# Art Style Classification Using Computer Vision

Team 1

Bokai Lai, Alex Griffith, Danish Siddiqui, Gabrielle Li, Keqing Wang

## 1. Introduction and Motivation

Over the past decade, digital art collections have grown at an unprecedented scale. Platforms such as WikiArt now host tens of thousands of paintings spanning different movements, time periods, and cultural traditions. These collections are invaluable for curators, educators, researchers, and recommendation systems, but their size also creates a practical problem. Without reliable organization, much of this material becomes difficult to search, study, or reuse.

For proper organization, accurate artistic style labeling is crucial. However, unlike object labels, which describe what appears in an image, style labels describe how a painting was made and how it fits into a broader artistic movement. This makes them especially useful for tasks such as cataloging collections, supporting art historical research, building educational tools, and enabling style-based recommendations. However, style labels are costly and inconsistent to obtain. Annotation is slow and subjective, and even trained art historians often disagree. Visual differences between movements are subtle, and stylistic cues emerge from texture, brushwork, color palette, and composition rather than from clearly identifiable objects.

In this project, we ask a concrete question. Given a large but imperfect dataset of paintings labeled by artistic movement, can modern deep learning models reliably predict the style of a new, unseen artwork? To investigate this, we implement and compare three approaches: a baseline convolutional neural network trained from scratch, a transfer learning model based on InceptionV3, and a Vision Transformer (ViT-Base). Naturally, each model illustrates its own clear benefits to this task. CNNs emphasize local visual patterns, transfer learning reuses representations learned from large natural image datasets, and Transformers model global relationships through self-attention.

It is important to note that our goal was not solely focused on discovering which model performs best, but also to understand why certain approaches succeeded or struggled in this task and how that informs us about the architectural choices in a more holistic manner. This report was written for an audience of peers who would have prior exposure to CNNs, transfer learning, and Vision Transformers. The goal is to present the analysis in a way that allows readers to follow the reasoning and apply the ideas and insights to their own related projects.

## 2. Background

The dataset used in this project is derived from WikiArt. After cleaning the data and merging closely related categories, we work with approximately 42.5k images spread across 13 art movements: Academic_Art, Art Nouveau, Baroque, Expressionism, Japanese_Art, Neoclassicism, Primitivism, Realism, Renaissance, Rococo, Romanticism, Symbolism, and Western_Medieval. Each movement contains at least 1,150 images. All images are first normalized to a 512×512 canvas, then downsampled to 224×224 to keep training feasible. Duplicate images are removed to prevent information leakage and reduce the risk of memorization, and the dataset is partitioned using standard 70%/15%/15% train–validation–test split.

This classification setting is fundamentally different from a standard object recognition task such as differentiating between a cat and a dog. Artistic style boundaries are inherently fuzzy, and the dataset contains unavoidable label noise. Many works naturally sit between movements such as Romanticism and Symbolism, while others represent transitional periods, for example between Baroque and Rococo. Consequently, stylistic differences are subtle, class boundaries overlap, and labels are imperfect, requiring models to learn nuanced visual patterns rather than rely on explicit semantic cues.

From the modeling perspective, the three approaches we explore bring different inductive biases. CNNs slide learned filters over small patches of the image and combine them hierarchically. They naturally pick up edges, textures, and local patterns, which is a good fit for brushwork and surface detail. Transfer learning takes a strong CNN backbone such as InceptionV3, pretrained on ImageNet, and reuses its filters as generic visual feature extractors, adding and training a new head for our specific task. Vision Transformers (ViTs) go in another direction: they chop the image into patches, view those patches as tokens, and use self-attention layers to model interactions between all patches. This contributes to ViT's strong advantage of modeling global structure at the cost of higher data and computational requirements.

## 3. Description of Methods

### *3.1.Data Pipeline and Preprocessing*

All three models share the same underlying data pipeline. Images are read from disk, resized to 224×224×3, converted to floating-point, and the three RGB channels are rescaled into the [0,1] range. For the training set, we applied light data augmentation: random horizontal flips, small rotations, and modest zooms. These transformations exposed the models to slight variations of the same painting without fundamentally altering stylistic characteristics. The validation and test sets were only normalized, not augmented, so that evaluation reflects realistic input.

The data were then batched and fed to the models using a streaming input pipeline, ensuring GPU time is spent on learning rather than waiting for I/O. The choice of a 224×224 resolution reflects a practical trade-off: it is low enough to keep memory usage and computation manageable, while still preserving most stylistic cues such as texture, color transitions, and brushwork. Original images are first normalized at a higher resolution and then downsampled, allowing the models to retain stylistic structure while remaining computationally feasible.

### *3.2.Baseline Convolutional Neural Network*

The baseline CNN model served as our point of reference. Architecturally, it is a conventional stack of convolutional layers with nonlinearities, each followed by max pooling. After several such blocks, the feature maps are flattened and passed into a large fully connected layer, and finally into a softmax output over the 13 classes. The model was trained using sparse categorical cross-entropy, which is standard for multi-class classification with integer labels, and optimized using an Adam-based optimizer. Inputs are fixed at a resolution of 224×224, and the network is initialized with ImageNet weights before being fine-tuned on the dataset.

The baseline CNN contains approximately 44.4 million parameters, with the majority concentrated in the final fully connected layer. Max pooling is applied after every convolutional block, which rapidly reduces the spatial resolution of the feature maps to roughly 26×26 in the final convolutional layer. While this aggressive downsampling simplifies optimization and reduces computational cost, it also discards fine-grained spatial information. As a result, the model becomes heavily reliant on the large fully connected layer, potentially limiting its ability to capture subtle stylistic details distributed across the image.

Training the baseline CNN requires several hours to complete. When evaluated on the test set, the final model achieved a test accuracy of approximately 78%. The CNN performed particularly well on Japanese_Art and Western_Medieval, where the visual vocabulary is more distinctive, while it struggles on categories such as Academic_Art and Symbolism,

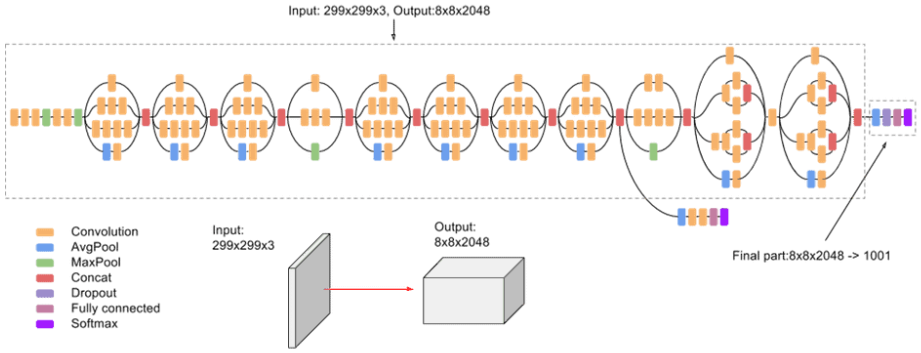
which are less clearly separated and more affected by label noise. Precision is highest for Primitivism and Rococo, while recall is highest for Japanese_Art and Romanticism.

These patterns align with the strengths and limitations of CNNs. Convolutional architectures are effective at capturing repeated textures and localized stylistic elements, making styles with strong, consistent visual signatures easier to identify. In contrast, when stylistic boundaries are subtle or overlapping, the model has greater difficulty separating classes based on local cues alone.

### *3.3.Transfer Learning with InceptionV3*

The second model switches out our custom CNN backbone for InceptionV3, a deeper, more sophisticated convolutional architecture that has been heavily optimized on ImageNet. InceptionV3 was chosen for a few reasons highlighted in the deck: it is a strong, production-tested CNN with high ImageNet accuracy; it has rich filters for edges, colors, textures, and curvature; and its multi-scale "Inception modules" can pick up both fine brush strokes and larger compositional patterns.

**Exhibit 3.1: Transfer Learning Hierarchy**




In our setup, the original InceptionV3 classification head is removed and replaced with a task-specific dense layer and softmax that output probabilities over the same 13 art movements. The pretrained Inception backbone is kept frozen during training, so only the newly added layers are updated. This choice is driven by both data and computational considerations. Training InceptionV3 from scratch would require far more data and compute than are available, whereas freezing the backbone allows us to leverage strong generic visual representations learned from ImageNet while training only a relatively small number of parameters on the art dataset.

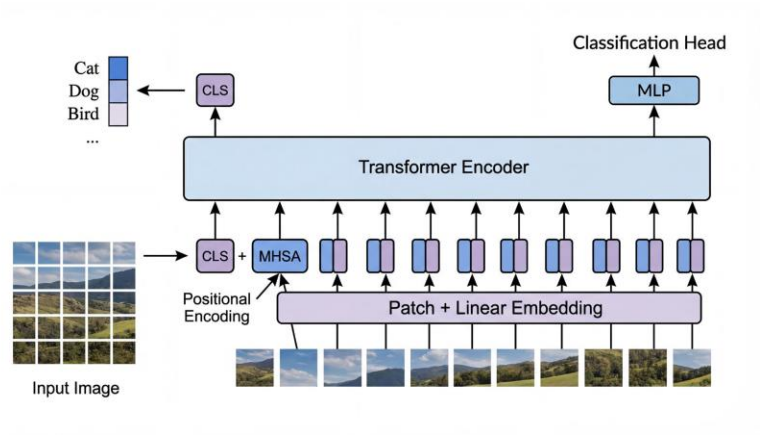
The resulting model trains quickly and avoids severe overfitting. Compared to an early CNN trained from scratch, which achieved accuracy near chance levels for many classes, the transfer learning approach yields a substantial improvement. In its final configuration, the model reaches training accuracy around 90% and achieves a test accuracy of approximately 70%, with F1 scores in the 0.65–0.80 range for many of the major art movements. These results indicate that visual features learned on ImageNet transfer reasonably well to the art style classification task.

At the same time, the model exhibits a clear performance ceiling. Because the InceptionV3 backbone is frozen, the network cannot adapt its lower- and mid-level filters to art-specific textures or to the idiosyncrasies of the WikiArt dataset. This limits its ability to handle class imbalance, label noise, and overlapping stylistic boundaries. A natural extension would be to unfreeze and fine-tune the deeper Inception blocks while keeping early layers fixed, allowing the model to specialize higher-level representations of style without losing the robustness of generic edge and color filters.

*3.4.Vision Transformer (ViT-Base)*

The third model takes a different route: instead of convolutions, it uses a Vision Transformer. A ViT starts by cutting an image into fixed-size patches, embedding each patch into a vector, and then feeding the sequence of patch embeddings into a stack of Transformer encoder blocks. A special CLS token attends to all other patches and is used to summarize the image for classification. In this project, we used a ViT-Base configuration with 224×224 input resolution, 16×16 patch size, twelve encoder blocks, twelve attention heads, 768-dimensional embeddings, and 3,072-dimensional MLP layers, followed by a classifier for the 13 art movements.

**Exhibit 3.2: ViT Hierarchy**




As with InceptionV3, the ViT backbone is loaded with pretrained ImageNet weights and then frozen. A new classification head is added and trained on our art labels for a modest number of epochs. The same preprocessed 224×224 images are used as input. This setup is a pragmatic compromise: ViT-Base contains tens of millions of parameters and fully fine-tuning it on 42.5k images would be expensive and prone to overfitting. Freezing the backbone allowed us to reuse a strong generic representation while keeping training stable and manageable.

In practice, however, this configuration does not perform as well as we might hope. Training and validation accuracy plateaus between 57– 60%, substantially below both the baseline CNN and the transfer learning model. Several classes, including Symbolism, Academic_Art, and Rococo, exhibit low recall and F1 scores despite relatively high precision. In other words, when the ViT predicts these styles, it is often correct, but it predicts them infrequently, missing many true examples.

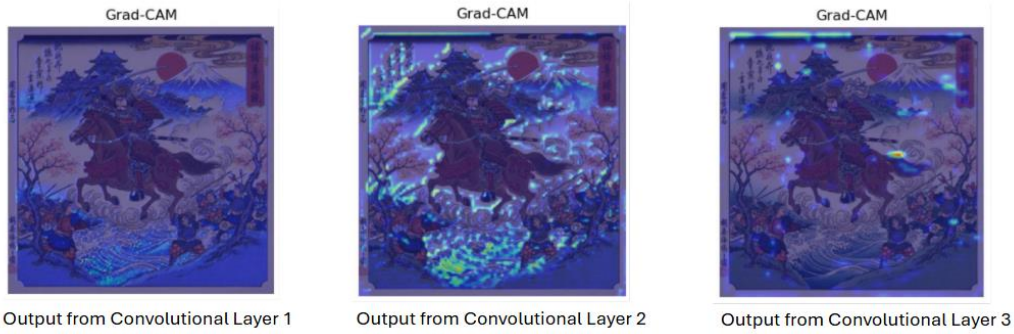
This behavior is consistent with underfitting. With the backbone frozen, the ViT effectively acts as a fixed feature extractor trained on generic photographic images rather than paintings. In addition, the 224×224 input resolution combined with a 16×16 patch size means that each patch covers a relatively large image region, which can obscure fine-grained textural cues that are important for distinguishing artistic styles. Finally, only the top head is trained, and only for fifteen epochs, which further restricts the model's ability to adapt. Together, these factors limit the ViT's effective capacity at the layers where art-specific structure would need to be learned.

Addressing these limitations would likely require unfreezing and fine-tuning deeper Transformer blocks and increasing input resolution, allowing the model to better capture stylistic detail and close the performance gap with convolutional approaches.

### *3.5.Model Interpretability with Grad-CAM*

To gain intuition into what the CNN is using to make predictions, we apply Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM highlights image regions that most influence a given class prediction by measuring the sensitivity of the class score to spatial locations in a convolutional feature map. When overlaid on the original painting, the resulting heatmaps provide a coarse but informative view of where the model is focusing its attention.

**Exhibit 3.3: CNN Results when Testing on an AI Generated, Japanese-Inspired Artwork**




Across layers, the Grad-CAM visualizations show a clear progression in representation. Earlier convolutional layers respond broadly to low-level features such as edges and simple textures, producing diffuse activations. Intermediate layers exhibit stronger and more structured responses, highlighting regions with repeated textures, ornamentation, and stylized patterns that are most informative for distinguishing artistic styles. In the deepest convolutional layers, activations become more selective and sparser due to increased abstraction and spatial downsampling, which explains why these layers often show weaker Grad-CAM intensity than intermediate ones. For correctly classified artworks, the highlighted regions align with intuitive stylistic cues such as brushwork, architectural detail, and characteristic color regions, suggesting that the CNN is learning meaningful style-related features rather than relying on spurious artifacts

## 4. Experiments

All three models are trained and evaluated using the same train–validation–test split. We report overall test accuracy as the primary metric, and also examine per-class precision, recall, and F1 scores, along with confusion matrices and training curves, which are provided in the appendix.

The baseline CNN performs best overall, achieving a test accuracy of approximately 78% and outperforming both the transfer learning model and the Vision Transformer. Performance is strongest for visually distinctive styles such as Japanese_Art and Western_Medieval, where both precision and recall are high. In contrast, Academic_Art and Symbolism remain more difficult to classify, consistent with their fuzzier stylistic boundaries and higher label noise. The CNN also exhibits high precision for Primitivism and Rococo, and high recall for Japanese_Art and Romanticism, reflecting its strength in capturing consistent, localized stylistic patterns.

**Exhibit 4.1 and 4.2: Results of Baseline CNN & Results of Transfer Learning**

| Baseline CNN | | | |
|---|---|---|---|
| Art Style | Precision | Recall | F1-Score |
| Academic Art | 0.76 | 0.65 | 0.70 |
| Art Nouveau | 0.78 | 0.74 | 0.76 |
| Baroque | 0.75 | 0.80 | 0.78 |
| Expressionism | 0.73 | 0.73 | 0.73 |
| Japanese Art | 0.80 | 0.85 | 0.82 |
| Neoclassicism | 0.81 | 0.78 | 0.79 |
| Primitivitism | 0.86 | 0.77 | 0.81 |
| Realism | 0.77 | 0.75 | 0.76 |
| Renaissance | 0.81 | 0.80 | 0.80 |
| Rococo | 0.84 | 0.74 | 0.78 |
| Romanticism | 0.73 | 0.83 | 0.79 |
| Symbolism | 0.77 | 0.63 | 0.69 |
| Western Medieval | 0.87 | 0.78 | 0.82 |
| Accuracy | | | 0.78 |
| Macro Avg | 0.79 | 0.76 | 0.77 |
| Weighted Avg | 0.78 | 0.78 | 0.78 |

| Transfer Learning | | | |
|---|---|---|---|
| Art Style | Precision | Recall | F1-Score |
| Academic Art | 0.66 | 0.48 | 0.56 |
| Art Nouveau | 0.70 | 0.74 | 0.72 |
| Baroque | 0.68 | 0.71 | 0.69 |
| Expressionism | 0.69 | 0.72 | 0.70 |
| Japanese Art | 0.81 | 0.82 | 0.81 |
| Neoclassicism | 0.70 | 0.74 | 0.72 |
| Primitivitism | 0.72 | 0.79 | 0.75 |
| Realism | 0.79 | 0.56 | 0.66 |
| Renaissance | 0.71 | 0.85 | 0.77 |
| Rococo | 0.54 | 0.68 | 0.60 |
| Romanticism | 0.66 | 0.67 | 0.67 |
| Symbolism | 0.55 | 0.22 | 0.31 |
| Western Medieval | 0.80 | 0.88 | 0.84 |
| Accuracy | | | 0.70 |
| Macro Avg | 0.69 | 0.68 | 0.68 |
| Weighted Avg | 0.70 | 0.70 | 0.69 |

The transfer-learning InceptionV3 model performed slightly worse but is still respectable. It achieved an average test accuracy of around 70%, with F1 scores in the 0.65–0.80 range for many major movements. Training accuracy climbed to about 90%, but because the backbone is frozen, there is a natural ceiling on how well the model can adapt to the specific quirks of the art domain. While the pretrained ImageNet filters provided strong generic visual features, they limited the model's ability to adapt to art-specific textures, label noise, and overlapping stylistic boundaries.

The ViT model, in its frozen-backbone configuration, performed the worst among the three models. Its train and validation accuracy plateau in the high-50% range, roughly 57–60%. Several styles show low recall and F1, which indicates that the model is underfitting and not confidently recognizing many examples even when they are present. With the backbone frozen and input resolution limited to 224×224, the ViT effectively acts as a fixed feature extractor trained on photographic images. Without fine-tuning its attention layers or increasing input resolution, it cannot fully leverage its capacity to model the long-range relationships that are central to artistic style.

If we put the three models side by side, a simple ordering emerges. The baseline CNN sits at the top with around 0.78 accuracy. The transfer-learning InceptionV3 model's accuracy was observed at around 0.70. Finally, the frozen-backbone ViT trailed at around 0.57–0.60. All three models handle the most visually distinct movements reasonably well and all three struggle with abstract or overlapping styles, such as Symbolism and the cluster of Renaissance, Baroque, Rococo, and Neoclassicism, where boundaries are inherently blurry.

**Exhibit 4.3: Results of Vision Transformer**

| Art Style | Precision | Recall | F1-Score |
|---|---|---|---|
| Academic Art | 0.65 | 0.33 | 0.44 |
| Art Nouveau | 0.63 | 0.62 | 0.62 |
| Baroque | 0.56 | 0.44 | 0.49 |
| Expressionism | 0.56 | 0.77 | 0.65 |
| Japanese Art | 0.83 | 0.69 | 0.76 |
| Neoclassicism | 0.69 | 0.44 | 0.54 |
| Primitivitism | 0.64 | 0.54 | 0.59 |
| Realism | 0.64 | 0.56 | 0.60 |
| Renaissance | 0.52 | 0.77 | 0.62 |
| Rococo | 0.49 | 0.44 | 0.46 |
| Romanticism | 0.49 | 0.60 | 0.54 |
| Symbolism | 0.25 | 0.14 | 0.18 |
| Western Medieval | 0.81 | 0.80 | 0.81 |
| Accuracy | | | 0.57 |
| Macro Avg | 0.60 | 0.55 | 0.56 |

The qualitative Grad-CAM visualizations gave some insight into the CNN's strong performance. They illustrated that, for successful predictions, the model was indeed focusing on stylistically meaningful regions rather than trivial areas. Combined with the quantitative metrics, this painted a consistent picture: in this domain, a relatively straightforward CNN with a good inductive bias and careful preprocessing can outperform more complex architectures that are not fully adapted.

**Exhibit 4.4: F1 Score Comparison**

| F1-Score | | | |
|---|---|---|---|
| Art Style | CNN | TL | ViT |
| Academic Art | 0.70 | -0.14 | -0.26 |
| Art Nouveau | 0.76 | -0.04 | -0.14 |
| Baroque | 0.78 | -0.09 | -0.29 |
| Expressionism | 0.73 | -0.03 | -0.08 |
| Japanese Art | 0.82 | -0.01 | -0.06 |
| Neoclassicism | 0.79 | -0.07 | -0.25 |
| Primitivitism | 0.81 | -0.06 | -0.22 |
| Renaissance | 0.76 | -0.10 | -0.16 |
| Rococo | 0.80 | -0.03 | -0.18 |
| Romanticism | 0.78 | -0.18 | -0.32 |
| Realism | 0.79 | -0.12 | -0.25 |
| Symbolism | 0.69 | -0.38 | -0.51 |
| Western Medieval | 0.82 | 0.02 | -0.01 |
| AVG Difference | | -0.09 | -0.21 |
| Macro Avg Acc. | 0.77 | 0.68 | 0.57 |

## 5. Conclusion

This project set out to classify paintings into art movements using deep learning and, in the process, to understand how different model families behave on a subtle, noisy task. While the models had varying levels of success, predominantly

due to computational constraints, each model demonstrated its strengths and frailties in the complex task of Art Style Classification.

The baseline CNN emerged as the strongest performer, reaching a test accuracy of around 78%. Its success reflects the fact that CNNs are well matched to style classification: they are very good at capturing local texture and brushwork, which are precisely the signals that differentiate many movements. The Grad-CAM visualizations further confirmed that the network was attending to intuitively meaningful areas of each painting.

The InceptionV3 transfer-learning model showed the strength of reusing pretrained backbones. It achieved a solid test accuracy of around 70% with relatively little training time and benefited from robust generic visual features. At the same time, freezing the backbone limited how far it could go: the filters could not be adapted to the artistic domain, and performance seemed capped below the CNN. This naturally suggests a next step: selectively unfreeze deeper Inception layers and fine-tune them on art, while keeping early layers fixed to retain stable edge and color detectors.

The Vision Transformer experiment was a useful reminder that more complex does not automatically translate to better in every setting. ViT-Base with a frozen backbone and relatively shallow head training underfit the task and plateaued around 57–60% accuracy. This does not mean that Transformers cannot work for art; rather, it shows that in this setting, a modest dataset size, low resolution, frozen backbone, they do not have enough room to thrive. To fully explore the potential of ViTs in this context, we would likely need higher-resolution inputs, more aggressive fine-tuning, stronger regularization, and possibly more or better curated data.

For practitioners and classmates, three takeaways are invaluable. Firstly, align your architecture with the dominant signal in your data: when the task is driven by texture and local style details, a well-designed CNN can outperform more complex alternatives. Secondly, use transfer learning thoughtfully. Frozen backbones are a strong starting point but consider unfreezing parts of the network if the target domain differs significantly from ImageNet. Finally, quantitative metrics are necessary but not sufficient; tools like Grad-CAM allow for sanity-checks as to what the network is relying on and can reveal whether it is truly capturing style or just latching onto artifacts.

Overall, the project shows that deep learning can make meaningful progress on art-style classification, but also that the success of a model is tightly tied to how well its assumptions line up with the structure of the problem. That is a theme that carries well beyond art, and it is one we expect to see again in future projects.

## 6. References and Future Reading

Cetinic, E., Lipic, T., & Grgic, S. (2018). Fine-tuning convolutional neural networks for fine art classification. *Expert Systems with Applications*, *114*, 107–118.

Rodriguez, C., Lech, M., & Pirogova, E. (2018). Classification of Style in Fine-Art Paintings Using Transfer Learning and Weighted Image Patches. 2018 12th International Conference on Signal Processing and Communication Systems Icspcs 2018 Proceedings. https://doi.org/10.1109/ICSPCS.2018.8631731

Boesch, G. (2025, September 30). Vision Transformers (VIT) in image recognition. *viso.ai*. https://viso.ai/deep-learning/vision-transformer-vit/

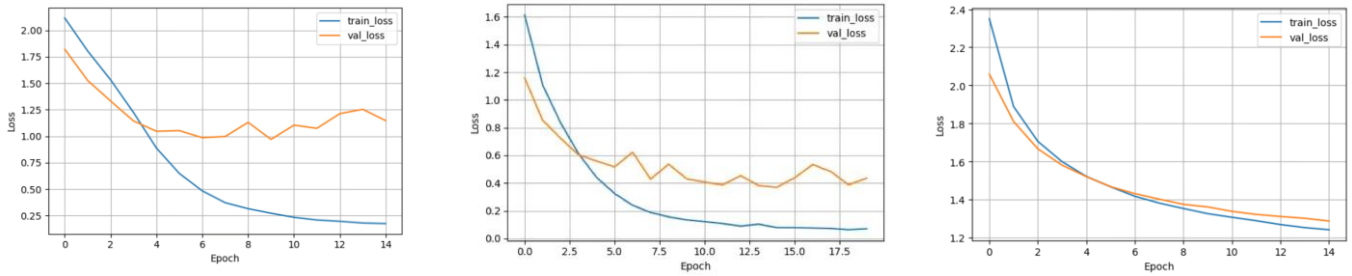
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020, October 22). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv.org. https://arxiv.org/abs/2010.11929

Cao, J., Yan, M., Jia, Y. *et al.* Application of a modified Inception-v3 model in the dynasty-based classification of ancient murals. *EURASIP J. Adv. Signal Process.* 2021, 49 (2021). https://doi.org/10.1186/s13634-021-00740-8

Alquarizm. "Transfer Learning – Using Inception V3 for Developing Image Classifier (Part 2)." Alquarizm, January 15, 2020. https://alquarizm.wordpress.com/2019/03/11/transfer-learning-using-inception-v3-for-developing-image-classifier-part-2/

## 7. Appendix

**Loss Graphs of Baseline CNN, Transfer Learning, and ViT**




**Accuracy Graphs of Baseline CNN, Transfer Learning, and ViT**