MuVieCAST: Multi-View Consistent Artistic Style Transfer

Nail Ibrahimli, Julian F. P. Kooij, Liangliang Nan
Delft University of Technology

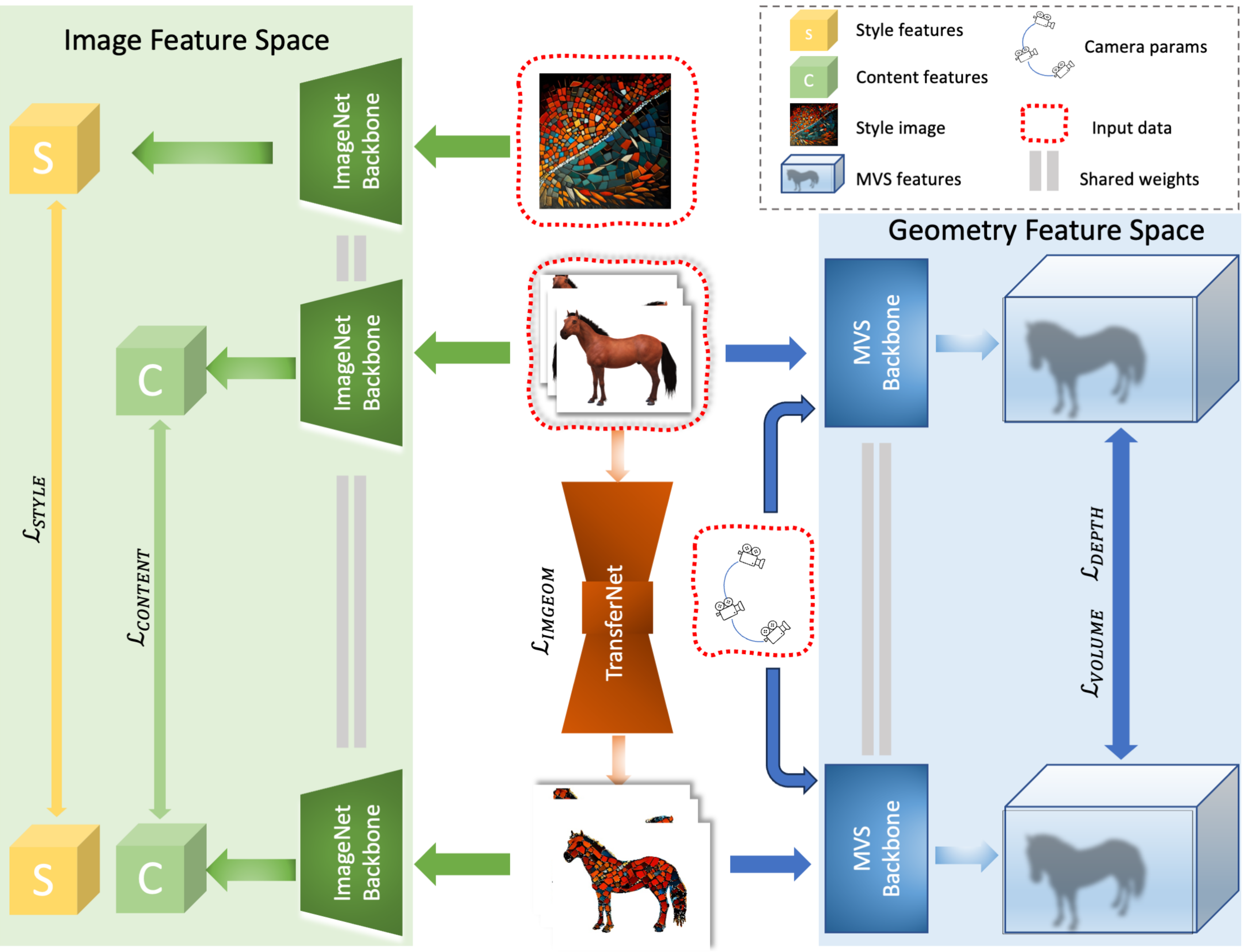
3D geoinformation
Department of Urbanism
Faculty of Architecture and the Built Environment
Delft University of Technology

3DV 2024
International Conference on 3D Vision
Davos, Switzerland
March 18-21, 2024

SCAN ME

Introduction and Methodology

MuVieCAST is a novel architecture that ensures multi-view consistent style transfer across diverse datasets and applications. Unlike other 3D style transfer methods, it can generate consistent stylized images directly from calibrated views, eliminating the need for explicit 3D scene representations. Our method is fast, flexible, and robust for tasks like novel-view synthesis, point cloud and neural mesh reconstruction.



MuVieCAST has three main components:

- **Content-style feature extraction** operates on the image feature space to preserve the content and style.
- **TransferNet** performs image transformation.
- **Geometry learning module** operates on the geometry feature space to preserve the geometry.

Different network configurations tested in the experiments

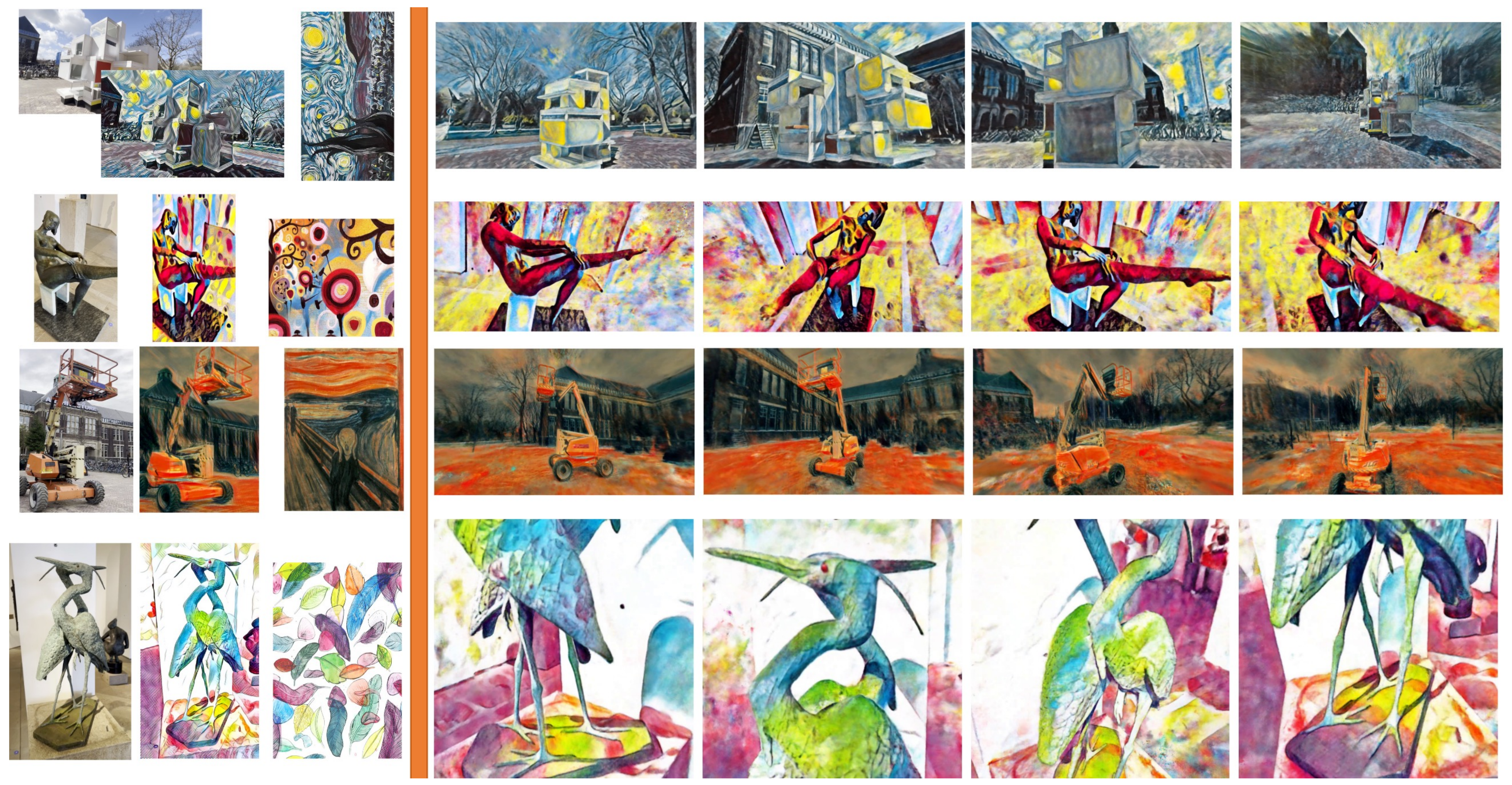
Naming	Geometry	ImageNet	TransferNet	Style loss	Total params	Trainable params
CasMVSNet_UNet	CasMVSNet [1]	VGG16	UNet	Gram [4]	10.2 M	1.7 M
CasMVSNet_AdaIN	CasMVSNet	VGG19	AdaIN [3]	IN statistics [3]	7.9 M	3.5 M
PatchMatchNet_UNet	PatchMatchNet [2]	VGG16	UNet	Gram	9.5 M	1.7 M
PatchMatchNet_AdaIN	PatchMatchNet	VGG19	AdaIN	IN statistics	7.2 M	3.5 M

Total loss function is a weighted combination of loss terms:

$$\mathcal{L}_{total} = \lambda_{content}\mathcal{L}_{content} + \lambda_{style}\mathcal{L}_{style} + \lambda_{imgeom}\mathcal{L}_{imgeom} + \lambda_{volume}\mathcal{L}_{volume} + \lambda_{depth}\mathcal{L}_{depth}$$

Qualitative Results

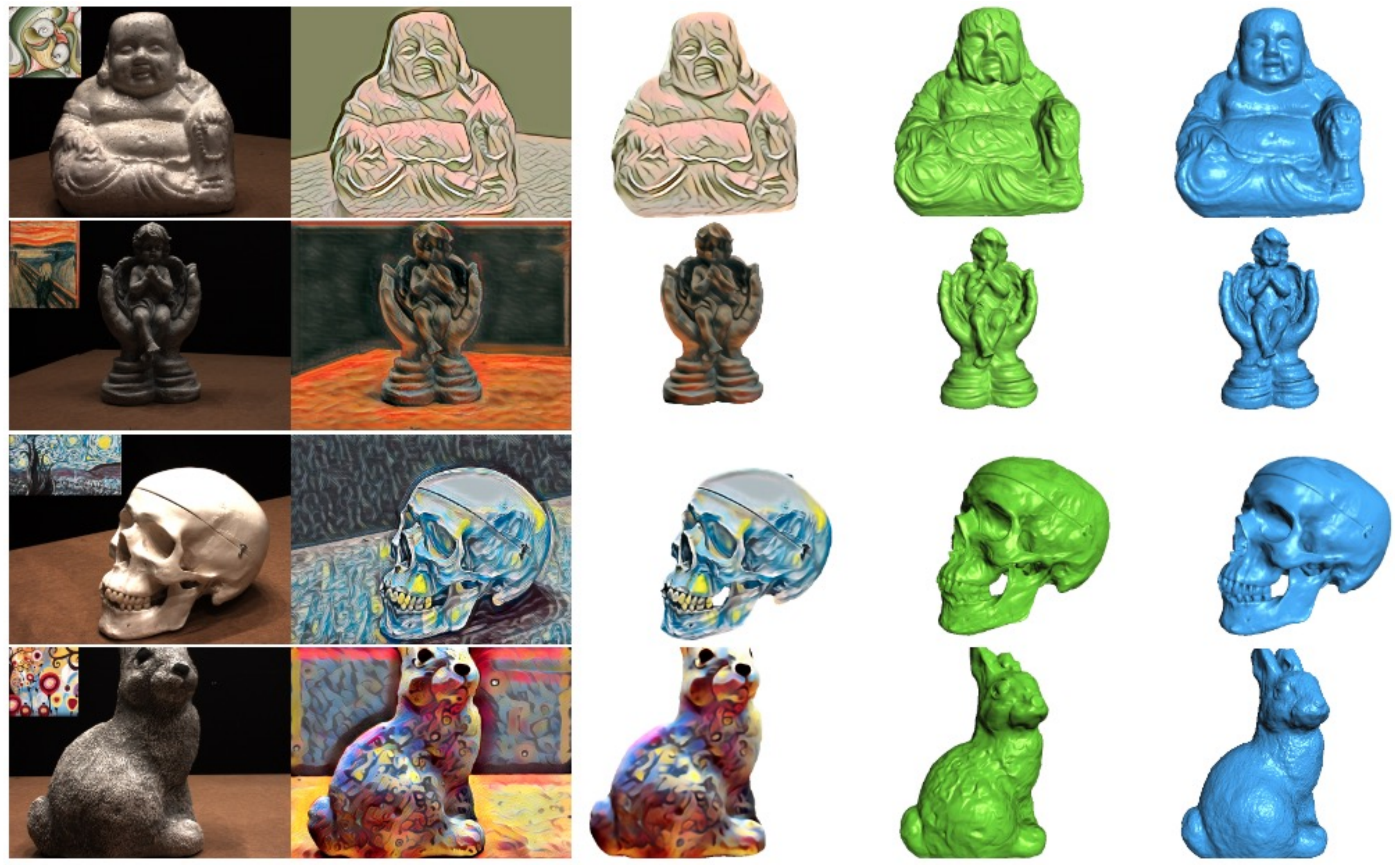
Novel View Synthesis



Input sample, stylized image and style image

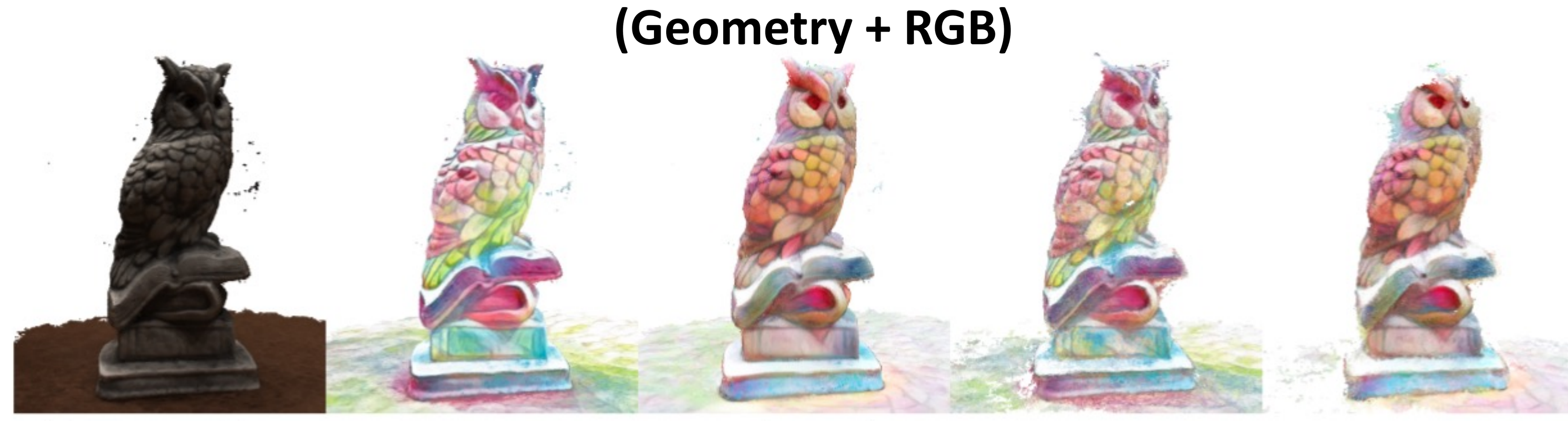
NeRF renderings reconstructed from stylized images

Neural Surface Reconstruction



(a) Input image (b) Stylized image (c) Stylized Surface rendering (d) Reconstruction from stylized images (e) Reconstruction from input images

MVS-based Pointcloud Reconstruction (Geometry + RGB)



Input + Input   Input + UNet   Input + AdaIN   UNet + UNet   AdaIN + AdaIN

UNet backbone performed better than the AdaIN backbone in terms of geometric consistency.

User Study

The left column shows the scene samples and style images shared with user study participants. Frames from our results are presented on the top rows, while frames from the **ARF** [5] method are displayed on the bottom rows. The charts indicate the preferences of the 40 participants.

Training Time

Using pretrained backbones accelerates training by solely addressing multi-view image style transfer. The training time for DTU scans with 49 images, a resolution of 640 × 480, a neighbouring view window size of 3, and a batch size of 1 per GPU on *dual RTX 2080 Ti* was measured. Training times for 10 epochs and backbone information are as follows:

Backbone information

Modules	Options	Pretrained	Trainable
Image learning	VGG16	ImageNet	No
	VGG19	ImageNet	No
Geometry Learning	CasMVSNet	DTU	No
	PatchMatchNet	DTU	No
TransferNet	UNet	MS COCO	Yes
	AdaIN	MS COCO	Yes

Training time for network configurations

Network Architecture	Training Time (seconds)
CasMVSNet_UNet	174.44
CasMVSNet_AdaIN	174.52
PatchMatchNet_UNet	153.03
PatchMatchNet_AdaIN	155.00

References

[1] Gu, Xiaodong et al. "Cascade cost volume for high-resolution multi-view stereo and stereo matching." CVPR 2020
[2] Wang, Fangjinhua et al. "Patchmatchnet: Learned multi-view patchmatch stereo." CVPR 2021
[3] Gatys, Leon et al. "Image style transfer using convolutional neural networks." CVPR 2016
[4] Huang, Xun et al. "Arbitrary style transfer inreal-time with adaptive instance normalization." ICCV2017
[5] Zhang, Kai et al. "Arf: Artistic radiance fields." ECCV 2022