


Comparison of Augmentation Techniques with Stable Diffusion for Aircraft Identification

Ryan P. O'Shea ^{*}, Gurpreet Singh, Ari B. Goodman, Thomas J. Keane, Michael A. Brenner, Christopher J. Jaworowski, Tushar A. Patel, and James T. Hing

Naval Air Warfare Center, Aircraft Division, Lakehurst, Lakehurst, NJ 08757, USA

Email: ryan.p.oshea3.civ@us.navy (R.P.O.); gurpreet.singh25.civ@us.navy (G.S.); ari.b.goodman.civ@us.navy (A.B.G.); thomas.j.keane12.civ@us.navy (T.J.K.); michael.a.brenner10.civ@us.navy (M.A.B.); christopher.j.jaworowski.civ@us.navy (C.J.J.); tushar.a.patel2.civ@us.navy (T.A.P.); james.t.hing.civ@us.navy (J.T.H.)

^{*}Corresponding author

Abstract—Many machine learning applications are constrained by limited quantity, quality, and variance of the collected real-world datasets used for training and evaluation. In this work, authors leveraged generative artificial intelligence techniques to extend the amount of data available to train and evaluate Convolutional Neural Networks (CNNs) for object detection and classification. Stable Diffusion was used as the core augmentation algorithm alongside traditional machine vision techniques. A variety of augmentation techniques were compared in terms of their impact on training and evaluating CNNs. The augmented images were used in part to train CNNs to improve the performance of detection models when evaluated on real-world images. Additional experiments were conducted which quantified the prediction performance on real-world data by measuring the performance on similar synthetic data. Within these experiments, various ratios of synthetic and real-world images were used to train networks which were then evaluated on real-world and synthetic holdout datasets.

Keywords—computer vision, synthetic data, generative models, validation and verification, neural style transfer

I. INTRODUCTION

When developing computer vision systems, domain coverage in the training dataset is paramount to ensure the deployed system can operate in the target environment. Unfortunately for many applications, relevant training data from the target environment does not exist or is not plentiful enough to support the creation of an effective computer vision system. Traditionally, synthetic data, usually created in a simulation environment, is used to bolster the training dataset with both common and rare operational examples. While effective for capturing a wide range of target scenarios, the effectiveness of synthetic data is diminished by the simulation to reality gap that separates the two domains. Within the visual domain, modern high-fidelity simulators are often used to generate synthetic data that closely resembles reality. While these solutions help to shrink the simulation to reality gap, they do not close it

entirely and often require the creation of high-fidelity assets to achieve results that appear realistic.

Both traditional machine vision and generative model-based techniques are frequently implemented to augment synthetic image data with the properties of other domains. The latest generation of generative image models, Latent Diffusion Models (LDMs) [1], have shown great promise in text to image generation and image-to-image translation tasks. This work focuses primarily on the use of these generative models for image-to-image translation with the overall goal of performing neural style transfer [2]. We expand upon existing work in the field of neural style transfer by applying it to the problem space of detecting and tracking aircraft across large shifts in the visual domain. A pipeline for image augmentation was developed to perform neural style transfer on synthetic images of aircraft on an aircraft carrier deck. The output of the pipeline is an augmented dataset with the “style” of images from a real-world camera system applied to it. In particular, the contributions of this paper are as follows:

- Comparison of the effectiveness of real, synthetic, and augmented data for the training of an object detection model that is robust to visual domain shift.
- A unique augmentation pipeline for performing localized style blending of synthetic objects inserted into an image of the target operating domain.
- Preliminary results on the effectiveness of using synthetic and augmented data to predict performance of an object detection model on unseen domains.

This paper is organized as follows: Section I covers the context for the problem space and motivation for the work, Section II provides an overview of relevant research, Section III explains data generation and evaluation methods, Section IV presents the results of the data augmentation experiments in the context of the paper's hypothesis, Section V concludes the paper by summarizing the work and identifying future work.

II. LITERATURE REVIEW

A. Synthetic Data

The need for large amounts of labeled data needed to train effective deep learning computer systems has led many to adopt synthetic data to supplement their training sets. As a result, a wide array of tools for generating automatically labeled synthetic data have been created for a variety of application domains [3–6]. While effective for simply increasing the number of training samples, synthetic data often falls victim to the reality gap where the differences between synthetic and real data prevent systems from learning highly generalizable object representations.

B. Style Transfer

To help bridge the reality gap between synthetic and real data, various methods for augmenting a synthetic image with features of a real image have been developed. This method of applying the style of one image to another falls under the domain of neural style transfer [7]. Gatys *et al.* [8] originally used convolutional neural networks to encode the style information of one image and apply it another image to create a new image in the encoded style. Other methods explore more traditional signal processing methods for perturbing specific components within an image to improve domain generalization [9].

C. Generative Models

The creation of Generative Adversarial Networks (GANs) [10] led to a significant amount of derivative work that utilizes adversarial generation techniques to apply desired styles and properties to an image. Adversarial generation based style transfer techniques in [11–21] utilize a variety of GAN based methods for transferring a target style onto both real and synthetic images. Isola *et al.* [22] utilize Conditional GANs to perform image to image translation on common conditional images like canny edge and segmentation maps. Recently, denoising diffusion models [23, 24], have shown impressive results in neural style transfer

tasks. Large text to image models build upon image diffusion models by adding textual input as a way to condition and control the denoising diffusion process to generate a specific image output [1, 25–27]. Zhang *et al.* [28] introduce the use of conditional control maps as another form of guidance for controlling the image diffusion process. As in Ref. [22], conditional generation significantly improves object fine detail saliency during the style transfer process which is critical for training object detection models.

III. MATERIALS AND METHODS

A. Synthetic Data Generation

The Unity game engine [29] was utilized to create a synthetic representative of an aircraft carrier deck environment. Examples from the synthetic environment and the real-world carrier that it was modeled after, are shown in Fig. 1. Three distinct visual domains in the target environment were identified and subsequently replicated in the synthetic environment. These visual domains represent our target domains for performing style transfer. Lower fidelity assets were intentionally used to quickly mockup equivalent viewpoints and asset locations from the real-world images while not focusing on making the simulation appear realistic. Three distinct camera views were used to provide coverage of the entire deck in the simulation to replicate the coverage generally provided by a potential real world camera system. Within each view, synthetic F/A-18Es were spawned at random locations at least partially in view of the camera with their yaws randomized.

Exactly 500 images were generated for each view in a given domain resulting in 1500 images per domain split equally across the three different camera views. In addition to generating the RGB images used for training the object detection model, the simulator also outputs an aircraft segmentation image and scene depth map which are used during the data augmentation process. The maps, shown alongside other inputs to the diffusion model in Fig. 2, were generated using shaders in the Unity Engine to produce different rendering types.

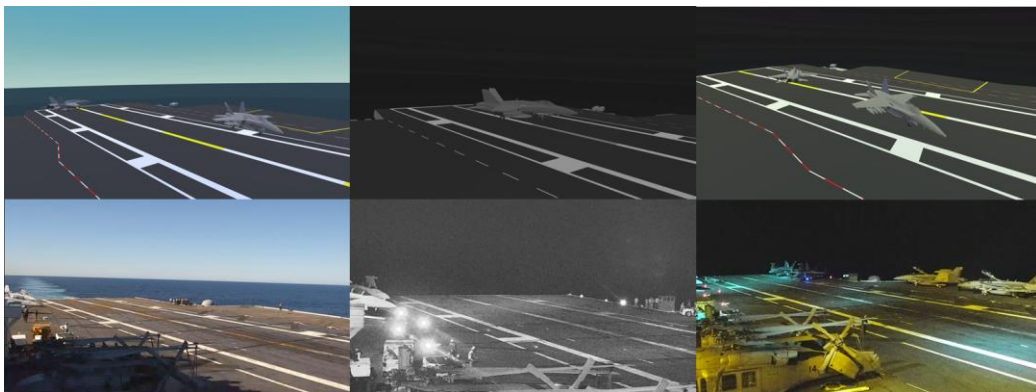


Fig. 1. A comparison of the synthetic carrier deck environment (Top) and the carrier deck that it was modeled after (Bottom). These images show representatives from left to right of the “Day Time”, “Gray Night” and “Yellow Night” domains which will form the basis of the style transfer experiments.

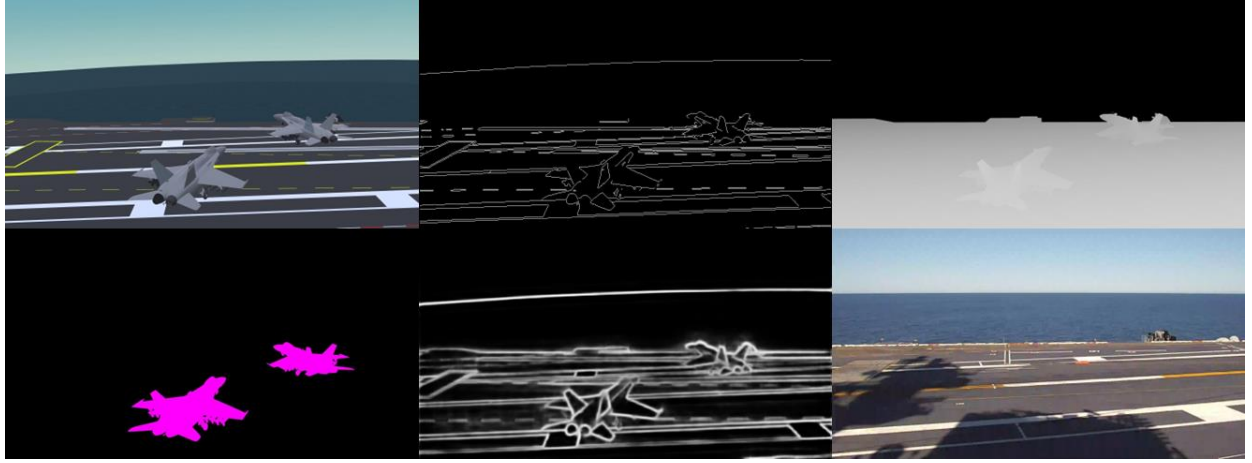


Fig. 2. Inputs into Stable Diffusion and ControlNet. Synthetic image (Top Left), Canny edges (Top Center), Unity generated depth map (Top Right), Unity generated aircraft segmentation map (Bottom Left), Soft Edges (Bottom Center), Reference control image (Bottom Right).

B. Synthetic Aircraft Augmentation

Two types of datasets were sent through the augmentation pipeline:

1. Full synthetic images directly from the simulator
2. Synthetic aircraft inserted into real world image of the flight deck

The first type required no other forms of preprocessing and was used directly by the data augmentation pipeline. The second method drew inspiration from inpainting techniques [30–32] and attempts to perform localized style blending on a synthetic aircraft inserted into a real-world image. Synthetic insert images, seen in Fig. 3, were generated using logical image operator functions provided by the OpenCV library [33] to transplant synthetic aircraft into the real-world image by using the generated aircraft segmentation map as a logical mask.

C. Generation Methods

This paper presents the results from two different stable diffusion based style transfer techniques:

1. Full image style transfer using a synthetic image
2. Targeted image style blending using an inpainting mask on images of synthetic aircraft inserted into real world images

All augmentations were made using the Stable Diffusion [1] image to image translation functionality in conjunction with ControlNet [28] for adding conditional controls to the generation. This involved using images from one of the two types of datasets described in the previous section as input directly into the diffusion model. The custom stable diffusion 1.5 model, Realistic Vision

V5.1, was used as the specific model for the generations. Images were translated one at a time with no batch processing due to limitations in the open-source API at the time. The open-source AUTOMATIC1111 Stable-Diffusion-Webui API was leveraged to access the image-to-image translation capabilities of stable diffusion and ControlNet in order to perform style transfer on all of the tested datasets. Generation parameters, besides denoising strength and inpainting use, were kept unchanged across all dataset augmentations. The key parameters used within the API can be found in Table I.

TABLE I. STABLE DIFFUSION INPUT PARAMETERS FOR BOTH GENERATION METHODS

Parameter	Augmentation Method 1	Augmentation Method 2
Prompt	(grainy:1.5) picture of a military (aircraft:1.4) on an ship deck, (F-18 super hornet:1.4), water in the distance, realistic, photorealistic, real word, military	
Negative Prompt	canvas frame, video game, bad art, bad anatomy, 3d render, signature, copywrite, text, shiny, distorted, bright colors, vibrant, red, trees, bushes, plants, treetops, forest, woods	
Image Size (Width x Height)	(960×540)	
Sampler	Euler A	
Sampling Steps	30	
Cfg Scale	7	
Denoising Strength	0.30	0.10
Inpainting	False	True
ControlNets	Canny, Depth, Segmentation, Softedge [34, 35]	



Fig. 3. Synthetic aircraft from each of the three domains, Day (left), Yellow Night (Center), and Grayscale Night (Right), inserted into real-world images for their respective domains.

Prompt parameters were determined experimentally and by using the CLIP interrogate [34] function to generate text prompts from both the input synthetic image and the target real world style image. Token weightings, (token: weight), in the prompts were used to highlight the key features within the desired output image. The ControlNets were chosen based on availability of the control inputs output by the simulator and what generally provided the best style transfer results based on visuals alone. Example inputs into the five ControlNet units can be seen in Fig. 2. The Canny edge and Soft edge images are generated by the built in ControlNet preprocessors while the depth and segmentation maps come from the Unity synthetic data generation process.

Method 1, full image style transfer, takes in an entirely synthetic image alongside the relevant conditional controls and outputs an image in the style of the reference control image. For this method, a higher, yet still relatively low, denoising strength is used to perform style transfer on the synthetic input image. As with other parameters, this value was experimentally determined based on how well generation result matched the target

style and maintained key aircraft features. Lower denoising strength values generally resulted in low levels of style transfer but high levels of object fidelity while high denoising generally result in high levels of style transfer, but poor level of object fidelity as shown in Fig. 4. While overall aircraft structure is maintained, finer details like wheels, underwing equipment, the cockpit, and general texture are lost or heavily altered in the augmentation process.

Method 1 was tested across all views for each domain with mixed results. As shown in Fig. 5, style transfer into the daytime domain performed well across all views with preservation of aircraft structure in most cases. Failures generally occurred on aircraft that were farther away from the camera and only represented a very small portion of the image. Even with the use of conditional controls [28] the diffusion process struggles with maintaining fine details on small objects. Based on visuals alone, the process also struggles to transfer low luminosity styles onto synthetic images like those of the gray night and yellow night domains.

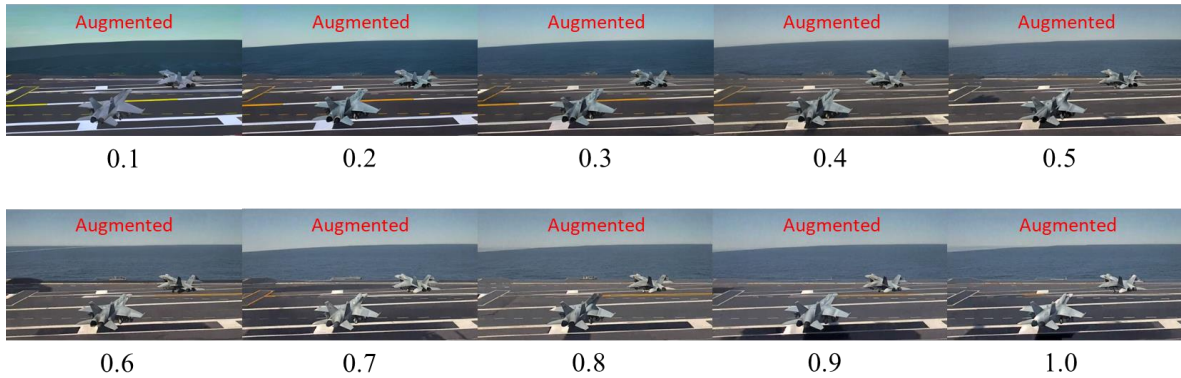


Fig. 4. Comparison of style transfer quality and aircraft fidelity across ten levels of denoising strength. The inputs into Stable Diffusion and ControlNet are the images shown in Fig. 3 with the denoising strength being the only variable.

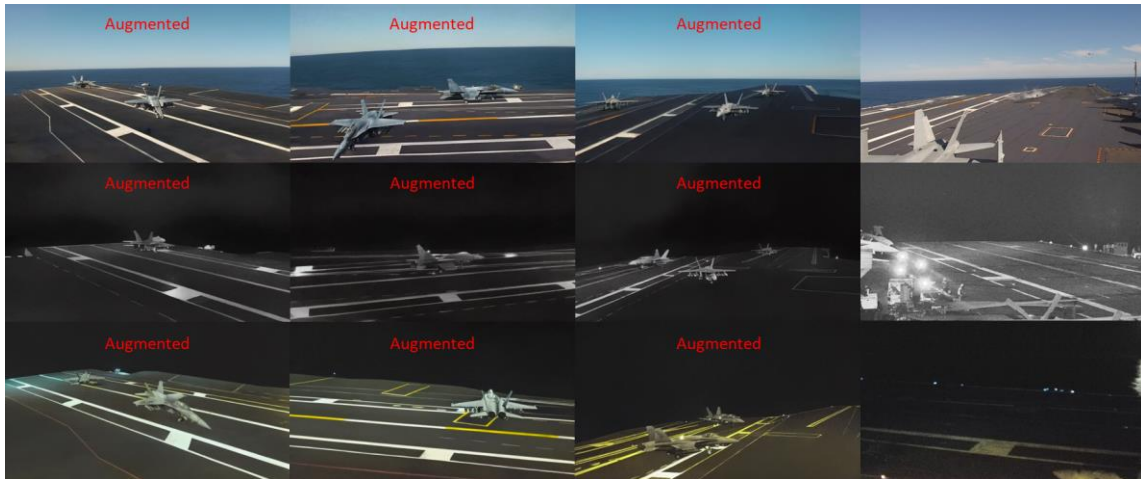


Fig. 5. Select method 1 style transfer results for the three visual domains (Day, Gray Night, Yellow Night) and three camera views (Back, Mid, Front). The final column shows representative images from the three real world visual domains.

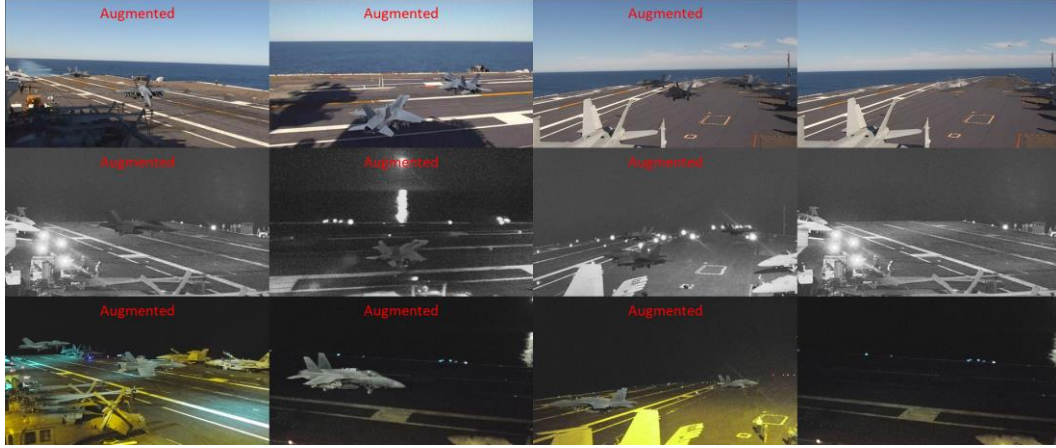


Fig. 6. Select synthetic aircraft insert style transfer results for the three visual domains (Day, Gray Night, Yellow Night) and three camera views (Back, Mid, Front). The final column shows representative images from the three real world visual domains.

Method 2, targeted style blending using inpainting, aims to perform the diffusion process over just the synthetic aircraft after it has been inserted into a real-world image using the process mentioned in the Synthetic Object Insertion section. An inpainting mask, a dilated binary version of the aircraft segmentation mask, is passed into Stable Diffusion along with the synthetic insert image and the conditional controls used in Method 1. The result is an augmented image with the surrounding style of the real-world image transferred onto the inserted aircraft without having to run the diffusion process across the entire image. This route was explored for several reasons, the primary one being that naturally including as much of the target domain directly in the image as possible should theoretically improve performance in that domain. Additionally, running the diffusion process across only a portion of the image lowers overall processing time required to perform style transfer on a dataset.

As in method 1, method 2 was tested across all combinations of camera views and visual domains as shown in Fig 6. Due to the lower denoising strength, finer details of the aircraft structure appear to be better maintained during the style transfer process in most cases. The method still struggles with smaller aircraft in the image and fails to maintain proper structure, texture, and orientation during the diffusion process. Additionally, the ability to match perceived illumination of the object seems to be heavily tied to the lighting level of the synthetic scene that the aircraft was taken from. This is particularly evident in the Fig. 6 Front Day time image, row one rightmost augmented image, where the synthetic aircraft were from a darker scene than the target domain. Evaluation on Object Detection Network

The YoloV8 object detection and classification network was used to evaluate the effectiveness of the presented methods for creating usable training data for the task of aircraft identification. During training, the same hyperparameters were kept for all trials with the only difference being the dataset used to train the model. The combinations of datasets and the subsequent performance of the model across all domains, real, augmented, and synthetic will be presented in the results

section. All networks and subsequent results are with respect to the task of identifying the F/A-18E aircraft in an image.

IV. RESULT AND DISCUSSION

Model performance was measured using two metrics, Average Precision (AP), and Average Recall (AR), for a given dataset. Holdout datasets were established for each of the nine unique dataset types which were formed from the combinations of an image type, real, augmented, or synthetic, and a visual domain, Day, Gray Night, or Yellow Night. Various proportions of mixtures of the different types of data were used to train models for different experiments with the current best performing mixtures shown in Table II. Each of the best performing mixtures involved using the full training datasets for the three real world domains as denoted by the “Full Real” dataset. The other best performing datasets consisted of adding the full sets of synthetic or augmented data to the Full Real dataset to supplement the training data. This was done in a naïve fashion by simply making them into one large dataset, splitting that into training and validation, and going through the standard training process using the new datasets.

TABLE II. CURRENT BEST AIRCRAFT DETECTION MODEL RESULTS

Dataset	Real Day (AP, AR)	Gray Night (AP, AR)	Yellow Night (AP, AR)
Full Real	(0.443, 0.194)	(0.296, 0.262)	(0.423, 0.363)
Full Real + Full Synth	(0.443, 0.196)	(0.304, 0.245)	(0.419, 0.355)
Full Real + Full Augment Method 1	(0.422, 0.2)	(0.309, 0.262)	(0.391, 0.348)
Full Real + Full Augment Method 2	(0.430, 0.21)	(0.312, 0.265)	(0.401, 0.356)

In addition to using the performance of an aircraft identification model to evaluate method performance, Fréchet Inception Distance (FID) [36] was used to quantify the style difference between image datasets. This was used as a direct measure to determine if the data augmentation methods were shrinking the domain gap between the synthetic and real data in a meaningful way and if that correlated with an increase in performance on

real world data. Table III shows the FID scores between a given real-world dataset and its corresponding synthetic and augmented datasets. The FID score is used here to quantify the style and domain distance between two given datasets where a smaller score signifies a higher similarity in datasets. According to this use case, Augment method 2, target image style blending, showed the smallest distance to the target real world dataset across all 3 target domains. This corresponded with the largest, albeit still marginal, increase in performance on the Gray Night domain but did not generalize across all domains.

TABLE III. FID DOMAIN DISTANCES BETWEEN DATASETS

Real World Domain	Dataset Type		
	Synthetic	Augment Method 1	Augment Method 2
Day	280.428	174.472	122.242
Gray Night	263.804	164.998	152.179
Yellow Night	207.684	177.66	118.556

Sections III and IV presented two methods for performing style transfer on a synthetic aircraft image as well as results on the performance of an aircraft identification network trained on the resulting images. The two methods both assume the existence of a target style image with a desirable viewpoint which is common in the type of fixed viewpoint environments that our research typically operates within. Both methods showed various levels of effectiveness on the task of style transfer. Across both methods, failure cases were primarily on small aircraft and when there were significant differences in illumination between source and target domains. One of the key challenges faced during style transfer experiments was intentionally creating “poor” quality noisy images. Traditionally, large generative image models such as Stable Diffusion, are trained with the intention of creating high quality visually aesthetic images [1]. The target style images used during this work had “undesirable” image features like noise and low contrast which are traditionally intentionally avoided in the image generation process. Using reference style conditional controls [28] helped to alleviate this limitation but there is room for improvement.

Overall, augmented data showed only marginal improvements in certain domains over synthetic data in terms of supplementing a small batch of real data to train a YoloV8 based aircraft identification network. The augmented data showed lower style distance to the real-world data both visually and according to the FID domain distance score. Several factors could contribute to the lack of improvements shown by augmented data. The authors believe the primary factor to be that the augmentation process produced poorly structured aircraft on a number of its generations which harm the ability of the network to learn a common representation for the aircraft. The other major factor considered by the authors is the presence of unlabeled real-world aircraft in the images from Method 2. The labels for these aircraft were not included in the training annotations for the networks which may have again harmed the ability of the network to learn the proper common representation for the aircraft.

Despite not achieving the desired result of creating augmented data that could improve and predict the performance of aircraft identification networks in other domains, the Stable Diffusion based style transfer shows promise for future work. Data scarcity is a common problem across many military applications which results in the use of synthetic data as a supplement when training object detection and classification systems. The authors plan to use the existing augmented data with other training methods such as pretraining on large amounts of augmented data and then fine tuning on the small batches of available real-world data. This would hopefully allow the system to learn a high-level structural representation of the target of interest from augmented data and then key finer details from the real-world data. With further fine tuning of the image augmentation process, the presented methods could provide a powerful tool for helping to bridge the sim to real gap.

V. CONCLUSION

In this paper, we introduce a diffusion model based image augmentation pipeline for performing style transfer on synthetic images of aircraft. The effectiveness of this data for training an aircraft identification network was testing with various mixtures of real, augmented, and synthetic data across three different visual domains. While the presented methods showed promising results on the task of style transfer for closing the sim to real gap, this did not translate into augmented data having an improved ability over synthetic data for training aircraft identification models. Potential issues in data generation and aircraft identifier processes have been identified and will be explored further in future works. Despite shortcomings on the desired aircraft identification results, the research contributes a style transfer technique that may find applicability in other low data regime use cases.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Ryan O’Shea performed background research, created the data augmentation pipeline, performed style transfer experiments, and wrote the paper. Gurpreet Singh defined experiment methodology, created synthetic data, and performed object detection model training experiments. Ari Goodman performed background research, guided research direction, and defined experiment methodology. Thomas Keane and Michael Brenner created domain distance tools and performed the relevant experiments with them. Christopher Jaworowski and Tushar Patel provided labeled real-world data for the experiments. James Hing provided synthetic data, background research, and supervisory research guidance to the team. All authors had approved the final version.

FUNDING

This work was supported and funded by the U.S. Naval Air Warfare Center—Aircraft Division’s Naval

Innovative Science and Engineering (NISE) program and the Office of Naval Research's (ONR) In-house Laboratory Independent Research (ILIR) program.

ACKNOWLEDGMENT

The authors wish to thank the NISE program and ONR for providing funding and research support for this effort.

REFERENCES

- [1] R. Rombach, A. Blattmann, and D. Lorenz *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.
- [2] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," arXiv preprint, arXiv:1508.06576, 2015.
- [3] A. Dosovitskiy, G. Ros, and F. Codevilla *et al.*, "CARLA: An open urban driving simulator," in *Proc. Conference on Robot Learning*, PMLR, October 2017, pp. 1–16.
- [4] K. M. Hart, A. B. Goodman, and R. P. O'Shea, "Automatic generation of machine learning synthetic data using ros," in *Proc. International Conference on Human-Computer Interaction*, Cham: Springer International Publishing, pp. 310–325, July 2021.
- [5] S. Borkman, A. Crespi, and S. Dhakad *et al.*, "Unity perception: Generate synthetic data for computer vision," arXiv preprint, arXiv:2107.04259, 2021.
- [6] M. Savva, A. X. Chang, and A. Dosovitskiy *et al.*, "MINOS: Multimodal indoor simulator for navigation in complex environments," arXiv preprint, arXiv:1712.03931, 2017.
- [7] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2414–2423. doi: 10.1109/CVPR.2016.265
- [8] J. E. Kyprianidis, J. Collomosse, and T. Wang *et al.*, "State of the 'Art': A taxonomy of artistic stylization techniques for images and video," in *Proc. IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 5, pp. 866–885, May 2013. doi: 10.1109/TVCG.2012.160
- [9] P. Chattopadhyay, K. Sarangmath, and V. Vijaykumar *et al.*, "Pasta: Proportional amplitude spectrum training augmentation for syn-to-real domain generalization," in *Proc. the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19288–19300.
- [10] I. Goodfellow, J. Pouget-Abadie, and M. Mirza *et al.*, "Generative adversarial nets. Advances in neural information processing systems," arXiv preprint, arXiv:1406.2661, 2014.
- [11] J. Y. Zhu, T. Park, and P. Isola *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.
- [12] Y. Burad and K. Burad, "Leveraging unpaired image to image translation for generating high quality synthetic data," in *Proc. 2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*, Pune, India, 2021, pp. 98–101. doi: 10.1109/ESCI50559.2021.9396865
- [13] H. Tang, H. Liu, and D. Xu *et al.*, "Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 4, pp. 1972–1987, 2021.
- [14] B. Ren, H. Tang, and N. Sebe, "Cascaded cross mlp-mixer gans for cross-view image translation," arXiv preprint, arXiv:2110.10183, 2021.
- [15] T. Park, A. A. Efros, R. Zhang, and J. Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020*, Springer International Publishing, Part IX, 16, pp. 319–345.
- [16] J. Hoffman, E. Tzeng, and T. Park *et al.*, "Cycada: Cycle-consistent adversarial domain adaptation," in *Proc International Conference on Machine Learning*, July 2018, pp. 1989–1998.
- [17] M. Y. Liu, X. Huang, J. Yu, and T. C. Wang *et al.*, "Generative adversarial networks for image and video synthesis: Algorithms and applications," in *Proc. the IEEE*, 2021, vol. 109, no. 5, pp. 839–862.
- [18] D. A. Hudson and L. Zitnick, "Generative adversarial transformers," in *Proc. International Conference on Machine Learning*, July 2021, pp. 4487–4499.
- [19] M. J. Chong, H. Y. Lee, and D. Forsyth, "Stylegan of all trades: Image manipulation with only pretrained stylegan," arXiv preprint, arXiv:2111.01619, 2021.
- [20] W. Liu, B. Luo, and J. Liu, "Synthetic data augmentation using multiscale attention CycleGAN for Aircraft detection in remote sensing images," in *Proc. IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022, pp. 1–5. doi: 10.1109/LGRS.2021.3052017
- [21] Y. T. Shen, H. Lee, and H. Kwon *et al.*, "Progressive transformation learning for leveraging virtual images in training," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 835–844.
- [22] P. Isola, J. Y. Zhu, and T. Zhou *et al.*, "Image-to-image translation with conditional adversarial networks," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.
- [23] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," arXiv preprint, arXiv:2010.02502, 2020.
- [24] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [25] C. Saharia, W. Chan, and S. Saxena *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36479–36494, 2022.
- [26] P. Dhariwal and A. Nichol, "Diffusion models beat GANS on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.
- [27] C. Meng, Y. He, and Y. Song *et al.*, "Sdedit: Guided image synthesis and editing with stochastic differential equations," arXiv preprint, arXiv:2108.01073, 2021.
- [28] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.
- [29] A. Juliani, V. P. Berges, and E. Teng *et al.*, "Unity: A general platform for intelligent agents," arXiv preprint, arXiv:1809.02627, 2018.
- [30] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. the ACM SIGGRAPH Conference on Computer Graphics*, 2000, pp. 417–424.
- [31] J. Yu, Z. Lin, J. Yang, and X. Shen *et al.*, "Generative image inpainting with contextual attention," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.
- [32] M. Yan, Q. Sun, and I. Frosio *et al.*, "How to close sim-real gap? transfer with segmentation!" arXiv preprint, arXiv:2005.07695, 2020.
- [33] G. Bradski, "The OpenCV library," *Dr. Dobb's Journal of Software Tools*, 2000.
- [34] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.
- [35] Z. Su, W. Liu, and Z. Yu *et al.*, "Pixel difference networks for efficient edge detection," in *Proc. the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5117–5127.
- [36] M. Heusel, H. Ramsauer, and T. Unterthiner *et al.*, "Gans trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

Copyright © 2025 by the authors. This is an open access article distributed under the Creative Commons Attribution License ([CC-BY-4.0](https://creativecommons.org/licenses/by/4.0/)), which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.