**Gradient-based** **3D mesh style transfer using 2D supervision**

|  |  |  |  |
| --- | --- | --- | --- |
| Abhishek Lalwani  University Of Massachusetts Amherst  alalwani@umass.edu | Himanshu Gupta  University Of Massachusetts Amherst  hgupta@umass.edu | Rushikesh Dudhat  University Of Massachusetts Amherst  rdudhat@umass.edu |  |

**Abstract**

*The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word “Abstract” as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. The abstract may be up to 3 inches (7.62 cm) long. Leave two blank lines after the Abstract, then begin the main text.*

# **Introduction**

Understanding the 3D world from 2D images is one of the fundamental problems in computer vision. Humans model the 3D world in their brains using images on their retinas, and live their daily existence using the constructed model. The machines, too, can act more intelligently by explicitly modeling the 3D world behind 2D images.

The process of generating an image from the 3D world is called rendering. Because this lies on the border between the 3D world and 2D images, it is crucially important in computer vision.

In recent years, convolutional neural networks (CNNs) have achieved considerable success in 2D image understanding [7, 13]. Therefore, incorporating rendering into neural networks has a high potential for 3D understanding.

What type of 3D representation is most appropriate for modeling the 3D world? Commonly used 3D formats are voxels, point clouds and polygon meshes. Voxels, which are 3D extensions of pixels, are the most widely used format in machine learning because they can be processed by CNNs [2, 17, 20, 24, 30, 31, 34, 35, 36]. However, it is difficult to process high resolution voxels because they are regularly sampled from 3D space and their memory efficiency is poor. The scalability of point clouds, which are sets of 3D points, is relatively high because point clouds are based on irregular sampling. However, textures and lighting are difficult to apply because point clouds do not have surfaces. Polygon meshes, which consist of sets of vertices and surfaces, are promising because they are scalable and have surfaces. Therefore, in this work, after experimenting with different 3D representation, we use the polygon mesh as our 3D format.

One advantage of polygon meshes over other representations in 3D understanding is its compactness. For example, to represent a large triangle, a polygon mesh only requires three vertices and one face, whereas voxels and point clouds require many sampling points over the face. Because polygon meshes represent 3D shapes with a small number of parameters, the model size and dataset size for 3D understanding can be made smaller.

Can we train a system including rendering as a neural network? This is a challenging problem. Rendering consists of projecting the vertices of a mesh onto the screen coordinate system and generating an image through regular grid sampling [16]. Although the former is a differentiable operation, the latter, referred to as rasterization, is difficult to integrate because back-propagation is prevented by the discrete operation.

In this project, we worked on an application of neural renderer presented by the paper Neural 3D Mesh Renderer [17], i.e gradient-based 3D mesh editing with 2D supervision. This basically includes 3D version of style transfer[6]. This task cannot be realized without a differentiable mesh renderer because voxels or point clouds have no smooth surfaces and textures can’t be captured in them.

The major contributions of this project can be summarized as follows:

* We perform 3D style transfer using voxel, point-cloud and mesh based 3D representation and compare the results.
* We perform 3D style transfer using Multiview and single view object reconstruction and present the findings that reconstruction does not capture the essence of texture as good as 3D mesh deformation.
* We finally present PyTorch implementation and optimization of gradient-based 3D mesh editing pipeline presented in the paper Neural 3D Mesh Renderer.
* radient-based 3D mesh editing with 2D supervision. This includes a 3D version of style transfer [6] and DeepDream [18]. This task cannot be realized without a differentiable mesh renderer because voxels or point clouds have no smooth surfaces.

## Language

All manuscripts must be in English.

## Dual submission

Please refer to the author guidelines on the CVPR 2021 web page for a discussion of the policy on dual submissions.

## Paper length

Papers, excluding the references section, must be no longer than eight pages in length. The references section will not be included in the page count, and there is no limit on the length of the references section. For example, a paper of eight pages with two pages of references would have a total length of 10 pages. **There will be no extra page charges for CVPR 2021**.

Overlength papers will simply not be reviewed. This includes papers where the margins and formatting are deemed to have been significantly altered from those laid down by this style guide. Note that this LATEX guide already sets figure captions and references in a smaller font. The reason such papers will not be reviewed is that there is no provision for supervised revisions of manuscripts. The reviewing process cannot determine the suitability of the paper for presentation in eight pages if it is reviewed in eleven.

## The ruler

The LATEX style defines a printed ruler which should be present in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document using a non-LATEX document preparation system, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LATEX users may change \documentclass command with the “final” option in the document preamble.) Reviewers: note that the ruler measurements do not align well with lines in the paper—this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (e.g. this line is 189.5), although in most cases one would expect that the approximate location will be adequate.

## Mathematics

Please number all of your sections and displayed equations. It is important for readers to be able to refer to any particular equation. Just because you didn’t refer to it in the text doesn’t mean some future reader might not need to refer to it. It is cumbersome to have to use circumlocutions like “the equation second from the top of page 3 column 1”. (Note that the ruler will not be present in the final copy, so is not an alternative to equation numbers). All authors will benefit from reading Mermin’s description of how to write mathematics: <http://www.pamitc.org/documents/mermin.pdf>. Every equation should be numbered, even if you don't refer to it!

## Blind review

Figure 1: Example of a figure with caption. Captions are set in roman, 9 point. Use a Drawing area to make space for figures.

Many authors misunderstand the concept of anonymizing for blind review. Blind review does not mean that one must remove citations to one’s own work—in fact it is often impossible to review a paper unless the previous citations are known and available.

Blind review means that you do not use the words “my” or “our” when citing previous work. That is all. (But see below for techreports)

Saying “this builds on the work of Lucy Smith [1]” does not say that you are Lucy Smith, it says that you are building on her work. If you are Smith and Jones, do not say “as we show in [7]”, say “as Smith and Jones show in [7]” and at the end of the paper, include reference 7 as you would any other cited work.

An example of a bad paper:

An analysis of the frobnicatable foo filter.

In this paper we present a performance analysis

of our previous paper [1], and show it to be inferior

to all previously known methods. Why the

previous paper was accepted without this analysis

is beyond me.

[1] Removed for blind review

An example of an excellent paper:

An analysis of the frobnicatable foo filter.

In this paper we present a performance analysis of

the paper of Smith et al. [1], and show it to be

inferior to all previously known methods. Why the

previous paper was accepted without this analysis

is beyond me.

[1] Smith, L and Jones, C. “The frobnicatable

foo filter, a fundamental contribution to human

knowledge”. Nature 381(12), 1-213.

If you are making a submission to another conference at the same time, which covers similar or overlapping material, you may need to refer to that submission in order to explain the differences, just as you would if you had previously published related work. In such cases, include the anonymized parallel submission [4] as additional material and cite it as

[1] Authors. “The frobnicatable foo filter”, Face and Gesture 2014 submission ID 324, Supplied as additional material efg324.pdf.

Finally, you may feel you need to tell the reader that more details can be found elsewhere, and refer them to a technical report. For conference submissions, the paper must stand on its own, and not require the reviewer to go to a techreport for further details. Thus, you may say in the body of the paper “further details may be found in [5]”. Then submit the techreport as additional material. Again, you may not assume the reviewers will read this material.

Sometimes your paper is about a problem which you tested using a tool which is widely known to be restricted to a single institution. For example, let’s say it’s 1969, you have solved a key problem on the Apollo lander, and you believe that the CVPR70 audience would like to hear about your solution. The work is a development of your celebrated 1968 paper entitled ”Zero-g frobnication: How being the only people in the world with access to the Apollo lander source code makes us a wow at parties”, by Zeus *et al.*

You can handle this paper like any other. Don’t write “We show how to improve our previous work [Anonymous, 1968]. This time we tested the algorithm on a lunar lander [name of lander removed for blind review]”. That would be silly, and would immediately identify the authors. Instead write the following:

We describe a system for zero-g frobnication. This system is new because it handles the following cases: A, B. Previous systems [Zeus et al. 1968] didn’t handle case B properly. Ours handles it by including a foo term in the bar integral.

...

The proposed system was integrated with the Apollo lunar lander, and went all the way to the moon, don’t you know. It displayed the following behaviours which show how well we solved cases A and B: ...

As you can see, the above text follows standard scientific convention, reads better than the first version, and does not explicitly name you as the authors. A reviewer might think it likely that the new paper was written by Zeus et al, but cannot make any decision based on that guess. He or she would have to be sure that no other authors could have been contracted to solve problem B.

FAQ

**Q:** Are acknowledgements OK?

**A:** No. Leave them for the final copy.

**Q:** How do I cite my results reported in open challenges?

**A:** To conform with the double blind review policy, you can report results of other challenge participants together with your results in your paper. For your results, however, you should not identify yourself and should not mention your participation in the challenge. Instead present your results referring to the method proposed in your paper and draw conclusions based on the experimental comparison to other results.

## Miscellaneous

When citing a multi-author paper, you may save space by using “*et alia*”, shortened to “*et al*.” (not “*et. al.*” as “*et*” is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct:

“Frobnication has been trendy lately. It was introduced

by Alpher [3], and subsequently developed by Alpher and Fotheringham-Smythe [1], and Alpher *et al.* [2].”

This is incorrect: “... subsequently developed by Alpher et al. [1] ...” because reference [1] has just two authors. If you use the \etal macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher et al.

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [1, 3, 4] to [3, 1, 4].

# Formatting your paper

All text must be in a two-column format. The total allowable width of the text area is inches (17.5 cm) wide byinches (22.54 cm) high. Columns are to be 31/4 inches (8.25 cm) wide, with a 5/16 inch (0.8 cm) space between them. The main title (on the first page) should begin 1.0 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be inches (2.86 cm) from the bottom edge of the page for 8.5 × 11-inch paper; for A4 paper, approximatelyinches (4.13 cm) from the bottom edge of the page.

## Margins and page numbering

Figure 2: Short captions should be centred.

All printed material, including text, illustrations, and charts, must be kept within a print area  inches (17.5 cm) wide by inches (22.54 cm) high. Page numbers should be in footer with page numbers, centered and .75 inches from the bottom of the page and make it start at the correct page number rather than the 4321 in the example (how to do that depends on your version of word or open office. Failure to use the correct page number, or place it properly, could result in the paper not being included in Xplore, (even if it passes PDF express (which does not check page number)

## Type-style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title 1-3/8 inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a twocolumn format.

MAIN TEXT. Type main text in 10-point Times, singlespaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Figure and table captions should be 9-point Roman type as in Figures 1 and 2. Short captions should be centred. Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, 1. Introduction) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. Should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

## Footnotes

Please use footnotes[[1]](#footnote-1) sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

## References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [4]. Where appropriate, include the name(s) of editors of referenced books.

## Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it.

You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic. When placing figures in LATEX, it’s almost always best to use \includegraphics, and to specify the figure width as a multiple of the line width as in the example below

\usepackage[dvips]{graphicx} ...

\includegraphics[width=0.8\linewidth]

{myfile.eps}

## Color

Please refer to the author guidelines on the CVPR 2021 web page for a discussion of the use of color in your document.

# Final copy

You must include your signed IEEE copyright release form when you submit your finished paper. We MUST have this form before your paper can be published in the proceedings. Please direct any questions to the production editor in charge of these proceedings at the IEEE Computer Society Press:

https://www.computer.org/about/contact

# References

FirstName Alpher, , and J. P. N. Fotheringham-Smythe. Frobnication revisited. Journal of Foo, 13(1):234–778, 2003.

FirstName Alpher, , FirstName Fotheringham-Smythe, and FirstName Gamow. Can a machine frobnicate? Journal of Foo, 14(1):234–778, 2004.

FirstName Alpher. Frobnication. Journal of Foo, 12(1):234–778, 2002.

Actual Author Name. The frobnicatable foo filter, 2014. Face and Gesture (to appear ID 324).

Actual Author Name. Frobnication tutorial, 2014. Some URL al tr.pdf.

1. **Gradient-based 3D mesh editing**

Gradient-based image editing techniques [6, 18] generate an image by minimizing a loss function L(x) on a 2D image x via gradient descent. In this work, instead of generating an image, we optimize a 3D mesh m consisting of vertices {vi}, faces {fi}, and textures {ti} based on its rendered image R(m|φi).

4.1 **2D-to-3D style transfer Method and Loss Functions**

In this section, we describe the method to transfer the style of an image x s onto a mesh mc, that is proposed in the neural Renderer paper [12].

For 2D images, style transfer is achieved by minimizing content loss and style loss simultaneously [6]. Specifically, content loss is defined using a feature extractor fc(x) and content image x c as Lc(x|x c ) = |fc(x) − fc(x c )| 2 2 . Style loss is defined using another feature extractor fs(x) and style image x s as Ls(x|x s ) = |M(fs(x)) − M(fs(x s ))| 2 F . M(x) transforms a vector into a Gram matrix.

In 2D-to-3D style transfer, content is specified as a 3D mesh mc . To make the shape of the generated mesh similar to that of mc , assuming that the vertices-to-faces relationships {fi} are the same for both meshes, content loss is redefined as Lc(m|mc ) = P {vi,v c i }∈(m,mc) |vi − v c i | 2 2 . Style loss is same as that in the 2D application. Specifically, Ls(m|x s , φ) = |M(fs(R(m, φ))) − M(fs(xs))| 2 F . A regularizer for noise reduction is also used alongside with the above mentioned loss. Let P denote the a set of colors of all pairs of adjacent pixels in an image R(m, φ). We define this loss as Lt(m|φ) = P {pa,pb}∈P |pa − pb| 2 2 .

The final objective function is L = λcLc + λsLs + λtLt. Initial solution of m is set as mc and L is minimized with respect to {vi} and {ti}.

4.2 **Architecture**

This section describes the complete network architecture of gradient based 3D mesh style transfer that we implemented in our work using PyTorch and Cuda frameworks. A neural renderer architecture is leveraged for this application and a style transfer component is built on top of that. The same is illustrated in the Fig. 2.

A 3D mesh object is fed to the neural renderer pipeline which outputs corresponding multiview images. A VGG16 network is used to extract features from those images which are compared with those extracted from stylized feature image that is also given as an input to the network. We have specifically used features from 2nd, 7th, 14th and 21st layer of VGG network to extract most meaningful textures. Now a style loss is calculated by summing over the differences for each feature map output of rendered and textured image. This loss signifies the texture captured by the mesh. Another loss is calculated which represents the degree by which original mesh has been deformed to capture the texture present in the style image. This is termed as content loss. Both the losses are used to counter balance themselves which leads to stable deformation in the mesh.

Now, a weighted sum of all the losses is taken which is backpropagated to train the network parameters i.e., texture and vertices of 3D mesh object. The deformed mesh is then again fed to the network along with the style image and this whole process is repeated for multiples epochs.

The result of the run is a mesh with deformations and textures that mimics the texture from the given style image. The results of our experiments and work can be seen in the next section.

1. This is what a footnote looks like. It often distracts the reader from the main flow of the argument. [↑](#footnote-ref-1)