based architecture, which we train by masking 75% of the data and reconstructing missing image and spectral tokens. We use this model to test three applications: spectral and image reconstruction from heavily masked data and redshift regression from images alone. It recovers key physical features, such as galaxy shapes, atomic emission line peaks, and broad continuum slopes, though it struggles with fine image details and line strengths. For redshift regression, the MMAE performs comparably or better than prior multi-modal models in terms of prediction scatter even when missing spectra in testing. These results highlight both the potential and limitations of masked autoencoders in astrophysics and motivate extensions to additional modalities, such as text, for foundation models.

1 Introduction

Next-generation astronomical surveys will image billions of galaxies, producing catalogs thousands of times larger than existing datasets to study how galaxies form and evolve across cosmic time [17, 19]. Galaxy spectra encode physically relevant information, such as redshift (z), a measure of how much a galaxy's light has been stretched by the universe's expansion. However, obtaining a spectrum can take roughly 100 times longer than capturing an image. While spectroscopic redshifts measured from a galaxy's spectrum are the most accurate, astronomers rely on photometric redshifts from images for this reason [22]. Since spectroscopy is infeasible at the scale of upcoming surveys, this motivates the development of models that leverage imaging to reconstruct missing spectra.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

Table 1: Data summary.

Dataset	Data Type	Source Count	z (90th pct.)	z (max)	i-mag (90th pct.)
DESI DR1	Spec, z	20,283,824	1.343	6.857	–
GalaxiesML	Img	286,401	1.155	4.000	22.171
GalaxiesML-Spectra	Spec, Img, z	134,533	1.581	4.119	20.635

Note: Disagreement between HSC and DESI spectroscopic redshifts causes a discrepancy in z (max).

Related Work Machine learning methods have been used in astronomy to address these challenges. Multi-layer perceptrons, convolutional, and Bayesian neural networks (MLP, CNN, BCNN) have been widely applied for photometric redshift estimation [e.g., 8, 7, 5, 4, 25, 24, 11, 26, 27, 18]. Additionally, QuasarNET applies CNNs to spectra for redshift estimation while AstroMAE uses masked autoencoders (MAEs) on galaxy images [6, 15]. Diffusion models have been used for generating galaxy images across redshifts [20, 14, 29]. Multi-modal efforts include AstroCLIP, which jointly embeds images and spectra for tasks such as classification and redshift estimation [23].

2 Contributions

We adapt a Multi-Modal Masked Autoencoder (MMAE) [2] for reconstructing spectra and images from incomplete data, an example of multimodal learning with missing modality [30]. We build a dataset of spectra and images for training and testing with partially or fully masked spectra, reflecting the real-world constraints of imaging surveys. Redshift regression serves as an auxiliary task to demonstrate the utility of the learned representations. We extend redshift regression to $z \sim 4$, substantially exceeding the redshift range ($z \lesssim 0.5$) explored in prior models [e.g. 23, 15].

Masked autoencoders have been applied in scientific contexts such as medical imaging [31], but their use for multi-modal astronomical data is underexplored. AstroMAE [15] employs the use of an MAE for redshift regression on galaxy images, but our work is, to our knowledge, the first instance of using an MAE for joint multi-modal reconstruction on images and spectra in astronomy.

3 Data

We assembled a multi-modal dataset[1] (referred to as GalaxiesML-Spectra) of 134,533 galaxies, each with 5-band images, 1D spectra, and spectroscopic redshifts. The maximum redshift is $z = 4.119$, providing greater coverage than previous datasets used for similar studies. The images are available in two size options (64×64 or 127×127 pixels), but our model was trained using the 64×64 version. All galaxy images in the dataset are contained in GalaxiesML, a machine learning image dataset built upon the Hyper-Suprime-Cam (HSC) Survey PDR2, covering g, r, i, z, y bands with a median seeing in the i-band of about 0.6 arcsec [12, 1]. GalaxiesML is biased towards lower redshifts and brighter magnitudes, as shown in Table 1. Spectra and redshifts were acquired from the DESI Data Release 1 (DR1), the largest spectroscopic galaxy redshift survey to date [9]. The spectra span a wavelength range of 3600–9824 Å with a spectral resolution of 2000 (at 3600 Å) to 5500 (at 9800 Å). GalaxiesML was cross-matched with DESI DR1. Stars and sources with duplicates, inconsistent redshifts, or >80% missing spectral data were removed, reducing the cross-matched dataset from 136,939 sources to 134,533. Spectral artifacts were removed and replaced with the median flux. A summary of the datasets is available in Table 1. Our released dataset contains unaugmented spectra after quality cuts and artifact removal. Data augmentation is performed as described in Section 4.

4 Methods

Our MMAE model[2] is a transformer-based architecture that models images and spectra to perform image and spectrum reconstruction and redshift regression (see visual representation in Figure 1). A patch-based tokenization strategy is used for both modalities. We follow the Vision Transformer

[1] Available at Zenodo (access link)

[2] Code available with GitHub (access link)

(ViT) [13] formulation to convert images into patch tokens. With a 2D convolution, galaxy images of shape $64 \times 64 \times 5$ are divided into $8 \times 8 \times 5$ patches and projected into a 256-dimensional embedding. A 2D learnable positional embedding of dimension 256 is added to preserve spatial ordering. Before passing spectra into the model, we perform data augmentation by normalizing as described in [6] and downsampling from 7783 to 259 flux pixels. An analogous 1D patch embedding process is applied for spectra, using patches of length 8 and a linear projection.

A random 75% subset of image and spectra patch tokens is zeroed out for each galaxy. After masking, each modality is separately encoded using a 1D transformer encoder with depth 4, 8 attention heads, and a dropout rate of 0.1. For modality fusion, we employ four layers of cross-attention blocks in which image features query spectral features and vice versa. This allows spectral structure to guide the interpretation of morphological features, informing the model as it reconstructs missing tokens from unmasked tokens in both modalities. After fusion, the tokens are aggregated via attention pooling to produce global image and spectrum embeddings which are concatenated into a joint representation.

The joint representation feeds into three task-specific heads: image and spectrum decoders, and a redshift regression head. The decoders are MLPs with GeLU activations [16] and dropout, while the regression head maps the fused embedding to a scalar redshift. Unlike the common practice of performing regression after MAE feature extraction, we integrate redshift prediction directly into masked autoencoding, which is novel in a multi-modal setting.

The dataset is randomly split into 70% training, 15% validation, and 15% testing. In training, 50% of the spectra are randomly zeroed entirely to avoid overreliance and simulate the real-life missing modality case. We optimize with AdamW [21] (weight decay 0.01, learning rate 0.0001), gradient clipping, and a loss function that is a weighted sum of three terms: mean-squared error reconstruction losses on masked regions for images and spectra and a redshift loss defined as $\mathcal{L}_z = 1 - \frac{1}{1+(dz/0.15)^2}$ where $dz = (z_{\text{pred}} - z_{\text{spec}})/(1 + z_{\text{spec}})$ and z_{pred} and z_{spec} are the predicted and spectroscopic redshifts [28]. The weights for the image, spectrum, and redshift losses are tunable parameters that have been set at 0.1, 0.01, and 1.0, respectively, to account for the difference in scale of the individual loss curves as identified through preliminary parameter tuning. In further studies, we will explore dynamic weighting with the incorporation of physics-aware terms such as line centers, widths, and ratios, continuum slope metrics, and perceptual or structure losses. The impact of masking probability itself will be explored through a study of model performance as a function of masking ratio, and we will explore physically-motivated masking including bandpass gaps and pixel-level instrument errors.

5 Results and discussion

5.1 Feature reconstruction

We evaluated the model's ability to reconstruct spectral features and images on a held-out test set of 20,181 galaxies. For image reconstruction, the model reproduces the shape and color of galaxies but struggles with fine morphological details in nearby galaxies ($z < 0.1$) and sky background noise. For spectral reconstruction, the model captures the broad spectral continuum shapes, even in cases where the entire spectrum is masked at test time, but fails to reproduce the random noise in the original spectra.

The model reproduces the locations of some common emission lines. For low-redshift ($z \sim 0.1$) galaxies, it roughly reproduces the position of the H-α (6563 Å rest frame) emission line, as in Figure 1. At high redshift ($z \sim 2$), the model reconstructs the location of Lyman-α (1215.67 Å rest frame) and, in some cases, C IV (1549 Å rest frame); see Figure 2. In all cases, line widths are systematically overestimated and the heights are underestimated. Reconstruction of other emission lines is poor, and the model occasionally wrongly generates Lyman-α emission for lower redshift, compact galaxies. The model does not capture the ratios of line fluxes which is a diagnostic tool used by astronomers to determine physical properties [3]. While the MMAE has learned to capture particularly common lines and global structure, spectrum generation remains challenging for weaker and less common emission lines or ambiguous morphology. Further studies will be conducted to assess model's ability to reproduce masked data using quantitative reconstruction metrics.

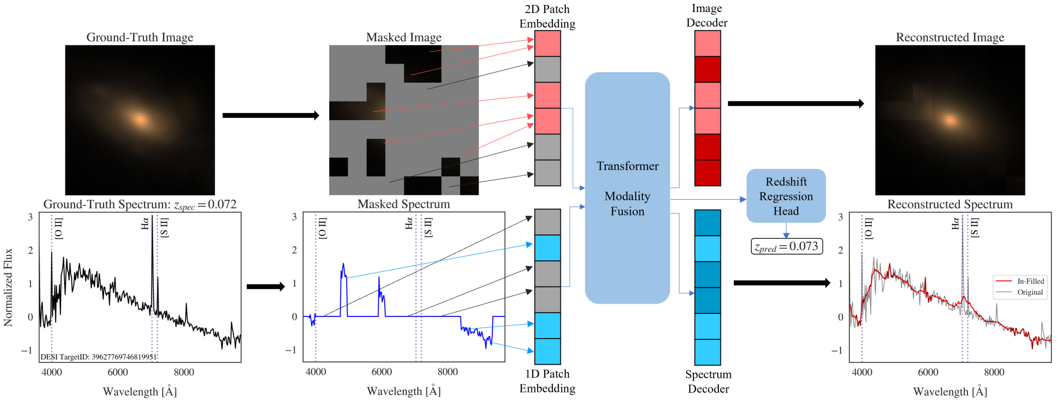

Figure 1: The model's architecture and reconstruction process are shown for a low redshift source with 75% masking of both modalities. We measure the peak location, amplitude, and width of H-α in the augmented and generated spectra. The H-α line has an observed center at 7042.8 Å with a height of 3.04 and a width of 34.5 Å, while the model reconstructed it at 7066.8 Å with a height of 0.62 and a width of 528 Å.

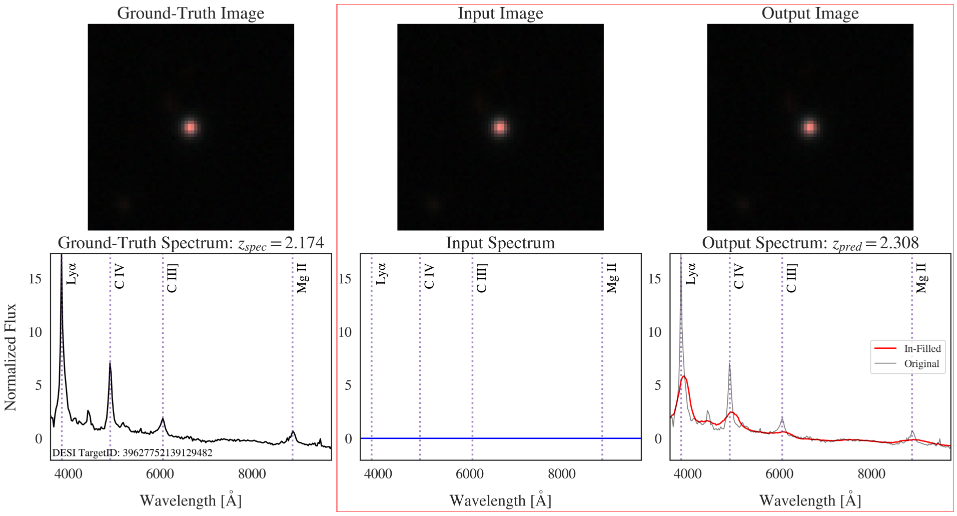

Figure 2: The model's reconstruction process is shown for a high redshift source with a fully masked spectrum and a fully unmasked image. We measure the peak location, amplitude, and width of Lyman-α and C IV in the augmented and generated spectra. The Lyman-α line has an observed center at 3851.6 Å with a height of 17.24 and a width of 48 Å, compared to a reconstructed center at 3923.6 Å, height 5.84, and width 312 Å. Similarly, the C IV line has an observed center at 4907.6 Å, height 7.07, and width 72 Å, while the reconstructed line is at 4931.6 Å, with height 2.48 and width 648 Å.

5.2 Redshift regression

Redshift regression was performed on the same test set. The model was provided with a 25% masked galaxy image and a fully masked spectrum, reflecting the observational constraints posed by upcoming surveys, which will not have spectra for the vast majority of galaxy images. Notably, masking 25% of the image produces better overall redshift regression results than supplying the model with the entire image. We suspect this is because slight masking serves as a form of regularization, preventing the model from over-fitting to small-scale features. The model achieves higher accuracy at lower redshift ($z \lesssim 1$) but degrades for higher redshift sources; we attribute this to the limited amount of high redshift data which restricts the model's ability to generalize. We intend to incorporate additional high redshift sources with imaging from the DESI Legacy Imaging Surveys to supplement this gap

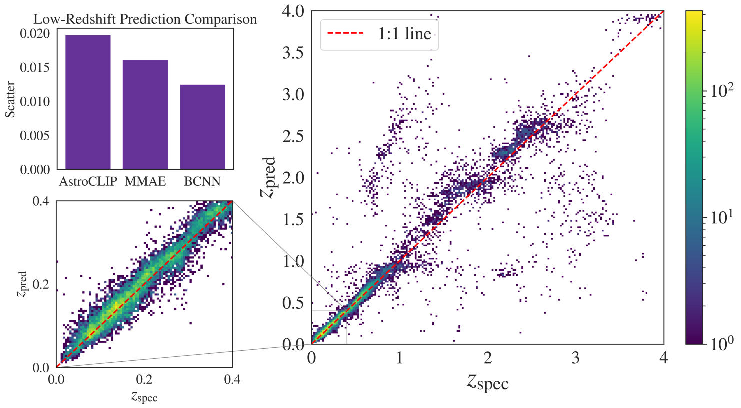

Figure 3: The model's redshift regression results for the entire redshift range are shown (right). The redshift predictions were obtained from test data that had 25% of the image masked and 100% of the spectrum masked. The low-redshift regime used for comparison to AstroCLIP [23] is shown in more detail in the bottom left. The top left panel shows the scatter of the MMAE compared to AstroCLIP and a BCNN model [18] for this low-redshift regime. Lower scatter corresponds to more precise predictions.

[10]. The predicted versus true redshift relation (see Figure 3) exhibits distinct step-like structure possibly corresponding to redshift intervals where strong spectral features shift into or out of the spectrograph's wavelength range, such as Lyman-α at $z \sim 2$.

We find that our model achieves better or comparable results to other models using images and spectra for redshift regression in terms of scatter, measured as the normalized median absolute deviation. The spectrum was fully masked in testing to resemble real-world survey conditions. Our MMAE model attains $\sigma_{\text{NMAD}} = 0.016$ when tested with 25% image masking, which is slightly better than AstroCLIP [23] ($\sigma_{\text{NMAD}} = 0.020$) in the low-redshift regime ($z \lesssim 0.4$, the limit of the AstroCLIP results available to the authors). This comparison is not entirely equivalent, as AstroCLIP was trained to a contrastive objective of aligning images and spectra and did not optimize for redshift prediction. However, we include this comparison as AstroCLIP is the strongest publicly available multi-modal model utilizing the same two modalities. Furthermore, evaluating on the shared task of image-only redshift prediction demonstrates how incorporating masked reconstruction influences downstream inference even with differing training objectives. Masking 25% of each image produces better redshift regression results than supplying the model with the entire image, in which case $\sigma_{\text{NMAD}} = 0.026$, no longer out-performing other methods. Additionally, a fine-tuned BCNN [18] achieves lower scatter overall ($\sigma_{\text{NMAD}} = 0.012$), indicating that our transformer-based approach remains less robust. This outcome is consistent with prior work suggesting that transformer-based architectures underperform relative to inception-style convolutional models for redshift prediction [15].

Conclusion This work demonstrates the potential of using a masked autoencoder on images and spectra for applications in galaxy evolution and cosmology. However, the current model has key limitations that limit its generalizability, as it struggles to reproduce fine details of images and spectra and shows degraded redshift regression accuracy at higher redshift. These limitations highlight the need for more physically motivated training objectives and robust architectures capable of capturing subtle but astrophysically meaningful features. Future work will incorporate physics-aware loss terms and observationally realistic masking strategies. In addition, this work can be easily extended to additional modalities such as such as textual metadata and natural language descriptions for foundation models for astronomy. Such models will be essential for scaling to the unprecedented data volumes and observational constraints of upcoming surveys.

Acknowledgments and Disclosure of Funding

Partial support for this work was provided by the Alfred P. Sloan Foundation and the NSF DGE-2034835.

This research used data obtained with the Dark Energy Spectroscopic Instrument (DESI). DESI construction and operations is managed by the Lawrence Berkeley National Laboratory. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High-Energy Physics, under Contract No. DE–AC02–05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract. Additional support for DESI was provided by the U.S. National Science Foundation (NSF), Division of Astronomical Sciences under Contract No. AST-0950945 to the NSF's National Optical-Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Humanities, Science and Technology of Mexico (CONAHCYT); the Ministry of Science and Innovation of Spain (MICINN), and by the DESI Member Institutions: www.desi.lbl.gov/collaborating-institutions. The DESI collaboration is honored to be permitted to conduct scientific research on I'oligam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation, the U.S. Department of Energy, or any of the listed funding agencies.

The Hyper Suprime-Cam (HSC) collaboration includes the astronomical communities of Japan and Taiwan, and Princeton University. The HSC instrumentation and software were developed by the National Astronomical Observatory of Japan (NAOJ), the Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU), the University of Tokyo, the High Energy Accelerator Research Organization (KEK), the Academia Sinica Institute for Astronomy and Astrophysics in Taiwan (ASIAA), and Princeton University. Funding was contributed by the FIRST program from the Japanese Cabinet Office, the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Society for the Promotion of Science (JSPS), Japan Science and Technology Agency (JST), the Toray Science Foundation, NAOJ, Kavli IPMU, KEK, ASIAA, and Princeton University. This paper makes use of software developed for the Large Synoptic Survey Telescope. We thank the LSST Project for making their code available as free software at http://dm.lsst.org. This paper is based [in part] on data collected at the Subaru Telescope and retrieved from the HSC data archive system, which is operated by the Subaru Telescope and Astronomy Data Center (ADC) at National Astronomical Observatory of Japan. Data analysis was in part carried out with the cooperation of Center for Computational Astrophysics (CfCA), National Astronomical Observatory of Japan. The Subaru Telescope is honored and grateful for the opportunity of observing the Universe from Maunakea, which has the cultural, historical and natural significance in Hawaii.

References

[1] H. Aihara et al. Second data release of the Hyper Suprime-Cam Subaru Strategic Program. *PASJ*, 71(6):114, Dec. 2019. doi: 10.1093/pasj/psz103.

[2] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir. MultiMAE: Multi-modal Multi-task Masked Autoencoders. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, volume 13697, pages 348–367. Springer Nature Switzerland, Cham, 2022. ISBN 978-3-031-19835-9 978-3-031-19836-6. doi: 10.1007/978-3-031-19836-6_20. URL `https://link.springer.com/10.1007/978-3-031-19836-6_20`. Series Title: Lecture Notes in Computer Science.

[3] J. A. Baldwin, M. M. Phillips, and R. Terlevich. Classification parameters for the emission-line spectra of extragalactic objects. *PASP*, 93:5–19, Feb. 1981. doi: 10.1086/130766.

[4] R. Beck, L. Dobos, T. Budavári, A. S. Szalay, and I. Csabai. Photometric redshifts for the SDSS Data Release 12. *Monthly Notices of the Royal Astronomical Society*, 460(2):1371–1381, Aug. 2016. ISSN 0035-8711. doi: 10.1093/mnras/stw1009. URL `https://doi.org/10.1093/mnras/stw1009`.

[5] C. Bonnett. Using neural networks to estimate redshift distributions. An application to CFHTLenS. *Monthly Notices of the Royal Astronomical Society*, 449 (1):1043–1056, May 2015. ISSN 1365-2966, 0035-8711. doi: 10.1093/mnras/stv230. URL http://academic.oup.com/mnras/article/449/1/1043/1302387/Using-neural-networks-to-estimate-redshift.

[6] N. Busca and C. Balland. QuasarNET: Human-level spectral classification and redshifting with Deep Neural Networks, Aug. 2018. URL http://arxiv.org/abs/1808.09955. arXiv:1808.09955 [astro-ph].

[7] M. Carrasco Kind and R. J. Brunner. TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests. *Monthly Notices of the Royal Astronomical Society*, 432(2):1483–1501, June 2013. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stt574. URL http://academic.oup.com/mnras/article/432/2/1483/1029454/TPZ-photometric-redshift-PDFs-and-ancillary.

[8] A. A. Collister and O. Lahav. ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. *PASP*, 116(818):345–351, Apr. 2004. doi: 10.1086/383254.

[9] DESI Collaboration et al. Data Release 1 of the Dark Energy Spectroscopic Instrument, Mar. 2025. URL http://arxiv.org/abs/2503.14745. arXiv:2503.14745 [astro-ph].

[10] A. Dey et al. Overview of the DESI Legacy Imaging Surveys. *The Astronomical Journal*, 157 (5):168, May 2019. doi: 10.3847/1538-3881/ab089d.

[11] B. Dey, B. H. Andrews, J. A. Newman, Y.-Y. Mao, M. M. Rau, and R. Zhou. Photometric Redshifts from SDSS Images with an Interpretable Deep Capsule Network. *Monthly Notices of the Royal Astronomical Society*, 515(4):5285–5305, Aug. 2022. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stac2105. URL http://arxiv.org/abs/2112.03939. arXiv:2112.03939 [astro-ph].

[12] T. Do, B. Boscoe, E. Jones, Y. Q. Li, and K. Alfaro. GalaxiesML: a dataset of galaxy images, photometry, redshifts, and structural parameters for machine learning, Sept. 2024. URL http://arxiv.org/abs/2410.00271. arXiv:2410.00271 [astro-ph].

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

[14] X. Fan, H. Tang, Y. Zeng, M. B. N. Kouwenhoven, and G. Zeng. Category-based galaxy image generation via diffusion models, 2025. URL https://arxiv.org/abs/2506.16255.

[15] A. D. Fathkouhi and G. C. Fox. AstroMAE: Redshift Prediction Using a Masked Autoencoder with a Novel Fine-Tuning Architecture, Sept. 2024. URL http://arxiv.org/abs/2409.01825. arXiv:2409.01825 [cs].

[16] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus), 2023. URL https://arxiv.org/abs/1606.08415.

[17] Z. Ivezic et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *The Astrophysical Journal*, 873(2):111, Mar. 2019. ISSN 0004-637X, 1538-4357. doi: 10.3847/1538-4357/ab042c. URL https://iopscience.iop.org/article/10.3847/1538-4357/ab042c.

[18] E. Jones, T. Do, Y. Q. Li, K. Alfaro, J. Singal, and B. Boscoe. Redshift Prediction with Images for Cosmology Using a Bayesian Convolutional Neural Network with Conformal Predictions. *The Astrophysical Journal*, 974(2):159, Oct. 2024. ISSN 0004-637X, 1538-4357. doi: 10.3847/1538-4357/ad6d5a. URL https://iopscience.iop.org/article/10.3847/1538-4357/ad6d5a.

[19] R. Laureijs et al. Euclid definition study report, 2011. URL https://arxiv.org/abs/1110.3193.

[20] A. Lizarraga, E. Hanchen Jiang, J. Nowack, Y. Q. Li, Y. Nian Wu, B. Boscoe, and T. Do. Understanding Galaxy Morphology Evolution Through Cosmic Time via Redshift Conditioned Diffusion Models. *arXiv e-prints*, art. arXiv:2411.18440, Nov. 2024. doi: 10.48550/arXiv.2411.18440.

[21] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019. URL `https://arxiv.org/abs/1711.05101`.

[22] J. A. Newman and D. Gruen. Photometric Redshifts for Next-Generation Surveys. *Annual Review of Astronomy and Astrophysics*, 60(1):363–414, Aug. 2022. ISSN 0066-4146, 1545-4282. doi: 10.1146/annurev-astro-032122-014611. URL `http://arxiv.org/abs/2206.13633`. arXiv:2206.13633 [astro-ph].

[23] L. Parker, F. Lanusse, S. Golkar, L. Sarra, M. Cranmer, A. Bietti, M. Eickenberg, G. Krawezik, M. McCabe, R. Ohana, M. Pettee, B. R.-S. Blancard, T. Tesileanu, K. Cho, and S. Ho. AstroCLIP: A Cross-Modal Foundation Model for Galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011, June 2024. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stae1450. URL `http://arxiv.org/abs/2310.03024`. arXiv:2310.03024 [astro-ph].

[24] J. Pasquet, E. Bertin, M. Treyer, S. Arnouts, and D. Fouchez. Photometric redshifts from SDSS images using a convolutional neural network. *Astronomy & Astrophysics*, 621:A26, Jan. 2019. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201833617. URL `https://www.aanda.org/10.1051/0004-6361/201833617`.

[25] I. Sadeh, F. B. Abdalla, and O. Lahav. ANNz2 - photometric redshift and probability distribution function estimation using machine learning. *Publications of the Astronomical Society of the Pacific*, 128(968):104502, Oct. 2016. ISSN 0004-6280, 1538-3873. doi: 10.1088/1538-3873/128/968/104502. URL `http://arxiv.org/abs/1507.00490`. arXiv:1507.00490 [astro-ph].

[26] S. Schuldt, S. H. Suyu, R. Cañameras, S. Taubenberger, T. Meinhardt, L. Leal-Taixé, and B. C. Hsieh. Photometric redshift estimation with a convolutional neural network: NetZ. *Astronomy & Astrophysics*, 651:A55, July 2021. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202039945. URL `https://www.aanda.org/articles/aa/abs/2021/07/aa39945-20/aa39945-20.html`. Publisher: EDP Sciences.

[27] J. Soriano, S. Saikrishnan, V. Seenivasan, B. Boscoe, J. Singal, and T. Do. Using different sources of ground truths and transfer learning to improve the generalization of photometric redshift estimation, 2024. URL `https://arxiv.org/abs/2411.18054`.

[28] M. Tanaka, J. Coupon, B.-C. Hsieh, S. Mineo, A. J. Nishizawa, J. Speagle, H. Furusawa, S. Miyazaki, and H. Murayama. Photometric redshifts for Hyper Suprime-Cam Subaru Strategic Program Data Release 1. *PASJ*, 70:S9, Jan. 2018. doi: 10.1093/pasj/psx077.

[29] T. Vičánek Martínez, N. Baron Perez, and M. Brüggen. Simulating images of radio galaxies with diffusion models. *Astronomy & Astrophysics*, 691:A360, Nov. 2024. ISSN 1432-0746. doi: 10.1051/0004-6361/202451429. URL `http://dx.doi.org/10.1051/0004-6361/202451429`.

[30] R. Wu, H. Wang, H.-T. Chen, and G. Carneiro. Deep Multimodal Learning with Missing Modality: A Survey, Oct. 2024. URL `http://arxiv.org/abs/2409.07825`. arXiv:2409.07825 [cs].

[31] L. Zhou, H. Liu, J. Bae, J. He, D. Samaras, and P. Prasanna. Self pre-training with masked autoencoders for medical image classification and segmentation, 2023. URL `https://arxiv.org/abs/2203.05573`.